



Scaffold

Version 5.0

User's Manual

Release Information The following release information applies to this version of the *Scaffold*. This document is applicable for Scaffold, Release 5.0 or greater, and is current until replaced.

Copyright © 3/11/21. Proteome Software, Inc., All rights reserved.

The information contained herein is proprietary and confidential and is the exclusive property of Proteome Software, Inc.. It may not be copied, disclosed, used, distributed, modified, or reproduced, in whole or in part, without the express written permission of Proteome Software, Inc.

Limit of Liability Proteome Software, Inc. has made its best effort in preparing this guide. Proteome Software, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this guide and specifically disclaims any implied warranties of merchantability or fitness for a particular purpose. Information in this document is subject to change without notice and does not represent a commitment on the part of Proteome Software, Inc. or any of its affiliates. The accuracy and completeness of the information contained herein and the opinions stated herein are not guaranteed or warranted to produce any particular results, and the advice and strategies contained herein may not be suitable for every user.

The software described herein is furnished under a license agreement or a non-disclosure agreement. The software may be copied or used only in accordance with the terms of the agreement. It is against the law to copy the software on any medium except as specifically allowed in the license or the non-disclosure agreement.

Trademarks The name *Proteome Software*, the Proteome Software logo, *Scaffold*, *Scaffold Q+*, *Scaffold Q+S*, *Scaffold LFQ*, *Scaffold PTM*, *Scaffold DIA* and *Scaffold Elements* logos are trademarks or registered trademarks of Proteome Software, Inc. All other products and company names mentioned herein may be trademarks or registered trademarks of their respective owners.

Customer Support Customer support is available to organizations that purchase *Scaffold*, *Scaffold Q+*, *Scaffold Q+S*, *Scaffold LFQ*, *Scaffold PTM*, *Scaffold DIA* or *Scaffold Elements* and that have an annual support agreement. Contact Proteome Software at:

Proteome Software, Inc.

1340 SW Bertha Blvd

Suite 10

Portland, OR 97219

1-800-944-6027 (Toll Free)

1-503-245-4910 (Fax)

www.proteomesoftware.com

www.proteomesoftware.com

Table of Contents

Preface	7
Chapter 1: Getting Started with Scaffold	9
Chapter 2: Identifying Proteins with Scaffold	25
Chapter 3: Loading Data in Scaffold	41
Chapter 4: Scaffold's Main Window	65
Chapter 5: Load Data View.....	123
Chapter 6: Samples View	133
Chapter 7: Proteins View.....	157
Chapter 8: Similarity View.....	169
Chapter 9: Quantify View	177
Chapter 10: Publish View	189
Chapter 11: Statistics View	191
Chapter 12: Protein Grouping and Clustering	203
Chapter 13: Quantitative Methods and Tests	215
Chapter 14: Precursor Intensity Quantitation	235
Chapter 15: Reports.....	245
Appendix.....	263

Appendix A. Algorithm References	264
Appendix B. Terminology	267
Appendix C. Terminology comparison between Scaffold 4 and Scaffold 3	270
Appendix D. Context Menus Right Click Commands	271
Appendix E. Holm's and Hochberg's Techniques to Control the Familywise Error Rate	273

Preface

Welcome to the Scaffold User's Manual. Its purpose is to answer users' questions and guide them through the procedures necessary for using Scaffold efficiently and effectively.

Using the manual

A Table of Contents and an Index are provided in this manual for the user's convenience. This Preface also provides a brief discussion of each chapter to further assist users in locating needed information.

Special information about the manual

This User's Manual has a dual-purpose design. It can be distributed electronically and printed on an as-needed basis, or it can be viewed on-line in its fully interactive capacity. If users print the document, for best results it is recommended that they print it on a duplex printer; however, single-sided printing is also possible. When the document is viewed on-line, a standard set of bookmarks appears in a frame on the left side of the document window for navigation through the manual. For better viewing, users can decrease the size of the bookmark frame and use the magnification box to adjust the display according to their viewing preferences.

If users do print the document using a single-sided printer, they might see a single blank page at the end of some chapters. This blank page has been added solely to ensure that the next chapter begins on an odd-numbered page. This blank page in no way indicates that the book is missing information.

Conventions used in the manual

The User's Manual uses the following conventions:

- Information that can vary in a command—variable information—is indicated by alphanumeric characters enclosed in angle brackets; for example, <ProteinAnalyte Name>.
- A new term, or term that must be emphasized for clarity of procedures, is *italicized*.
- Page numbering is “on-line friendly.” Pages are numbered from 1 to x, *starting with the cover* and ending on the last page of the index.
- This manual is intended for both print and on-line viewing.
- If information appears in [blue](#), it is a hyperlink. Table of Contents and Index entries are also hyperlinks. Click the hyperlink to advance to the referenced information.
- A sample set of Demo data, available for download from <http://www.proteomesoftware.com/products/demo-data> is used as the basis for most screen captures, examples, and data manipulations that are shown in the manual.

Assumptions in the manual

The Scaffold User's Manual assumes that:

- The user is familiar with Windows operating systems, and basic Windows navigational elements, content formatting and layout tools.
- The user has the appropriate licensing to run Scaffold.

Chapter 1

Getting Started with Scaffold

System Requirements

For information about the system requirements for Scaffold, see:

<https://www.proteomesoftware.com/system-requirements>

Installing Scaffold

Scaffold runs on Windows, MAC or Linux systems. Follow these instructions to install the application on your system:

Request an evaluation by filling in the form found at . You will receive download instructions and a license key to activate the software via email.

2. Download and launch the installation executable.
3. Carefully follow the instructions provided in the installation wizard, accepting the user agreement when prompted and moving through the screens by clicking Next.

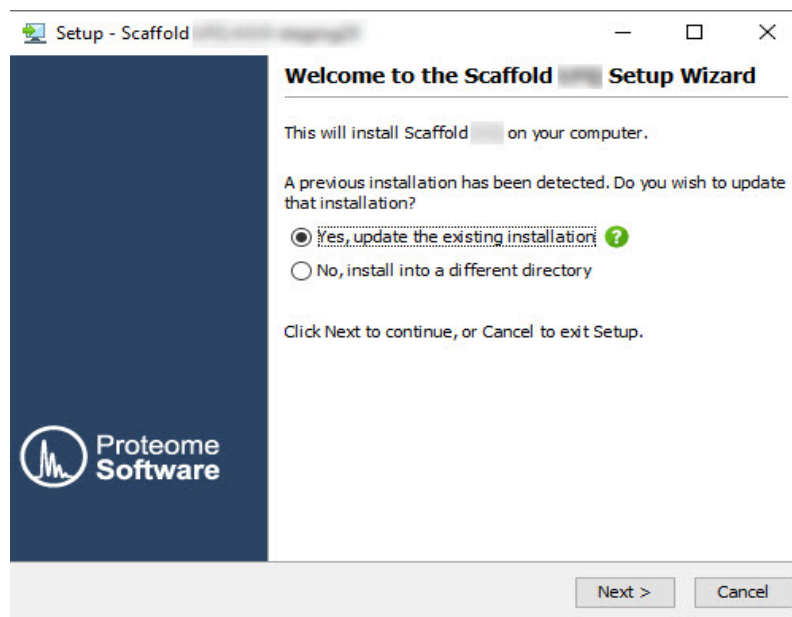


Figure 1-1: Scaffold installation Setup Wizard

4. The installer will then provide you an opportunity to allocate memory to Scaffold. We recommend that you set the Maximum Memory to approximately 80% of the amount of physical RAM on your system. Click “Next”.
5. You may then select a Start Menu Folder for the application and choose whether or not to create shortcuts for all users of the system. The next screen allows you to set a file association between SF3 files and Scaffold, and the following screen allows creation of desktop icons. Clicking “Next” begins the installation.
6. Finally, Scaffold allows you to select the option to have the program open at the closing of the wizard. Click “Finish”..



*For better performance you should allocate as much RAM as possible to Scaffold. The memory setting can be adjusted after installation by selecting the menu option **Edit > Preferences - Memory tab**. You must close Scaffold and restart the program in order for the new memory setting to take effect.*

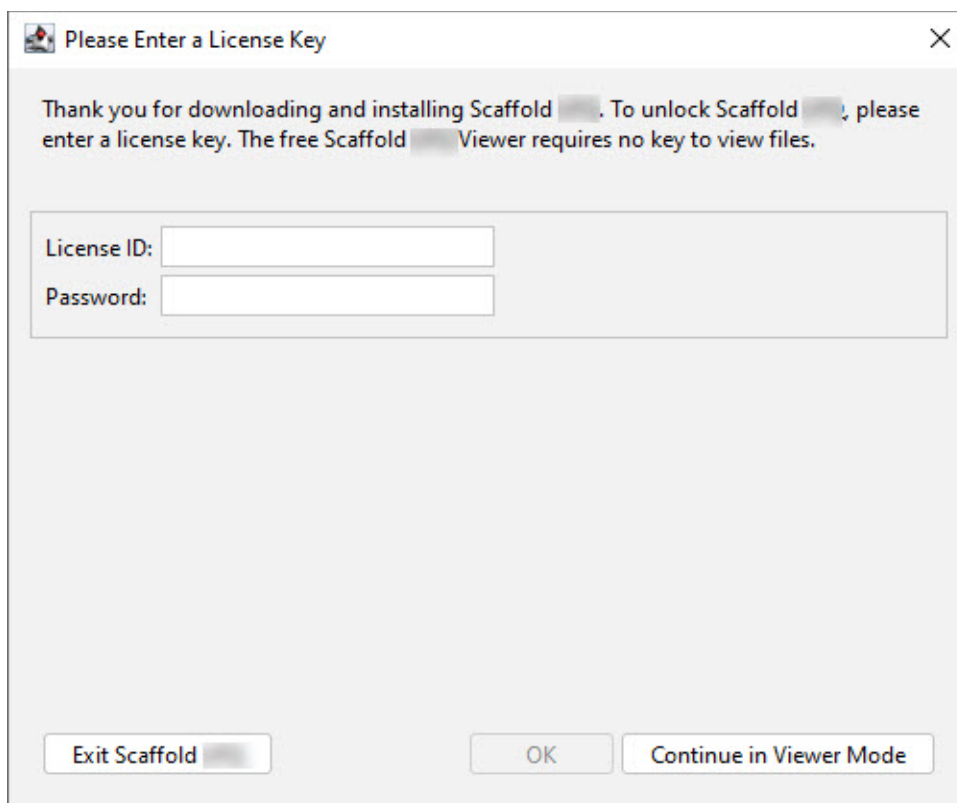
After Scaffold has been installed on a computer, a shortcut icon for the application is placed on the desktop. An option is also available from the Start menu. Double-clicking the desktop icon launches Scaffold, as does, for Windows computers, selecting the option from the Start menu (**Start > All Programs > Scaffold > Scaffold**)

Licensing

The first time Scaffold opens after installation, the Enter License Key dialog box opens.

Keys and passwords may be typed, pasted or dragged into the appropriate fields. Both items may be pasted or dragged together.

Figure 1-2: Scaffold License Key messages



Please Enter a License Key

Thank you for downloading and installing Scaffold [redacted]. To unlock Scaffold [redacted], please enter a license key. The free Scaffold [redacted] Viewer requires no key to view files.

License ID:

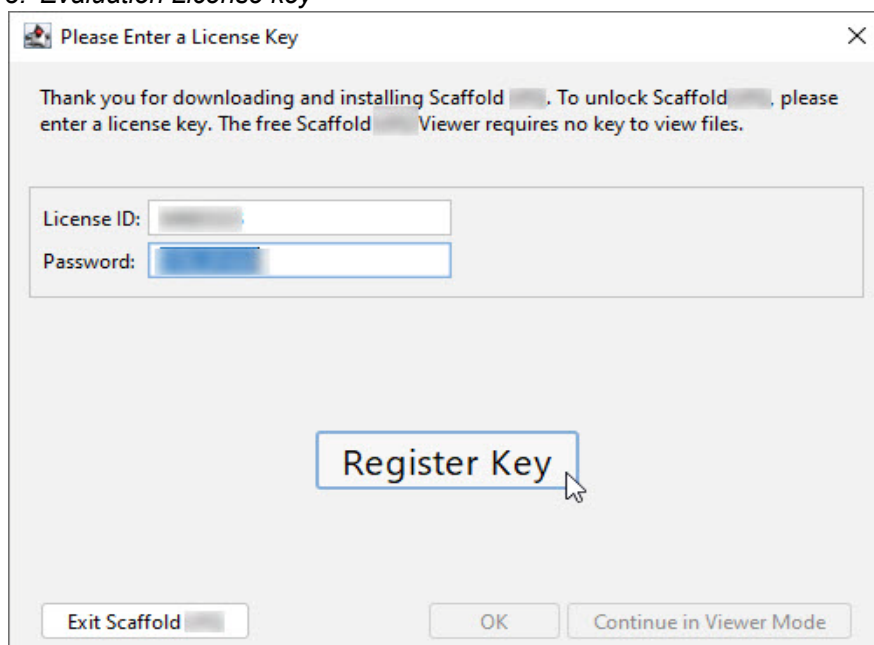
Password:

Exit Scaffold [redacted] OK Continue in Viewer Mode

Two kinds of keys are available to activate the software:

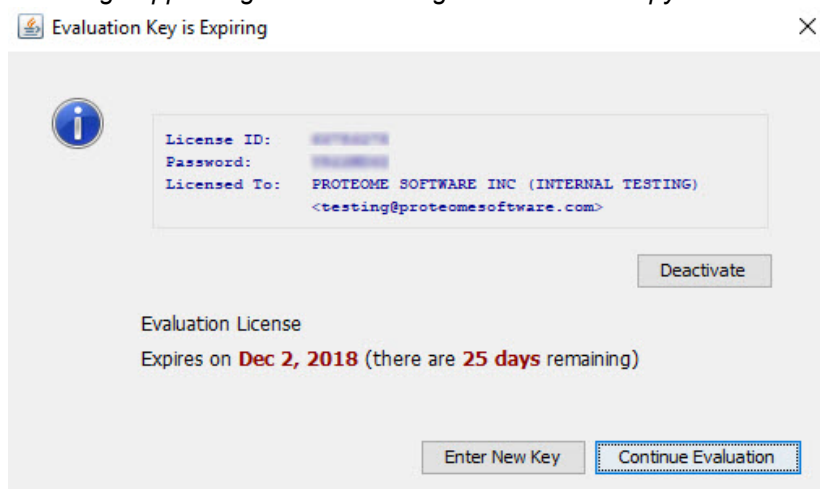
Evaluation key - An Evaluation key is valid for a limited period. A free evaluation key for Scaffold may be obtained through www.proteomesoftware.com. An evaluation key may be used on two computers. Once the key and password have been copied and pasted into the license key dialog box, a message will appear below it, displaying confirmation of the key registration. Pressing OK starts the application.

Figure 1-3: Evaluation License key



Every time Scaffold is launched in evaluation mode, a message appears showing the remaining time available for evaluation and offering the option to enter a new key.

Figure 1-4: Message appearing when launching an evaluation copy of Scaffold

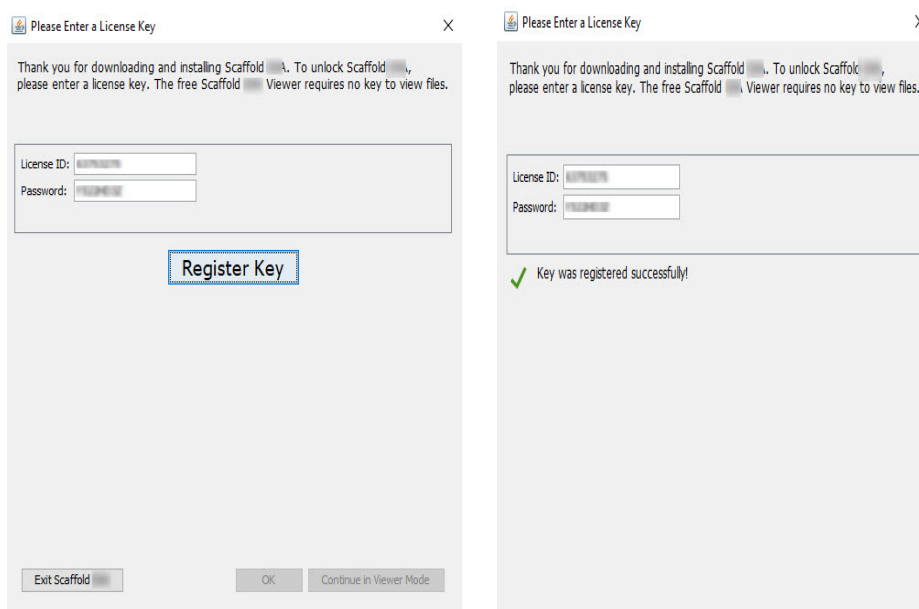


Time-Based License key—a Time-Based License key allows the user to access all features of the software permanently. It only allows upgrades within a certain time limit, however. The time tracks the length of the support contract. Once expired, Scaffold will continue to work beyond the expiration date, but no upgrades are allowed unless the support contract is renewed.

Contact sales@proteomesoftware.com to purchase the appropriate key.

A Time-Based License key is valid only for a single computer. If it is necessary to move the Scaffold installation to a different computer, see for instructions to transfer the key at no charge.

Figure 1-5: Time-Based License key



When the Time-Based License key and password are entered, pressing **Register Key** verifies their validity and a message appears describing the status of the key.

Once the key is successfully registered, pressing OK closes the dialog box and a Scaffold Welcome message opens.



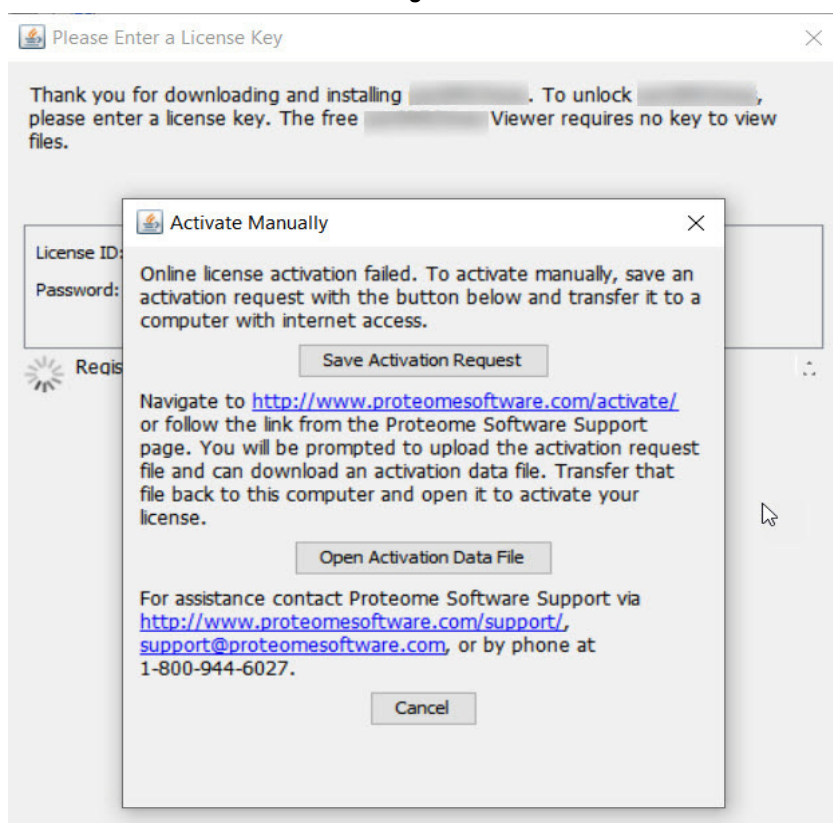
If the user is using an evaluation copy of Scaffold, then an Evaluation message opens, indicating the number of days left in the evaluation period. The user must click OK to close this message and then the Scaffold Welcome message opens.

From this window, the user may create a new experiment, open an existing experiment (*.SF3 file), or work with the demonstration data that is provided in the Scaffold installation.

Registering a Time-Based License key with no INTERNET connection

When a Time-Based License key is entered and the Register Key button is pressed, but no INTERNET connection is available, a dialog appears, providing instructions for manual activation.

Figure 1-6: Manual or offline activation dialog



To activate Scaffold without an internet connection:

1. First, use the Save Activation Request button to create an activation request file.
2. Transfer this file to a computer with internet access (e.g. using a USB drive).
3. On the connected computer, navigate to <http://www.proteomesoftware.com/activate/> This link is also accessible from the Proteome Software Support page (<http://www.proteomesoftware.com/support/>) to make it easier to access from the internet-connected computer.
4. The License Portal will open. The Portal provides two different options for activating your software. Use the Browse button in the Upload Request File section on the right, and select the activation request file that was transferred from the offline computer (See [Figure 1-7](#) below).

Figure 1-7: The Proteome Software License Portal

LICENSE PORTAL

[License Portal Home](#) > Manual Request [Log In](#)

Manual Request

This page may be used for processing manual requests, including activation, deactivation, and license refreshing and status checks. Please use the appropriate method of posting the request to retrieve a response.

Copy and Paste Request

Please copy the request from the application, right-click in the text box below and click paste, then click the submit button below.

Please paste the contents of the request here.

Submit

Upload Request File

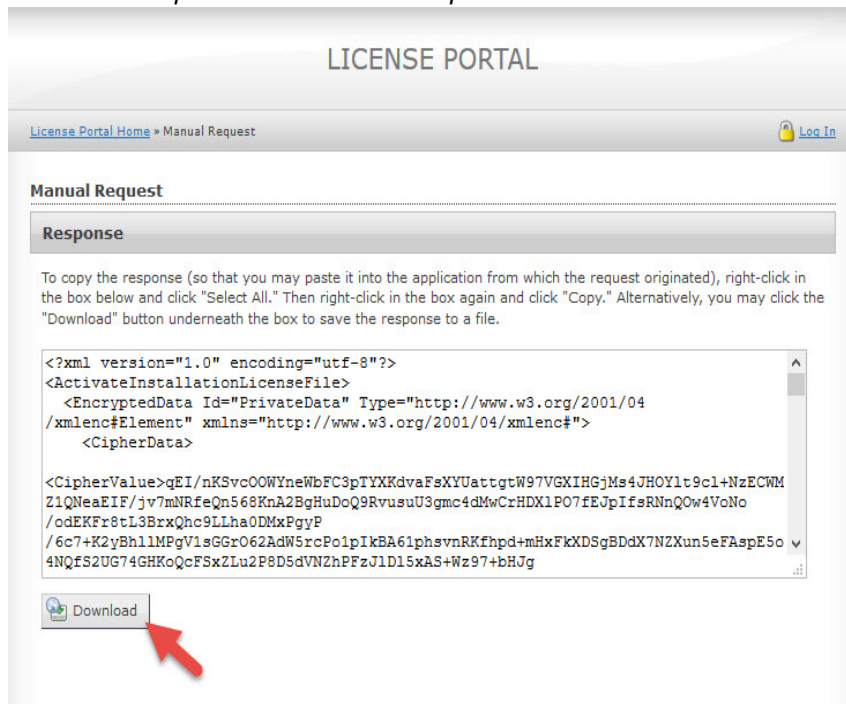
Please select the file you wish to upload below and click the submit button.

Browse... No file selected.

Submit

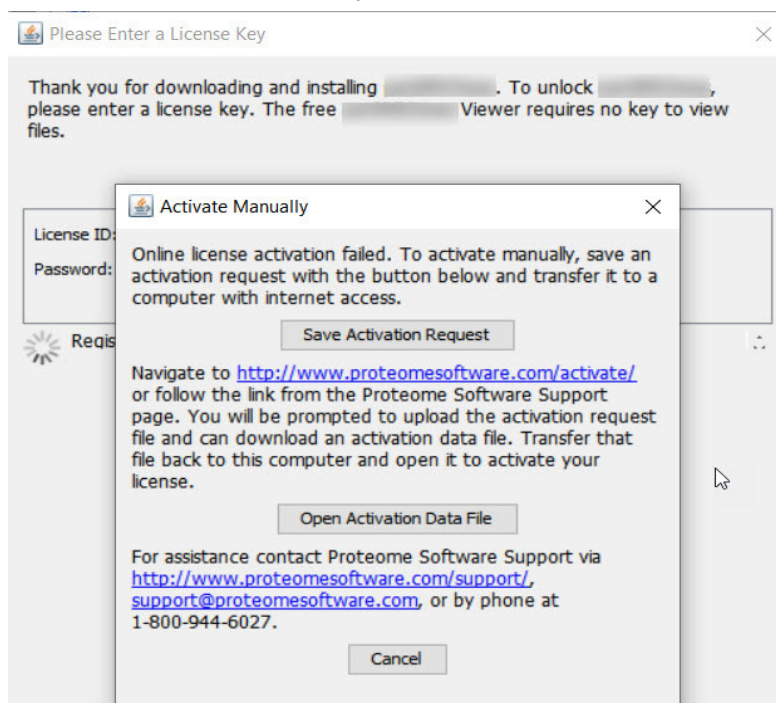
5. Click the Submit button just below the Browse button to upload the activation request file. The license portal will respond with a long text sequence(See [Figure 1-8](#) below).

Figure 1-8: .License Portal Response to Activation Request



6. Click the Download button to save the response to a file named response.xml, which will be downloaded to the default download location.
7. Transfer the response.xml file to the computer on which Scaffold has been installed.
8. Return to Scaffold on the disconnected computer. Select Open Activation Data File.

Figure 1-9: Select Activation File returned by the License Portal



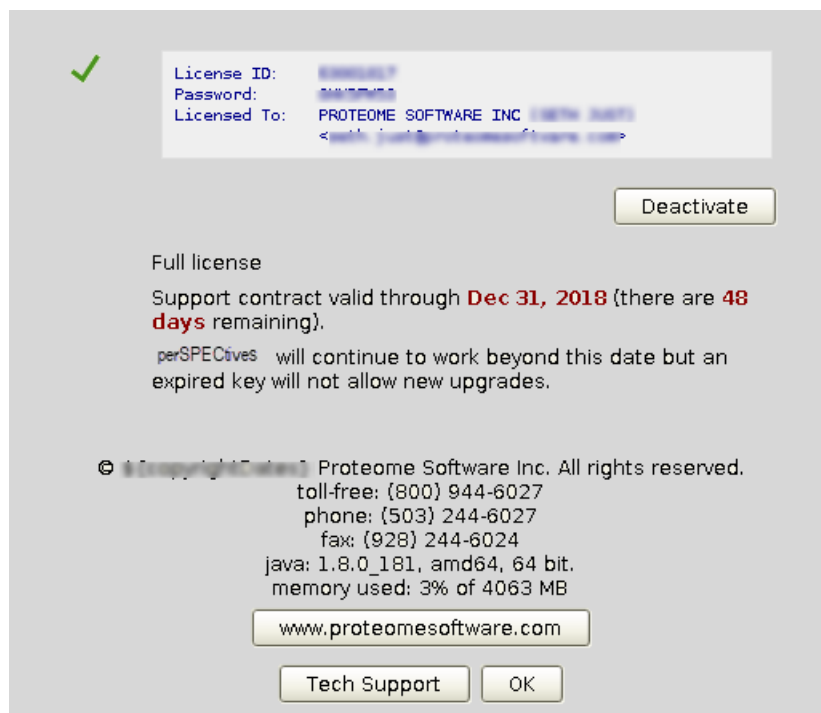
9. Browse to locate the response.xml file and click Open.
10. Scaffold should report that the key was registered successfully. If not, please contact Proteome Software Support for assistance.

Time based license key renewal

Time based license keys have time limits connected to the term of the user's support contract. When the support contract expires, Scaffold continues to work but upgrades are not allowed until the contract is renewed. The status of the Scaffold license key may be checked by selecting **Help > About Scaffold** from the main menu.

If the contract has expired and the user wishes to upgrade Scaffold, clicking the **Renew** button in the dialog opens the **Key reset Request** page on the Proteome Software website. The user should complete the request. A sales representative will promptly contact him/her providing further information.

Figure 1-10: About Scaffold dialog



Moving Scaffold to a different computer

Each permanent Scaffold key allows activation of the program on a single computer. If it becomes necessary to reinstall the program either on a different computer or on the same computer following an operating system upgrade or hardware replacement, the user may deactivate the key and then reactivate it on the new system. This may be done once per support contract period. If additional reinstallations are required within the same period, please contact Proteome Software Support.

To deactivate a key:

1. Be sure you have a record of your key and password. These were sent via email at the time of purchase, or may be copied from the Help>>About Scaffold dialog.
2. Select Help>>Update License Key and click the Deactivate button.

To reinstall Scaffold:

1. Download the program from the Proteome Software website to the new system and run the installation program.
2. Paste in the key and password and register as described in [Installing Scaffold](#).

Scaffold Viewer

A free Scaffold Viewer may be downloaded from www.proteomesoftware.com. The Viewer can open and display any *.SF3 file created by Scaffold, and allows users to distribute Scaffold results to colleagues, collaborators or reviewers.

The Viewer may be installed on any number of computers, and multiple instances of the Viewer may be run on a single computer simultaneously. It performs most of the functions of the full Scaffold program, but it cannot load search results files and analyze data.

With the Viewer, the user (or his/her collaborators) can view the data, just as in Scaffold, by samples, proteins, peptides or spectra. The user may apply thresholds, change summarization, adjust legend colors, move columns and hide rows. The Viewer user may also validate the peptide/spectrum matches.

Only a single fully-licensed instance of Scaffold may be run on a computer at one time. Additional instances will function as Viewers.

Scaffold Batch

Scaffold Batch is a command line version of Scaffold. It can load and analyze the same data that Scaffold does, but in a batch rather than in an interactive environment. Batch mode means that Scaffold Batch can be run on the command line or called from a batch script. Scaffold Batch is an extended version of Scaffold that can be used as one component of an automated Proteomics work flow.

When the user installs Scaffold Elements, a copy of the interactive version of Scaffold is automatically installed as well. Like Scaffold, Scaffold Elements is locked to one computer by a software key. As a command line program, Scaffold Elements is intended to be called from a batch script. In the Microsoft world this might be a *.BAT file. In the Linux world this might be a *.SH file.

Scaffold Elements is driven by an XML file (*.SCAFML) that specifies all the needed operations to create a *.SF3 Scaffold file experiment.

If Scaffold is installed with a Scaffold Batch license, the user may generate basic SCAFML files and run the analyses they specify through a Batch Job Queue Dialog (or Wizard) available through the regular Scaffold Graphical User Interface. A Batch license activates an additional icon in the main menu bar



For more detailed technical information about how to install and run Scaffold, consult the Scaffold Elements manual at: www.proteomesoftware.com/pdf/scaffold_batch_users_guide.pdf.

How Scaffold structures data

Scaffold stores all the data related to an experiment in one single file. Each experiment file (*.SF3) can hold a large number of spectra and associated data. The user can create, name, and save as many experiment files as disk space permits, but only one at a time unless Scaffold is run in batch mode with a Scaffold Batch license. Multiple experiments can be opened in the Viewer mode.

Experimenters frequently categorize biological samples into larger groups to compare, for example, diseased with control, treated with control, day1 with day2, pregnant with not pregnant, etc. To capture this, Scaffold associates a sample category with each biological sample.

Data associated with a biological sample (abbreviated in Scaffold as BioSample) comes from a sample taken by a doctor, medical researcher, or biologist, such as a drop of blood or biopsy from a patient, or a tissue sample from a model organism or cell line. Using such techniques as 2D gels or liquid chromatography, proteins or peptides from these biosamples are then separated from each other. Each resulting individual band, spot, or LC fraction processed by a mass spectrometer is one mass spectrometry sample (abbreviated in Scaffold as MS sample).

One BioSample is therefore typically made up of more than one MS sample — sometimes many more.

Scaffold can also process data from MuDPIT experiments, in which case the analysis combines peptides from all fractions into one MS sample for protein identification.

Data Loading

Scaffold can load data in two different modes. Traditionally, Scaffold loaded only search engine results. It allowed the user to run confirmatory searches using X! Tandem, but did not allow processing of raw files. Beginning with Scaffold 5, however, the program also supports searching raw files using MSFragger¹.

Loading Raw Data

In order to search raw data through Scaffold, the user must download, install and license MSFragger on the system on which Scaffold is installed, then point Scaffold to the executable file. Scaffold then provides an interface for entering search parameters, running the search and processing the results. Currently, spectral counting is the only quantitative method supported for raw searches through Scaffold. To perform quantitation based on precursor intensity or isobaric labeling with MSFragger, the user should run FragPipe and load the results into Scaffold.

1. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., & Nesvizhskii, A. I. (2017). MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nature Methods*, 14(5), 513-520.

Linking Search Engines

Two search engines, MSFragger and X! Tandem, may be linked to Scaffold. MSFragger allows Scaffold to search raw data files or to confirm results from other search engines. X! Tandem may only be used for confirmatory searches when loading results from other search engines.

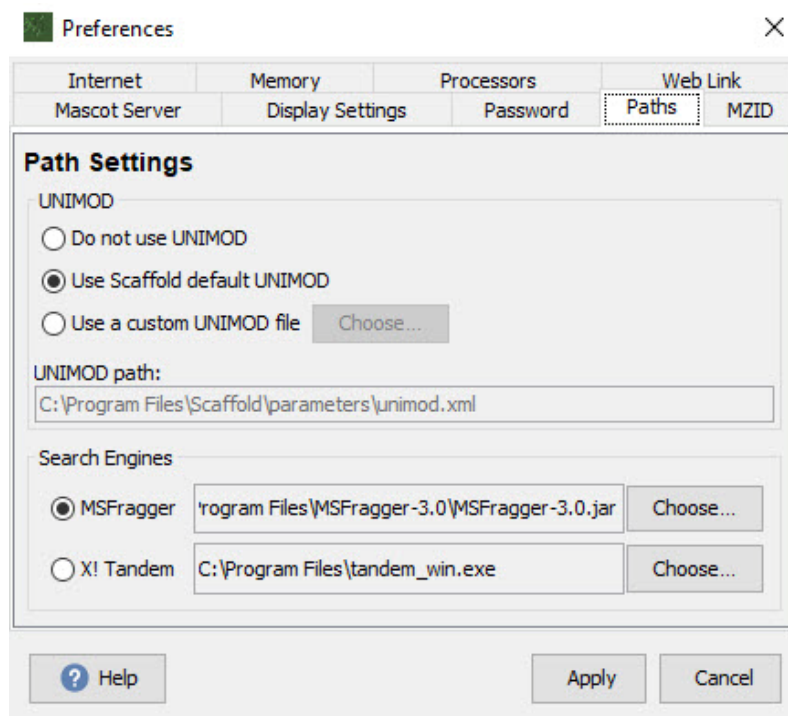
Please note that X! Tandem is no longer provided as part of the Scaffold installation and must be downloaded, installed and linked. X! Tandem may be downloaded from <https://www.proteomesoftware.com/xtandem>.

The user must also download MSFragger (see <https://www.proteomesoftware.com/msfragger>), install it and point Scaffold to the executable jar file. It is the responsibility of the user to ensure that all licensing requirements are observed.

Before running searches it is necessary to provide Scaffold with the location of the MSFragger executable jar and/or the X! Tandem executable file using the Edit>Preferences>Paths dialog.

To link a search engine, click the Choose button for that program, browse to the location of the MSFragger jar file or the X! Tandem executable file and click Open. Scaffold must be set to use either MSFragger or X! Tandem. The user may change from one to the other at any time, but may not use both in a single experiment. To select a search engine for Scaffold to use, click the radio button next to the program's name.

Figure 1-1: Linking Search Engines to Scaffold



Loading Search Engine Results

Scaffold imports data generated by a large variety of search engines, including Mascot, SEQUEST Spectrum Mill, OMSSA, Phenyx, X!Tandem, MaxQuant, Proteome Discoverer and FragPipe. It also supports those search engines that are able to export their results in the mzIdentML format. Results of different search engines can be freely combined in one experiment. Each SEQUEST folder is imported as one “file,” as is each Mascot or X!Tandem file.

Importing files requires access rights from the computer where Scaffold is installed to the location where the data files containing the search results reside.

BioSamples are defined as part of the import process. The user names and categorizes each BioSample and specifies the files to be included, and the BioSample is created when data is loaded and analyzed. Then names and categories of BioSamples may be edited after loading.

The user initially loads data files into Scaffold using the Scaffold Loading Wizard, which is launched by selecting **New** from the [Main menu commands](#). Files may be added to an existing experiment by using options available in the [Load Data View](#).

Documents on the Proteome Software website provide detailed information on search engine data files compatible with Scaffold, Scaffold Q+ or Scaffold Q+S:

[File_compatibility_matrix.pdf](#) as well as how to set up searches with the most popular search engines to optimize compatibility with Scaffold:
[loading_search_engine_results_into_scaffold.pdf](#).

Quantitative Data

Scaffold Q+ and Q+ S are Proteome Software's software packages for doing labeled quantitation.

- Q+ loads iTRAQ (Applied Biosystems) and Tandem Mass Tagged (TMT, Thermo Scientific) labeled data.
- Q+S can also load stable isotope labeled samples.

Quantitative data is loaded into Scaffold Q+ or Scaffold Q+S through Scaffold's file loading wizard.

Quantitative Data File compatibility

Please, check the file compatibility matrix at the following link:

[File_compatibility_matrix.pdf](#)

Characterizing data

The data files imported into Scaffold are the results of a previous search against a specific FASTA protein database, using a particular search engine (SEQUEST, Mascot, X! Tandem or others), on a particular set of data.

When the user imports these data files, Scaffold needs to know certain characteristics of the specific search, so the user in the loading phase will be asked to:

- [Specifying the Database](#)
- Specify the parameters used for the search

Specifying the Database

As part of the loading process, the user needs to specify the particular FASTA database that the initial search engine used to identify proteins in the analyzed samples. This database must also be stored on a location accessible to the system where Scaffold is installed. It is important to specify the correct database. Without this information, Scaffold cannot display the full sequence of amino acids in a peptide, nor can it calculate, the sequence coverage.

All search engines, such as SEQUEST, Mascot, or X! Tandem, store the name of the database they have searched with their results. If the user is uncertain of the database used, he/she can use a text editor to search the search engine output files for the database name. It is possible, though, that the correct database resides on a local network under a different name.

It is best to use the same database for all search results loaded into a specific Scaffold experiment. This permits Scaffold to accurately align proteins found in different samples.

Chapter 2

Identifying Proteins with Scaffold

Scaffold is a tool designed to help mass spectrometrists and medical researchers confidently identify proteins in biological samples. Using output data from most of the search engines currently available, such as: SEQUEST®, Mascot®, MaxQuant, X! Tandem, MSFragger and many others, Scaffold validates, organizes, and interprets mass spectrometry data, so that a user can easily manage large amounts of data, compare samples, and search for protein modifications.

Scaffold makes it easier to search data repeatedly, using various search methods to find results that might otherwise be missed. For example, it enables the user to export unidentified spectra, which can then be searched against larger databases to find additional proteins.

Alternatively, Scaffold can export a new FASTA database consisting of only those proteins found in the loaded BioSamples to allow searching of unidentified spectra against the subset database using different parameters — for example, specifying other variable modifications.

Whether the aim of the user is broadening or deepening a search, Scaffold can then import the new data and bring to bear its tools for compiling, comparing, and analyzing the results.

This chapter covers the following topics:

- [“Scaffold Flexible Workflow” on page 26](#), which provides a brief description of possible work-flows to improve the analysis of the data sets loaded in a Scaffold experiment.
- [“Increased Confidence Using Peptide and Protein Validation Algorithms” on page 28](#), which describes the statistical validation methods used in Scaffold.
- [“Scaffold Views” on page 32](#), which provides an overview of the different data views available in the Scaffold window.

Scaffold Flexible Workflow

Scaffold supplements database search engines; it does not replace them. The user continues to analyze the output of his/her mass spectrometry experiments through SEQUEST, Mascot, MaxQuant, X!Tandem or another search engine compatible with Scaffold, as usual. Results are then imported into Scaffold.

- [Simple Workflow including Scaffold](#)
- [Broadening the Search](#)
- [Deepening the Search](#)



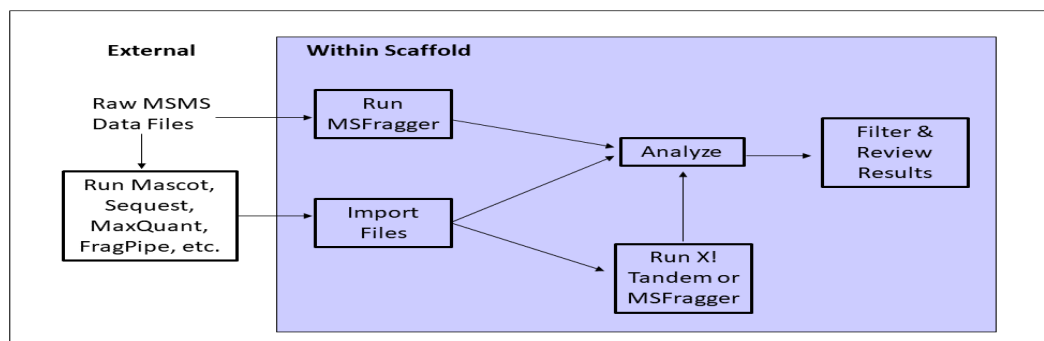
- For more information about compatibility with Scaffold check the following document: http://www.proteomesoftware.com/pdf/file_compatibility_matrix.pdf
- X!Tandem and MSFragger may be linked to Scaffold and run from within the program. MSFragger may be used to search raw data and either program may be run to increase the confidence of identifications from other search engines.

Simple Workflow including Scaffold

Database search engines match MS/MS spectra to theoretical spectra for peptides generated from theoretical digestion of proteins whose sequences are contained in a FASTA database. This results in files containing scored peptide-spectrum matches (PSMs). Scaffold uses scientifically validated, probabilistic methods to combine, evaluate and analyze these results, assigning probabilities to the PSMs to identify peptides that are likely to be contained in the samples and grouping the peptides to identify likely proteins. For more information about how Scaffold analyzes data, see the references in the appendix [Algorithm References](#).

Once the data has been loaded and analyzed by Scaffold, results are saved in special Scaffold files that bear the extension *.SF3. These Scaffold files can be closed and then reopened at a later time either through a full or a viewer version of Scaffold, see [Scaffold Viewer](#).

Figure 2-1: Data Analysis Workflow including Scaffold



Broadening the Search

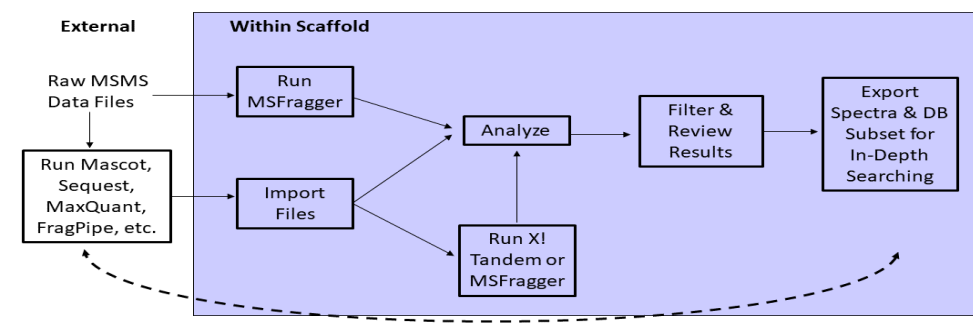
For further investigations the user can export the results to an Excel spreadsheet, to mzIdentML files or to an .SFDB file, which may be further analyzed with Scaffold LFFQ Spectral libraries may also be exported as .BLIB files, which may be used in Skyline¹ or Scaffold DIA.

Unidentified spectra may also be exported in a format that allows further searching by SEQUEST, Mascot, MaxQuant or by any other compatible search engine. The original searches can be repeated with different parameters, or the results can be searched using another search engine, or against a different database. Importing the new results into Scaffold augments the initial Scaffold analysis.

Deepening the Search

If, rather than casting a wider net, a deeper look is preferred, the user can export the unmatched spectra and a subset database consisting only of those proteins already found in the BioSamples included in the Scaffold experiment. The spectra can then be searched against this subset database, allowing a relatively fast search of these known proteins using variations on search parameters or looking for additional modifications.

Figure 2-2: Deepening the search



1. Pino, Mass Spectrometry Reviews 2017 article (cited by 38)
The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics

Increased Confidence Using Peptide and Protein Validation Algorithms

For validating peptide and protein identifications Scaffold offers several different scoring systems:

Peptide Scoring

- Percolator - Percolator² is a semi-supervised learning algorithm for peptide evaluation developed by Lukas Käll. It is widely used in a variety of proteomics software. Beginning with Scaffold 5, an implementation of Percolator is now bundled into Scaffold.
- [PeptideProphet](#) scoring algorithm with **High Mass accuracy**- This is a Bayesian statistical algorithm developed by the Institute for Systems Biology and available in Scaffold since its first version with the addition of high mass accuracy analysis.
- Prefiltered Mode - This option allows the user to load data which has already been evaluated and filtered by the search engine. All of Scaffold's peptide scoring algorithms require the full distribution of psm's for their probability models to succeed. In prefiltered mode, Scaffold relies on the scores assigned by the search engine and does not compute peptide probabilities. This mode should not be used when combining results from different search engines.
- [LFDR-based scoring system](#) - Developed in Scaffold 4, this algorithm was designed for high mass accuracy data (including data acquired on QExactive instruments). It has now been deprecated and it is recommended that Percolator be used instead.
- [PeptideProphet](#) scoring algorithm without **High Mass accuracy**- Bayesian statistical algorithm developed by the Institute for Systems Biology and available in Scaffold since its first version with high mass accuracy analysis included.

Protein Scoring

- [ProteinProphet](#) scoring algorithm - Bayesian statistical algorithm developed by the Institute for Systems Biology and available in Scaffold since its first version.

Implementations of the last two algorithms have been widely distributed under the names PeptideProphet™ and ProteinProphet®. Scaffold uses an independent implementation of these algorithms, for more information see [Algorithm References](#), [Keller \(2002\)](#) and [Nesvizhskii \(2003\)](#).

Percolator

Percolator² is a widely used and well-accepted algorithm for evaluating peptide-spectrum

2. Lukas Käll, Jesse Canterbury, Jason Weston, William Stafford Noble and Michael J. MacCoss.
Semi-supervised learning for peptide identification from shotgun proteomics datasets
Nature Methods 4:923 – 925, November 2007

matches. It requires decoy matches. Scaffold provides search engine scores and mass accuracy information to Percolator and it calculates posterior error probabilities (PEP) for the PSMs. Scaffold converts these scores to peptide probabilities.

-

PeptideProphet

When using PeptideProphet, Scaffold determines the distributions of the scores assigned by a search engine like SEQUEST, Mascot, MaxQuant or others. These depend on the size of the database used for the search and the specific characteristics of the analyzed sample, see [Keller \(2002\)](#). From these distributions, Scaffold translates the search engine scores into probabilities that each identification is correct. Scaffold's probabilities can then be used as threshold filters, allowing the identifications to be viewed at various confidence levels.

Scaffold's method contrasts with SEQUEST's, which uses an XCorr cut-off that depends on neither database size nor sample characteristics, frequently requiring ad hoc corrections for these parameters. Scaffold's statistical approach yields more reliable estimates of the probability of a correct identification.

Scaffold's method also supplements Mascot's. Mascot provides a probability estimate based on database size, but not on sample characteristics. By incorporating sample-specific distributions, Scaffold provides better estimates of the probability of correct identifications.

Two versions of PeptideProphet scoring are offered. In the generally preferred method, A further refinement in scoring is performed based on the mass accuracy of the PSMs. For use with lower mass accuracy instruments, the traditional PeptideProphet approach is offered without the high mass accuracy adjustment.

-

Prefiltered Mode

In some cases, it may be better to rely on the search engine's evaluation of peptide identifications rather than on Scaffold's probability models. For this reason, Scaffold provides the option to load search results in Prefiltered Mode. There are two cases in which this is the preferred mode, when loading:

- **Proteome Discoverer results with Percolator scores** - Prefiltered Mode is preferred if peptide identifications have been assessed by Percolator³ in Proteome Discoverer. If Proteome Discoverer results loaded into Scaffold in Prefiltered Mode contain PEP values calculated by Percolator, Scaffold uses these scores to compute probabilities. This allows the user to filter results using probability or FDR thresholds as usual.
- **Data from a single search engine that has already been thresholded or filtered prior to loading into Scaffold** - In this case, some of the data needed by Scaffold's probability models will be missing, affecting the results. In general, if a user chooses Prefiltered

3. Lukas Käll, John D. Storey, Michael J. MacCoss and William Stafford Noble
Assigning confidence measures to peptides identified by tandem mass spectrometry
Journal of Proteome Research, 7(1):29-34, January 2008

Mode upon loading, no peptide or protein probabilities will be displayed in Scaffold or in any reports exported from Scaffold. No probability or FDR-based thresholds may be applied, but the user may create [Custom Peptide Thresholds](#) based only on search engine scores to further filter the data.

In Prefiltered Mode, Scaffold displays all peptides as reported by the search engine that produced the data, but it runs its own protein analysis, so the identified protein list may differ from the original analysis by the search engine.

LFDR-based scoring system

In this method, peptide identifications are validated by discriminant scoring using a Naïve Bayes classifier generated through iterative rounds of training and validation to optimize training data set choices. Peptide probabilities are assessed using a Bayesian approach to local FDR (LFDR) estimation.

LFDR incorporates all psm scores reported by a search engine, as well as mass accuracy. Instead of PeptideProphet's LDA or Percolator's SVM classifier, LFDR uses log-likelihood ratios generated by Naïve Bayes classifiers to discriminate between target and decoy hits. Naïve Bayes was chosen specifically for robustness to over-fitting, a frequently occurring problem when training a classifier on a subset of testing data.

Training data is selected by iteratively testing three sets of ten classifiers to home in on the optimal number of spectra to avoid training with incorrect identifications assigned to target proteins. Then the posterior peptide probabilities are derived using LFDR estimates in a Bayesian framework. Instead of considering LFDR bins of discrete score distances, Scaffold uses variable width bins keeping the number of values in each bin constant. This gives more refined assessments of probability in score areas with more values, while simultaneously ensuring that LFDR estimates stay reasonable with fewer values. Finally, a Bayesian algorithm is used to confirm peptide probabilities based on likelihoods calculated using parent mass accuracy.

After the calculations have been done, a statistical test is applied to insure that LFDR will produce a good probability model. If the results of this test indicate that it would not, Scaffold falls back to using PeptideProphet with delta mass correction.

ProteinProphet

ProteinProphet groups the peptides by their corresponding protein(s) to compute probabilities that those proteins were present in the original sample, see [Nesvizhskii \(2003\)](#).

Three options are offered for protein grouping: experiment wide protein grouping, independent sample protein grouping, and protein cluster analysis. In experiment wide or independent sample grouping, each peptide is assigned to a single protein.

In protein cluster analysis mode, peptide weights are calculated based on how many proteins share the peptide and these weights are used in the ProteinProphet algorithm to more accurately model protein probabilities with shared peptides. When clustering is applied, the Similarity View reports the peptide weights used as percentages.

Comparisons to increase confidence in protein identification

To increase the confidence in protein identifications, Scaffold offers a number of instructive comparisons:

- Run X!Tandem, see [Validation with X!Tandem](#) or MSFragger, see [Searching with MSFragger](#), on data from another search engine. Peptides found to match with two different search engines are more likely to be valid than those peptides for which the two search engines disagree.
- Compare replicates of a biological sample to see if the same proteins are identified in each. The [Samples View](#) facilitates this with a direct, side-by-side view of all samples.
- Compare replicate proteins from the same spots on different gels to see if the same proteins are identified in each of them. The [Proteins View](#) allows sorting of spots according to the proteins they contain, the gel they came from, or other labels the user chooses.
- Compare the peptide patterns seen in each replicate. The sequence coverage shown in the [Proteins View](#) enables the user to determine at a glance whether the same peptides appear repeatedly in various samples.
- For each protein, compare the number of peptides identified. For each peptide, compare its scores from various search engines. The [Proteins View](#) lists the peptides and associated statistics.
- For peptide identifications of interest, examine the spectrum in the [Proteins View](#). Long ladders that match ion peaks with the amino acids in the peptide sequence strongly indicate valid results, whereas large numbers of unidentified peaks do not.
- For increased confidence in Scaffold's statistical analysis examine the [Statistics View](#) to insure that statistical assumptions are met.
- For information on combining multiple searches, see [Searle \(2008\)](#) and [improving-sensitivity-by-combining-MS-MS-results](#)

Scaffold Views

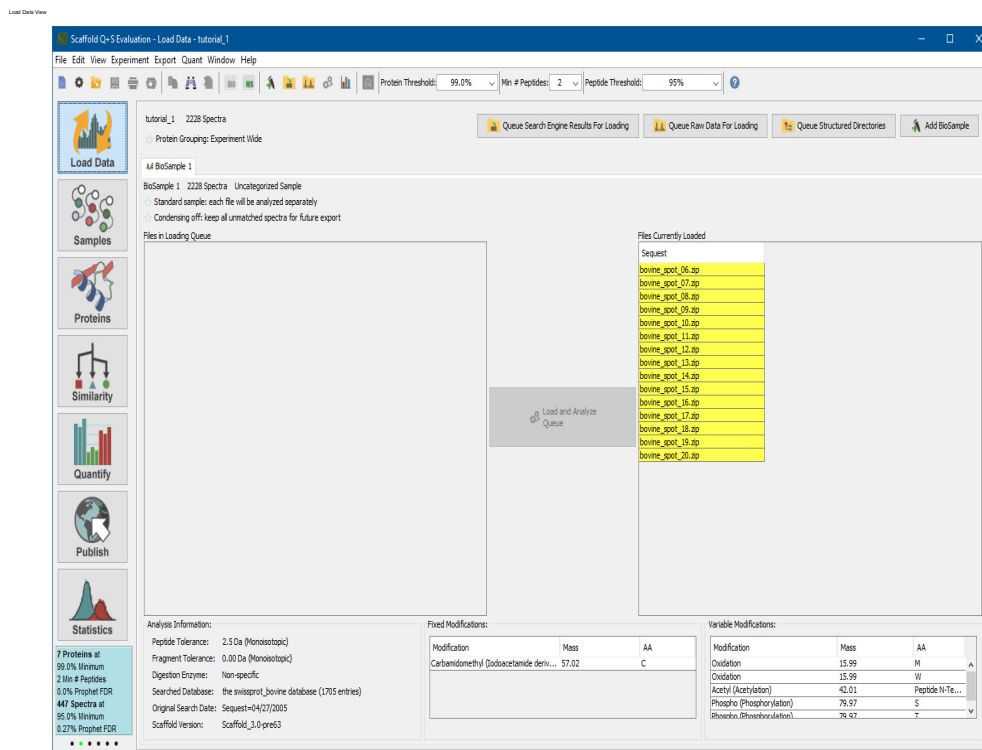
Scaffold offers both a high-level overview of the imported search results and a detailed look at supporting data, facilitating both top-down and bottom-up analysis. Scaffold presents the more detailed levels in a coherent structure, helping the user in verifying critical findings.

Scaffold Views:

- [Load Data View](#)
- [Samples View](#)
- [Proteins View](#)
- [Similarity View](#)
- [Quantify View](#)
- [Publish View](#)
- [Statistics View](#)

Load Data View

The [Load Data View](#) allows the user to load additional data, review the list of files loaded in each BioSample, edit BioSample's information, or delete already loaded MS Samples.



Samples View

The [Samples View](#) provides overviews that help the user make direct comparisons among categories of samples, BioSamples and MS samples. It lists and summarizes the proteins identified in each MS sample.

The list of proteins is shown summarized in two levels of hierarchy:

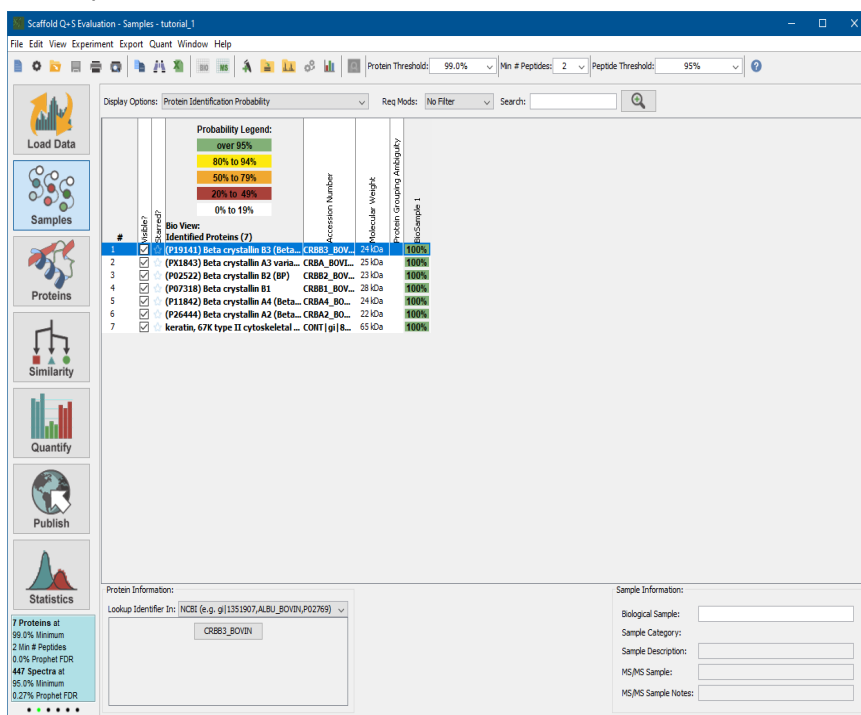
- **Protein groups** - groups of proteins that have been identified based on an identical set of peptides. They are shown collapsed and, by default, represented by the protein that has the highest probability and the largest associated number of spectra. The proteins included in the group are shown in the “[Protein Information pane](#)” on [page 154](#).
- **Protein Clusters** - sets of proteins or proteins groups created using a hierarchical clustering algorithm similar, but more stringent, to the Mascot's family clustering algorithm. Proteins or protein groups members of the cluster share some peptides but not all of them. They are by default represented by the protein that shows the highest associated probability. Clusters can be collapsed or expanded directly in the protein list.

For each protein or protein group that Scaffold identifies various Display Options are available providing different statistical information and counting options, see [Display Options](#).

For the highest-level overview, MS samples are grouped into BioSamples and results can be viewed collapsed in a single column summary for the entire group of MS samples, for further information see [MS Sample vs BioSample summarization levels](#).

To better focus on the most useful results, confidence thresholds allow setting minimum standards for identification probability, or for the other available Display Options. It is possible in this way to screen out less significant findings for a shorter, higher confidence list. Or by relaxing the thresholds it is possible to find less-confident identifications that might be more promising for further investigation, for more information go to [Sorting](#).

Figure 2-3: Samples View

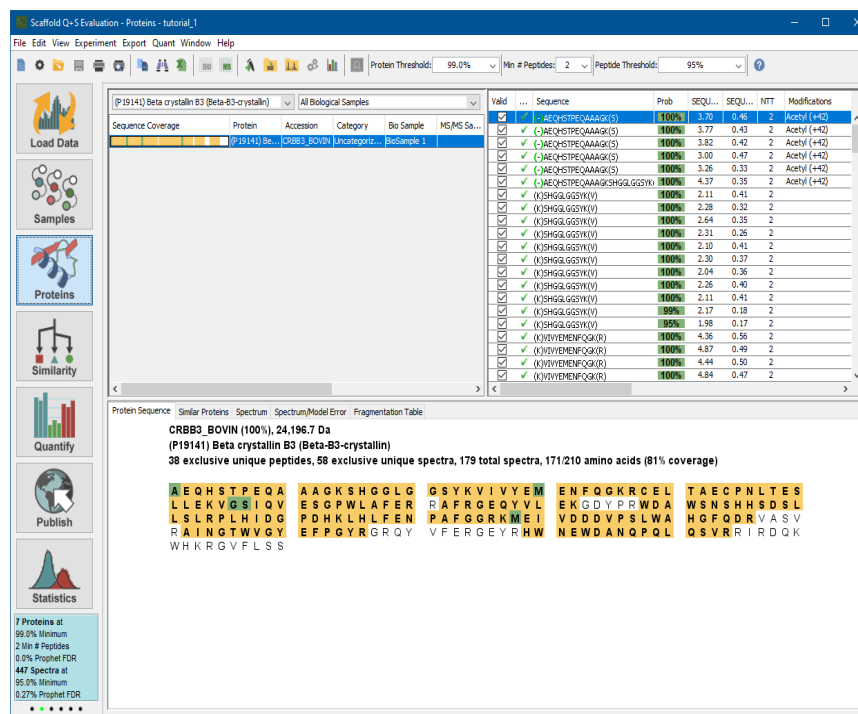


Proteins View

The **Proteins View** includes a large amount of detailed information about a protein organized in different tables and graphs:

- Sequence coverage for this and similar proteins;
- The peptide sequence, with identified peptides highlighted in yellow and modifications highlighted in green;
- The spectra used to identify each peptide, with associated error measurements;
- The fragmentation table listing the ion fragments along with their associated peaks.

Figure 2-4: Proteins View



For each peptide, the user can also see:

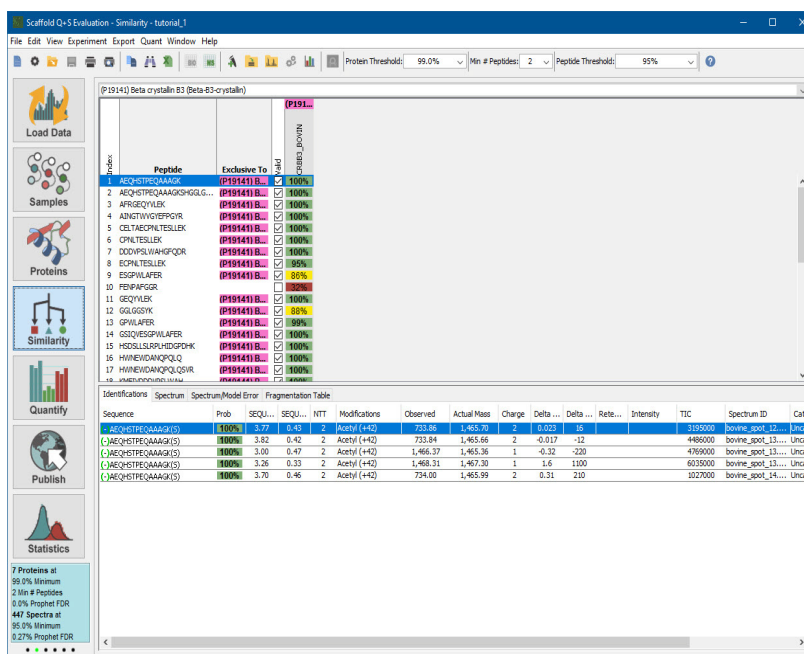
- *its charge, mass, and position in the peptide sequence;*
- *associated confidence scores from other search engines;*
- *modifications, if any.*

Having all of this information available at a glance or at a mouse click, provides confidence in the results.

Similarity View

The **Similarity View** allows to analyze in detail how the different identified peptides in a protein are shared with other protein groups.

Figure 2-5: Similarity View



For each peptide, the corresponding proteins to which it could belong are listed on the right.

- The user can “check” or “uncheck” the valid box for a peptide sequence. Unchecking the box removes that peptide from Scaffold's probability calculations.
- Peptides identified in particular protein groups are color coded to match their protein group

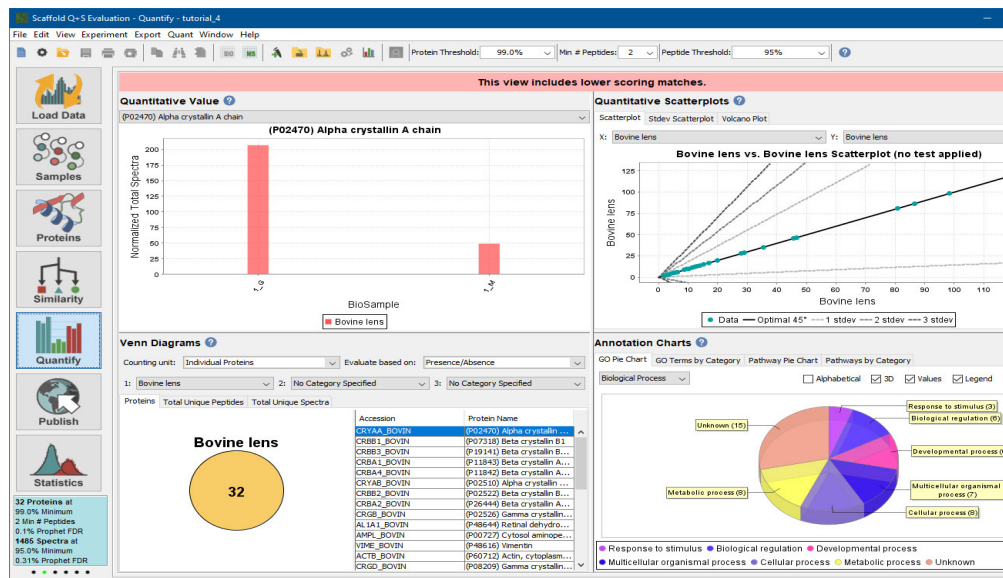
Quantify View

The **Quantify View** provides graphical tools to help the user visualize experiments and draw conclusions about the quantitative relationships demonstrated in the data. From the Quantify View, the user can compare quantitative values between samples and categories, analyze the biological functions of the proteins identified in the experiment, and assess the reliability of the statistical analysis of the data.

The information is organized in the following panes:

- **The Quantitative Value pane** - provides information about quantitative values of a specific protein and allows comparisons among biosamples and categories. A drop down list allows the user to choose which protein's quantitative values are displayed.
- **The Quantitative Scatterplots pane** - it includes tools to analyze differences and correlations among protein quantitative values in the different categories.

- Figure 2-6: Quantify View*

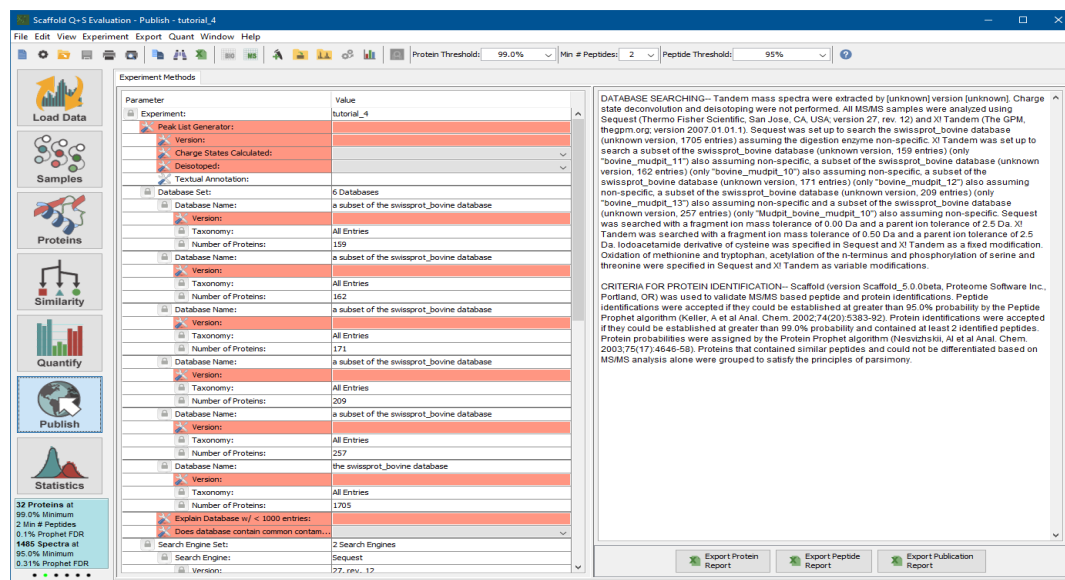


Publish View

The Publish View provides detailed information about the experiment in general.

- **The Experimental Methods tab** - which describes the parameters used when performing the experiment. The information listed is about experimental data typically required by the major proteomics journals such as:
 - Molecular & Cellular Proteomics,
 - Proteomics
 - Journal of Proteomics Research

Figure 2-7: Publish View

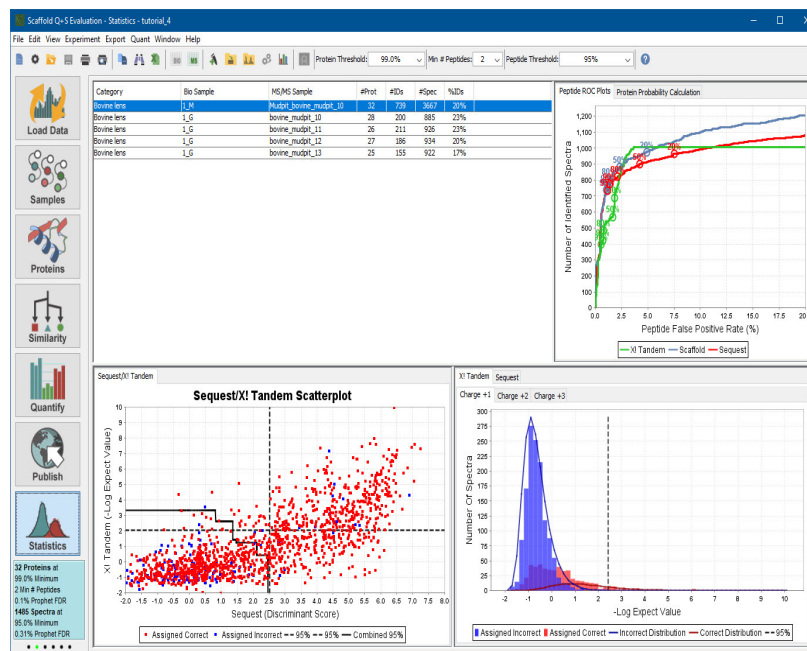


Using the mzIdentML exports, it is now possible to submit data analyzed through Scaffold to the PRIDE public data repository.

Statistics View

The [Statistics View](#) provides tools to assess the validity of peptides identified in every MS sample included in an experiment. It allows the user to check in details how the selected validation algorithm is applied to the loaded data.

Figure 2-8: Statistics View



The Statistics view provides a way to verify that the underlying assumptions of Scaffold are met. If data meets Scaffold's assumptions the user can have confidence in the analysis results.

Chapter 3

Loading Data in Scaffold

Scaffold can import and analyze data produced by a variety of search engines, and can also run MSFragger for full or supplementary searches and X! Tandem as a supplementary search. Results from different search engines can be combined in a single experiment and in a single BioSample if the samples were digested with the same enzyme, the raw files were processed with the same peak-picking software, and they were searched against the same database.



When multiple search engine results for the same sample are included in a BioSample, Scaffold recognizes this and groups the scores for each spectrum, combining the results to produce one MS sample.

Data files, whether raw files or search engine results, are loaded into Scaffold through the Scaffold Loading Wizard.

- **The Scaffold loading Wizard** can be started in one of the following ways:
 - Selecting **File > New** from the main menu.
 - Clicking the icon **New**
 - From the Load View by clicking the **Add BioSample** button

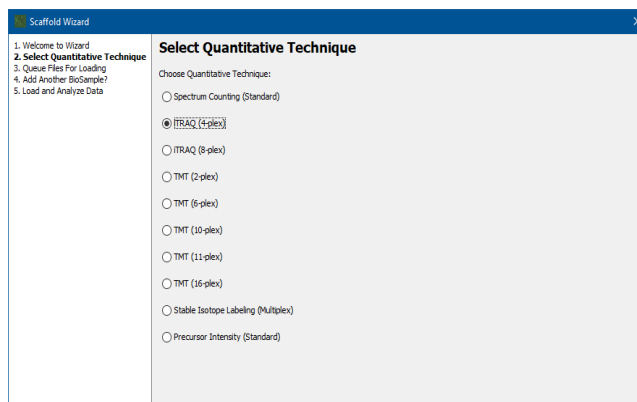
Once the Wizard is started, the Welcome to Wizard page opens. The user goes through a series of pages that guide the loading of the data.

Using The Scaffold Loading Wizard

- [Select Quantitative Technique](#)
- [New BioSample](#)
- [Queue files for loading](#)
- [Queue more files for loading](#)
- [Add another BioSample](#)
- [Load and Analyze Data](#)

Select Quantitative Technique

Figure 3-1: Scaffold Wizard, New Quantitative Technique page in Scaffold Q+S



The **Select Quantitative Technique** page appears only when running Scaffold Q+ or Scaffold Q+S. It offers a list of available quantitative techniques for analyzing the loaded samples. Spectrum Count is the only option available in standard Scaffold, although Precursor Intensity Quantitation is available through the Quantitative Value display option (see [Label Free Quantitative Methods](#)) if intensity values have been calculated by the search engine.

After data has been loaded, the user may change the specification of the quantitative method by selecting the menu option **Quant > Edit Quantitative Method/Purity Correction**.

When running standard Scaffold the following option is available:

- **Spectrum Counting** (Standard) - The sample will be treated exactly as in regular Scaffold.

When running Scaffold Q+ the following options are available:

- **iTRAQ(4-plex)** - Samples are treated as 4 multiplexed samples labeled with iTRAQ.
- **iTRAQ(8-plex)** - Samples are treated as 8 multiplexed samples labeled with iTRAQ.
- **TMT(2-plex)** - Samples are treated as 2 multiplexed samples labeled with TMT.
- **TMT(6-plex)** - Samples are treated as 6 multiplexed samples labeled with TMT.
- **TMT(10-plex)** - Samples are treated as 10 multiplexed samples labeled with TMT.
- **TMT(11-plex)** - Samples are treated as 11 multiplexed samples labeled with TMT.
- **TMT(16-plex)** - Samples are treated as 16 multiplexed samples labeled with TMT.
- **Precursor Intensity** (Standard) Label free quantitation using area under the MS1 peaks. Supported data files are listed here [Scaffold can import and analyze data produced by a variety of search engines, and can also run MSFragger for full or supplementary searches](#)

and X! Tandem as a supplementary search. Results from different search engines can be combined in a single experiment and in a single BioSample if the samples were digested with the same enzyme, the raw files were processed with the same peak-picking software, and they were searched against the same database.

When running Scaffold Q+S the following additional option is included as well all the choices listed above:

- **Stable Isotope Labeling** (multiplex)- Q+ S will load SILAC results from SILAC analysis performed with different types of software typically used in the scientific community. Q+S will load Dimethyl data analyzed using Mascot Distiller.
4. Click **Next**, the Scaffold Wizard, New BioSample page opens.

Figure 3-2: Scaffold Wizard, New BioSample page

New BioSample

The process of entering data in Scaffold begins with the definition of a BioSample. Scaffold uses the **BioSample** name in labeling and sorting analysis results. Group of BioSamples can be defined by specifying a **Category or Sample Category**

- **Sample Name** - Short descriptive name for the BioSample. Longer descriptions can go in the description field.
- **Sample Category** - Category names can be used to group BioSamples. Typical names might include control, diseased, treated; or might reflect different dates of a time course. Note: To see BioSamples listed in the same category, the category names must be

completely identical. The easiest way to accomplish this is by selecting the category name from the drop down list whenever available.

- **Description** - Include any longer description of the BioSample

In the lower part of the page specific loading options are available:

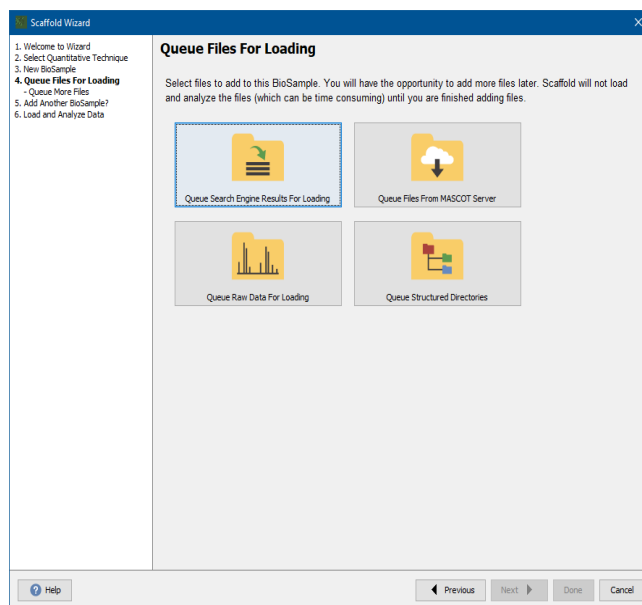
- **MuDPIT Experiment** - Check this box if samples were directly digested before separation. Scaffold then combines all the MS samples in the current BioSample

It is possible to select **Done** to exit the **Wizard** and define the MS samples later using the Queue Files for Loading button in Load Data View and will be asked for more information about the MS samples making up the BioSample at that time.

The usual path is to click **Next** to proceed to the selection of the MS samples to be loaded in this **BioSample**.

Queue files for loading

Figure 3-3: Scaffold Wizard, Queue Files for Loading page



- **Queue Search Engine Results** - opens a file browser to allow selection of search engine results files for loading into Scaffold.
- **Queue Files from Mascot Server** - Users with access to a licensed Mascot Server may link the server to Scaffold and queue Mascot search results directly from the Search Queue. To configure Scaffold to access a Mascot Server, use Edit>>Preferences>>Mascot Server. For instructions, see [Mascot Server](#).
- **Queue Raw Data for Loading** - opens a file browser for selection of raw files to be searched with MSFragger. If MSFragger is not the selected search engine a message

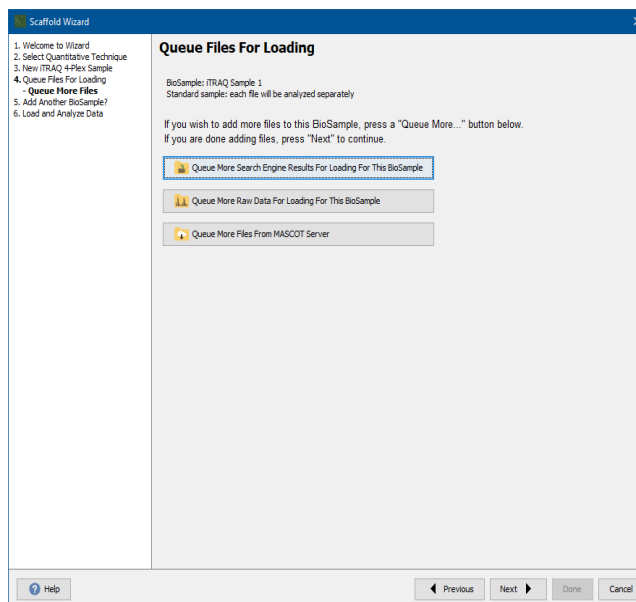
appears prompting the user to change the selection through **Edit>Preferences>Paths**. If MSFragger is properly configured and selected and raw files are queued, the Wizard will prompt for MSFragger search parameters at a later stage (see [Searching Raw Data with MSFragger](#))

- **Queue Structured Directories** - this method allows loading of multiple MS samples into multiple BioSamples at the same time. Samples are assigned to BioSamples based on the arrangement of the data files in a directory structure (see [Queue Structured Directory for Loading](#)).

Queue more files for loading

Once files have been queued through any of these paths, Scaffold offers the opportunity to queue additional files for the current BioSample. This makes it possible to add files from a different directory or through a different path. This dialogue comes up each time more files are added. When all files for the current BioSample have been queued, clicking Next causes the Wizard to move on.

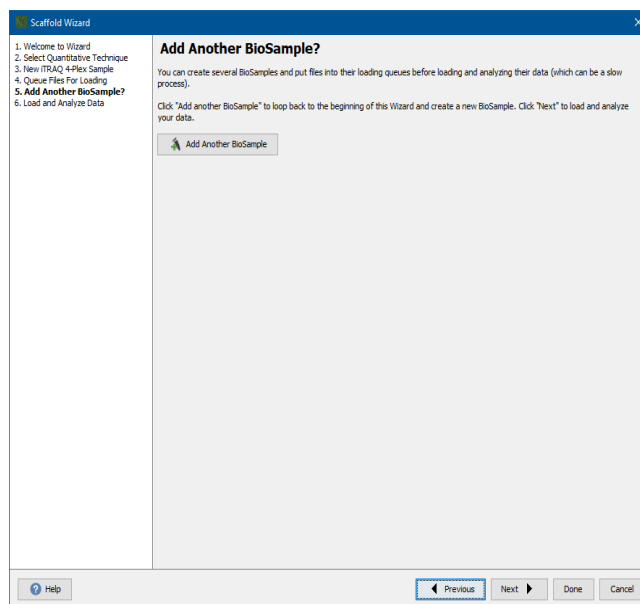
Figure 3-4: Scaffold Wizard, Queue More Files for Loading page



Add another BioSample

- If more BioSamples are included in an experiment the Add Another Biosample page allows the user to loop back to the beginning of the Wizard and add different files to a new BioSample. This operation can be repeated as many times as necessary to queue all of the BioSamples included in the experiment.

Figure 3-5: Scaffold Wizard, Add Another BioSample? page



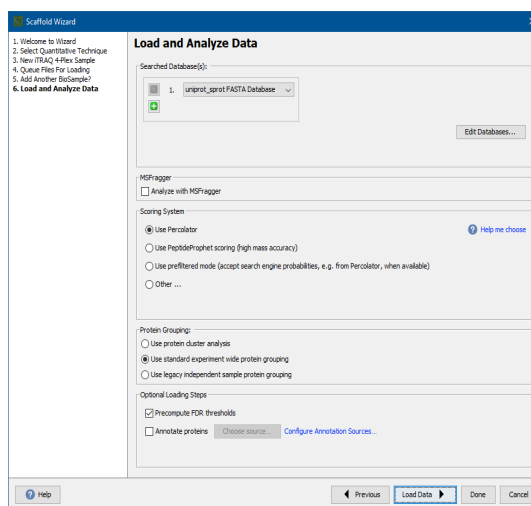
Clicking Next proceeds to the Load and Analyze Data page.

Load and Analyze Data

The Scaffold Wizard Load and Analyze Data page opens after files have been queued for all BioSamples. The page is divided into various panes. Each pane contains options for customizing the way Scaffold analyzes the data during the loading phase:

- [Searched Database Pane](#)
- [Analyze with MSFragger or X!Tandem Pane](#)
- [Scoring System Pane](#)
- [Protein Grouping Pane](#)
- [Optional Loading Steps](#)

Figure 3-6: Scaffold Wizard, Load and Analyze Data page



- Continue to [“Specifying analysis options and analyzing the data” on page 47.](#)

Specifying analysis options and analyzing the data

Searched Database Pane

This pane allows the user to select or import the FASTA database(s) used to create the data files loaded into Scaffold. The databases used in the last experiment appear in the list at the top of the pane. To change which database is selected on a line, click the pull-down to the right of the database name. If the data being loaded is the result of a search against more than one database, additional databases may be added by clicking on the green plus sign below the selected databases. To remove a database, click the red X beside its name. At least one database must be selected in order to load data. If the desired database does not appear in the dropdown list, it can be imported by choosing the **Set up new database...** option in the dropdown or by using the **Edit Databases....** button.

To add new databases the user can click the Add New Database button. For more detailed information continue to [“Specifying the FASTA database” on page 52..](#)

Analyze with MSFragger or X!Tandem Pane

When loading search engine results, selecting this option runs an additional database search on the loaded data with variable modifications chosen by the user. Depending on which search engine has been selected in Edit>Preferences>Paths, the search will either use MSFragger or X! Tandem. This operation improves protein identifications but significantly increases analysis times. Additional options for the supplementary search are set in the next dialog, see [“Validation with X!Tandem” on page 56.](#)

When loading raw files, the option to analyze with MSFragger is preselected and the next dialog allows the user to select parameters for the search.

Scoring System Pane

This pane offers different post processing scoring algorithms to apply when Scaffold analyzes the imported data:

- **Use Percolator** - Passes search engine scores and mass accuracy information to Percolator which computes posterior error probabilities which Scaffold converts to peptide probabilities.
- **Use PeptideProphet Scoring (high Mass Accuracy)** - This option will use the standard PeptideProphet algorithm with a high mass accuracy adjustment.
- **Use Prefiltered Mode** - For data that has already had a threshold or filter applied. In this mode, Scaffold does not calculate peptide or protein probabilities, and simply accepts the peptide identifications reported by the search engine. It can be used only if loading data from a single search engine as Scaffold cannot collate spectra without doing probability calculations. Scaffold's probability models rely on having the full distribution of PSM's, good and bad, and are unreliable if applied to prefiltered data, so this is the preferred option if an FDR or probability threshold has already been applied. Note that Scaffold still does protein inference and grouping, so protein identifications may differ from those reported by the search engine even in this mode.
- **Other** - Clicking this radio button brings up two additional options, which are included for historical compatibility:
 - **Use Legacy LFDR Scoring** - The LFDR algorithm has been deprecated beginning in Scaffold 5. Algorithm for assessing the confidence level of the identified peptides. Based on a Bayesian approach to LFDR (Local False Discovery Rate), this algorithm, was designed for QExactive and high mass accuracy data, see [LFDR-based scoring system](#).



When mzIdentML files are selected for loading, “Use Percolator or Use LFDR Scoring” are the only options available for selection. If the mzid files do not include decoys the files are automatically processed with PeptideProphet with delta mass correction.

When loading ProteinPilot AB Sciex mzid files, Scaffold directly utilizes the scores produced by Paragon, the search engine algorithm included in ProteinPilot, and it does not process the data through its probability models.

- **Use Legacy PeptideProphet Scoring (Standard)** - Standard PeptideProphet with no high mass accuracy.

For references and information about the scoring algorithms used in Scaffold see [Algorithm References](#).



When the data set to be analyzed was not searched using the decoy option or against a decoy concatenated database, the Legacy PeptideProphet Scoring high Mass accuracy option will be automatically selected.

Protein Grouping Pane

This pane shows the available grouping analysis options performed by Scaffold over the list of identified proteins.

- **Use protein cluster analysis** - Clustering provides an additional hierarchical grouping level above the standard Scaffold protein grouping. While similar to Mascot's hierarchical family clustering, Scaffold 4 clusters are created in a way that often succeeds in separating proteins into sets of biologically meaningful isoforms. Each cluster showed in the Samples View can be expanded or collapsed. The clusters sub-menu contains options for expanding or collapsing all of the protein clusters displayed in the Samples Table.



The initial default option is "Use Protein Cluster Analysis"

- **Use standard experiment wide protein grouping** - When selected, Scaffold groups proteins across all MS samples and BioSamples.
- **Use legacy independent sample protein grouping** - When selected, Scaffold groups proteins only within each MS sample. Each MS samples appears as if it was loaded independently.

For more information on the grouping and clustering algorithms used in Scaffold see ["Protein Grouping and Clustering" on page 203](#).

Optional Loading Steps

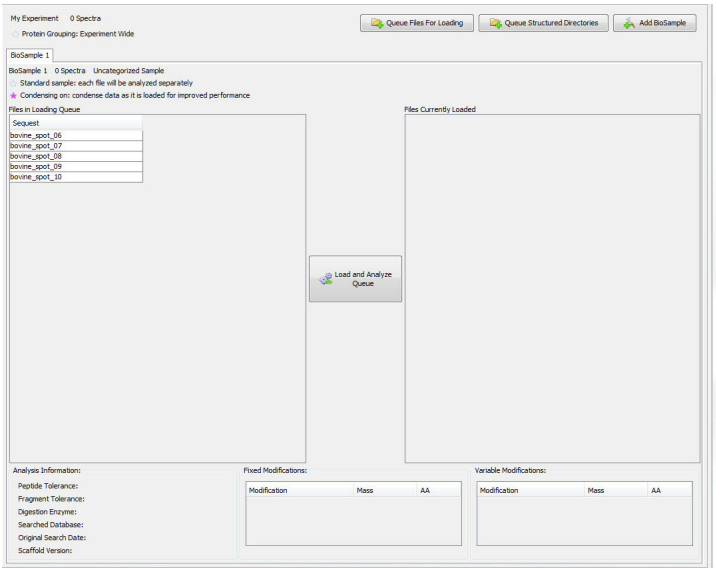
Included options for precomputing FDR thresholds and searching the Gene Ontology annotations, GO terms, during loading:

- **Do not precompute FDR thresholds** - if this box is checked, the FDR landscape will not be precomputed. This can result in faster loading for large experiments. If the user subsequently decides to apply an FDR threshold, Scaffold will compute the FDR landscape at that time.
- **Annotate Proteins**- Opens the Add Annotations dialog, see [Add or Edit Annotations](#). this allows the user to specify all necessary parameters so that annotations may be added during loading and displayed when the file opens. If the desired annotation source has not yet been configured, the user may click on the Configure Annotation Sources link to open the **Go Term Configuration** dialog.

Modify the make up of BioSamples

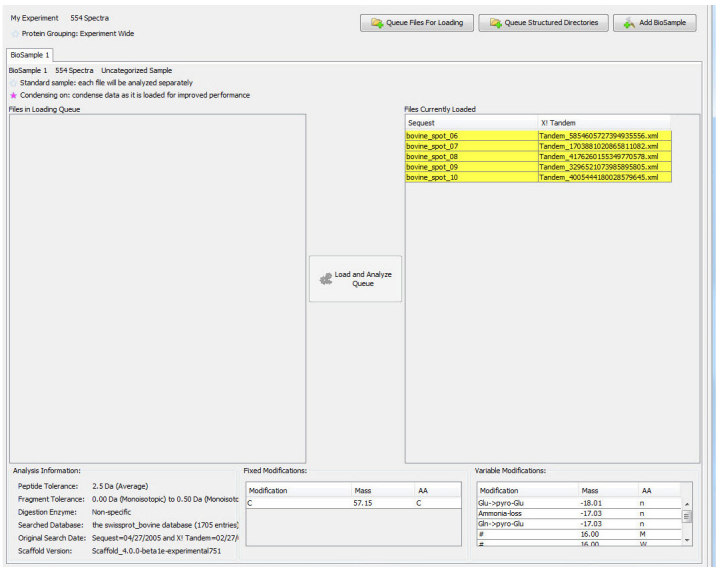
When the files have been loaded, they move from the import queue on the left side of the Load data View table to the ready pane on the right.

Figure 3-7: Load Data View Files in Import Queue



As they are analyzed, Scaffold highlights them in yellow and when all have been analyzed, Scaffold switches to the Samples view

Figure 3-8: Load Data View Files in Import Queue





At any time the user can remove files from either the Loading Queue or the Ready Pane.

If a file has been loaded by mistake, or to change the make-up of a BioSample:

1. Click the Load Data icon to go to the Load Data View
2. The files loaded and analyzed appear on the right. Click on one to select it.



Note that you select the entire MS sample, including runs from all search engines.

3. Click the right mouse button. A single menu item appears —Remove Selected Samples. Click it.
4. A second dialog asks you to confirm the removal. For now, click Don't Remove

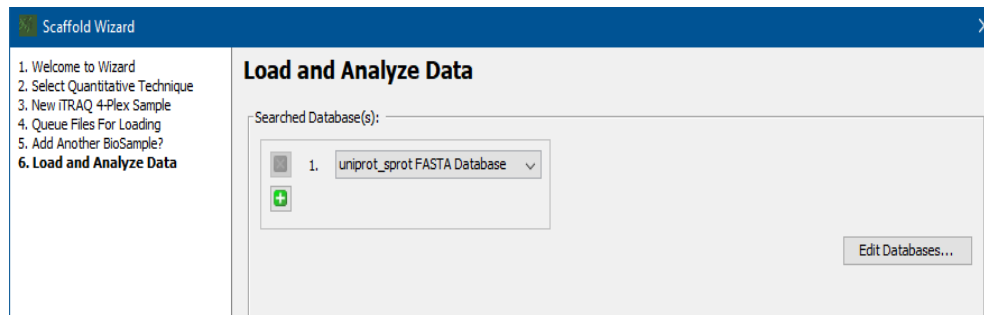
Save the Experiment

To Save the experiment:

1. Go to the File menu and select Save
2. Navigate to the folder in which you wish to save these tutorials, and enter the filename tutorial_3seq. Scaffold appends the suffix.SF3 to your experiment files.

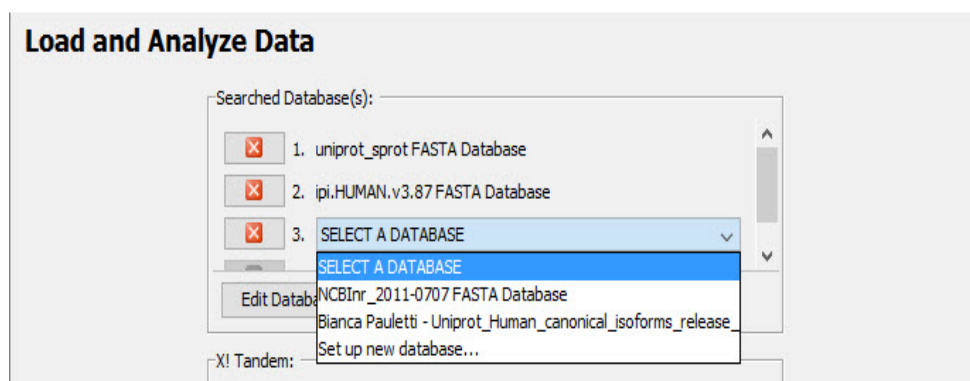
Specifying the FASTA database

Figure 3-9: Searched Database Pane



1. Specify the FASTA database(s) that is (are) associated with these sample files. You can select from a list of existing FASTA databases shown in the pull down menus, or you can add a new FASTA database.
 - Select a database from the existing list. Click + to add another.

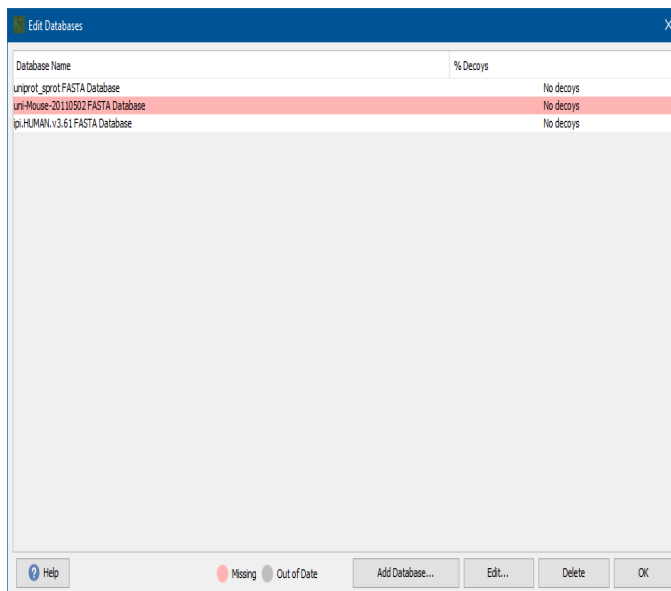
Figure 3-10: Searched Database Pane: pull down list



If the FASTA database selected is not identical to the external protein database, including the version, that you used for searching your experimental data, the protein sequences and molecular weights of some proteins might not be available later in the Samples and Proteins Views

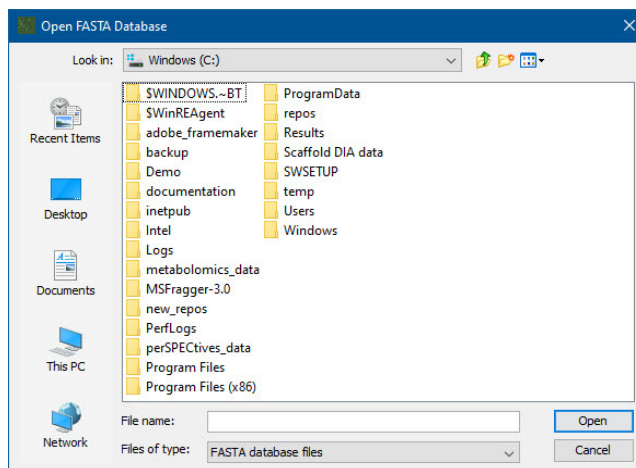
2. Click Set up New Database in the dropdown.

Figure 3-11: The Edit Databases dialog box opens. *Edit Databases dialog box*



3. On the Edit Databases dialog box, click New Database.
 - The Open FASTA Database dialog box opens.

Figure 3-12: Open FASTA Database dialog box



4. Navigate to the directory in which you saved your sample data set and FASTA database, select the FASTA database, and then click Open.



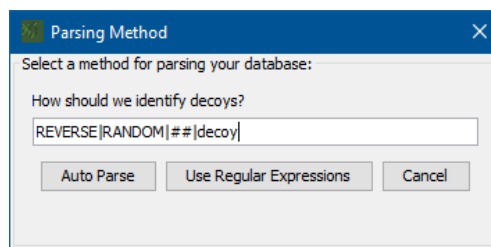
If you are carrying out this procedure using the sample tutorial_3seq data provided by Proteome Software, navigate to where you have saved the subset FASTA data base swissprot_bovine.fasta.



If you are carrying out this procedure using the sample tutorial_3mas data provided by Proteome Software, navigate to where you have saved the subset FASTA data base control_sprot.fasta.

The Parsing Method dialog box opens. You use the options on this dialog box to select the parsing rules that display protein accession numbers and protein descriptions in the correct format.

Figure 3-13: Parsing Method dialog box



5. Do one of the following:

- Click Auto Parse to have Scaffold decide on the parsing rules to use. If you are parsing a database that contains decoys make sure the decoy identification tag is included in the How should we identify Decoys? list shown in the dialog box.



Auto Parse is the preferred method for parsing the database. If you are carrying out this procedure using the sample tutorial_3seq data or the sample tutorial_3mas data provided by Proteome Software, then select Auto Parse.

- Click Use Regular Expressions to open the Configure Database Parser dialog box and select a specific pre-configured parsing rule to use, or create your own parsing rule. See [Figure 3-14 on page 55](#).

Figure 3-14: Configure Database Parser dialog box

Configure Database Parser

Parsed Accession Numbers and Protein Names:

Name Database Parser: uniprot_sprot FASTA Database

Reset File Location: F:\uniprot_sprot.fasta

Database Version:

Accession	Protein Description	Decoy	Alternate ID
Q6GZX4	Putative transcription factor 001R OS=Frog virus 3 (isolate Goorha) OX=654924 GN=...		FV3-001R
Q6GZX3	Uncharacterized protein 002L OS=Frog virus 3 (isolate Goorha) OX=654924 GN=F...		FV3-002L
Q197F8	Uncharacterized protein 002R OS=Invertebrate iridescent virus 3 OX=345201 GN=...		IV3-002R
Q197F7	Uncharacterized protein 003L OS=Invertebrate iridescent virus 3 OX=345201 GN=...		IV3-003L
Q6GZX2	Uncharacterized protein 3R OS=Frog virus 3 (isolate Goorha) OX=654924 GN=FV3...		FV3-003R
Q6GZX1	Uncharacterized protein 004R OS=Frog virus 3 (isolate Goorha) OX=654924 GN=F...		FV3-004R
Q197F5	Uncharacterized protein 005L OS=Invertebrate iridescent virus 3 OX=345201 GN=F...		IV3-005L
Q6GZX0	Uncharacterized protein 005R OS=Frog virus 3 (isolate Goorha) OX=654924 GN=F...		FV3-005R
Q91G88	Putative KIA-N domain-containing protein 006L OS=Invertebrate iridescent virus 6...		IV6-006L
Q6GZW9	Uncharacterized protein 006R OS=Frog virus 3 (isolate Goorha) OX=654924 GN=F...		FV3-006R
B2ZDY1	RING finger protein Z OS=Whitewater Arroyo mammarynavirus (isolate Rat/United ...		Z
A9JR22	RING finger protein Z OS=Tamiami mammarynavirus (isolate Rat/United States/W 1...		Z
Q88470	RING finger protein Z OS=Tacaribe virus (strain Franze-Fernandez) OX=928313 G...		Z
P08105	Putative uncharacterized protein Z OS=Ovis aries OX=9940 PE=4 SV=1		Z
Q6UY62	RING finger protein Z OS=Sabia mammarynavirus (isolate Human/Brazil/SPH114202...		Z

UniProt/Swiss-Prot Alt. Accession (UniProtKB)

Accession Number Parse Rule: >[?:(?:sp)|(?tr)]\{[^\}]*\}

Description Parse Rule: >[^\}]*\{[^\}]*\{[^\}]*\}

Decoy Protein Parse Rule: REVERSE[RANDOM]##[decoy]

Alternate ID Parse Rule: >.*GN=([^\}]+)

☒ Use Magic Matching

Help Export... Apply Cancel

- After the parsing rules are applied, the Edit Databases dialog box returns, with the correct database selected. Click OK.
- Continue to [“Load and Analyze Data”](#) on page 46.

Validation with X!Tandem

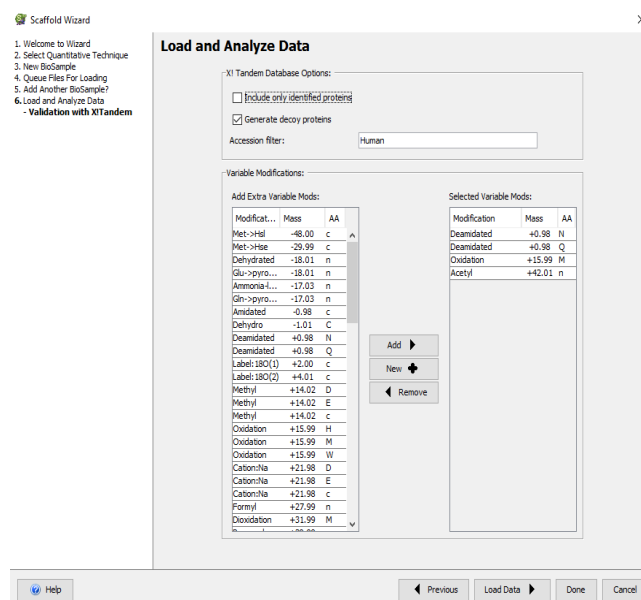
Each search engine uses its own algorithms to identify proteins. Identification confidence is higher when multiple algorithms find the same protein. Likewise, knowing which protein IDs are not confirmed by a second search engine lets the user screen out some false positives. This means that adding X! Tandem results to previous output files gives more confident protein identifications. It will, however, increase processing time.

X! Tandem runs quickly as compared to SEQUEST or Mascot, but the user should expect searching Swiss-Prot with several hundred thousand spectra to take considerable time and searching a huge database, such as the NR database, will take even much longer.

To include X! Tandem results as part of a Scaffold experiment, make sure that X! Tandem is linked and selected through Edit>Preferences>Paths, then check the box labeled Analyze with X!Tandem in the [Load and Analyze Data](#) page of the Loading Wizard. X! Tandem runs with the same parameters as the original loaded search files but variable modifications might be added to the ones already included.

1. The Scaffold Wizard, Load and Analyze - Validation with X!Tandem page opens

Figure 3-15: Scaffold Wizard, Load and Analyze - Validation with X!Tandem



The Wizard **Validation with X! Tandem** page includes two panes:

- **X! Tandem Options pane**

Since for large databases the X! Tandem search can take a long time the option **Include only identified proteins** was added to minimize its execution time. Checking this box means that X! Tandem searches only the subset of the proteins that were previously found with the original search engine.

Another option is to **Generate decoy proteins**. When this box is checked, Scaffold generates reversed sequences from each protein that is to be included in the database that is supplied to X! Tandem and includes the decoys in the search.

Finally, an **Accession filter** box allows the user to type in a character string which is used to filter the database. This allows the user to specify, e.g. a specific taxonomy to be searched, provided that some substring of the database entries identifies the taxonomy.

- **Variable Modifications pane**

From the input files Scaffold reads the parameters used to search a database by the search engine that produced the files. The parameters include instrument mass error tolerances, digestion enzymes, and fixed and variable modifications. Scaffold then passes these parameters to X! Tandem when it is run.

Of all the parameters passed onto the X! Tandem search the user can only modify the list of variable modifications. The variable modifications already present in the original search files readily appear listed in the **Selected Variable Mods** table located on the right side of this pane. The **Add Extra Variable Mods** table shows a list of standard UNIMOD variable modifications that can be added to the **Selected Variable Mods** table.

Between the two tables there are three functional buttons which allow to **Add** or **Remove** a variable mod from the **Selected Variable Mods** table or create a **New+** custom variable mod in the **Add Extra Variable Mods** table, see [Build A Modification](#).

Selecting more variable modifications may increase the number of peptides identified. It will surely increase the run time for X!Tandem's analysis. If many modifications are chosen, it will take many times longer to execute.

Note: When a peptide starts with E or Q, X!Tandem automatically checks for the formation of pyroglutamic acid, i.e., the loss of water or ammonia, respectively. This modification happens spontaneously in solution and failure to test for it can result in missing significant peptide hits. The analogous reaction for iodoacetimide blocked cysteine (loss of ammonia) is also considered. This modification is considered to be an N-terminal modification only, so it does not affect any potential modifications specified for Q, E or C. More information is available at <http://thegpm.org/TANDEM/api/rpmm.html>.

The modification tables can be sorted by clicking on the header for any column.

2. Add or remove modifications to the list already included in the original search data by using the arrows between the two lists.
3. Click Load data to start the analysis.
4. Continue to [“Modify the make up of BioSamples” on page 50](#).

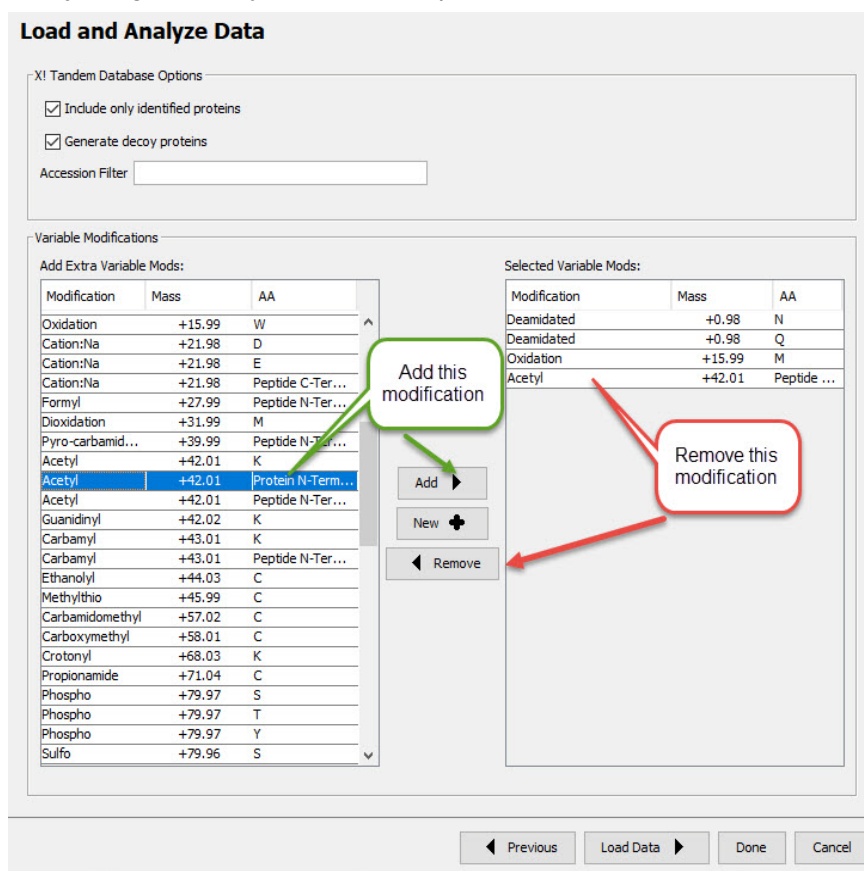
Protein N-Terminal Acetylation

Scaffold does not recognize protein n-terminal modifications when reading search results. Instead, these modifications are classified as peptide n-terminal. This is not really a problem,

as peptides exhibiting the modifications have already been identified by the search engine, which presumably applied the modification correctly. When the data is to be searched by X! Tandem, however, it is important to distinguish between protein n-terminal and peptide n-terminal modification.

When data files containing protein n-terminal acetylation are loaded into Scaffold and the user wishes to run X! Tandem, it is important that the variable modification list be adjusted. Scaffold will show Peptide n-terminal Acetyl as a selected variable modification. This should be removed from the selected list, and Protein n-terminal Acetyl, found in the list on the left, should be added instead. If this modification is included, Scaffold will set the X! Tandem “protein, quick acetyl” parameter to “yes”.

Figure 3-16: Adjusting the Acetyl modification specification:

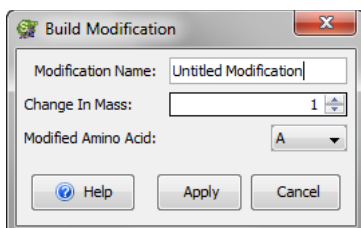


Build A Modification

If the user wants to search with X!Tandem using a variable modification not included in the **Add Extra Variable Mods** table he/she can define the new mod choosing **New+** on the Validation with X!Tandem Wizard page and bring up the Build Modification dialog.

Figure 3-17: Build A Modification dialog

- **Modification name** - Name that will be used in the Proteins View, Peptide pane and in the Spectrum Report for this modification. This name is saved with the Scaffold file



- **Change in Mass** - The mass difference in AMU due to this modification. Even though the modifications are only displayed as with one place after the decimal point on the modifications list, the mass is stored with the accuracy that was entered when defined.
- **Modified Amino Acid** - Pull down list with possible amino acid choices. Custom defined modifications can only apply to one amino acid. If the defined modification applies to several amino acids, the user has define several modifications accordingly.

Searching with MSFragger

MSFragger may be used to search raw data from within Scaffold or it may be used in conjunction with another search engine to identify additional peptides or to re-search the identified peptides in order to increase confidence in their correctness.

The user must first download MSFragger (see <https://www.proteomesoftware.com/msfragger>), install it and point Scaffold to the executable jar file. It is the responsibility of the user to ensure that all licensing requirements are observed. To link Scaffold to MSFragger, use the menu option Edit>Preferences>Paths. To use MSFragger, the corresponding radio button in that dialog must be selected.

Searching Raw Data with MSFragger

MSFragger can load Thermo .raw files, Bruker .d directories, and mzML files. Queue raw files using the Queue Raw Files option as described in [Queue files for loading](#). This causes the Analyze with MSFragger checkbox to be preselected in the Load and Analyze page of the Loading Wizard.

Figure 3-18: The Load and Analyze page with MSFragger preselected

Load and Analyze Data

Searched Database(s):

1. uniprot_sprot FASTA Database

Edit Databases...

MSFragger

☒ Analyze with MSFragger

Scoring System

☒ Use Percolator [Help me choose](#)

☐ Use PeptideProphet scoring (high mass accuracy)

☐ Use prefiltered mode (accept search engine probabilities, e.g. from Percolator, when available)

☐ Other ...

Protein Grouping:

☒ Use protein cluster analysis

☐ Use standard experiment wide protein grouping

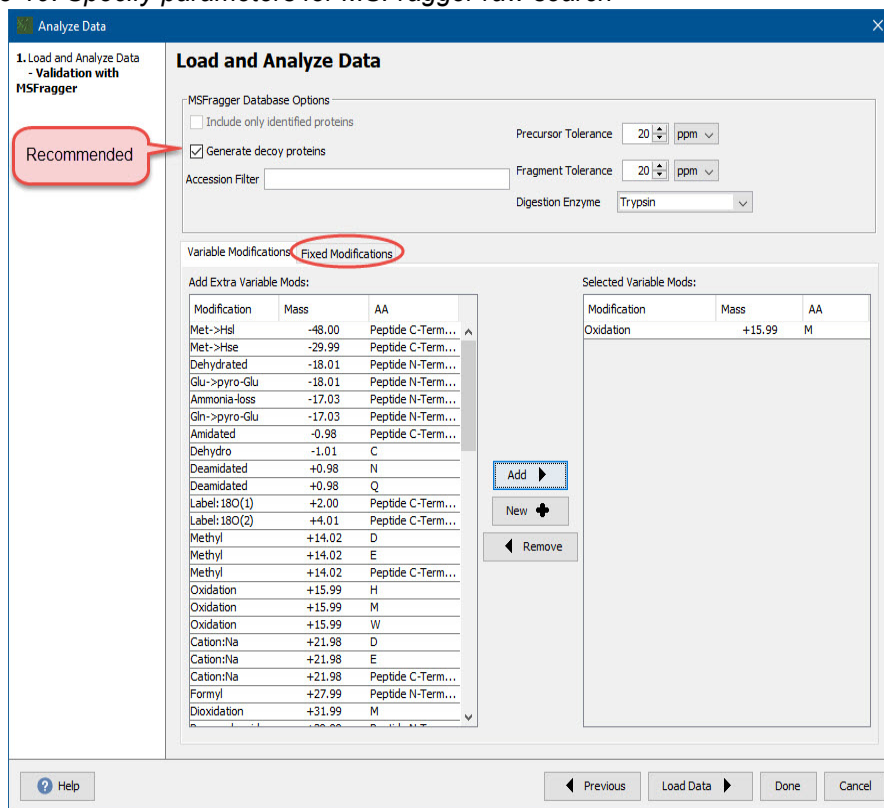
☐ Use legacy independent sample protein grouping

Optional Loading Steps

☒ Precompute FDR thresholds

Click Next and the Configure MSFragger dialog will open. Specify all requested search parameters. Be careful to specify fixed modifications through the fixed modifications tab as well as any desired variable modifications and proceed with the Wizard. It is not possible to do a confirmatory search within Scaffold when searching raw files with MSFragger.

Figure 3-19: Specify parameters for MSFragger raw search



It is recommended that Scaffold be allowed to create a decoy database for the search, as MSFragger has very strict requirements for the format of decoy entries in the FASTA and Scaffold will handle the necessary formatting. If the user wishes to provide a concatenated decoy FASTA, it is important to consult the MSFragger documentation to be sure that the decoy entries begin with an acceptable prefix and to adjust Scaffold's decoy parse rule for the FASTA (see [Specifying the FASTA database](#)).

Currently, Scaffold only supports spectral counting and TIC quantitation in MSFragger searches.

Supplementary Searches with MSFragger

To use MSFragger as a confirmatory search engine, make sure that it is linked and selected through Edit>Preferences>Paths, then check the box labeled Analyze with MSFragger in the [Load and Analyze Data](#) page of the Loading Wizard.

Figure 3-20: Select the option to do a second search with MSFragger

Load and Analyze Data

Searched Database(s):

1. uniprot_sprot FASTA Database

Edit Databases...

MSFragger

☒ Analyze with MSFragger

Scoring System

☒ Use Percolator

☐ Use PeptideProphet scoring (high mass accuracy)

☐ Use prefiltered mode (accept search engine probabilities, e.g. from Percolator, when available)

☐ Other ...

[Help me choose](#)

The Scaffold Wizard, Load and Analyze - Validation with MSFragger page opens

Figure 3-21: Scaffold Wizard, Load and Analyze - Validation with MSFragger

Scaffold Wizard

1. Welcome to Wizard
2. Select Quantitative Technique
3. New BioSample
4. Queue Files For Loading
5. Add Another BioSample?
6. Load and Analyze Data
- Validation with MSFragger

Load and Analyze Data

MSFragger Database Options

☒ Include only identified proteins

☒ Generate decoy proteins

Accession Filter

Variable Modifications

Fixed Modifications

Add Extra Variable Mods:

Modification	Mass	AA
Met->Hsl	-48.00	Peptide C-Ter...
Met->Hse	-29.99	Peptide C-Ter...
Dehydrated	-18.01	Peptide N-Ter...
Glu->pyro-Glu	-18.01	Peptide N-Ter...
Ammonia-loss	-17.03	Peptide N-Ter...
Gln->pyro-Glu	-17.03	Peptide N-Ter...
Amidated	-0.98	Peptide C-Ter...
Dehydro	-1.01	C
Deamidated	+0.98	N
Deamidated	+0.98	Q
Label:18O(1)	+2.00	Peptide C-Ter...
Label:18O(2)	+4.01	Peptide C-Ter...
Methyl	+14.02	D
Methyl	+14.02	E
Methyl	+14.02	Peptide C-Ter...
Oxidation	+15.99	H
Oxidation	+15.99	M
Oxidation	+15.99	W
Cation:Na	+21.98	D
Cation:Na	+21.98	E
Cation:Na	+21.98	Peptide C-Ter...
Formyl	+27.99	Peptide N-Ter...
Dioxidation	+31.99	M

Selected Variable Mods:

Modification	Mass	AA
Dehydrated	-18.01	Peptide N...
Ammonia-loss	-17.03	Peptide N...
Oxidation	+15.99	M
Carbamyl	+43.01	K
Carbamyl	+43.01	Peptide N...

Add

New

Remove

Help

Previous

Load Data

Done

Cancel

Chose whether to do a full new search or just research the identified peptides

Preset to be same as primary search

Recommended

The Wizard Validation with MSFragger page includes two panes:

- **MSFragger Options pane**

The option **Include only identified proteins** is offered to minimize execution time. Checking this box means that the FASTA used in the MSFragger search only contains the proteins that were previously identified by the original search engine.

Another option is to **Generate decoy proteins**. When this box is checked, Scaffold generates reversed sequences from each protein that is to be included in the FASTA database that is supplied to MSFragger. It is highly recommended that they user select this option.

Finally, an **Accession filter** box allows the user to type in a character string which is used to filter the database. This allows the user to specify, e.g. a specific taxonomy to be searched, provided that some substring of the database entries identifies the taxonomy.

Search parameters, such as parent and fragment tolerances, enzyme etc. are set to the values used in the original search.

- **Variable Modifications pane**

Scaffold reads the parameters used to search a database by the original search engine from the input files. The parameters include instrument mass error tolerances, digestion enzymes, and fixed and variable modifications. Scaffold then passes these parameters to X! Tandem when it is run.

The only parameter that can be adjusted for the MSFragger search is the list of variable modifications. The variable modifications already present in the original search files appear in the **Selected Variable Mods** table located on the right side of this pane. The **Add Extra Variable Mods** table shows a list of standard UNIMOD variable modifications that can be added to the **Selected Variable Mods** table.

Between the two tables there are three functional buttons which allow to **Add** or **Remove** a variable mod from the **Selected Variable Mods** table or create a **New+** custom variable mod in the **Add Extra Variable Mods** table, see [Build A Modification](#).

Selecting more variable modifications may increase the number of peptides identified. It will surely increase the run time for MSFragger's analysis.

The modification tables can be sorted by clicking on the header for any column.

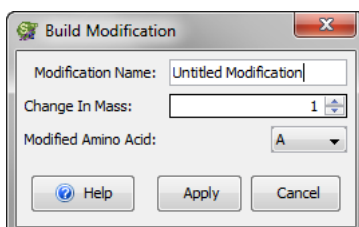
5. Add or remove modifications to be included in the search by using the arrows between the two lists.
6. Click Load data to start the analysis.

Build A Modification

To use a variable modification not included in the **Add Extra Variable Mods** table in the MSFragger search, define the new mod choosing **New+** on the Validation with MSFragger Wizard page and bring up the Build Modification dialog.

Figure 3-22: Build A Modification dialog

- **Modification name** - Name that will be used in the Proteins View, Peptide pane and in the Spectrum Report for this modification. This name is saved with the Scaffold file



- **Change in Mass** - The mass difference in AMU due to this modification. Even though the modifications are only displayed as with one place after the decimal point on the modifications list, the mass is stored with the accuracy that was entered when defined.
- **Modified Amino Acid** - Pull down list with possible amino acid choices. Custom defined modifications can only apply to one amino acid. If the defined modification applies to several amino acids, the user has define several modifications accordingly.

Chapter 4

Scaffold's Main Window

Scaffold's Main Window

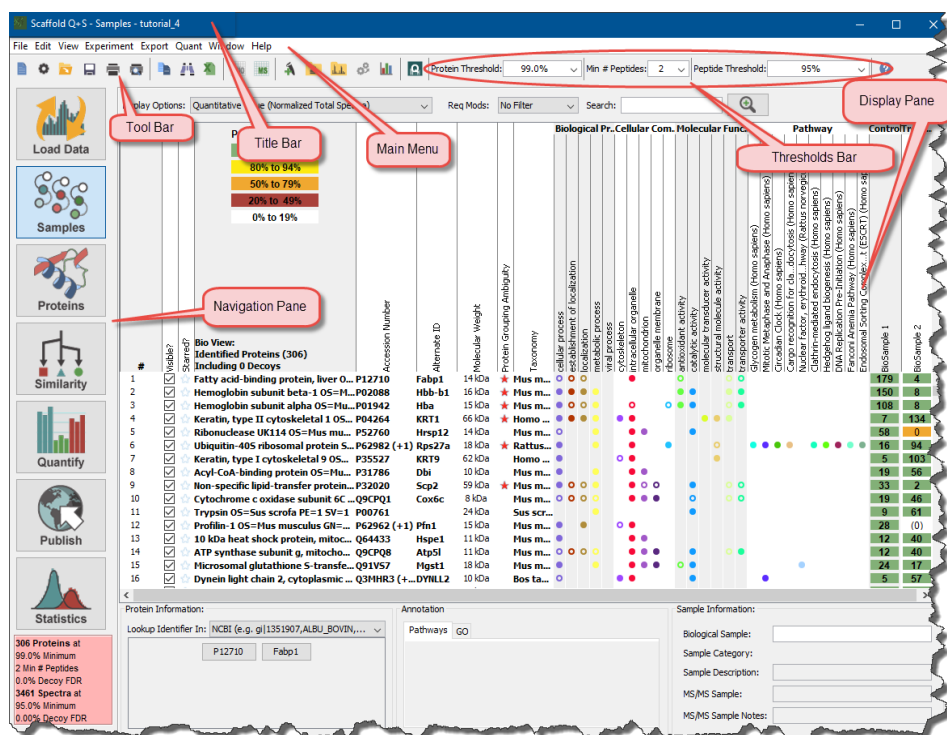
The Scaffold application is built around a main general window which includes a series of selectable views. Each view provides a particular perspective for looking at the loaded experiment. There are a number of tools common to all views and specific tools that help navigate within a selected view.

The following sections describe the major components of the Scaffold's main window:

- The [“Title bar” on page 68](#)
- The [“Main menu commands” on page 69](#)
- The [“Tool-bar” on page 112](#)
- The [“Thresholding pane” on page 113](#)
- The [“Navigation pane” on page 113](#)
- The [“Display pane” on page 117](#).
- [Title bar](#)
- [Main menu commands](#)
- [Tool-bar](#)
- [Thresholding pane](#)
- [Navigation pane](#)
- [Display pane](#)

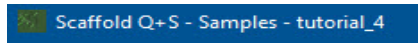
See also [The Samples View](#)

Figure 4-1: Scaffold window



Title bar

Figure 4-2: Title bar



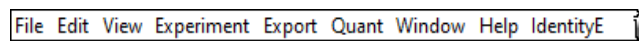
Depending on the type of license acquired either “Scaffold”, “Scaffold Q+” or “Scaffold Q+S” is always shown in the title bar at the top of the Scaffold window, together with the Scaffold icon. Additional text is displayed depending on the actions that the user is currently carrying out in Scaffold. For example, if the user has opened a file, then - <Experiment name> is also displayed in the title bar.



*The version of Scaffold in use is not displayed in the Title bar. The user must go the **Help > About** option in the main menu to determine the version number. See [“Main menu commands”](#) below.*



Main menu commands


















Figure 4-3: Scaffold's Main Menu

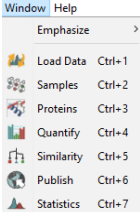
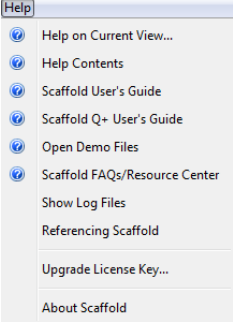
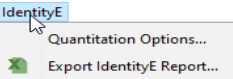


The Scaffold main menu is set up in a standard Windows menu format with menu commands grouped into menus (File, Edit, View, Experiment, Export, Quant, and Help) across the menu bar. When loading Waters IdentityE type of data an extra menu IdentityE appears after the Help menu. Some of these menu commands are available in other areas of the application as well.

Menu	Menu Commands
File File Edit View Experiment Export Qu New Ctrl+N Batch Job Queue... Ctrl+J Open Ctrl+O Merge... Ctrl+M Close Save Ctrl+S Save As... Save Condensed Data... Ctrl+T Print... Ctrl+P Print Preview... Exit	<ul style="list-style-type: none"> • New—Initializes a Wizard which guides the user through the loading phase of the search data files in Scaffold. See Loading Data in Scaffold. Scaffold can import and analyze data produced by a variety of search engines, and can also run MSFragger for full or supplementary searches and X! Tandem as a supplementary search. Results from different search engines can be combined in a single experiment and in a single BioSample if the samples were digested with the same enzyme, the raw files were processed with the same peak-picking software, and they were searched against the same database. • Batch Job Queue - opens a dialog that allows the user to set up and run Scaffold Batch experiments. • Open—Opens a saved Scaffold experiment file, *.SF3, through a file browser. • Merge—Merges *.SF3 files in one single Scaffold experiment. See Loading Data into ScaffoldMerge. • Close—Closes the current experiment, standard Windows behavior. • Save—Saves the current experiment, standard Windows behavior. • Save As—Saves the current experiment offering the option to use a different name, standard Windows behavior. • Save Condensed Data—Save Condensed Data • Print—Prints the current view. • Print Preview—Previews the current view with the option of printing the document. • Exit—Closes the Scaffold window.

Menu	Menu Commands
Edit Edit View Experiment Export Quant Wi  Copy Ctrl+C  Find Ctrl+F Edit FASTA Databases... Ctrl+D Edit Peptide Thresholds... Edit Annotation Options... Bulk Operations... > Preferences... Advanced Preferences...	<ul style="list-style-type: none"> • Copy—For each View copies to the clipboard the first table appearing at the top of the View. From there, the user can paste it into a third-party program such as Excel or Microsoft Word. • Find—Opens a find dialog box that searches the first table present in the Current View • Edit FASTA Database...—See Edit FASTA Databases. • Edit Peptide Threshold...—See Custom Peptide Thresholds • Edit Annotation Options...—See GO Annotations and Pathways in Scaffold • Bulk Operation...—Tag, show and hide proteins. Expand and collapse clusters. Options also available on right click. See Protein List. • Preferences...—See Preferences • Advanced Preferences...—See Advanced Preferences
View View Experiment Export Quant Window I <input checked="" type="checkbox"/> Navigation Pane Ctrl+O Switch Sample View > Switch Display Options > Show Entire Protein Clusters Show Lower Scoring Matches Show <5% Probabilities <input checked="" type="checkbox"/> Show Sample Notes Show Hidden Proteins <input checked="" type="checkbox"/> Show Alternate ID <input checked="" type="checkbox"/> Show GO Annotations Navigate >	<ul style="list-style-type: none"> • Navigation Pane—Toggles the view of the Navigation pane. • Switch Sample View—Selects level of summarization in Samples View, see MS Sample vs BioSample summarization levels • Switch Display Options—See Display Options • Show Entire Protein Clusters—See Clusters in the Samples Table • Show Lower Scoring Matches—See Show Lower Scoring Matches • Show <5% Probabilities—See Show <5% Probabilities • Show Sample Notes—Toggles the view of the Information Panes • Show Hidden Proteins—Toggles the view of Hidden Proteins • Show Alternate ID—Toggles the display of the Alternate ID column if Alternate IDs have been applied see . • Show GO Annotations—Toggles the view of the GO terms in the The Samples TableThe Samples Table, see Configure Annotation Sources • Navigate—Navigates through tabs when present in a dialog or pane.

Menu	Menu Commands
Experiment Experiment Export Quant Window Help Edit Experiment... Edit BioSample... Ctrl+E  Add BioSample... Ctrl+B  Delete Biological Sample  Queue Search Engine Results For Loading Ctrl+R  Queue Raw Data For Loading  Queue Files From MASCOT Server... Ctrl+L  Queue Structured Directories For Loading... Apply New Database... Extract Alternate IDs... Apply Protein Annotation Preferences...  Load and Analyze Queue Ctrl+A Reset Peptide Validation... Add or Edit Annotations...  Quantitative Analysis...	<ul style="list-style-type: none"> • Edit Experiment—See • Edit BioSample—See Edit BioSample • Add BioSample—Initializes The Loading Wizard, see Using The Scaffold Loading Wizard • Delete BioSample—Deletes a loaded biosample from the current Scaffold experiment. Particularly useful in the Load Data view. • QueueSearch Engine Results for Loading—See Queue Search Engine Results for Loading. • Queue Raw Data for Loading - See Queue Raw Data for Loading • Queue Files From Mascot Server For Loading—See Queue Files From Mascot Server for Loading.... • Queue Structured Directories For Loading—See Queue Structured Directory for Loading. • Apply New Database—See Apply New Database • Apply Protein Annotation Preferences—See Apply Protein Annotation Preferences • Load and Analyze Queue—Available only when there are files listed in the loading Queue in the Load Data ViewThe Load Data View waiting to be loaded in Scaffold. When selected it opens the Load and Analyze Data page of the Loading Wizard. • Reset Peptide Validation—See Reset Peptide Validation • Add or Edit Annotations—Applies imported annotations to the Samples Table.To import GO databases see GO Annotations and Pathways in Scaffold • Quantitative Analysis— See Quantitative Analysis...
Export Export Quant Window Help Subset Database... Spectra... BLIB... ProtXML... mzIdentML... SFDB... Scaffold Batch... Scaffold Batch Archive...  Publication Report...  Samples Report...  Spectrum Reports >  Peptide Reports >  Protein Reports >  PSEA-Quant Report...  Current View...  Complete...	<ul style="list-style-type: none"> • Subset DATABASE— See Subset Database • Spectra— See Spectra • ProtXML— See ProtXML report • mzIdentML — See mzIdentML • Scaffold Batch— See Scaffold Batch... • Scaffold Batch Archive— See Scaffold Batch Archive... • Export To Excel: <ul style="list-style-type: none"> • Publication Report— See Publication report • Samples Report—Generates a tab-delimited Samples table appearing in the Samples View, see Samples report • Spectrum Reports— See Spectrum report • Peptide Reports—Generates a tab-delimited Peptide table for all proteins appearing in the Samples View, see Peptide report • Protein Reports—Opens the SQL dialog box see Protein report • Current View— See Current View report • Complete— See Complete report
Quant  Launch Q+ Quantitation Module... Edit Quantitative Method/Purity Correction...	<ul style="list-style-type: none"> • Launch Q+ Quantitation Browser— When using Scaffold Q+ or Scaffold Q+S this command is available for switching to the Q+/Q+S quantitation window. • Edit Quantitative method/purity Correction— See Edit Quantitative Samples

Menu	Menu Commands
<p>Window</p> 	<ul style="list-style-type: none"> • Emphasize - In Views which contain multiple panes, such as the Proteins or Quantify View, this function allows the user to select a subset of panes to display in larger format. • Switch to a different Scaffold view, which is equivalent to clicking the buttons located in the Navigation pane
<p>Help</p> 	<ul style="list-style-type: none"> • Help on Current View...—Opens the Online Help that is specific for the currently displayed topic. • Help Contents—Opens the Contents page for the Online Help. • Scaffold User's Guide—Opens the current Scaffold User's Guide. • Scaffold Q+ User's Guide—Opens the current Scaffold Q+S User's Guide. • Open Demo Files—Opens the folder where Scaffold demo files are stored. The user can choose any of the pre-loaded files to test Scaffold capabilities. • Scaffold FAQs/Resource Center—Opens the user's default web browser to the Home page of the Proteome Software resource center. • Show Log Files—Opens a folder containing Scaffold error_log and output_log files • Referencing Scaffold— See Referencing Scaffold • Upgrade License Key - opens a dialog through which a new license key may be entered if the user has chosen to upgrade to a key with more functionality. • About Scaffold—Provides the release information for the current version of Scaffold, license information, contact information for Proteome Software, Inc.. It also reports information about the system where Scaffold is installed, the amount of memory available to the software and the percentage of memory used by the application.
<p>IdentityE</p> 	<ul style="list-style-type: none"> • Quantitation Options—Quantitation Option • Export IdentityE report—Generates a tab-delimited report containing the list of peptides used to calculate the intensities assigned to each protein in the list of identified proteins shown in the Samples view.

File menu

The command **File > Merge** allows the user to combine different Scaffold experiments into one single *.SF3 file. It is active only when an existing Scaffold experiment has already been created or opened. Selecting this command calls the **Import Scaffold File** file chooser for selection of a *.SF3 file to be merged with the current opened Scaffold experiment. Once a file is selected the [Queue Scaffold Files for Merging](#) window opens allowing the user to add more files to the list of *.SF3 files to be merged.

When Scaffold merges the experiments in the list, the BioSamples in each of them are treated as separate samples. If they have duplicate names, numbers are appended to the original names to distinguish them.

Queue Scaffold Files for Merging

After the first file to be merged is selected, the dialog **Queue Scaffold Files for Merging** appears to allow additional files to be added to the list to be merged together into the same experiment.

The button **Add More Files** opens a file chooser which allows the user to locate, select and add files to the list. The **Merge** button merges the list of files to the original Scaffold experiment creating a new BioSample for each BioSample in each file in the list.

Caution: It is not possible to delete a specific file from this list.

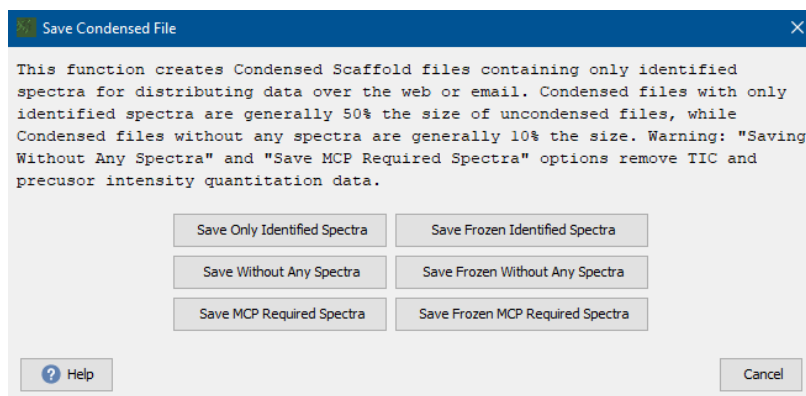
Once the files are merged the Delete Biological Samples operation can be used to delete any undesired BioSample.

Save Condensed Data

The **File > Save Condensed Data** menu option reduces the size of the *.sf3 file Scaffold saves.

Caution: Executing the Save Condensed Data changes the data in the running copy of Scaffold. Once the user saves condensed there is no undo that will restore all the data.

Figure 4-4: Save condensed dialog



There are six options for condensing data while it is saved:

- **Save Only Identified Spectra** [Save Only Identified Spectra](#)—This option saves all the data that can be seen in the Scaffold Viewer. It does not save the spectra that were not matched to peptides. Saving with this option generally cuts the size of the *.sf3 file in half.
- **Save Frozen Only Identified Spectra** [Save Frozen Only Identified Spectra](#)— This command condenses the saved output file just like the **Save Only Identified Spectra** option does except it also freezes the data in the files.
- **Save Without Any Spectra** [Save Without Any Spectra](#)—This option saves all the peptides and their scores but does not save any of the spectra.
- **Save Frozen Without Any Spectra** [Save Frozen Without Any Spectra](#)—This command condenses the saved output file just like the **Save Without Any Spectra** option does except it also freezes the data in the file.

- **Save MCP Required Spectra** [Save MCP Required Spectra](#)— This saves only those spectra required by the proteomics journal.
- **Save Frozen MCP Required Spectra** [Save Frozen only MCP Required Spectra](#)—This command condenses the saved output file just like the **Save MCP Required Spectra** option does except it also freezes the data in the file.

Since the spectra are 90% of the bulk of the data, an *.SF3 file saved without spectra will be reduced to only about 10% of the size of the uncondensed file.

Save Only Identified Spectra

This option helps reduce the size of the Scaffold *.sf3 file without losing any protein/peptide information, or any of the spectra used to validate the peptide identifications. If the data was loaded without the condense option, then this option generally reduces the size of the *.sf3 file by about 50%. Smaller files are easier to share with colleagues.

If the data was loaded with the condense option, the data size of the *.sf3 is slightly but not dramatically smaller. The **Condense On Load** routine makes a guess about which spectra will match peptides before the data is really loaded. This guess has to be conservative so that no good spectra is lost.

The option Save Only Identified Spectra can be selected only after Scaffold knows which spectra will be matched to peptides so it knows which spectra to save and which will never be seen in Scaffold. The save option saves all the spectra that are matched to any peptide with a probability greater than 5% and a protein probability greater than 20%. When this save option is run there is no way to undo the operation.

Save Frozen Only Identified Spectra

This command may be invoked from:

- **File > Save Condensed Data**
- MCP Submission tab in step 3 Package for Submission on the **Publish View** where it is labeled as **Save Only Identified Spectra**.

This command condenses the saved output file just like the **Save Only Identified Spectra** option does except it also freezes the data in the file.

An *.sf3 file that is frozen can be opened in Scaffold or the Scaffold Viewer, but it can't be changed. This means that you can be sure that people looking at the file are seeing it just as you intended. They can't hide proteins or peptides or change thresholds.

Save Without Any Spectra

Many people, particularly people who have their samples analyzed by a core lab, do not ever look at the spectra to confirm the protein identifications. For these people, the size of the Scaffold *.sf3 file is significantly smaller if the file is saved using the menu option **File > Save Condensed Data** followed by the **Save Without Any Spectra** button.

Caution: There is no undo. If the user wants to view the spectra again without reloading all the data into Scaffold, then he /she needs to save the file without the condense option first.

*.sf3 files without the spectra are roughly 10% of the size of the *.sf3 file with the spectra. This save option does not change the protein or peptide identifications. It does however prevent the user from exporting the spectra with the menu option **Export > Spectra**.

Save Frozen Without Any Spectra

This command condenses the saved output file just like the **Save Without Any Spectra** option does except it also freezes the data in the file.

This will significantly reduce the size of the Scaffold *.sf3 file, but only spectra matching modified peptides and spectra belonging to proteins identified by only one or two spectra will be visible in the Spectrum tab of the Proteins Page.

A *.sf3 file that is frozen can be opened in Scaffold or the Scaffold Viewer, but it can't be changed. This means that the creator of the file can be sure that the people looking at the file are seeing it just as he/she intended. Viewers can't hide proteins or peptides or change thresholds.

Save MCP Required Spectra

This command saves only those spectra required by the Proteomics Journal, which are:

- Spectra for all proteins identified on the basis of ONE OR TWO exclusive unique peptides
- Spectra for ALL modified peptides

All these spectra must be submitted to a public repository

The MCP Submission tab has a checklist with the different requirements for properly submit data and information for publication. The third item in the list, Package for submission, has a button **Save Only MCP Required Spectra**. Since submissions to the journals are **Frozen**, using this option produces a file that can't be changed.

Save Frozen only MCP Required Spectra

This command may be invoked from **File > Save Condensed** or from the MCP tab third item of the list of requirements **Package for Submission** button **Save Only MCP Required Spectra**.

This command condenses the saved output file just like the **Save Only MCP Required Spectra** option does except it also freezes the data in the file.

This will significantly reduce the size of the Scaffold *.sf3 file, but only spectra matching modified peptides and spectra belonging to proteins identified by only one or two spectra will be visible in the Spectrum tab of the Proteins Page.

A *.sf3 file that is frozen can be opened in Scaffold or in the Scaffold Viewer, but it cannot be changed. This means that the creator of the file can be sure that the people looking at the file are seeing it just as he/she intended. Viewers cannot hide proteins or peptides or change thresholds.

FASTA databases in Scaffold

When loading data in Scaffold it is preferable to import the same database used for protein identification by the search engine. Scaffold derives the sequence information from the fasta database provided by the user. If the database is not exactly the same one used for the search it might not contain all the identification information needed by Scaffold to derive the protein molecular weight and sequence shown in the Samples and Proteins view respectively.

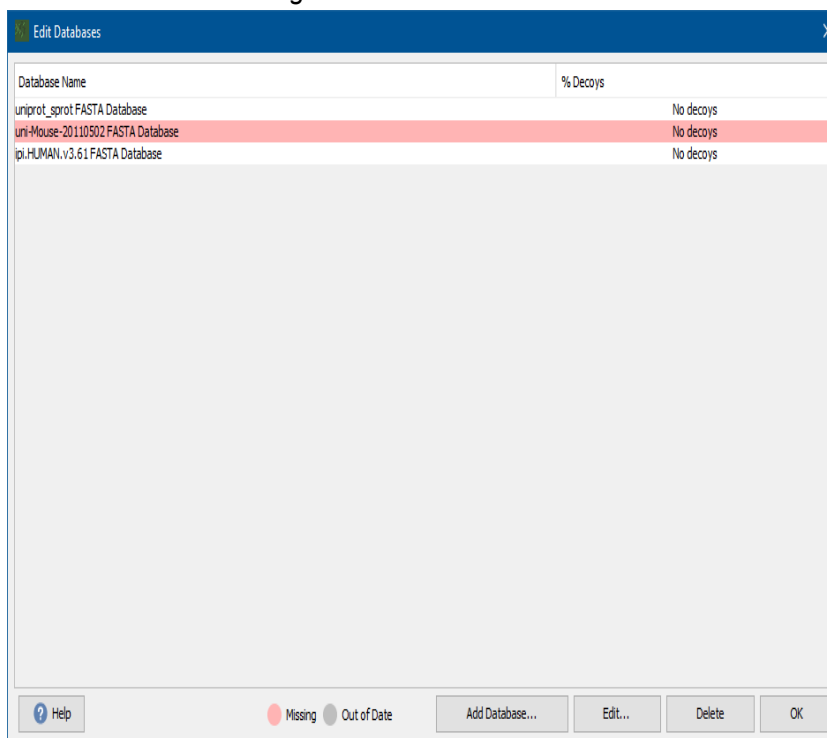
Edit

- **Copy**—For each View copies to the clipboard the first table appearing at the top of the View. From there, the user can paste it into a third-party program such as Excel or Microsoft Word.
- **Find**—Opens a find dialog box that searches the first table present in the Current View
- **Edit FASTA Database...**—See [Edit FASTA Databases](#).
- **Edit Peptide Threshold...**—See [Custom Peptide Thresholds](#)
- **Edit Go Terms Options...**—See [Configure Annotation Sources](#)
- **Bulk Operation...**—See [Bulk Operations...](#)
- **Preferences...**—See [Preferences](#)
- **Advanced Preferences...**—See [Advanced Preferences](#)

Edit FASTA Databases

To add and parse databases the user needs to open the **Edit Databases** dialog either selecting the menu option **Edit > Edit FASTA Databases** or clicking the button **Add New Database** located in the Search Database pane in the [Load and Analyze Data](#) page in the Scaffold loading Wizard. The selection opens the **Edit Databases** dialog which contains a table listing the databases already available in Scaffold and a number of functional buttons.

Figure 4-5: Edit Database dialog



All data loaded in the same experiment should have been searched against the same database or same set of databases

- **Loaded Databases table** - The table lists all the databases already available in Scaffold with information about the percent of decoys included in each of them. A database in the list might appear highlighted in various colors as a warning.
 - Pink highlight: *Missing database*. Scaffold cannot connect with the database using the current stored information.
 - Grey highlight: *Out of date database*. The database indexing is not updated with the Scaffold version in use.

When selecting a highlighted database the button **FIX** appears. Depending on the issue, clicking fix either calls the parsing method directly to build or rebuild the related index file or asks for a location where to find the database if it was moved.

- **Add Database button** - The button adds a new database to Scaffold. When selected a file browser appears allowing the user to point Scaffold to the location where the database he/she wants to load is stored. Once the FASTA file is selected the [Parsing Method Dialog](#) appears.
- **Edit button** - To edit one of the existing databases the user has to select a name from the Loaded Databases table and click **Edit**. The [Parsing Method Dialog](#) appears.

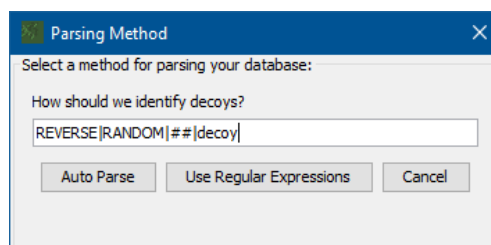
- **Delete button** - To delete one of the existing databases the user has to select a name from the Loaded Databases table and click **Delete**.

After parsing rules are applied or databases are deleted the user can click **OK**.

See also [FASTA databases in Scaffold](#)

Parsing Method Dialog

Figure 4-6: Database Parsing methods dialog



This dialog appears when a new database is added or when an already existing database is reedited. It allows the user to select one of the two parsing methods Scaffold uses to align protein names and accession numbers:

- **Auto Parse**— This option provides an automatic way of searching for the optimal accession numbers between the database and the type of data loaded into Scaffold. It initially identifies the type of parsing rule that better fits both the data and the selected database. It then matches the rule protein by protein while ensuring uniqueness. If a protein does not include the type of rule initially selected, **Auto Parse** looks for other rules more compatible with the specific protein accession number and defaults to a more general accession number if everything fails.
- **Use Regular Expressions**— This option opens the [Configure Database Parser](#) window which allows the verification of the protein names and accession numbers alignment for Scaffold's indexing and parsing. It also gives the possibility to modify the parsing rules according to the user's needs.
- **How should we identify Decoys?**— This box contains typical tags used to label decoy proteins in a database. When parsing a database that contains decoys the user should make sure that the decoy identification tag used in the his/her database is included in the list.



When loading FragPipe data or when searching MSFragger data with a reverse-concatenated FASTA (rather than allowing Scaffold to generate decoys), it is necessary to change the Decoy parse rule to recognize a prefix identifying the decoys. Generally, this is: “^REV_”.

Configure Database Parser

Figure 4-7: Configure Database Parser dialog

Configure Database Parser

Parsed Accession Numbers and Protein Names:

Name Database Parser:

Reset File Location:

Database Version:

Accession	Protein Description	Decoy	Alternate ID
Q6GZX4	Putative transcription factor 001R OS=Frog virus 3 (isolate Goorha) OX=654924 G...		FV3-001R
Q6GZX3	Uncharacterized protein 002L OS=Frog virus 3 (isolate Goorha) OX=654924 GN=F...		FV3-002L
Q197F8	Uncharacterized protein 002R OS=Invertebrate iridescent virus 3 OX=345201 GN=...		IIV3-002R
Q197F7	Uncharacterized protein 003L OS=Invertebrate iridescent virus 3 OX=345201 GN=...		IIV3-003L
Q6GZX2	Uncharacterized protein 3R OS=Frog virus 3 (isolate Goorha) OX=654924 GN=FV3...		FV3-003R
Q6GZX1	Uncharacterized protein 004R OS=Frog virus 3 (isolate Goorha) OX=654924 GN=F...		FV3-004R
Q197F5	Uncharacterized protein 005L OS=Invertebrate iridescent virus 3 OX=345201 GN=...		IIV3-005L
Q6GZX0	Uncharacterized protein 005R OS=Frog virus 3 (isolate Goorha) OX=654924 GN=F...		FV3-005R
Q91G88	Putative K1A-N domain-containing protein 006L OS=Invertebrate iridescent virus 6 ...		IIV6-006L
Q6GZX9	Uncharacterized protein 006R OS=Frog virus 3 (isolate Goorha) OX=654924 GN=F...		FV3-006R
B2ZDY1	RING finger protein Z OS=Whitewater Arroyo mammarynavirus (isolate Rat/United ...		Z
A9JR22	RING finger protein Z OS=Tamiami mammarynavirus (isolate Rat/United States/W 1...		Z
Q88470	RING finger protein Z OS=Tacaripe virus (strain Franze-Fernandez) OX=928313 G...		Z
P08105	Putative uncharacterized protein Z OS=Ovis aries OX=9940 PE=4 SV=1		
Q6UY62	RING finger protein Z OS=Sabia mammarynavirus (isolate Human/Brasil/SPH114202...		Z

UniProt/Swiss-Prot Alt. Accession (UniProtKB)

Description Parse Rule:

Decoy Protein Parse Rule:

Alternate ID Parse Rule:

☒ Use Magic Matching

Help Export... Apply Cancel

This dialog contains tools to help the user describe and edit the location of the selected database:

- **Name Database Parser**—Through this text box the user can change the name assigned to the database when loaded
- **Reset file location**— By clicking this button the user can point Scaffold to a different location where the database is stored.
- **Database version**— This text box can be used to define the Database version

The dialog also contains tools to help the user parse the database as needed:

- **Parsed Accession numbers and protein names table** — This table lists a sample of the protein accession numbers and descriptions contained in the database. The list includes proteins selected from the top and the bottom of the database file to give an idea of the type of accession numbers used in the database. The accession numbers and protein descriptions are shown parsed according to the rules selected from the pull down list of parsing rules, see below.
- **Parsing Rules Pull down list**— The list includes a number of standard parsing methods for different types of databases and their related accession numbers format (like Swiss-prot, Uniprot-Sprot etc). Once a particular rule is selected the different related parsing strings are shown in the test boxes located on the right hand side of the list. The user's specified selection in the list allows editing of the rules appearing in the text boxes and when clicking in a different text box Scaffold automatically verifies the validity of the

specified rule. In addition to rules to parse the protein name and description, the user may specify a rule to recognize decoy proteins and a rule to extract Gene IDs or other alternate identifiers for the proteins from the database.

- **Magic Matching check box**- This tool optimizes the accession numbers available for the proteins to better match databases and loaded data. Once a parsing rule has been selected by the user through the pull down list, Magic Matching checks protein by protein if that type of accession number properly matches the protein and finds alternatives if it does not. If no alternatives are available it defaults to a generic accession number.

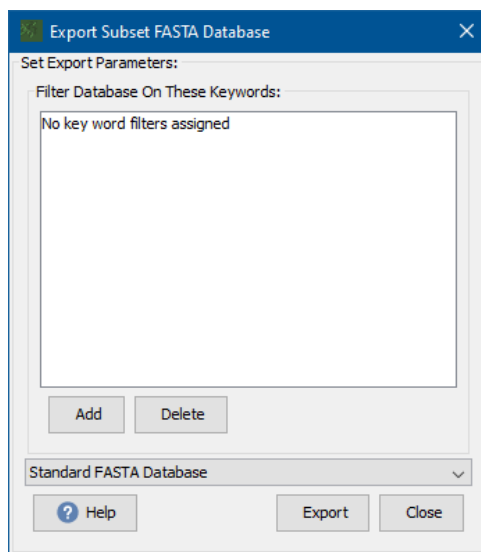
Magic Matching works by breaking the accession number provided by the search engine into tokens, splitting on spaces, bars and other delimiters, so that, e.g. sp|A9JR22|Z_TAMVU is broken into three tokens, “sp”, “A9JR22”, and “Z_TAMVU”. The entries in the FASTA are also broken into tokens. Scaffold then moves through the tokenized FASTA entries, trying to find an entry that has three consecutive tokens that match the accession tokens.

In the bottom left corner of the dialog there are two buttons one that calls for the Online-help and another one, the **Export** button, that calls the [Export Subset FASTA Database](#) dialog to create a subset database or a decoy database.

Export Subset FASTA Database

This dialog is called by the **Export** button located in the [Configure Database Parser](#) dialog. It provides tools to create a new filtered FASTA database or decoy database starting from the one selected in the Configure Database Parser dialog. It contains a list and pull down menu.

Figure 4-8: Export Subset FASTA Databases dialog



- **List of Filter Keywords** - Any of the keywords in the list is used to filter the original database for accession numbers that contains them. Note that the keywords are not case sensitive, do not have to be complete words, and can be multiple word phrases. Key words

can be added to and deleted from the list using the buttons present at the bottom of the list. The button Add opens the Add Keywords Filter dialog where the user can type in a new word.

This option is most often used to create a FASTA file for a specific species from a huge database. The taxonomy of the protein is listed in different ways in different databases, so the user needs to choose keywords appropriately. For example, to select only bovine proteins from the complete UniPROT database, enter the keyword "_BOVIN" or to select rat proteins, enter the keyword "_Rattus".

- **Database type pull down menu-** It shows the list of possible types of databases that can be created through this function:
 - *Standard FASTA database* - This option is used when the user wants to filter a large database with specific keyword to reduce its size.
 - *Reverse FASTA Database* - Each accession number has a "-R" appended to it. The protein description is unchanged. The protein sequence is reversed.
 - *Random FASTA Database* - Each accession number has a "-R" appended to it. The protein description is unchanged. The protein sequence is scrambled in a random manner.
 - *Reverse Concatenated FASTA Database* - Each protein in the original FASTA file appears unchanged, but it is preceded in the FASTA file by the reverse protein ("-R" appended to accession number and sequence reversed). This database is twice as long as the original.
 - *Random Concatenated FASTA Database* - Each protein in the original FASTA file appears unchanged, but it is preceded in the FASTA file by the randomly scrambled protein ("-R" appended to accession number and sequence scrambled). This database is twice as long as the original.

After selecting the appropriate options the user can click Export to save the new database.

GO Annotations and Pathways in Scaffold

Scaffold can search for GO terms through the NCBI website or by importing GO annotation databases through the main menu command **Edit > Edit Annotation Options**. This menu command opens the **Configure Annotation Sources** dialog.

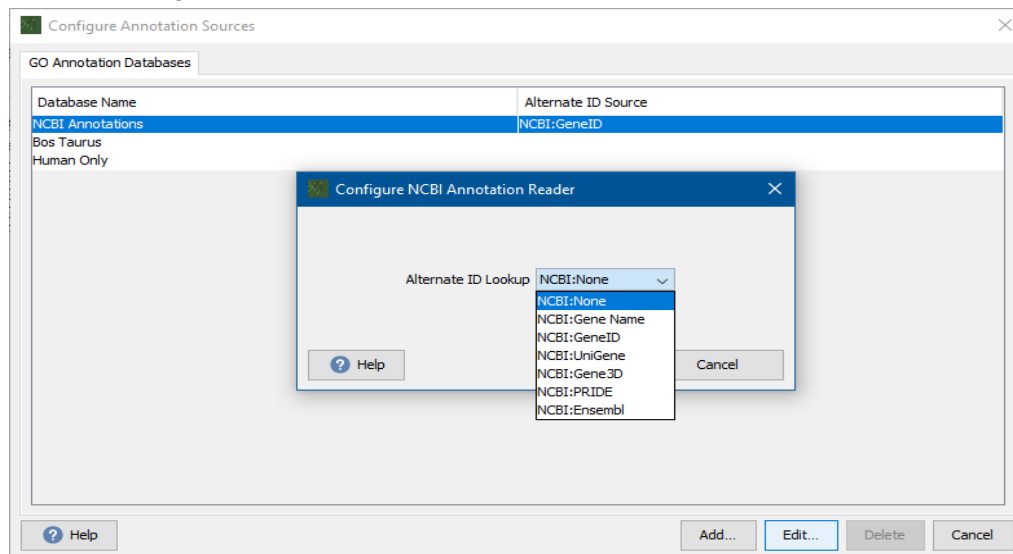
Configure Annotation Sources

This dialog contains a table which lists all of the GO annotation databases already configured in Scaffold and ready to apply. The option **NCBI Annotations** is always listed, as it fetches GO terms from the NCBI website and does not require that any information is imported by the user. Nevertheless, the user may wish to configure some options for NCBI annotation.

Configuring NCBI Annotation

When querying the NCBI database, the user may request that Gene IDs, Gene Names or other identifiers be returned by the database. These identifiers will be displayed in a column titled “Alternate ID” in the Samples View. To configure this feature, select NCBI Annotations then click **Edit** at the bottom of the pane. The dialog that opens contains a dropdown that allows the user to select the type of alternate identifier to be displayed.

Figure 4-9: Configure NCBI Annotation Reader



GO Annotation Files

Other GO sources may be added to the table by importing GO annotation files, which may be obtained in various ways.

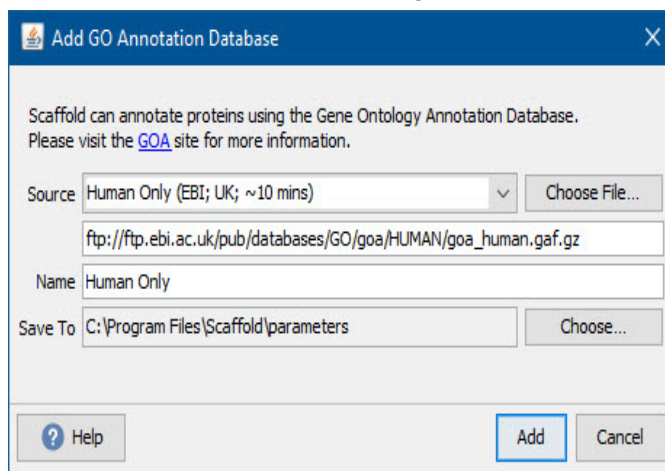
To add a new GO source, click the **Add** Button to open the **Add GO Annotation Database** dialog. This dialog offers options to download GOA files from the Internet, or import GOA files downloaded externally, obtained from colleagues or custom-created and stored on the computer. It is also possible to access GOA databases previously created and saved by Scaffold.

The GO Annotations Tab also includes a search box and the [Add annotation sources](#) button to import GO databases.

Add annotation sources

The **Add** button opens a dialog through which the user can import GO databases. A pull-down list directs Scaffold to different web locations from which specific GO Database files can be automatically downloaded and imported. Note that each GO database only needs to be imported once, and then can be used repeatedly.

Figure 4-10: Add GO Annotations Database dialog



The pull-down list includes the following items:

- **Human Only** - provides a download of the human subset. It takes about 10 minutes to download.
- **Other Website** - the user can type in a website address from which a GO Database can be downloaded.
- **Other File** - the user can direct Scaffold to a location on his/her computer where the GO database is stored.
- **A series of species-specific GOA files** - Because the All Proteomes file is so large, Scaffold supplies links to import a series of species-specific GOA databases from geneontology.org.

After one of the options is selected, clicking **Add** starts the operation of importing the GO annotation database into Scaffold. A new row appears in the list of already loaded databases, showing the name of the newly added database and the number of annotations included in it.

Close the dialog, and Scaffold is ready to annotate the protein list in the Samples table. The user can start this process by choosing **Experiment > Add or Edit Annotations**.

Bulk Operations...

Bulk operations include tagging, showing and hiding proteins appearing in the Proteins list, see [Protein List](#). When the clustering function is applied this menu allows the user to expand or collapse all clusters present in the proteins list.

These options are also available when right clicking over a selected protein in the proteins table.

Preferences

The Preferences dialog provides a series of modifiable options organized in a number of different tabs. Through this dialog the user can modify parameters and settings to customize the way Scaffold experiments appear and run.

Selecting the menu item **Edit > Preferences** opens the **Preferences** dialog which contains the following tabs:

- [Internet](#)
- [Memory](#)
- [Processors](#)
- [Web Link](#)
- [Mascot Server](#)
- [Display Settings](#)
- [Password](#)
- [Paths Settings](#)
- [MZID](#)

Internet

In the Internet Settings dialog the user can enter a Proxy server name or IP address and a proxy port number. Through check boxes in this dialog box, the user may:

- **Allow Scaffold to connect to the Internet** If this box is unchecked, then Scaffold cannot access the Internet. Users may want to have this box unchecked if their organization prevents connections to the Internet.
- **Use HTTP Proxy Server**
 - **Proxy Server name (or IP address)**
 - **Proxy port number**

Proxy servers may be used by an organization's IT departments to filter communications to and from the Internet. If that is the case, then users need to set the Proxy Server Name and Port Number. Users can check if there is any need to use proxy server settings by looking at how their web browser is connected to the web.

Back to [Preferences](#)

Memory

This tab allows the user to set the maximum amount of memory that Scaffold is allowed to use. Scaffold is a memory-intensive program and needs a large amount of RAM to be able run at a decent speed.

When setting the amount of memory Scaffold should use it is important leave enough memory for other programs to run.



- *The new memory setting will take effect only after the application has been closed and restarted.*

Back to [Preferences](#)

Processors

This tab provides information to Scaffold about the maximum number of processors available for threading computations. The default value is the maximum number of processors available in the system where the application is installed.

The Scaffold application in itself uses only two threads. Assigning more than two threads to Scaffold mainly affects how fast the X! Tandem version bundled with Scaffold executes thus optimizing the throughput.

Back to [Preferences](#)

Web Link

Through this tab the user can add, change or delete the on-line protein lookup databases links. These web sites appear in the **Lookup Accession Number in:** pull down list found in the [Protein Information pane](#) in the Samples View.

When selecting these databases it is important to note that they need to support lookup of proteins using either the type of accession numbers displayed in the Samples View, or identifiers displayed in the Alternate ID column.

Clicking New Database opens the **Configure Web Link** [Configure Web Link](#) dialog where information for a new Online database can be added.

The linked database can be either a public database, or an internal one. It just has to have a URL that queries the database. This link could also be to a web site that performs a calculation or does a BLAST.

The user might want to set up several links to the same database that query it in using different types of accession numbers.

The **Edit** button allows the user to modify a web link to adjust the URL if it has changed, or a better URL has become available.

Back to [Preferences](#)

Configure Web Link

This dialog creates a new web link Scaffold can use to search protein information on the web through the Protein Information Pane. When created the link is added to the **Lookup Accession Number in:** pull down list found in the Protein Information pane located in the Samples View.

To create a new link the user needs to input the URL for a website where the database is accessible in the **Link To:** text box. Note that "http://" must be included in the URL address.

The intent here is to use the URL to query a protein database that is on the web, for example

SWISS-PROT. To execute a query, the web database needs to know how to look up the protein. These queries are frequently done by embedding the accession number in the URL address. When entering the URL in this box, we suggest using "%s" to represent the accession number.

Each web site is indexed in a different way. For example the EXPASY site can find proteins given either the accession number (P02769) or the SWISS-PROT protein entry name (ALBU_BOVIN), but WikiPathways can only access proteins by the gene name.

The **Name Web Link** text box assigns a name to the link displayed in the **Lookup Identifier In:** pull down list found in the Protein Information pane located in the Samples View. The convention is to include in this name an example of the type of accession number the web link is able to use. For example, the IPI link has the name "IPI (ie:IPI00574195)".

Through the button **Test Link** it is possible to verify if the input information reaches the correct website. If no Internet connection is established, the user should check his/her Internet Settings. If a website comes up but the protein is not found, then he/she should check the link in the **Link To:** text box.

Clicking **Apply** finalizes the creation of the web link.

Back to [Web Link](#)

Mascot Server

Scaffold can load data directly from a Mascot Server. This tab contains a text box where the user can set up the connection to the server by writing the web address of the available Mascot server.

The button **Test Connection**, located on the lower right corner of the tab page, provides a quick way to check if the connection works properly.

- If no security is implemented, Scaffold connects directly to the Mascot sever. When the **Test Connection** button is clicked, a message appears stating whether the connection was successful or not.
- If security is enabled on the Mascot Server and the **Test Connection** button is clicked a login window pops up asking for an account name and password. The user has to make sure that the account he/she is using in Mascot has administrative rights.



Scaffold does not download files from a Mascot Server if the user is logged on as a GUEST and an error is shown.

Back to [Preferences](#)

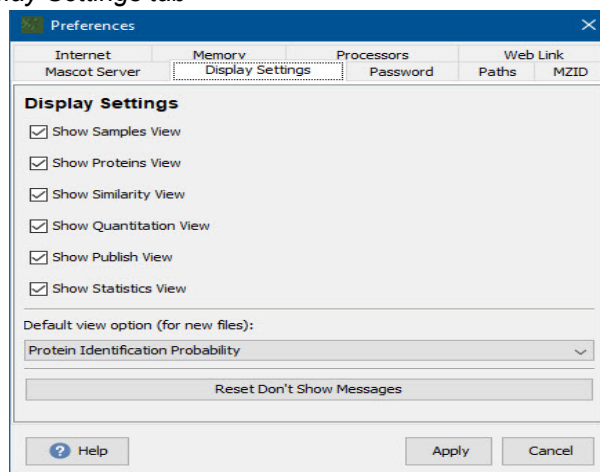
Display Settings

Scaffold provides different ways to look at the data included in an experiment through different views: Load Data, Samples, Proteins, Similarity, Quantify, Publish and Statistics. The Display Settings tab allows the user to decide which of the available Views is visible.

Through this tab the user can also define which default Display Options is selected when a new experiment is created and reset messages that were selected not to show anymore, like the initial repetitive dialogs that appear when the Wizard is opened.

The tab includes a check box list of all the views available in Scaffold, a pull down list of the Display Options to select the appropriate default value and a reset button.

Figure 4-11: Display Settings tab



Check box list

If one of the views in the list is not checked it will not display and the corresponding button in the Navigation pane will also not be visible.

These settings are saved with the experiment when the *.sf3 file is created. For example, if the user turns off the Statistics and Proteins views and saves the file, when the file is reopened, only the Load Data, Samples and Publish views are visible.

This feature can be useful when sending the results to someone who doesn't need to see certain details or to check the validity of the statistical analysis.

Access to Display Settings may be controlled by a password by checking the appropriate box on the [Password](#) tab of the Preferences dialog. This means that the user can control which pages his/her collaborators can view. The user can use this in conjunction with the password protection on the filters to control which proteins can be viewed.

Default View Options (for new files) pull down list

The bottom portion of the window stores the preferred Display Options for the Samples View when a new experiment is created. A pull down list allows the user to select what information Scaffold will initially show when the loading phase is completed and the Samples view is initially shown.

Reset button

The **Reset Don't Show Messages** button restores the messages that were checked to not show again when requested.

Back to [Preferences](#)

Password

Through this tab the user can select to use a password to protect certain views and operations available in a Scaffold experiment once it is saved in a *.sf3 file. For example, the user can set filter thresholds to display only data with above 90% confidence or restrict access to only the Samples and Publish pages, in this way hiding the messy details on the Proteins and Statistics pages. The user can also prevent anyone from reanalyzing his or her data by locking the export of the spectra.

A password gives the user control, control of what the people viewing data can see and do.

- **Use Password...** - Turns on and off the password protection
- **Protect Exporting Spectra** - Password required to export spectra.
- **Protect Resetting Thresholds** - Password required to change Min Peptide filter or define custom filters.
- **Protect Changing Display Settings** - Password required to hide or display hidden pages.
- **Protect Hidden Proteins** - Password required to access **View > Show Hidden Proteins**

Back to [Preferences](#)

Set Password

When this dialog opens the user can define a password that controls what data is displayed in the various Scaffold views, how much detail is available and what data can be exported out of a saved Scaffold experiment.

Back to [Password](#)

Paths Settings

The Scaffold Installation comes with a generic UNIMOD database typically used to unify modifications naming among different search engines when their results are loaded in the same Scaffold experiment.

This Tab allows the user to select alternative UNIMOD databases when loading data:

- **Do not use UNIMOD** - This option tells Scaffold to retrieve the information about modifications directly from the search engine results that are being loaded.
- **Use Scaffold default UNIMOD** - This options tells Scaffold to use the default UNIMOD database.
- **Use a custom UNIMOD file** - This option allows the user to direct Scaffold to a location where a custom UNIMOD database is available and retrieve the modification information from the selected database.

Back to [Preferences](#)

MZID

A checkbox that determines whether Scaffold should read protein sequences from mzIdentML files when they are loaded, if they are available in the files. The alternative is to

read the sequences from the FASTA that is applied during loading.
The default setting is to read sequences from the FASTA.

Advanced Preferences

Scaffold includes several validation algorithms used to assess how probable a peptide or protein identification is, such as the PeptideProphet algorithm, see [“Increased Confidence Using Peptide and Protein Validation Algorithms” on page 28](#).

One of the parameters included in these algorithms is the Discriminant Score, a search engine score or a calculation based on one or more search engine scores which Scaffold uses as a primary score for data from a particular search engine.

Depending on how the search engine parameters are set, in some cases Scaffold’s usual Discriminant Score for a search engine may not provide a good basis for computing peptide probabilities. When this happens, the peptide probability assignments, and consequently the protein probabilities computed by Scaffold are unreliable. This behavior is particularly evident in loading data searched using Sequest or Sequest HT in Proteome Discoverer when the Delta Cn value cannot be computed for many of the peptide matches.

To compensate for this problem Scaffold provides a dialog reached through the **Edit > Advanced Preferences** dialog through which the user can choose which scoring function Scaffold should use to validate the peptide and protein identifications.

For instructions in setting parameters in Proteome Discoverer and other search engines, see: <https://www.proteomesoftware.com/data-formats>



Proteome Software now recommends using a workflow in Proteome Discoverer that includes Percolator and loading the results into Scaffold in prefiltered mode. With this workflow there will be no need to use the Advanced Preferences tab.

The Advanced Preferences dialog is reached through the **Edit > Advanced Preferences...** menu command. It includes two separate tabs, one for each of the following search engines:

Sequest tab

When running Sequest searches in Proteome Discoverer using the default settings suggested by Thermo, the MSF output files do not include records of unassigned spectra and lower scoring Peptide Spectra Matches (PSMs). The unassigned spectra and the PSMs are needed for calculating the Delta Cn score included in the formula Scaffold applies to compute the Sequest discriminant score used in its validation algorithms.

Scaffold uses a different definition of delta Cn than the original Sequest definition. The equation Scaffold uses for deltaCn for the Nth best peptide is:

$$\text{deltaCnN} = ((\text{XCorr1} - \text{XCorr2}) / \text{XCorr1}) - ((\text{XCorr1} - \text{XCorrN}) / \text{XCorr1})$$

When there is only one peptide reported for a spectrum:

XCorr1 = the XCorr of the reported peptide

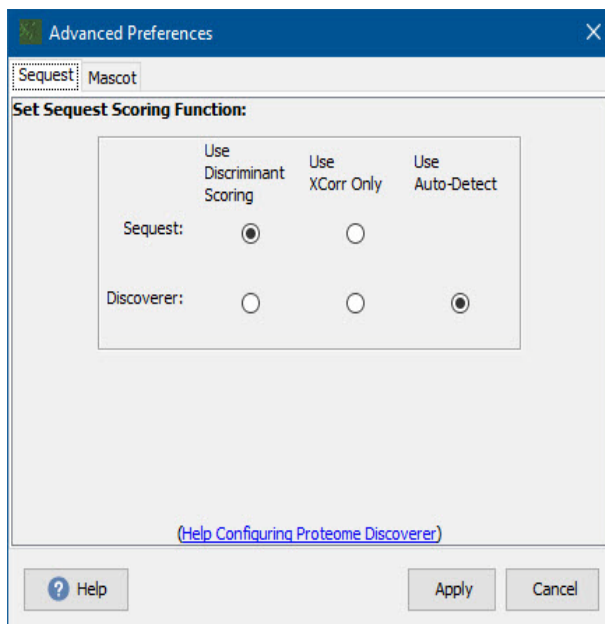
XCorr2 = min(XCorr1, max(peptideWithoutProteinXCorrThreshold, absoluteXCorrThreshold))

In the original equation, $\Delta Cn1$ is always 0 and all other ΔCn s are less than 1. In Scaffold, this is adjusted so that $\Delta Cn1 \geq 0$, $\Delta Cn2 = 0$, and all others < 0 . Thus, $\Delta Cn1$ becomes a useful number.

The Sequest Discriminant Score formula use by Scaffold is a normalized version of the Sequest XCORR score and depends upon the charge state of a peptide. For example, for charge +2 the discriminant function is: For instructions in how to configure searches in Proteome Discoverer for best compatibility with Scaffold, see <https://www.proteomesoftware.com/data-formats>.

If these instructions are not followed, or if in spite of following these instructions, Scaffold's probability models do not properly assess peptide probabilities as a result of faulty ΔCn calculations, the user may adjust the Discriminant Score used by Scaffold through the Sequest tab in the Advanced Preferences dialog.

Figure 4-13: Setting Sequest Scoring Function



Mascot tab

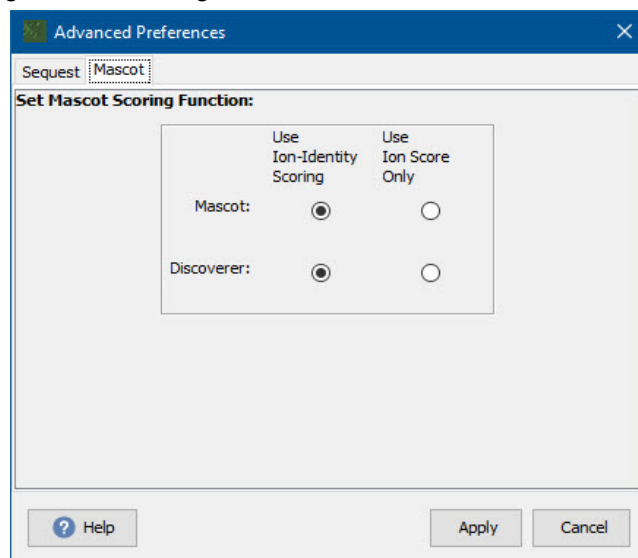
When loading Mascot data, Scaffold's validation algorithms calculate the Discriminant Score as the Mascot Ions Score minus the Mascot Identity Score.

The Identity Score, which provides a significance threshold for the Ions Score, as defined by Matrix Science, records the level at which an identification has a 5% probability of being due to a random match. Mascot's concept of probability is somewhat different than Scaffold's, but roughly speaking when setting Scaffold's **Min Peptide** probability to 95%, the black vertical line in the Mascot's Histogram shown in Scaffold's Statistical View should be close to zero on the discriminant scale.

Depending on the parameters set for the searches, Scaffold may be unable to properly calculate the Identity Score. This affects the values of the calculated peptide probabilities

and consequently the probability assigned to the list of identified proteins. When this happens, the user can select “Ion Score Only” as the scoring option used by the validation algorithms through the **Advanced Preferences dialog > Mascot tab**.

Figure 4-14: Setting Mascot Scoring Function



The Mascot tab includes a table listing the different programs producing Mascot search results and radial buttons for selecting which scoring function Scaffold uses in the selected validation algorithm:

Suggestions concerning proper settings for Mascot in Proteome Discoverer may be found in: <https://www.proteomesoftware.com/data-formats>

Note that selecting **Use Ion Score Only** provides a list of proteins different in length than selecting **Use Ion-Identity Scoring**.

Show Lower Scoring Matches

The command **View > Show Lower Scoring Matches** toggles the option to display a protein in a sample even if it does not meet the current filters and thresholds in that sample if it does meet thresholds in at least one other sample.

In some cases several samples may identify a protein at very different confidence levels. For example, sample 1 may identify protein A with 95% probability and sample 2 may only identify it with 60% probability.

- If the option **View > Show Lower Scoring Matches** is selected, then the filters and thresholds affect only which protein rows are shown and both the 95% and the 60% values would be displayed, even if the protein threshold was set at 90%.
- If the **View > Show Lower Scoring Matches** option is not selected, then the sample values that do not meet the filter values are suppressed. This means that the 95% value for sample 1 would be shown, but no value would be shown for sample 2.



*It is particularly important to be aware of the status of the **View > Show Lower Scoring Matches** option since it affects the counts shown in the table and varies the quantitative values in graphs and figures throughout the program.*

Note: When probabilities are lower than 5% values will not be shown unless [Show <5% Probabilities](#) is selected.

Show <5% Probabilities

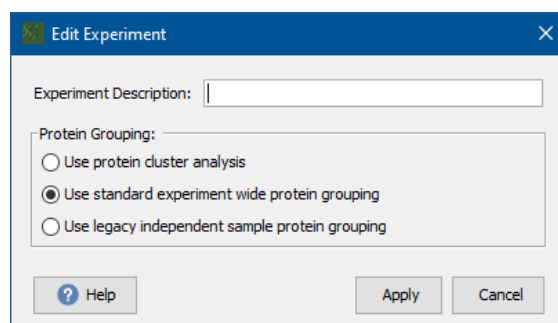
When Show lower Scoring Matches is not selected values for proteins that have probabilities less than 5% are not shown. This option allows those values to be visible.

-
- **Queue Raw Data for Loading**[Queue Raw Data for Loading](#)
-
- **Extract Alternate ID**
-

Edit Experiment

The menu option **Experiment > Edit Experiment** opens a dialog through which the user can add or edit a description of the experiment.

Figure 4-18: Edit Experiment menu option



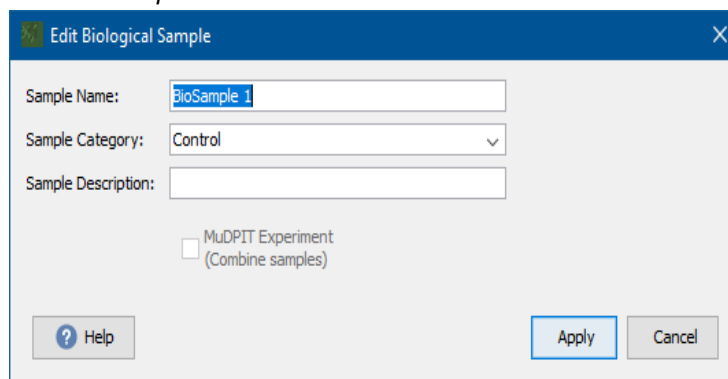
In the full version of Scaffold this dialog also contains the Protein Grouping pane where it is possible to toggle the various protein grouping options available in the program.

- **Use protein cluster analysis** - When selected, Scaffold uses [Shared Peptide Grouping and Protein Cluster Analysis](#) to group and pair the list of identified proteins.
- **Use standard experiment wide protein grouping** - When selected Scaffold uses [Standard Protein grouping](#) with no clustering.
- **Use legacy independent sample protein grouping** - Scaffold uses [Standard Protein grouping](#) with no clustering, but the grouping is done within biosamples and not across biosamples.

Edit BioSample

The menu option **Experiment > BioSample** opens a dialog in which the user can add or edit the name of the sample, its category and description, see [Organize Samples Into Categories](#).

Figure 4-19: Edit BioSample



Note that, defining these parameters in a concise and consistent manner is quite useful since Scaffold uses the Sample and Category names as column headers in the Samples View.

The dialog also shows whether the data was loaded using the MudPIT options.



- *While this option may be selected in any view, it is highly recommended to use it only from the Load Data view to facilitate the selection of the BioSample to be modified.*
- *To avoid unintended inconsistencies in category names, choose the appropriate name from the drop-down list whenever available.*

Organize Samples Into Categories

When BioSamples are defined in the Loading Wizard, they can also be organized into Categories. If Categories are not initially defined when BioSamples are created, the BioSamples can be organized later by selecting the menu option **Experiment > Edit Biological Sample**.

Categories are useful in two ways. The first is that the columns in [The Samples Table](#) group the BioSamples into categories. For example if the samples are put into categories “Treated” and “Control”, then the samples in the “Control” category will be grouped together to the left of the samples in the “Treated” category.

The second way categories are useful is to organize the samples in order to find which proteins among categories are differentially expressed. Scaffold offers several options for comparing the expression level of each protein between categories. The quantitative analysis terms, **Experiment > Quantitative Analysis...**, T-Test and ANOVA both measure the statistical probability of difference between categories. Likewise the Quantify View organizes data in categories.

Apply New Database

The menu option **Experiment > Apply New Database** allows the user to address the following situations:

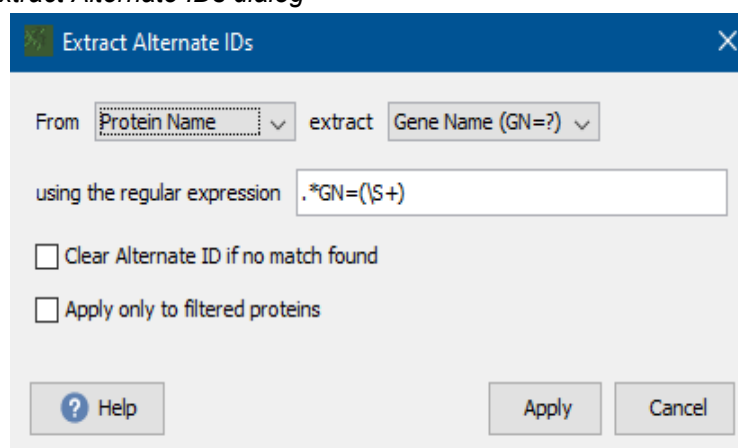
- **Incorrect parsing of protein accession numbers** - When this happens, [The Samples Table](#) reports question marks in the Molecular weight column while in the Proteins View the protein sequence is missing. Usually, this is related to an incorrect parsing of the database selected when loading the data into Scaffold. Either the database is not the same as the one used in the searches or Scaffold was not able to apply proper parsing rules to connect the accession numbers appearing in the search results to the database used when loading. Selecting another database or re-parsing the database used in the loading phase typically resolves the problem.
- **Loading data searched using multiple databases** - More than one FASTA database may be selected during loading, however if some of the databases used in the search are not specified during loading, they may be applied later. When this happens, proteins identified with the other databases are not correctly parsed and their molecular weights appear as a question marks. This problem can be resolved by selecting and applying the other databases used in the analyses. Scaffold picks up the unidentified proteins and resolves the question marks appearing in the molecular weight column and retrieves the protein sequences appearing in the Peptides View.

When selected, the menu option opens the Select Database dialog showing the list of FASTA databases currently loaded in Scaffold. The user can then choose a different database and apply it to the current list of proteins. This dialog to [Edit FASTA Databases](#).

Extract Alternate IDs

This option allows the user to create and populate an Alternate ID column by parsing out information contained in the protein names or accession numbers. When this option is selected, a dialog opens through which the user can specify parse rules for extraction of Gene Name, Gene ID or some other alternate ID that may be specified within the text in the Protein Name or Accession Number column.

Figure 4-20: Extract Alternate IDs dialog



- The first dropdown selects the column from which to extract the IDs.
- The second dropdown offers pre-specified regular expressions for some common alternate ID strings. For example, often the Gene Name is included in the Protein Name string parsed from UniProt style FASTAs, and can be pulled out using the Gene Name parse rule offered here. There is also an option to specify a custom parse rule. For assistance with creation of parse rules, please contact support@proteomesoftware.com.
- A checkbox allows the user to specify whether to clear existing alternate IDs if no new ID is found for a protein that already had an alternate ID specified. If a new Alternate ID is found for a specific protein, it will overwrite the existing alternate ID. The effect of this checkbox is to determine whether this operation will update already applied alternate IDs or replace them.
- A checkbox allows the user to specify whether Alternate IDs should be added only to the set of proteins currently displayed or to all proteins even if they are currently hidden because of thresholds or filtering.

Load and Analyze Queue

This command is available only when there are files present in the Loading Queue in the [Load Data View](#). When selected, it opens the [Load and Analyze Data](#) page of the Loading

Wizard where the user can select the proper loading parameters and load the data in Scaffold

Reset Peptide Validation

In the [Proteins View](#) Scaffold provides tools to manually inspect the identification of peptides. A validation check box records the status of a peptide; when selected a peptide is considered valid. The user, upon visually inspecting the related spectrum, can invalidate a peptide by manually deselecting its check box.

The menu option **Experiment > Reset Peptide Validation** provides a global tool to automatically validate all peptides above a specified probability by selecting the related check box and deselecting those below it. When a different probability is selected the command resets all previous user validations. The default probability is 0%.

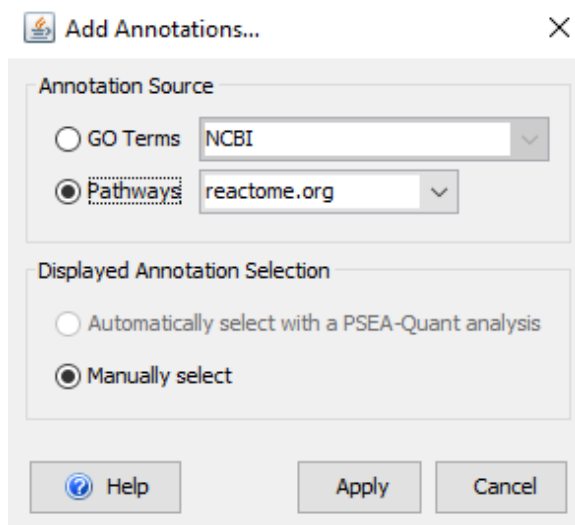
This function can be used in two possible scenarios of the peptide validation process:

- **Globally create a set of validated peptides based on probability assignments.**
Peptides are considered valid only if their probability is greater than the minimum amount set in the pull-down menu **Minimum Peptide Probability**, its default probability being 0%. Select a different value and click Apply. All peptides with probability less than the Minimum Peptide Probability will be shown unchecked in the Proteins View and not considered for analysis.
- **Reset the manually validated peptides to their initial status**, the initial status being identified by the minimum peptide probability recorded in the pull-down menu.

Add or Edit Annotations

This menu option opens a dialog which allows the user to begin a search for either GO terms or pathways associated with the proteins identified in the experiment.

Figure 4-21: Experiment>Add or Edit Annotations opens the Add Annotations dialog



Annotation Source

The first section of the dialog consists of two radio buttons that determine whether the

operation will search for GO Terms or Pathways. Only one of these options may be chosen at a time, but GO Terms and Pathways may be applied sequentially, resulting in the display of both sets of annotations in the Samples View simultaneously.

Each radio button has a pull-down menu associated with it. These list the possible sources which may be searched for each type of annotation. For GO terms, the list consists of NCBI, which is always available if the computer is connected to the Internet, and the names of any GOA databases configured through Edit>Edit Annotation Options (see [Configure Annotation Sources](#)). If GO Terms have already been added to the experiment, an additional option, Current Set, is also available. Choosing this option allows the user to adjust the displayed Go Terms without performing another GO Term search.

For Pathways, the following options are offered:

- **WikiPathways** - pathways are retrieved from the WikiPathways database¹. WikiPathways associates pathways with gene names, so to use this option it is necessary to have gene names in the Alternate ID column. Gene names may be added by applying a FASTA database with parsing rules that extract gene names, or by configuring NCBI to retrieve gene names and applying GO terms from NCBI.
- **Reactome** - pathways are retrieved from the Reactome database². Scaffold retrieves pathways from Reactome using the protein accession numbers. The user may select the source based on the type of accession number in the experiment. The default is Uniprot, and requires that the accession numbers be standard Uniprot accession numbers (e.g. P02769, not ALBU_BOVIN). It is important, therefore, to parse the FASTA database appropriately to produce this type of accession number before loading samples in order to use Reactome.
- **Current Set** - if pathways have already been added to the experiment, this additional option is offered. It allows the user to adjust the displayed pathway columns without conducting another pathway search.

When annotations are added to a Scaffold experiment, all GO terms or pathways that match the protein identifiers are retrieved and stored in the SF3 file. As a result, the user may choose to display different annotations without the need to re-search the annotation source. The one exception is that in a Reactome search, only pathways pertaining to the selected taxonomies are retrieved. Changing taxonomies requires a new search.

Displayed Annotation Selection

In this section of the dialog, the user specifies how the set of displayed annotations should be selected. There are two options:

- **Automatically select with a PSEA-Quant Analysis**³ - this option opens a dialog to allow the user to configure a PSEA-Quant analysis to identify annotations that are enriched in the current set of proteins and to select displayed annotations from the results.

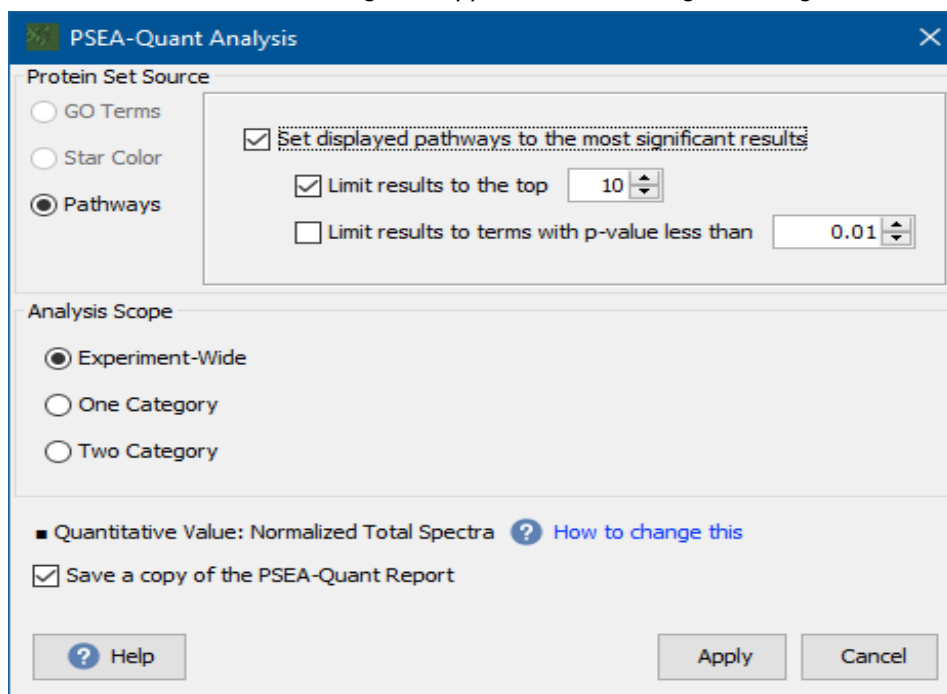
1. Kelder T, Pico AR, Hanspers K, van Iersel MP, Evelo C, Conklin BR. (2009) Mining Biological Pathways Using WikiPathways Web Services. PLoS ONE 4(7): (link to article)
2. Fabregat et al. 2017 PMID: 28249561

- **Manually Select** - this option allows the user to select the annotations that should be displayed from a list of all available choices (see [Manually Select Displayed Annotations](#)).

Select Displayed Annotations with PSEA-Quant

PSEA-Quant may be used to identify annotations which are likely to be of biological relevance. Scaffold offer the option to use the results of this analysis to automatically select which annotations should be displayed as columns in the Samples View.

Figure 4-22: The PSEA-Quant dialog as it appears when adding or editing annotations



The PSEA-Quant Analysis Dialog

The Protein Set Source has already been determined when this dialog opens as part of the process of adding annotations, and will either be GO Terms or Pathways. For more information, see

Set displayed annotations to the most significant results- A check box allows the user to limit the columns that will be displayed based on the results of the PSEA-Quant analysis. If this box is checked, the user has two more options, each selected by its own check box:

- **Limit results to the top ...** - Specify the number of annotation columns to be displayed. Annotations will be selected in order of lowest p-value, but the selected number will appear regardless of statistical significance.

3. Lavallée-Adam, M., Rauniyar, N., McClatchy, D. B., & Yates, J. R. (2014). PSEA-Quant: A protein set enrichment analysis on label-free and label-based protein quantification data. *Journal of proteome research*

- **Limit results to terms with p-value less than ...**- Specify a significance level and only those annotations with p-values below that value will be displayed.
- The two options may be combined by checking both boxes. In this case, only significant annotations will be displayed, but if more than the specified number are significant, the list will be trimmed to the selected number.

Analysis Scope

Three different types of PSEA-Quant analysis may be performed by Scaffold, if supported by the design of the experiment:

- **Experiment-wide** - assesses enrichment in all MS Samples in the experiment.
- **One Category** - assesses enrichment only in samples belonging to a specified category, e.g. a specific treatment or condition. When this option is selected, the dialog displays a pull-down list from which to select the category.
- **Two Category** - assesses for enrichment in one category as compared to another. This option uses a modification of the published PSEA-Quant algorithm, which compares the ratios of the quantitative values in the two categories. Scaffold compares $\log(1+\text{ratio})$ instead, in order to lessen the influence proteins within a group with unusually large ratios.



- *Note that the two category analysis of category A over B and of B over A are different. The first returns protein groups enriched in A, while the second returns protein groups enriched in B.*

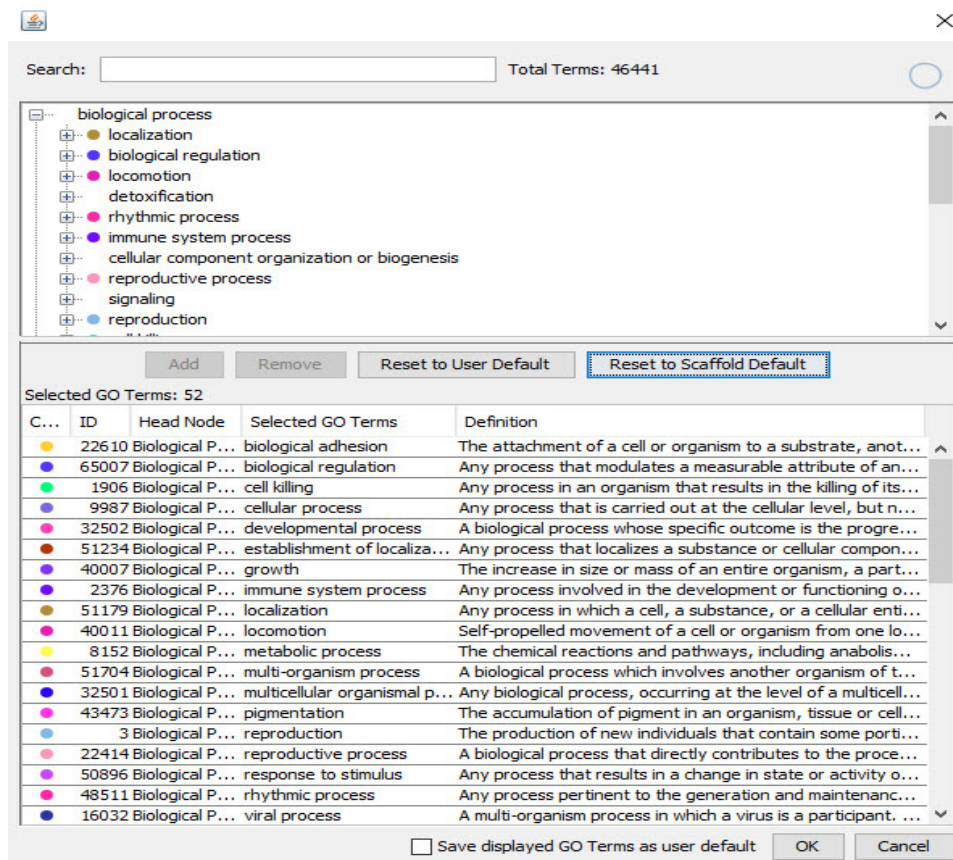
Save a copy of the PSEA report - if this box is checked, the user will be presented with a file browser to choose the location for the report.

Manually Select Displayed Annotations

This option allows the user to select the GO terms or Pathways that will be displayed. Manual selection is a good option when looking for changes in specific groups of proteins. The interface for manual selection varies according to the type of annotations.

Select Displayed GO Annotations

Figure 4-23: Displayed GO Term manual selection dialog



Using this dialog, the user can create and modify a custom list of GO terms. The terms in this list are then displayed as extra columns in the [Samples View](#) whenever they are present in the experiment.

The Tab is divided into sections:

- **Search Field** - Searches terms available in the GO terms database loaded in Scaffold.
- **GO Tree list** - Hierarchical list of all the terms present in the loaded GO database
- **Add and Remove GO terms** - Provides tools for creating the custom Display list
- **Display List** - List of GO terms selected by the user that will be visible in [The Samples Table](#).
- **Save and Apply**- Allows the user to save the current Display List if changed

To create a new custom GO terms Display List the user needs to follow these instructions:

1. If the **Display List** is not empty select all the rows and press delete.
2. Search and select any GO term of interest present in the loaded GO database either by typing a name in the **Search Field** or by selecting a row in the **GO Tree List**.

3. Click **Add**; the selected term or group of terms is added to the **Display List**. Terms may be selected individually or by domain or group. If a group or domain is selected, all terms in that group will be added to the **Display List**.
4. To remove terms from the **Display List**, select a term or group of terms to be discarded then click **Remove**.
5. To save the current selections as User Defaults check the box **Save displayed GO terms as user default**.

When a Scaffold experiment is saved, the displayed GO terms are saved within the *.SF3 file.

When a new file is created, or when Scaffold is closed, the list of displayed GO terms is unchanged. To reset the list to the defaults, the user may click the **Reset to User Default** or the **Reset to Scaffold Default** button.

Select Displayed Pathways

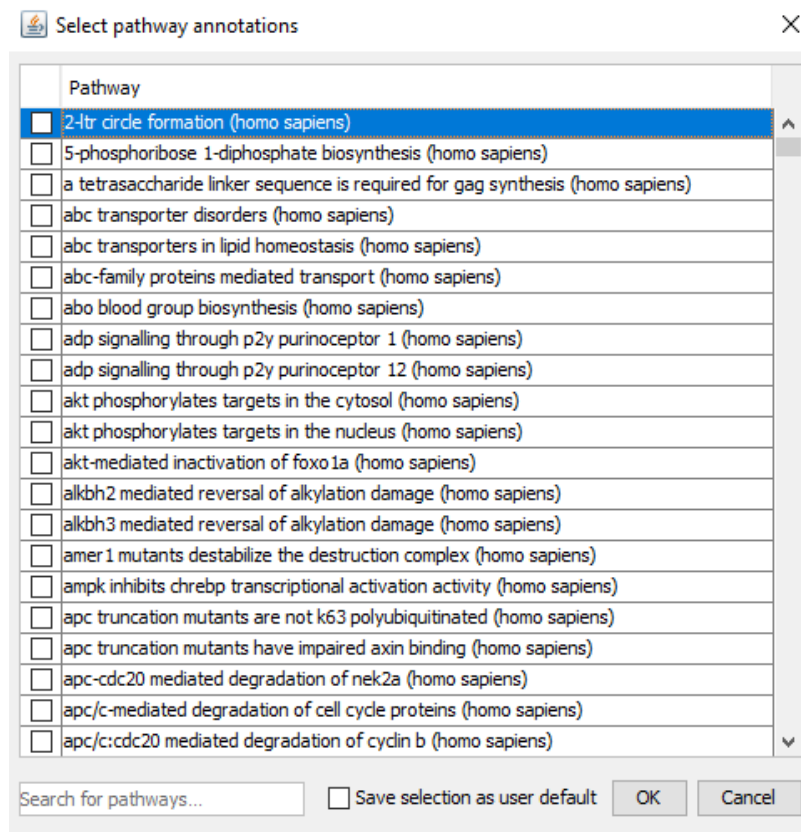
Select Taxonomy for Reactome

Figure 4-24: When The Annotation Source is set to Reactome and the user chooses to manually select the displayed pathways, the first step is to indicate the taxonomy or taxonomies of interest. A list of all available taxonomies in the Reactome database is displayed. A search function is supplied. Check one or more to query the database for pathways in those taxonomies. To check many, select the ones of interest using CTRL-click, Shift-click or CTRL A then right-click and choose Mark as Selected. The selection made will be saved as the user default and automatically checked the next time the dialog opens. *Select taxonomies for Reactome search*

Next, all pathways in the Reactome database pertaining to the selected taxonomies are retrieved, and presented to the user in a dialog box.

Select Displayed Pathways

Figure 4-25: Manually select displayed pathways



Check all pathways that should be displayed in the Samples View. A search function is supplied. To check many pathways, select the ones of interest using CTRL-click, Shift-click or CTRL-A then right-click and choose Mark as Selected.



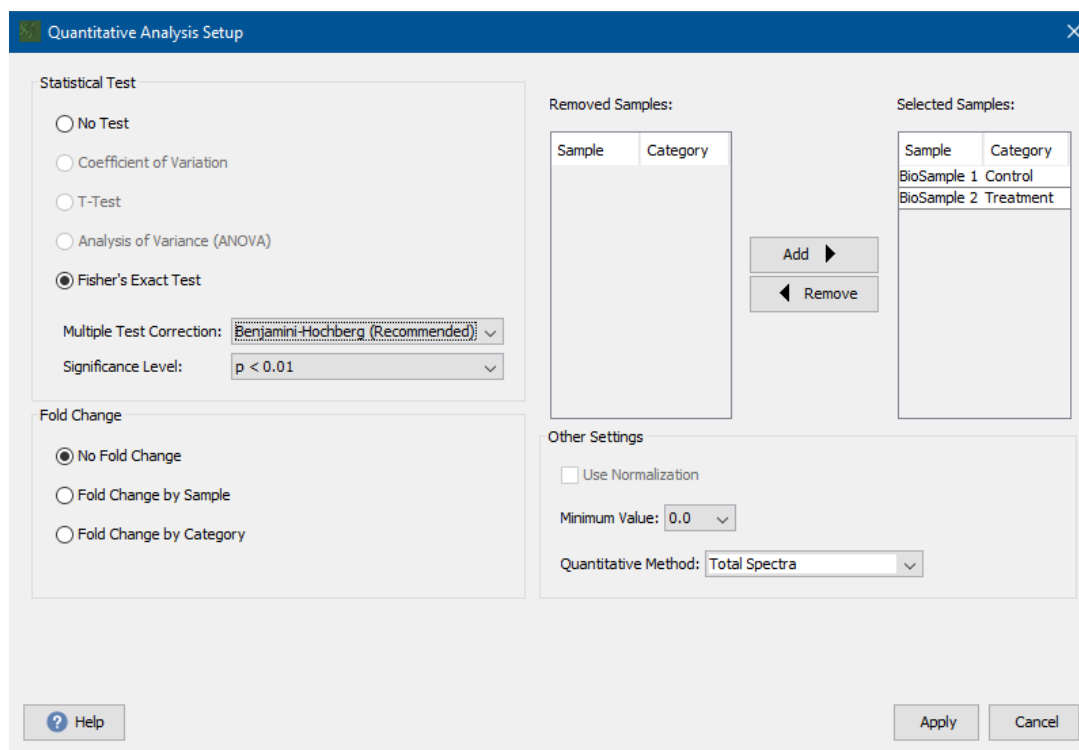
- When pathways are manually selected, Scaffold shows columns for all selected pathways even if no identified proteins are associated with some of them. This is different from GO term annotation, in which columns are displayed only for GO terms associated with at least one identified protein.

Quantitative Analysis...

Scaffold offers a number of statistical tests to determine the significance of trends in the data. These tests can be set up through the menu option **Experiment > Quantitative Analysis...** When this option is selected the dialog **Quantitative Analysis setup** opens, showing the list of statistical tests available, normalization and quantitative methods options and two lists from which the user can choose the different categories he/she wants to compare and apply inference tests to. There are up to six tests potentially available, depending on the number of

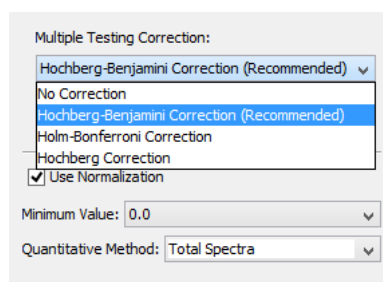
loaded samples and categories.

Figure 4-26: Quantitative Analysis Setup Dialog



Under the list of tests there are two pull-down menus that allow the user to define the Significance Level and apply multiple testing corrections, see [Multiple test significance levels and corrections](#).

Figure 4-27: Quantitative Analysis: Multiple Testing corrections



Other features:

- **Use Normalization** check-box, see [Normalization among BioSamples in Scaffold](#)
- **Minimum Value** pull-down list
- **Quantitative Methods** pull-down list, see [Label Free Quantitative Methods](#)

Note: When the **Use Normalization** check-box is not selected and Total Spectrum Count is the quantitative method chosen for the analysis, the values shown in the Samples View when the

display option **Quantitative Value** is chosen are going to be the same as the one reported when **Total Spectrum count** is the selected Display Option.

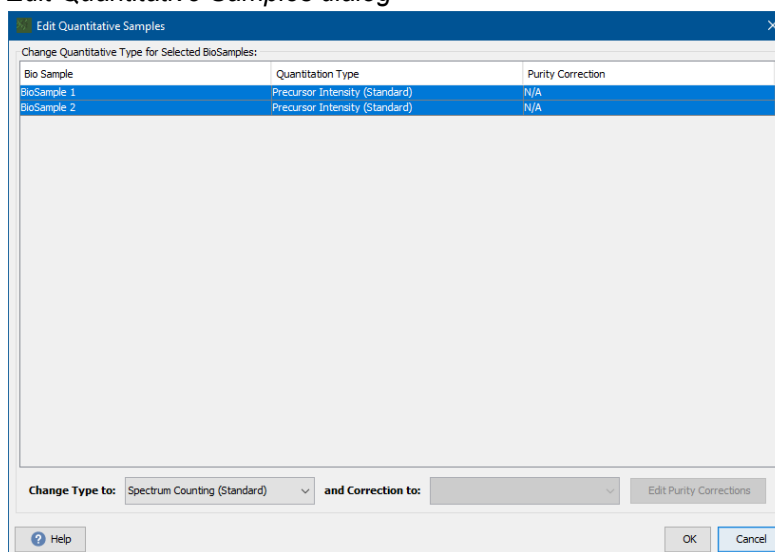
-
- BLIB^{BLIB} export

Quant

Edit Quantitative Samples

The menu option **Quant > Edit Quantitative Method/Purity Correction** opens the **Edit Quantitative Samples** dialog. The dialog allows the user to change the quantitative methods selected for each BioSample when loading data into Scaffold

Figure 4-28: *Edit Quantitative Samples dialog*



The dialog includes a table listing all the different BioSamples included in the current Scaffold experiment, the quantitative method selected when loading and if relevant the related purity correction.

Below the table there are a couple of pull-down lists and a button:

- **Change Type to:** - Pull down menu that lists the available selections for quantitation methods in Scaffold.
- **and Correction to:** - This pull-down menu is available only when iTRAQ and TMT as selected as quantitative methods. It lists the Purity corrections tables available.
- **Edit Purity Correction** - This button is available only when iTRAQ and TMT as selected as quantitative methods. When selected it opens the [Edit iTRAQ/TMT Purity Corrections](#) dialog.

Changing the Quantitative type for a specific BioSample:

1. From the **Change type to** pull-down list select a different quantitative method and then click **OK**.

2. When the quantitative type selected is either iTRAQ or TMT, the **and Correction to:** pull-down list and the **Edit Purity Correction** button become available.
 - Select a correction from the list if available or select **Other...** to open the [Edit iTRAQ/TMT Purity Corrections](#) dialog from where the user can create new purity corrections tables or edit existing ones. This dialog can also be reached by clicking the **Edit Purity Correction** button.

Edit iTRAQ/TMT Purity Corrections

Isotopic impurity of TMT or iTRAQ reagents can cause inaccurate calculation of the relative quantities of samples. Correction factors may be applied to adjust for this problem. The correction factors for iTRAQ 4-plex and 8-plex reagents generally do not vary from lot to lot, so Scaffold provides default purity correction tables which can be applied for these methods. Every batch of iTRAQ 16-plex or TMT reagents, however, contains different trace levels of isotopic impurities that need to be corrected. The correction factors, or purity values, are usually reported in a certificate of analysis that comes with the iTRAQ or TMT reagents kit. They indicate the percentages of each reporter ion that have masses differing by -2, -1, +1 and +2 Da from the nominal reporter ion mass due to isotopic variants.

Note: It is strongly recommended to add these correction factors into Scaffold.

The **Edit iTRAQ/TMT Corrections** dialog opens when the **Edit Purity Corrections** button or the **Other...** option in the **and Correction to:** pull-down list present in the [Edit Quantitative Samples](#) dialog are selected.

The dialog includes the Loaded Purity Corrections table which lists saved correction tables for the method selected in the **Change Type to:** control and a number of functional buttons appearing beneath the table:

- **New Correction** - Opens the dialog [Purity Corrections](#) so the user can define a new purity correction table.
- **Edit...** - Opens the dialog [Purity Corrections](#) with the currently selected purity correction table shown so the user can adjust the values already included in the table or add others.
- **Delete** - Deletes the selected entries from the table
- **Close** - Closes the dialog without applying the changes
- **Apply** - Chooses the selected Purity Correction table.

Purity Corrections

The **Purity Correction** dialog opens when selected from the [Edit iTRAQ/TMT Purity Corrections](#) dialog through the buttons **New Correction** and **Edit...**

Figure 4-29: Purity Correction dialog

	-2	-1	+0	+1	+2
iTRAQ-114	0	1	92.9	5.9	0.2
iTRAQ-115	0	2	92.3	5.6	0.1
iTRAQ-116	0	3	92.4	4.5	0.1
iTRAQ-117	0.1	4	92.3	3.5	0.1

The dialog contains a matrix into which the user may input the isotope correction factors for iTRAQ or TMT. TMT and iTRAQ corrections are reported differently. TMT concentrations are reported relative to the monoisotopic concentrations, while iTRAQ corrections are reported as a percentage of the total concentration. It is important to select the correct option in the **View relative to:** control.

When a new Purity Correction table is created the user must assign a name to the table by typing one in the **Name** text box located above the matrix. Whether editing an existing Purity Corrections table or creating a new one, clicking **Apply** finalizes the operation and closes the dialog.

Note: For more information about the way Scaffold calculates and applies iTRAQ corrections see the following publication: [Shadforth \(2005\)](#).

To enter iTRAQ 4-plex or 8-plex purity corrections:

Generally, it is not necessary to enter specific correction values for these reagents; the default corrections should be used. If, however, the user wishes to enter different correction values, a new purity correction may be created according to the following directions:

In the **View relative to:** control, select **Total concentration**.

Type the percentages for each iTRAQ reagent in the order listed in the Certificate of Analysis.

To enter TMT 2-plex, 6-plex, 10-plex or 11-plex purity corrections:

In the **View relative to:** control, select **Monoisotopic concentration**.

Type the percentages for each TMT reagent in the order listed in the Certificate of Analysis.

To enter TMT 16-plex purity corrections:

ThermoFisher Scientific™ provides purity correction tables for individual lots of TMTpro 16plex Label Reagents. Values from these tables should be entered into Scaffold as follows:

In the **View relative to:** control, select Monoisotopic concentration.

Enter only the values for which the adjusted masses correspond to other mass tags. These values are indicated in the table by the appearance of the corresponding mass tags in parentheses. These values should be entered into the -1 or +1 columns in the appropriate rows. A sample product data sheet with the values to be entered into Scaffold highlighted and the resulting correctly completed purity correction table in Scaffold are shown in the figures below.

Figure 4-30: Sample Product Data Sheet for TMTpro 16plex

ThermoFisher
SCIENTIFIC
The world leader
in serving science

PRODUCT DATA SHEET
Thermo Scientific™ TMTpro 16plex Label Reagent Set, 1 x 5mg

Product Number: A44520 Lot Number: UL297970

Form: TMTpro reagents are supplied dried, 5 mg/tube. Make a stock solution by reconstituting each tube with 200 µl dry acetonitrile.

Note: ** Reporter ion isotopic distributions (+1, +2) are primarily for natural carbon isotopes with reporter ion interference for each mass tag shown in parentheses. Incomplete stable isotope incorporation (-2, -1) are also reported for carbon and nitrogen isotopes. Reporter ion isotopic distributions can be used as isotope correction factors in TMTpro 16plex method template in Proteome Discoverer software version 2.3 and above.

****Reporter Ion Isotopic Distributions:**

Mass Tag	Reporter Ion Mass	-2x 13C	-13C -15N	-1	M+	+1	+2
TMTpro-126	126.127726	N/A%	N/A%	N/A%	100%	0.34%	9.31% (127C)
TMTpro-127N	127.124761	N/A%	N/A%	N/A%	100%	N/A%	8.98% (128N)
TMTpro-127C	127.131081	N/A%	N/A%	0.75% (126)	100%	0.33%	7.98% (128C)
TMTpro-128N	128.128116	N/A%	0.00%	0.95% (127N)	100%	N/A%	8.38% (129N)
TMTpro-128C	128.134436	0.00%	N/A%	1.47% (127C)	100%	0.34%	8.91% (129C)
TMTpro-129N	129.131471	0.00%	0.00%	1.46% (128N)	100%	N/A%	6.88% (130N)
TMTpro-129C	129.13779	0.51%	N/A%	2.74% (128C)	100%	0.36%	6.15% (130C)
TMTpro-130N	130.134825	0.49%	0.02%	2.76% (129N)	100%	N/A%	5.98% (131N)
TMTpro-130C	130.141145	0.04%	N/A%	3.10% (129C)	100%	0.42%	4.82% (131C)
TMTpro-131N	131.13818	0.04%	0.04%	3.09% (130N)	100%	N/A%	4.75% (132N)
TMTpro-131C	131.1445	0.06%	N/A%	3.81% (130C)	100%	0.40%	3.29% (132C)
TMTpro-132N	132.141535	0.04%	0.00%	2.84% (131N)	100%	N/A%	3.51% (133N)
TMTpro-132C	132.147855	0.10%	N/A%	4.14% (131C)	100%	0.40%	1.80% (133C)
TMTpro-133N	133.14489	0.36%	0.01%	3.64% (132N)	100%	N/A%	1.94% (134N)
TMTpro-133C	133.15121	0.88%	N/A%	4.70% (132C)	100%	0.40%	1.01% (134C)
TMTpro-134N	134.148245	0.40%	0.04%	4.92% (133N)	100%	N/A%	1.05% (135N)

Stability: Reagent is highly moisture sensitive. Product is guaranteed for one year from date of receipt in original, unopened packaging.

Storage: Store at -20°C

Figure 4-31: Entries in the Purity Correction dialog for TMTpro 16plex

Purity Correction

Name: TMT 16-Plex Purity Correction - lot UL297970

View relative to: Monoisotopic concentration

	-2	-1	+0	+1	+2
TMT-126	0	0	100	9.31	0
TMT-127N	0	0.71	100	8.98	0
TMT-127C	0	0.75	100	7.98	0
TMT-128N	0	0.95	100	8.38	0
TMT-128C	0	1.47	100	6.91	0
TMT-129N	0	1.46	100	6.86	0
TMT-129C	0	2.74	100	6.15	0
TMT-130N	0	2.76	100	5.98	0
TMT-130C	0	3.10	100	4.82	0
TMT-131N	0	3.09	100	4.75	0
TMT-131C	0	3.81	100	3.29	0
TMT-132N	0	2.84	100	3.51	0
TMT-132C	0	4.14	100	1.80	0
TMT-133N	0	3.64	100	1.94	0
TMT-133C	0	4.70	100	1.01	0
TMT-134N	0	4.92	100	1.05	0

Help Apply Cancel

Window

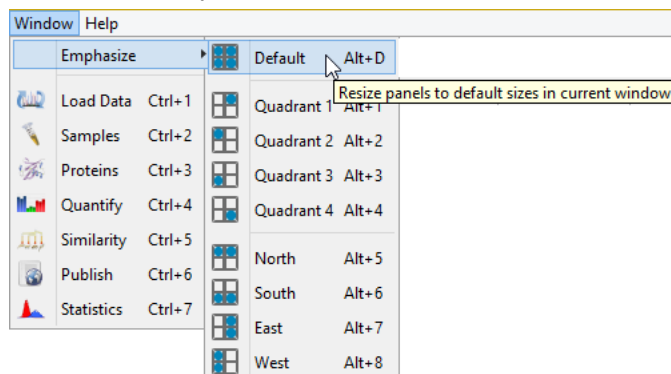
Through this menu the user can access the Emphasize window options or switch to a different Scaffold views, which is equivalent to clicking the buttons located in the [Navigation pane](#)

- Emphasize - see [Emphasize windows options](#).

Emphasize windows options

A number of Views are subdivided into panes that contain different tools to visualize and analyze the loaded data. When panes are present in a view the user can bring any of them to a prominent position in the Scaffold window by selecting one of the options available in the Emphasize sub menu or use the appropriate accelerator, see [Figure4-32](#).

Figure 4-32: Emphasize window options.



Once an option is selected it resizes the selected quadrant or cardinal section to fill the [Display pane](#).

Help

Referencing Scaffold

The user is free to copy, modify, and distribute the following examples for citing Scaffold in publications and reports.

Scaffold (Proteome Software, Inc., Portland, OR 97219, USA) was used to probabilistically validate protein identifications derived from MS/MS sequencing results using the X!Tandem ([Craig \(2003\)](#)) and ProteinProphet computer algorithms ([Nesvizhskii \(2003\)](#)).

Scaffold (Proteome Software, Inc., Portland, OR 97219, Oregon, USA) was used to validate protein identifications derived from MS/MS sequencing results. Scaffold verifies peptide identifications assigned by SEQUEST, Mascot or other search engines (list other search engines used to derive the imported data) using the X!Tandem database searching program ([Craig \(2003\)](#) and [Searle \(2008\)](#)). Scaffold then probabilistically validates these peptide identifications using PeptideProphet ([Keller \(2002\)](#)) and derives corresponding protein probabilities using ProteinProphet ([Nesvizhskii \(2003\)](#) and [Searle \(2010\)](#)).

IdentityE

Scaffold supports Waters' IdentityE (aka MS^E , aka hi-lo energy scanning). To be able to load data analyzed using PLGS into Scaffold, Proteome Software, in collaboration with Waters, developed a plug-in that comes with the PLGS installation. The plug-in specifically creates files compatible with Scaffold. Waters should have provided a Scaffold plug-in manual which guides the user through the Scaffold plug-in installation, but if this is not the case

there is a copy available on Proteome Software's website at [Scaffold4 PLGS plug-in](#).

The Scaffold plug-in only exports data that was searched in PLGS using MS^E.

Furthermore searches need to be run in PLGS with FDR set to 100% so that enough negative hits can be exported and available for Scaffold to be able to compute peptide and protein prophet probabilities in a statistically correct fashion.

When IdentityE data is loaded into Scaffold, an additional menu, IdentityE Menu, appears on the main menu bar after the Help menu. The menu provides options to configure absolute quantification.

Warning: Scaffold uses its own algorithms (Peptide/Protein Prophet, protein grouping) to determine both the list of proteins displayed in the Samples Table and their absolute quantities. While the intent is to reproduce Water's quantification strategy (top 3 peptides per protein), because of these algorithm differences, the list of proteins displayed and their quantities may differ somewhat from what's displayed in Protein Lynx. If you notice particularly large or confusing discrepancies, please do let us know.

Quantitation Option

Selecting the entry **IdentityE > Quantitation Option** opens the dialog **PLGS Quant Configuration**. The dialog contains:















- **Known Abundance Protein** pull-down list - The list is used to select among the proteins listed in the Samples table the protein that has a known abundance input the value and use it for quantitative purposes.
- **Use accession not name** check box - Used to toggle the way the proteins are shown in the above pull-down list.
- **How much?** text box - Used to input the quantitation normalization factor when the data is shown using weight or volume.
- **Select Unit for Showing Data** pull-down menu - Used to choose the units of measurement for the quantitation. The Default value is intensity, calculated using Water's quantitation strategy (top 3 peptides per protein).





Tool-bar

Figure 4-33: Scaffold Tool Bar



The Scaffold tool-bar contains icons that represent equivalent commands for frequently used main menu options.

Icon	Function
	New —Launches a Wizard which guides the user through the loading phase of the search data files in Scaffold. See Data files, whether raw files or search engine results, are loaded into Scaffold through the Scaffold Loading Wizard.
	Batch Job Queue Dialog - Only appears if Scaffold is installed with a Batch license. Opens a wizard that allows the user to generate and run Scaffold Batch experiments.
	Open —Opens a saved Scaffold experiment file, *.SF3, through a file browser.
	Save —Standard Windows behavior.
	Print —Prints the current view.
	Print Preview —Previews current view with the option to print the document.
	Copy —For each view copies to the clipboard the first table appearing at the top of the view. From there, the user can paste it into a third-party program such as Excel or Microsoft Word.
	Find —Opens a find dialog box that searches the first table present in the current view
	Excel —Exports the information that is contained in the current view to a tab-delimited text file that can be opened and viewed in Excel.
	BioSample Summarization level —See The BioSample Mode.
	MS/MS Sample Summarization level —See The MS/MS Sample Mode.
	Add BioSample —Not available in the Viewer version, it launches the Loading Wizard
	Queue Search Engine Results for Loading —Not available when running as a Viewer, see Queue Search Engine Results for Loading.
	Queue Raw Files for Loading —Not available when running as a Viewer, see Queue Raw Data for Loading

Icon	Function
	Load and Analyze Queue —Not available in the Viewer version and active only when there are files listed in the loading Queue in the Load Data View waiting to be loaded in Scaffold. When selected it opens the Load and Analyze Data page of the Loading Wizard.
	Quantitative Analysis —Opens the Quantitative Analysis... dialog
	Scaffold Q+/Scaffold Q+S —Available when running Scaffold Q+ or Scaffold Q+S, opens the Q+ Quantitation Browser window.
	Help —Opens the Scaffold Online Help.

Thresholding pane

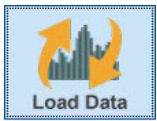
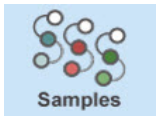
The Scaffold Thresholding pane, located on the right of the Tool-bar, contains filters and thresholding tools the user can adjust to increase or decrease the length of the displayed protein list in the Samples Table, see [Applying Thresholds](#).

Figure 4-34: Scaffold Thresholding Pane

Protein Threshold:	99.0% ▼	Min # Peptides:	2 ▼	Peptide Threshold:	95%
--------------------	---------	-----------------	-----	--------------------	-----


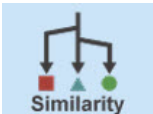
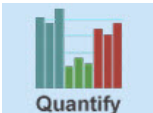


Navigation pane

The Scaffold Navigation pane is a vertical bar displayed on the left side of the Scaffold window. The bar contains buttons that toggle the seven different views available in the Scaffold main window.

	<ul style="list-style-type: none"> • See The Load Data View
	<ul style="list-style-type: none"> • See The Samples View

Chapter 4

Scaffold's Main Window

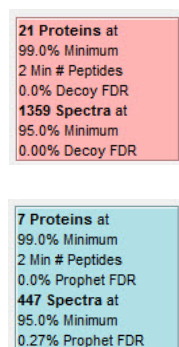
 <p>Proteins</p>	<ul style="list-style-type: none"> • See The Proteins View
 <p>Similarity</p>	<ul style="list-style-type: none"> • See The Similarity View
 <p>Quantify</p>	<ul style="list-style-type: none"> • See The Quantify View
 <p>Publish</p>	<ul style="list-style-type: none"> • See The Publish View
 <p>Statistics</p>	<ul style="list-style-type: none"> • See The Statistics View

FDR Dashboard

Scaffold calculates the False Discovery Rate (FDR) for both peptides and proteins and reports the values in the **FDR Dash Board** located underneath the navigation pane.

Protein and peptide FDR values are reported based on the specific protein and peptide thresholds selected in the [Thresholding pane](#).

Figure 4-35: FDR Info Box - Red background: searches run with decoy concatenated database; Blue background: searches run with target database



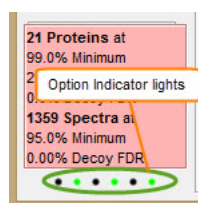
Depending on the type of database search that produced the data loaded into Scaffold, the FDR is calculated in the following ways:

- When the search is performed against a database without decoys, the FDR is calculated with proteins and peptides probabilities estimated using Peptide and ProteinProphet. The FDR dashboard where the values are reported appears with a blue background.
- When the search is performed against a decoy or reversed concatenated database, or when a decoys are generated by the search engine, the FDR is calculated using the count of decoys against target identification hits. If proteins and/or peptides are filtered based on FDR, then the dashboard reports the specific protein and peptide thresholds necessary to reach those specific FDR values based on the FDR Browser landscape. The FDR box where the values are reported appears with a red background.

Option Indicator Lights

The Option Indicator Lights are six multi colored dots located at the bottom of the Navigation Pane underneath the [FDR Dashboard](#) in the Scaffold Main window.

Figure 4-36: Option Indicator Lights



Their scope is to remind the user about the status of the following options:

- *View Menu Options* (green when selected):
 - Show less <5% probability
 - Show lower Scoring Matches

Chapter 4

Scaffold's Main Window

- Show entire protein Clusters
- *Load and Analyze Options* (green when selected):
 - Use Protein Cluster Analysis
 - Use Independent Sample Grouping strategy
- *Scoring Scheme*:
 - LFDR-green
 - PeptideProphet with Delta Mass correction - orange
 - PeptideProphet no mass correction -black

The user can always hover over each one of the colored dots to check their function; a tool tip appears providing a description of the selected dot.

Display pane

The information included in the different views appears in the Scaffold Display pane. Depending on the view, the type of information reported might appear framed in one or more tables or graphs included in one or more sub-panes. All panes and tables included in Scaffold share the following characteristics:

- [Tool-tips](#)
- [Resizing of columns and panes](#)
- [Moving columns](#)
- [Column sorting](#)
- [Multi selection of rows in the Samples Table](#)
- [Mouse Right Click Context Menus](#)
- [Graph Features](#)

Tool-tips

The user can view information about fields or columns in a View by just hovering the mouse pointer over the location of interest. This operation opens a collapsed tool-tip. Pressing F2 opens an expanded tool-tip. Pressing the Escape (ESC) key on the keyboard closes the expanded tool-tip,

Figure 4-37: Viewing information in a collapsed tool-tip

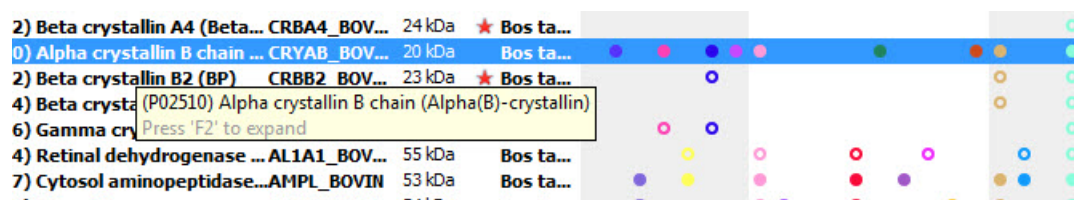
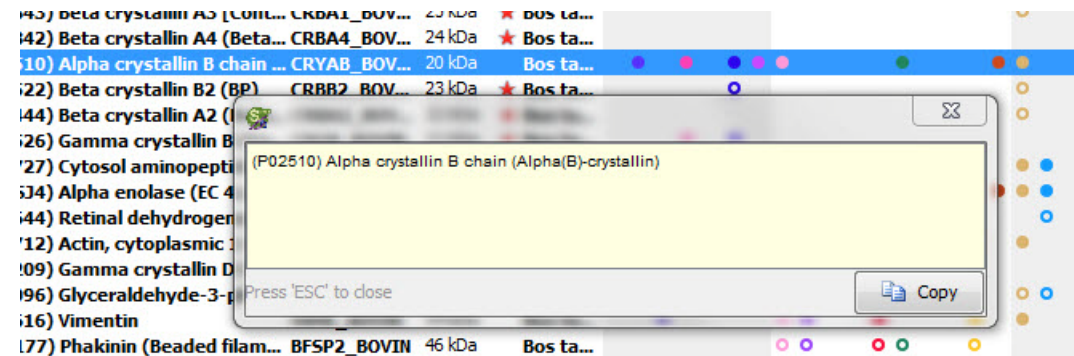


Figure 4-38: Viewing information in an expanded tool-tip



Resizing of columns and panes

The user can resize columns and different panes in each of the views to better suit his/her

working needs. For example, in [The Samples Table](#), the user can change the width of a column by resting the mouse pointer on the right side of a column heading until the pointer changes to a double-headed arrow, and then dragging the boundary until the column is the width that he or she wants.

Figure 4-39: Changing the width of a column in the Samples View



Moving columns

In all tables throughout Scaffold, but the Samples Table, every column can be moved around from one position to another for more comfortable access to the data that is summarized in them.

The user simply has to click on the header of the column that he/she desire to move and drag it to the location where he/she wants to place it. Switching to another view will keep the columns in the new positions.

Figure 4-40: Moving columns around in tables

Valid	...	Sequence	Prob	SEQU	NTT	Modifications
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(-)AEQHSTPEQAAAGK(S)	100%	0.43	70	2 Acetyl (+42)
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(-)AEQHSTPEQAAAGK(S)	100%	0.43	77	2 Acetyl (+42)
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(-)AEQHSTPEQAAAGK(S)	100%	0.42	82	2 Acetyl (+42)
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(-)AEQHSTPEQAAAGK(S)	100%	0.47	00	2 Acetyl (+42)
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(-)AEQHSTPEQAAAGK(S)	100%	0.33	26	2 Acetyl (+42)
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(-)AEQHSTPEQAAAGKSHGGLGGSYK	100%	0.35	37	2 Acetyl (+42)
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)SHGGLGGSYK(V)	100%	0.41	11	2
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)SHGGLGGSYK(V)	100%	0.32	28	2
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)SHGGLGGSYK(V)	100%	0.35	64	2
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)SHGGLGGSYK(V)	100%	0.26	31	2
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)SHGGLGGSYK(V)	100%	0.41	10	2
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)SHGGLGGSYK(V)	100%	0.37	30	2
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)SHGGLGGSYK(V)	100%	0.36	04	2
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)SHGGLGGSYK(V)	100%	0.40	26	2
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)SHGGLGGSYK(V)	100%	0.41	11	2

Column sorting

In all tables throughout Scaffold, the user can use the tri-state column sorting feature and sort the display by clicking on any column header. For example, to sort the proteins based on increasing molecular weight, the user can click the Molecular Weight column header once. To sort the proteins based on decreasing molecular weight, the user can click the Molecular Weight column header twice. To return to the default display, the user can click the Molecular Weight column header a third time.

Multi selection of rows in the Samples Table

In the Samples table the user can select multiple rows by using either the SHIFT or the

CTRL key, depending whether the desired selection has contiguous rows or not, and the click of the mouse in a pretty standard fashion. Other functions can then be applied, like assigning a star to the selected group of proteins in the Samples table, for example.

Figure 4-41: Rows multi-selection in the Samples View

#	Visible?	Bio View: Identified Proteins (31/32)	Accession Number	Molecular Weight	Protein Grouping Ambiguity	Taxonomy	Biological Process										Cellular Component										Molecular Function										M	C																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																						
							biological adhesion	biological regulation	cellular process	developmental process	metabolic process	mucellular organismal process	response to stimulus	cytoplasm	endosome	extracellular region	extracellular organelle	intracellular organelle	membrane	mitochondrion	nucleus	organelle part	plasma membrane	landing	catalytic regulator	enzyme regulator activity	molecular function	structural molecule activity	translation regulator activity																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																															
		Probability Legend:																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																										

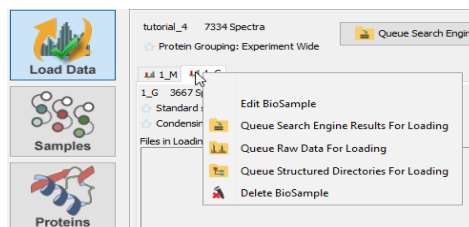
Mouse Right Click Context Menus

When the user right clicks the mouse while hovering over the Display Pane, a menu with various options appears close to the working arrow. Depending on the current view the list of options available in the menu varies. A description of the mouse right click commands is provided in [“Context Menus Right Click Commands” on page 271](#)

Load Data View

The following menu appears when the user right clicks on the BioSample name tab.

Right Click Menu A:



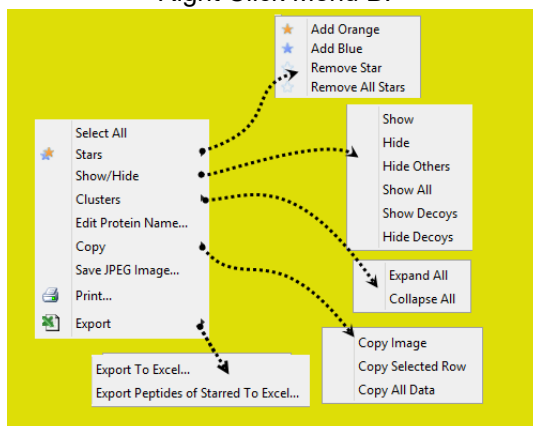
Samples View

When the user right clicks anywhere over the list of proteins the following menu appears which contains a number of sub-menus:

Chapter 4

Scaffold's Main Window

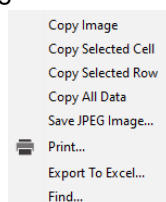
Right Click Menu B:



Proteins View

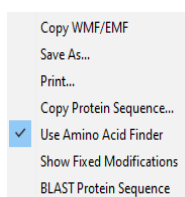
When the user right clicks over the Proteins View generally [Right Click Menu C](#) appears, but depending on the selected tab a different menu might be available.

Right Click Menu C:



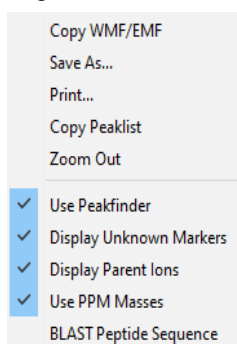
- Protein Sequence tab - Available right click menu:

Right Click Menu D:



- Spectrum Tab and Spectrum/Model Error - Available right click menu:

Right Click Menu E:



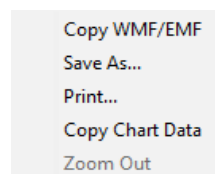
Similarity View

- Grouping Table - Available right click menu: [Right Click Menu C](#)
- Identifications Tab and Fragmentation Table tab - Available right click menu: [Right Click Menu C](#)
- Spectrum and Spectrum/Model Error tab - Available right click menu: [Right Click Menu E](#)

Quantify View

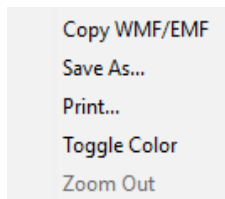
- [The Quantitative Value pane](#) available right click menu:

Right Click Menu F:



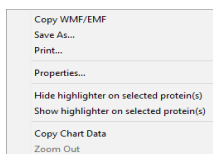
- [The Quantitative Scatterplots pane](#) available right click menu:

Right Click Menu G:



- [The Venn Diagrams pane](#) - Available right click menu:

Right Click Menu H:



When clicking on a Venn diagram set a list of the proteins or unique peptides or spectra will appear. When mousing on the list and right clicking on it [Right Click Menu C](#) appears.

- Gene Ontology Terms Pane

When mousing over in any of the tabs available in this pane [Right Click Menu C](#) appears.

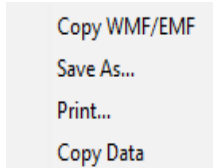
Publish View

When mousing over the Experiment Methods tab [Right Click Menu C](#) appears.

Statistics View

When mousing over the MS/MS samples table [Right Click Menu C](#) appears. When mousing over the various graphs found in the different panes [Right Click Menu F](#) appears. When mousing over the FDR browser the available right click menu is:

Right Click Menu I:



[Display pane](#)

Graph Features

Every graph appearing in any of the Scaffold's view contain the following tools:

- **Zoom Function** - Zooming in within a graph is done through holding down the left mouse button and dragging the pointer from left to right. A box, drawn from the upper left hand corner of the graph towards the lower right hand corner, is formed and highlighted in gray around the area that, after releasing the mouse button, is being enlarged for viewing within the graph plotting area. A single click of the mouse zooms out the image to the previous magnification. Clicking various times returns the graph to the initial 100% magnification.

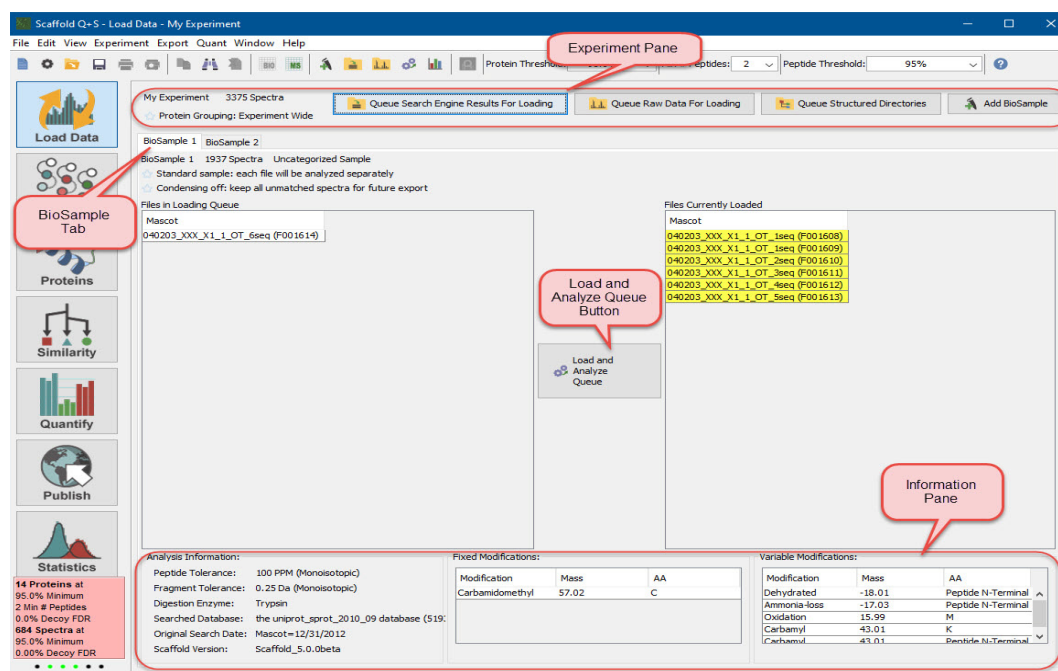
Context menu - The user can right-click on a graph to open a context menu. The type of context menu might depend on the view where the graph appears, check [Mouse Right Click Context Menus](#) and [Context Menus Right Click Commands](#) for more information.

Chapter 5

Load Data View

Scaffold's Load Data View provides an overview of the currently opened experiment along with tools for loading additional MS files or deleting them or adding or deleting BioSamples. Through this view the user can view the list of files loaded into each BioSample; add or delete BioSamples or MS samples; check their search parameters including the fixed and variable modifications or edit BioSample information.

Figure 5-1: The Load Data View



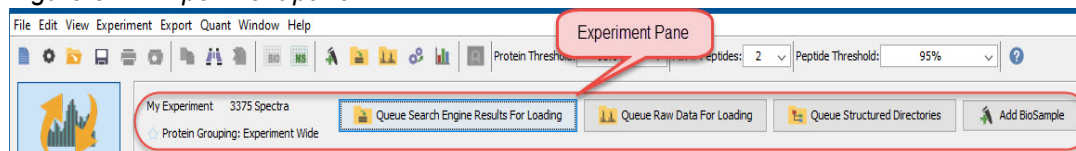
The elements that constitute the Samples View are:

- “[Experiment Pane](#)” on [page 124](#), which provides general information about the currently loaded experiment, and tools to add MS samples to a BioSample or create a new BioSample within the current experiment.
- “[BioSample tabs](#)” on [page 130](#), which contain the lists of already loaded MS files or files in queue for each BioSample together with specific loading information.
- “[Information pane](#)” on [page 131](#), which provides specific in depth information about the search files loaded in a specific BioSample.

Experiment Pane

The Experiment Pane provides general information about the currently loaded Scaffold experiment.

Figure 5-2: Experiment pane



On the left side of the pane Scaffold shows the name of the current experiment, which by default is called My Experiment, the total number of spectra currently loaded in the experiment and the type of grouping selected at the time of loading.

There are two types of grouping modes implemented in Scaffold, that can be selected when loading data:

- **Experiment wide** - Scaffold groups proteins across all MS samples and BioSamples.
- **Independent sample** - Scaffold groups proteins only within each MS sample. Each MS sample appears as if it was loaded independently.

In the top right portion of the pane, not appearing in the Viewer mode, there are four buttons:

- [Queue Search Engine Results for Loading](#) - Adds more MS samples to the selected BioSample
-
- [Queue Structured Directory for Loading](#) - Adds more MS samples organized in separate directories.
- **Add Biological Sample** - Adds a new BioSample to the experiment by starting the Loading Wizard.



*Files can also be added to a BioSample directly from the Mascot Server. The user can do so by selecting a BioSample and going to the menu option **Experiment** > **Queue Files from Mascot Server**, see [Queue Files From Mascot Server for Loading....](#)*

Queue Search Engine Results for Loading

Selecting this command opens a standard file browser. From there the user can navigate to the location where the data files to be loaded in Scaffold are stored. Once the files are selected Scaffold places them in the Loading Queue

The **Queue Files For Loading** command can be selected from the following locations in the program:

- The Experiment menu

- The Load data View
- The [Queue files for loading](#) page in the Wizard

The user must be able to reach the files of interest from the system where Scaffold is installed and should make sure that the format of the data files is supported by Scaffold by consulting the [Scaffold File Compatibility Matrix](#).

When this option is selected outside of the Loading Wizard with multiple biological samples already defined, the user should carefully select the right biological sample in the Load Data view before beginning to queue files.

Multiple files can be selected at the same time as long as the samples were digested with the same enzyme, the raw files were processed with the same peak-picking software, and they were searched against the same FASTA database. The user is asked to specify that database in the [Load and Analyze Data](#) page of the Wizard.

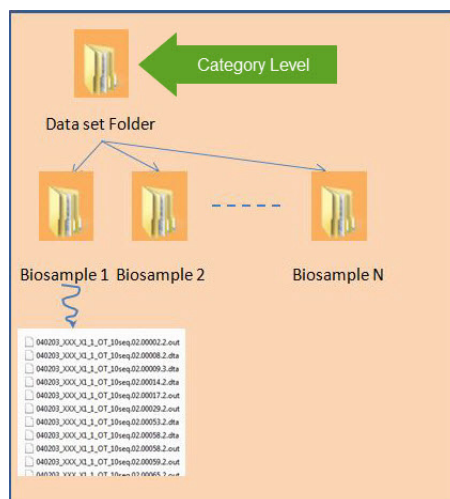
Queue Raw Data for Loading

Selecting this command opens a standard file browser. From there the user can navigate to a location containing raw data files to be loaded into the selected BioSample. Once the files are selected Scaffold places them in the Loading Queue and they can be searched using MSFragger.

Queue Structured Directory for Loading

This command allows the user to streamline the loading of groups of data files already organized into directories and subdirectories by BioSample into a single category. More categories may be added to the queue by repeatedly using the Queue Structured Directories option. The organization of the directories should be similar to the one depicted in [Figure 5-3](#). After files are queued by this method, the user must click the Load and Analyze button to actually load the data.

Figure 5-3: Organization of structured directories



Clicking the button **Queue Structured Directories** opens a file browser. The points Scaffold to the top level of the structured directory containing the data files to be loaded in a specific Category. Scaffold loads all the MS files found in one of the second level folders in a single BioSample, even when the files are contained in sub-sub-folders.

Once the top level folder is selected a dialog opens asking whether the files to be loaded will be search engine results or raw files. Next, the user is asked to define the name of the category. If Scaffold Q+ or Scaffold Q+Sis running, the dialog also asks the user to define the type of quantitation to perform with the data.

Figure 5-4: Choosing the file type

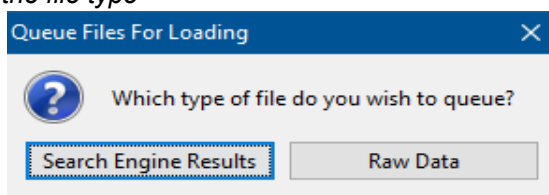
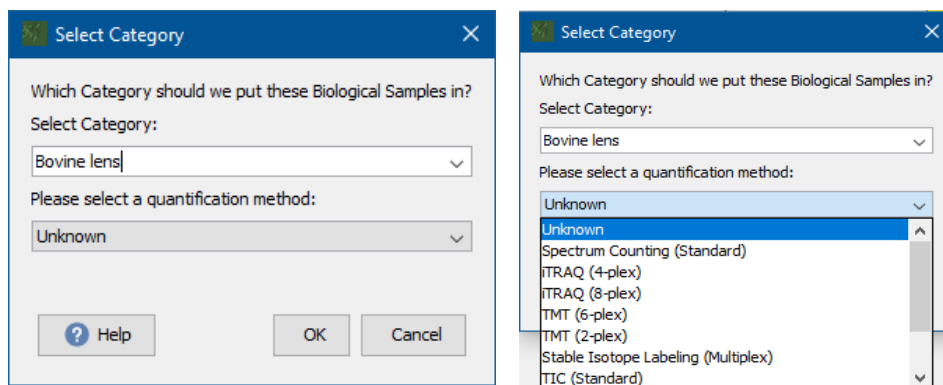


Figure 5-5: Defining Categories and quantitation methods



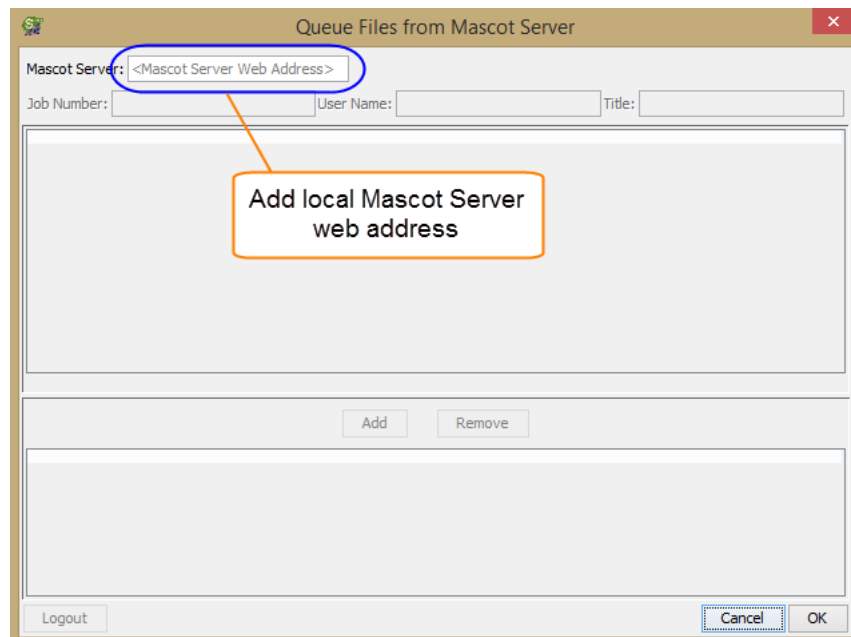
If the data is already organized by category into different folders, the user can load one category at a time by pointing Scaffold to one specific category folder. All of the sub-folders in the selected folder are assigned to biosamples with names corresponding to the sub-folder names and are grouped under that particular category.

Queue Files From Mascot Server for Loading...

The user can open this dialog either from the **Scaffold Wizard** on the Queue Files for Loading page or from the menu item **Experiment > Queue Files From Mascot Server for Loading**. The dialog contains tools to connect Scaffold to a Mascot Server, select and download searched data files directly from there into Scaffold.

When calling the dialog from the menu item **Experiment > Queue Files From Mascot Server for Loading** the user should make sure to have chosen the appropriate Bio/MS Sample before adding data.

Figure 5-6: Connecting to the Mascot Server from Scaffold

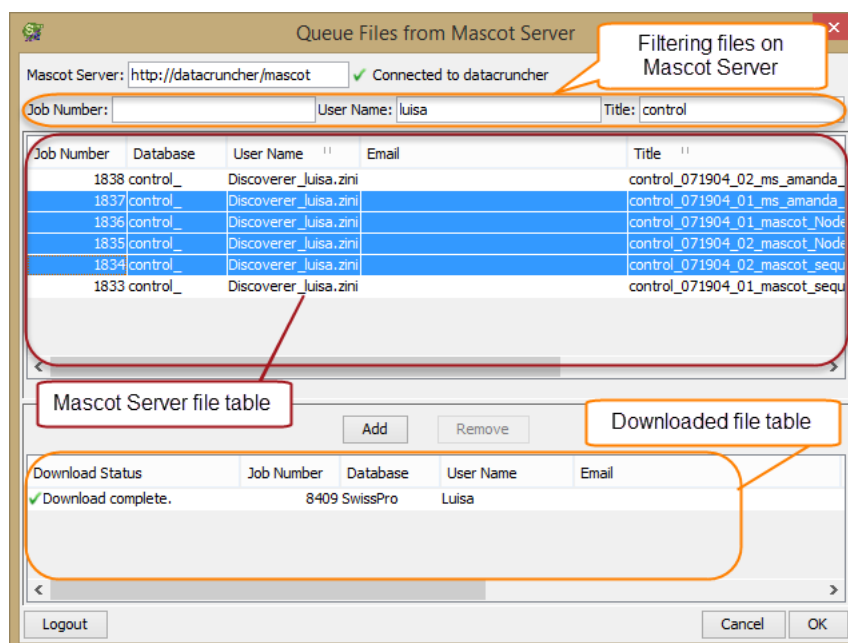


The first time this dialog is opened, the user must connect Scaffold to the local Mascot server. This is done by placing the web address of the local Mascot Server in the Mascot Server text box located in the top left corner of the dialog. If no security is implemented, Scaffold connects directly to the Mascot Server showing a list of files available for download.

If security is enabled on the Mascot server, a login window pops up asking for account name and password. The user should make sure that the specified account has administrative privileges in Mascot.

Note: **Edit > Preferences > Mascot Server** allows the user to create a default connection to a Mascot Server of choice. Scaffold automatically logs in to the server specified in the settings.

Figure 5-7: Queue files from Mascot Server for Loading



The dialog can be divided into 5 different panes containing a number of tools to help the user smoothly select and load data files into Scaffold.

- **Connection and filtering pane** - which contains information about the status of the connection to the Mascot Server spelled out in the **Mascot Server:** address text box. When the server is not connected this is the only text box available in the dialog, see [Figure 5-6](#). Below the connection information there are three different filters that can be applied to the data files shown in the Mascot Server file table. This helps the user quickly locate the files he/she wants to load into Scaffold. The available filter are:
 - *Job number*
 - *User name*
 - *Title*
- **Mascot Server file table**- When connected, the table lists the search data files saved on the server. The table shows the typical functionalities described in the [Display pane](#) section plus it accepts bulk operations like the standard windows multiple selection of files. The filtering pane acts on this table so that the user can easily locate the files he/she wants to load in Scaffold.
- **Action pane** - Contains two buttons. The **Add** button: active only when there are files selected from the **Mascot file table**. It starts the download of the chosen data files from the Mascot server to the computer where Scaffold is running and adds them to the loading queue in the Load Data View. The **Delete** button: active only when there are files selected in the Download file table. When clicked it deletes the highlighted files.

- **Downloaded file table** - The table lists the files downloaded from the Mascot Server to the computer where Scaffold is running. The table shows the typical functionalities described in the [Display pane](#) section plus it accepts bulk operations like the standard windows multiple selection of files. The status of the download is reported in the Download status column when completed a green check appears.
- **Completion pane** - Contains three buttons. The **Logout** button: the user can use this button to logout of the current Mascot Server and login to another one. The **Cancel** button: standard Windows functionality. The **OK** button: to finalize the lading of the downloaded files.

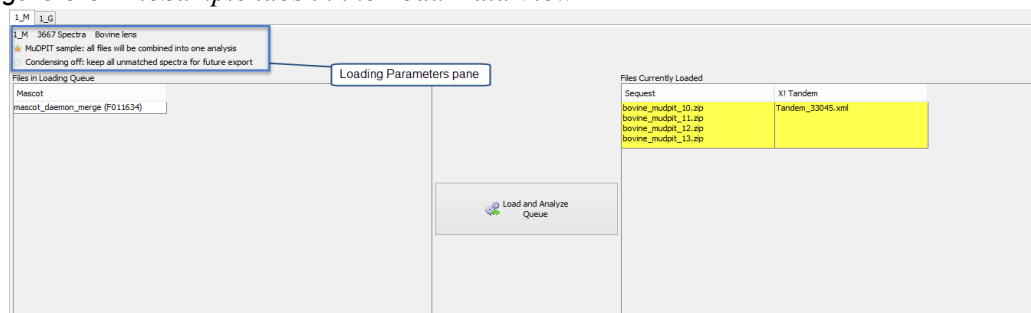


*If the user is logged into the Mascot Server as a Guest, Scaffold will not accept to download files and will show an error. When this happens, to be able to access the Mascot login window again, the user has to clear the address in the **Mascot Server:** address text box and then press enter. The Mascot Login dialog opens allowing the user to login with a different account.*

BioSample tabs

In the Samples View, each BioSample defined in the currently opened experiment, has a specific tab window assigned to it. The tab window is labeled with the same name as the BioSample and contains information about the loading status of the experiment, the MS experiments loaded or about to be loaded into the BioSample and which option they were chosen at the time of the load.

Figure 5-8: BioSample tabs in the Load Data View



The elements included in the BioSample window are the following:

- **Loading Parameters pane** - The information included in this pane reports the name of the BioSample, the number of spectra loaded and the name of the category. Underneath, it reports the settings selected during the loading phase on the New BioSample page of the Wizard. These include whether standard (separate processing for each MS sample) or MuDPIT (combined analysis for all samples) was applied and whether or not the samples were loaded using the condensing option. The user can change these settings by editing the BioSample, which can be done from the Experiment menu or by right clicking the BioSample tab.
- **Files in Loading Queue table** - Lists the files ready to be loaded in Scaffold. If the user has selected files from more than two search engines, he/she needs to scroll towards the right side of the table to see all the files.
- **Files Currently Loaded** - Lists the files already loaded in Scaffold. The loaded files are highlighted in yellow when the Scaffold analysis is completed. After analysis, files from the same MS sample run through multiple search engines, are aligned on a single row. Hovering the cursor over a file shows its full name and the database loaded.
- **Load and Analyze Queue button** - Opens the Load and Analyze Data page of the Loading Wizard, see [Load and Analyze Data](#).

The user can modify the name of the BioSample and categories by opening the Edit BioSample dialog. This dialog can be reached through the Experiment menu or by right-clicking the mouse over the BioSample tab, see Samples View in the [Context Menus Right Click Commands](#) section.

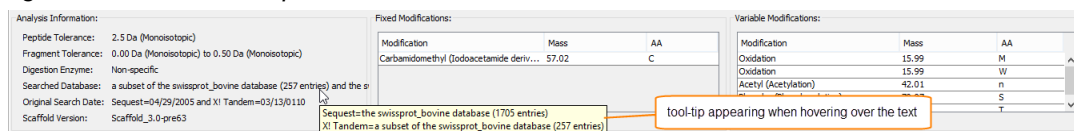
Note: The user can delete files before analysis, or MS samples after analysis, by right-clicking and selecting Remove Selected Samples.

Information pane

The bottom section of the Load Data view contains three information panes. Each of them provides information related to how the loaded data was analyzed by the search engine:

- [The Analysis Information pane](#)
- [The Fixed Modifications pane](#)
- [The Variable Modifications pane](#)

Figure 5-9: Information panes in the Load Data view



The information in these panes describes the data contained in the files listed in **Files Currently Loaded**. If specific files are highlighted in this table, then the information is restricted to the highlighted files. Otherwise it describes all files in the displayed BioSample.

In contrast to these information panes which describe only one sample at the time, the Publish view summarizes this analysis information for all the samples.

The Analysis Information pane

The Analysis Information pane lists the peptide and fragment mass tolerances, the digestion enzyme, the database searched and when that search was done. The Scaffold version is the version into which the data was loaded. If several files are selected and these files have different parameters, this box shows the range of values. When the cursor hovers over the data, a tool-tip shows further details.

The Fixed Modifications pane

This pane contains a table listing the fixed modifications, with their masses and the related modified amino acids, used during the searches in the loaded files belonging to a specific BioSample.

The Variable Modifications pane

This pane contains a table listing the variable modifications, with their masses and related modified amino acids, used during the searches belonging to a specific BioSample.

Note: When a peptide starts with E or Q, X!Tandem automatically checks for the formation of pyroglutamic acid, i.e., the loss of water or ammonia, respectively. The Pyro-Glu modification then automatically appears in the table, see [Analyze with MSFragger or X!Tandem Pane](#).



*The fixed and variable modifications are those used in the searches represented by the files in the **Files Currently Loaded** list. If different modification sets were used in different files, the modifications that were not searched for in all files are highlighted in red. The user can see to which files a red modification applies by hovering over it.*

Chapter 6

Samples View

Scaffold's Samples View offers overviews and tools to help the user make direct comparisons among BioSamples, MS Samples and Categories regarding identified protein probabilities and quantitative differences.

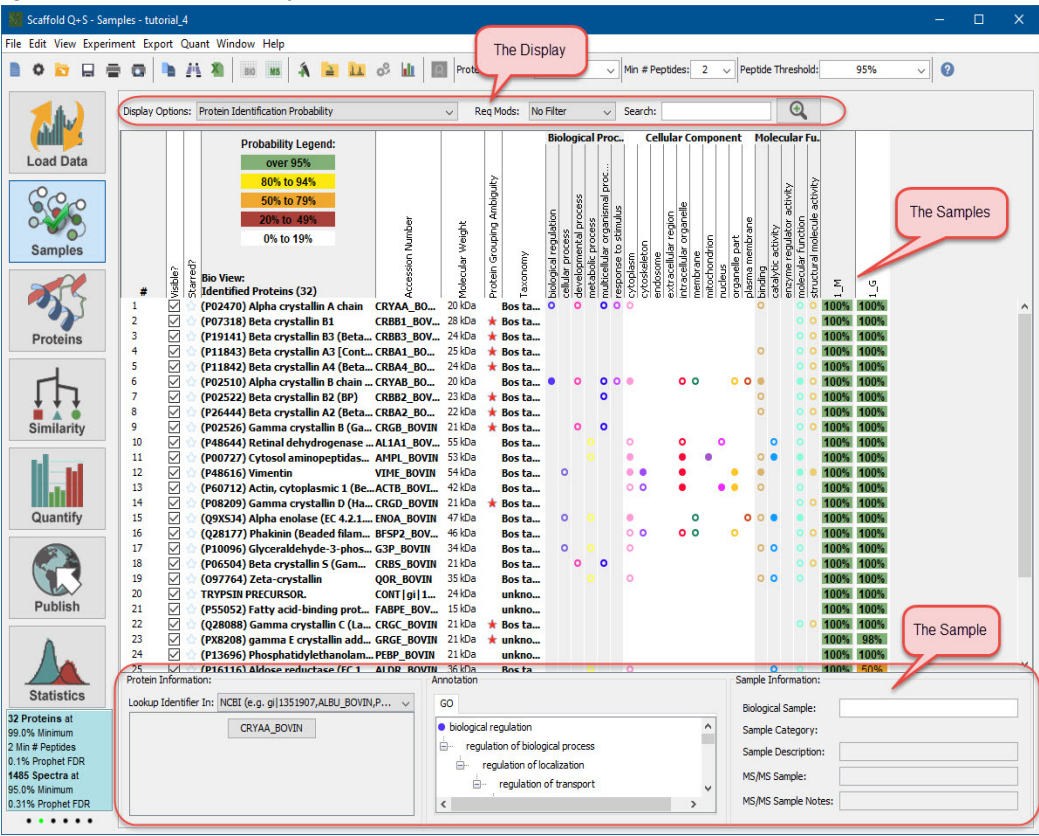
This chapter provides a detailed description of the features and tools available in this view:

- [“The Samples View is composed of the following elements:” on page 133](#)

The Samples View is composed of the following elements:

- [The Samples Table](#), which displays a summary of the experiment results.
- [Applying Thresholds](#), which describes how to increase or decreases the number of proteins listed in the Samples Table.
- [The Display pane](#), which provides options to view estimates of differential expression. Scaffold provides a multitude of statistical methods to filter results and provides both simple and advanced search options.
- [Information Panes](#), which specify useful protein, Gene Ontology, Pathway and sample information.

Figure 6-1: Scaffold Samples View



The Samples Table

The Samples Table provides a list of the protein groups and clusters identified in an experiment together with quantitative and qualitative information distributed among the different MS Samples or BioSamples and defined categories. It can be described as a collection of [Frequency Tables](#), one for each of the identified protein groups or protein clusters visible in the table. This type of table is typically referred to in statistics as a [Contingency table](#).

In the Samples table the list of identified protein groups and clusters appears as a series of rows while the various BioSamples or MS Samples appear as columns. The first seven columns in the table provide basic information about the identified proteins like their names, accession numbers and molecular weights together with columns containing tools to tag or manipulate the list like tagging [Proteins of Interest](#), select [Hidden Proteins](#) or check [Protein Grouping Ambiguity](#). When Quantitative analysis is selected other columns are added to the table just before the BioSamples or Ms Samples columns. Columns are also added when GO annotation are searched through the command [Add or Edit Annotations](#).

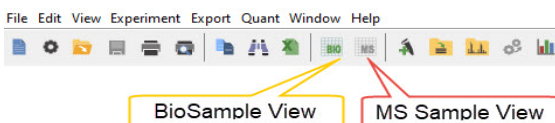
When the Samples View first opens, all protein groups that meet the default threshold settings are listed ordered by default as described in [Sorting](#).

The user may select from two levels of summarization at which to view the results in the Samples Table:

- [The BioSample Mode](#), which provides a single column overview of all the proteins groups or clusters in a given BioSample.
- [The MS/MS Sample Mode](#), which displays protein identifications in separate columns by mass spectrometry samples

The two summarization views can be toggled using the BIO and MS buttons located underneath the main menu bar.

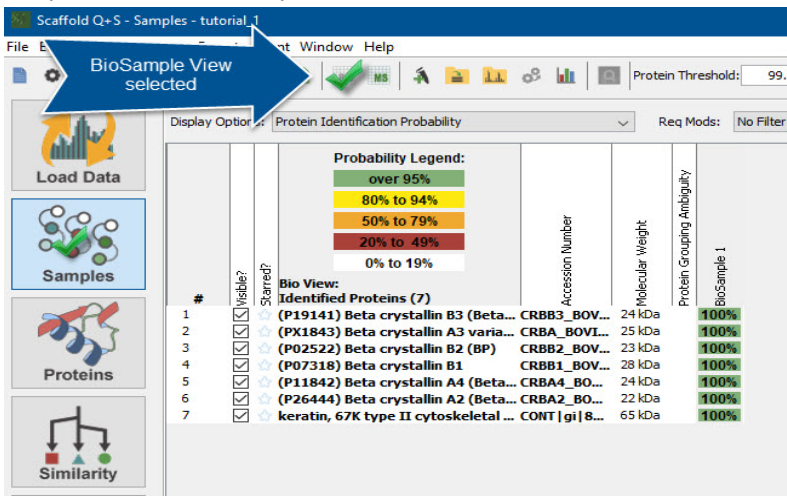
Figure 6-2: Scaffold Samples View - BioSample View/MS Sample View toggle buttons



The BioSample Mode

In this mode, Scaffold combines values from all MS Samples in each BioSample, summarizing the results at the BioSample level. Each [BioSample](#) is represented by a column. The columns are sorted first by category and then by BioSample name.

Figure 6-3: Samples View - BioSample View

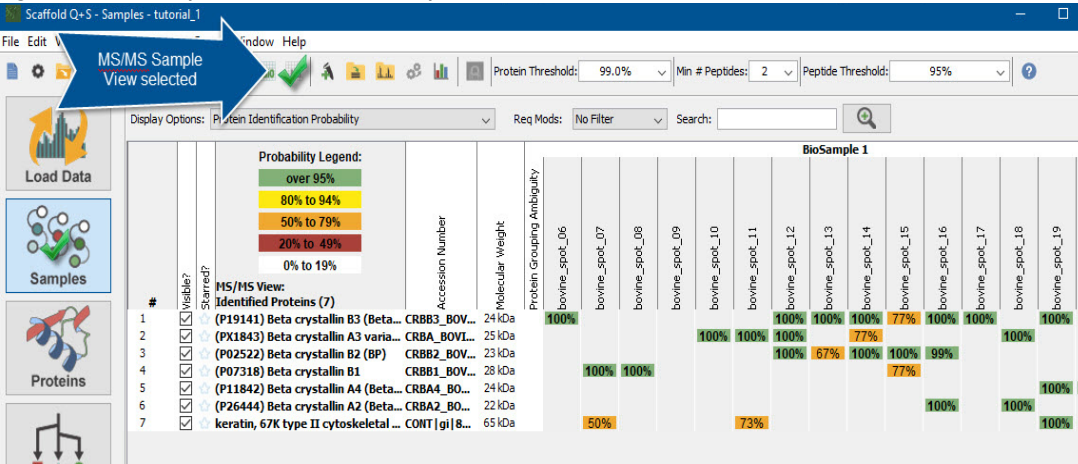


The MS/MS Sample Mode

When this mode is selected, the Samples Table shows one column for each MS Sample, sorted first by category, then by BioSample and finally by MS Sample. It is useful, for example, for analyzing samples processed with gels.

Scaffold displays a row in the table for each protein group or protein cluster which has at least one MS Sample identified that passes the selected filter and thresholds requirements. To display values for these proteins in MS Samples that do not meet the confidence requirements, the user can select [Show Lower Scoring Matches](#) from the View Menu.

Figure 6-4: Samples View - MS Samples View



MS Sample vs BioSample summarization levels

Data associated with a BioSample might come from a sample taken by a doctor, medical researcher, or biologist, such as a drop of blood or tissue from an organism.

Using such techniques as 2D gels or liquid chromatography, proteins or peptides from these BioSamples are then separated from each other. Each resulting individual band, spot, or LC fraction then processed by a mass spectrometer is one mass spectrometry sample (abbreviated as MS sample).

One BioSample is therefore typically made up of more than one MS sample - sometimes many more.

Scaffold allows the user to display results in either [The BioSample Mode](#) or [The MS/MS Sample Mode](#).

Protein List

To simplify the inspection of the identified proteins, Scaffold aggregates the protein list using two levels of hierarchy.

- **Protein Group** - a group of proteins with identical sets of peptides.

In the Protein list, the protein groups are always displayed collapsed. The number of proteins in the group minus 1 is shown in parenthesis close to the accession number of the protein representing the group, see B in [Figure 6-5](#). By default, the protein that has the highest probability and the highest number of associated spectra will represent the group in the list.

When the accession number for a protein group is clicked, a pull-down list becomes available. From this list, the user can select a different protein to represent the group in the protein list, see A in [Figure 6-5](#). It is also possible to change the default selection of proteins representing the groups by going to [Apply Protein Annotation Preferences](#).

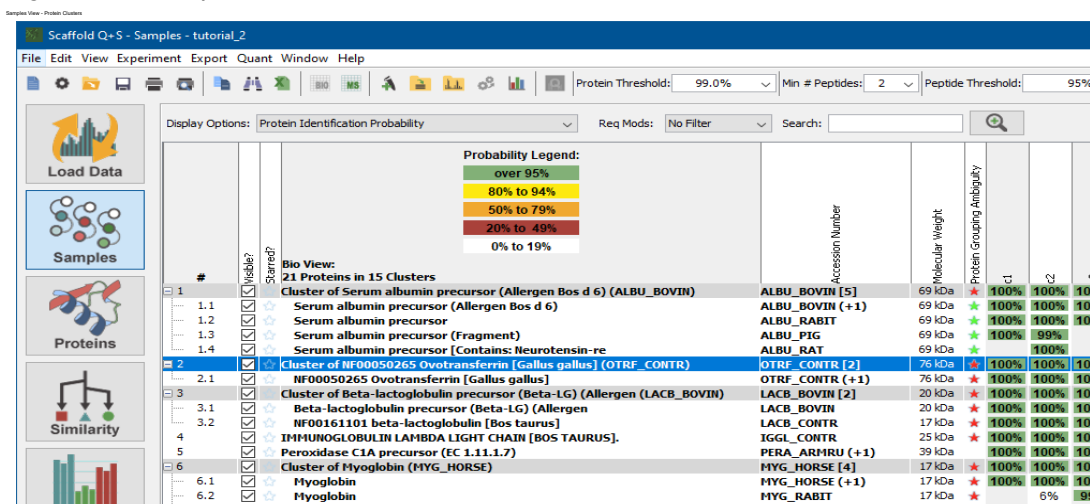
The proteins in the group are also listed in the [Protein Information pane](#) represented as buttons labeled with each protein's accession number. Clicking one of the buttons allows the user to gather further information by looking up the proteins on specific websites.

Figure 6-5: Samples View - A : Pull-down list of proteins in a group or cluster. B: Regular appearance of a group. C: Alternate ID (often a Gene ID or Gene Name). D: Pull-down list for selection of a website. E: Button to look up the protein on the selected website. F: Button to look up alternate ID on selected website.

The screenshot displays the Scaffold Samples View interface. At the top, there are controls for 'Display Options' (set to 'Protein Identification Probability'), 'Req Mods' (set to 'No Filter'), and a 'Search' field. A 'Probability Legend' is shown with color-coded bars for identification probability ranges: over 95% (green), 80% to 94% (yellow), 50% to 79% (orange), 20% to 49% (red), and 0% to 19% (white). The main table lists protein clusters with columns for '#', 'Visible?', 'Starred?', 'Bio View', 'Accession Number', 'Alternate ID', 'Molecular Weight', 'Protein Group', 'Taxonomy', and 'Biological Process'. A cluster of 32 proteins in 31 clusters is highlighted. Below the main table, a detailed view of a selected cluster (P08209) is shown, including a 'Protein Identification' section with a 'Lookup' button (E) and a 'Gene Ontology' section with a 'Lookup' button (F). The 'Protein Identification' section shows a pull-down list (A) for selecting a protein group (CRYGD_BOVIN) and a button (E) to look up the protein on the selected website. The 'Gene Ontology' section shows a pull-down list (D) for selecting a website (NCBI) and a button (F) to look up the alternate ID on the selected website. The 'Biological Process' column shows various processes like 'molecular function', 'structural molecule activity', and 'structural constituent of eye lens'.

- **Protein Cluster** - a set of protein groups created using a hierarchical clustering algorithm. The clustering algorithm is similar to the one used by Mascot to create protein families, but with more stringent grouping rules. Members of the cluster share some, but not all peptides. Protein clusters are by default represented by the protein group that shows the highest associated probability. Clusters can be collapsed or expanded directly in the protein list, see Figure 6-6. For more information about Scaffold clusters see Chapter 12, “Protein Grouping and Clustering,” on page 203.

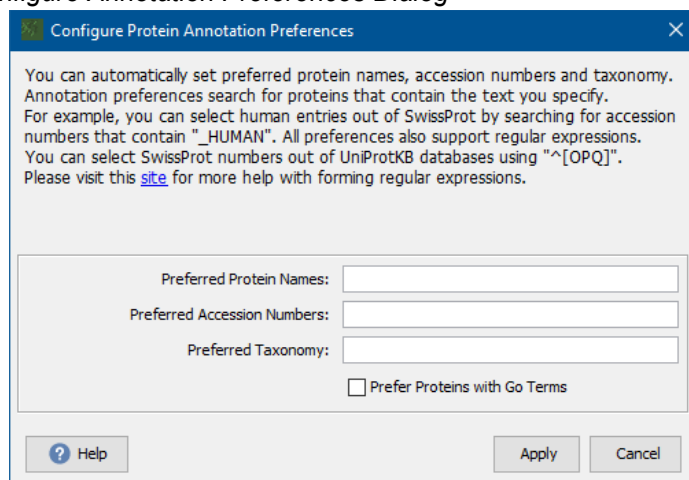
Figure 6-6: Samples View - Protein clusters



Apply Protein Annotation Preferences

The menu option **Experiment > Apply Protein Annotation Preferences** opens the dialog **Configure Protein Annotation Preferences** where the user can globally define which protein in a protein group is visible in the protein list in the Samples Table.

Figure 6-7: Configure Annotation Preferences Dialog



The dialog provides a number of text boxes through which the user can input text. Proteins that contain the specified text string will then be given preference in choosing the names of protein groups or clusters. If a text string entered is a regular expression, proteins that match the regular expression are given preference.

A checkbox also allows the user to select the option to give preference to proteins with associated GO annotations. Only the annotations associated with the visible protein name will be shown in the Samples View.

By default Scaffold automatically selects the visible protein by relying on the following five criteria in the order shown below:

1. Prefer proteins that contain sequences (user cannot modify this preference)
2. Prefer the accession number preference
3. Prefer the protein name preference
4. Prefer the taxonomy preference
5. Prefer proteins that contain GO terms

Probability Legend

To provide a measure of the likelihood that protein identifications are correct in any BioSample or MS Sample, Scaffold uses various validation algorithms to assign identification probabilities to the peptides, see [Increased Confidence Using Peptide and Protein Validation Algorithms](#). Then, using ProteinProphet, it groups the peptides by their corresponding protein(s) and computes probabilities that those proteins were present in the original sample, see [ProteinProphet](#). When loading data through the Wizard the user is presented with the following peptide validation scoring systems from which to choose:


- **PeptideProphet Scoring** - This scoring algorithm learns the distributions of search scores and peptide properties among correct and incorrect peptides and uses those distributions to compute for each peptide a probability that it is correct, see [PeptideProphet](#).
- **LFDR-based Scoring** - This is a novel scoring algorithm based on a Bayesian approach to local False Discovery Rate. It is especially effective for QExactive and other high mass accuracy data, see [LFDR-based scoring system](#).
- If **Prefiltered Mode** is selected, no probability values or probability legend are shown, and no probability or FDR-base thresholds may be applied.

The protein probability values are reported in the Samples Table when selected from the Display Options pull-down list. They are color coded to highlight significant differences in protein identification confidence. The coloring is kept even when another statistics is selected from the Display Options list. Located at the top of the view, the Probability Legend defines the color coding for the protein identification probability.

Sorting

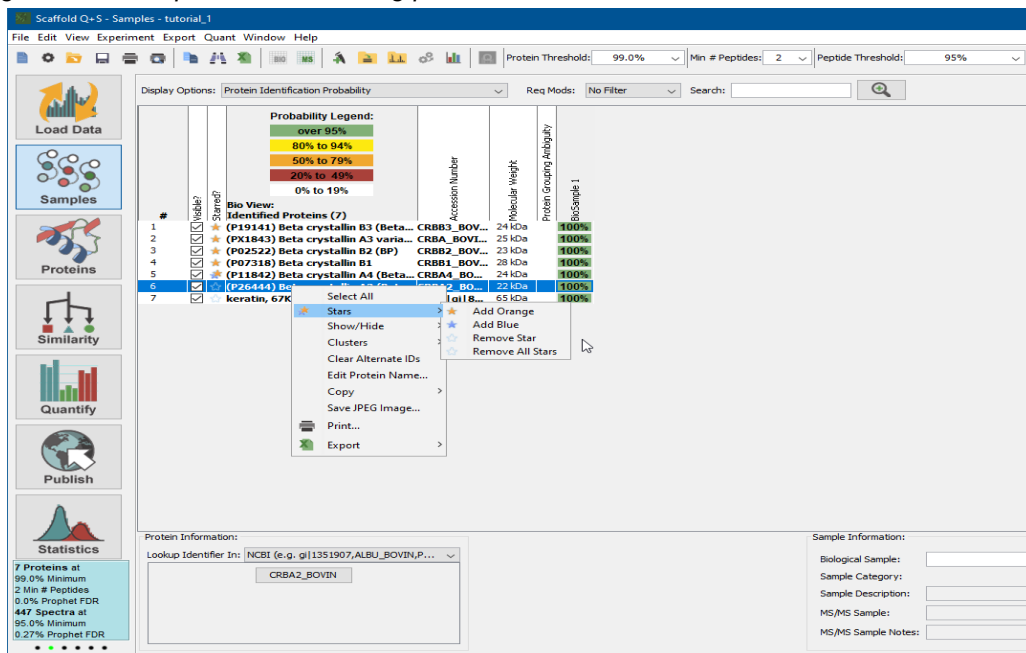
When the Samples View opens, the displayed proteins are initially sorted based on total spectrum count at a Protein Probability of 50%; if any proteins have the same count, then the proteins are sorted alphabetically based on their accession numbers. You can use the tri-state column sorting feature and sort the display by clicking on any column header. For example, to sort the proteins based on increasing molecular weight, click the Molecular Weight column header once. To sort the proteins based on decreasing molecular weight, click the Molecular Weight column header twice. To return to the default display, click the Molecular Weight column header a third time.

Proteins of Interest

It is often helpful to mark proteins in an experiment that are of special interest. By clicking the Star icon  in the **Starred?** column for a protein repeatedly, the user may apply a blue star, an orange star or a combination of both. It is also possible to apply stars by selecting one or more proteins, right-clicking and

selecting one of the star options from the context menu. By using a combination of different stars it is possible to create three different sets of proteins of interest. To bring these proteins to the top of the display, click the **Starred?** column header twice. To return to the default protein display, click the column header again.

Figure 6-8: Samples View - Starring proteins



Hidden Proteins

If proteins that are not of interest to the user are displayed in the Samples View, and/or contaminants are displayed, the user can remove these proteins from the table. To hide the entire protein entry in the Samples View, simply clear the Visible option for the protein. For example, to eliminate Trypsin products from the view, carry out a search for all proteins that contain “Trypsin” in their names, and then clear the Visible check boxes for all proteins that meet this search criteria. Only those proteins that do not have “Trypsin” in their names will be displayed.



To display the proteins that are hidden, go to the View menu and toggle the menu entry Show Hidden Proteins

Protein Grouping Ambiguity

In the Samples View, a star in the column Protein Grouping Ambiguity indicates that the protein in this row is associated with one or more other proteins that share some, but not all, of their peptides. This visual clue marks the proteins for which it may be worthwhile to examine the shared peptides in the Similarity View. The stars in the Protein Grouping Ambiguity column are red when Scaffold loads the data. The stars turn green as a reminder that the user has already examined the Similarity view for a specific protein. Double clicking on a red ambiguity star opens the Similarity View for the selected protein.

Applying Thresholds

There are three different filters that can be used to increase or decrease the length of the displayed protein list in the Samples Table. Their function is to set minimum characteristics for identification confidence:

- [Protein Threshold](#)
- [Minimum Number of Peptides](#)
- [Peptide Thresholds](#)

The protein and peptide thresholds filter the results based on probabilities or, if the samples were searched using decoys, FDR values. The drop-down lists includes one or both options, depending on the type of searches loaded into Scaffold. It is possible to type a custom FDR threshold directly into the box by adding “FDR” to the end of the string, e.g. “10.3% FDR”, for more information see [FDR Thresholding](#).

Figure 6-9: Scaffold Confidence Filters

Protein Threshold:	99.0% ▼	Min # Peptides:	2 ▼	Peptide Threshold:	95% ▼
--------------------	---------	-----------------	-----	--------------------	-------

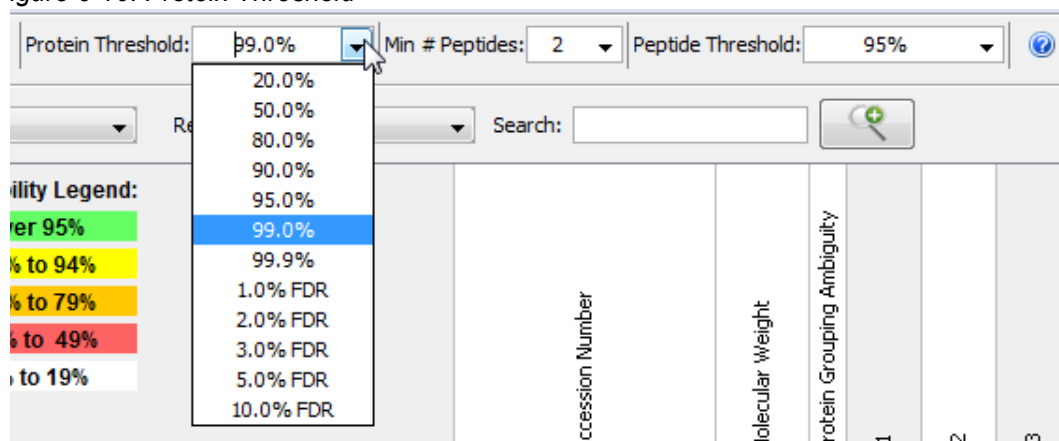


- *Proteins are displayed if each of the filter options is met by at least one sample.*
- *Filters can be locked with a password. When locked, the filters cannot be changed unless the password is entered. This allows you to control what proteins are displayed when you distribute a Scaffold file.*
- *Also note that Protein probability is derived in part from peptide probability, so setting the protein probability much lower than the peptide probability may not display any more results*

Protein Threshold

Through this pull-down list the user can set the minimum requirement for Scaffold's calculated probability of correct protein identification. When the data loaded in Scaffold has been searched against a decoy database, [FDR Thresholding](#) options become available as well, see [Figure 6-10](#).

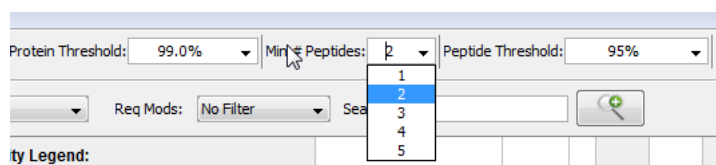
Figure 6-10: Protein Threshold



Minimum Number of Peptides

Through this pull-down list the user can set the number of unique peptides that must be found for one protein in order to consider the protein to be identified.

Figure 6-11: Minimum Number of Peptides



Peptide Thresholds

Through this pull-down list the user can set a threshold for the certainty of a peptide identification in order for it to be counted toward the minimum number of peptides. When the data loaded in Scaffold has been searched against a decoy database, [FDR Thresholding](#) options become available see [Figure 6-12](#).



This filter setting affects not only which proteins are shown but also the reported values shown for number of Exclusive Unique Peptides, number of Total Unique Spectra, Number of Exclusive Unique Spectra, and Percent of Total Spectra.

Among the entries for this filter, shown in the drop-down list, the selection **Custom...** allows defining peptide filters based on the underlying search engines scores. See [Custom Peptide Thresholds](#) for more information regarding this option.



If data has been loaded in Prefiltered Mode, unless probabilities can be computed from Percolator scores, by default no peptide threshold is applied. Only custom peptide thresholds with no probabilities or FDR rules may be used.

Protein Threshold: Min # Peptides: Peptide Threshold:

Req Mods: Search:

Quality Legend:

- 95%
- to 94%
- to 79%
- to 49%
- to 19%

Accession Number

Molecular Weight

Protein

1

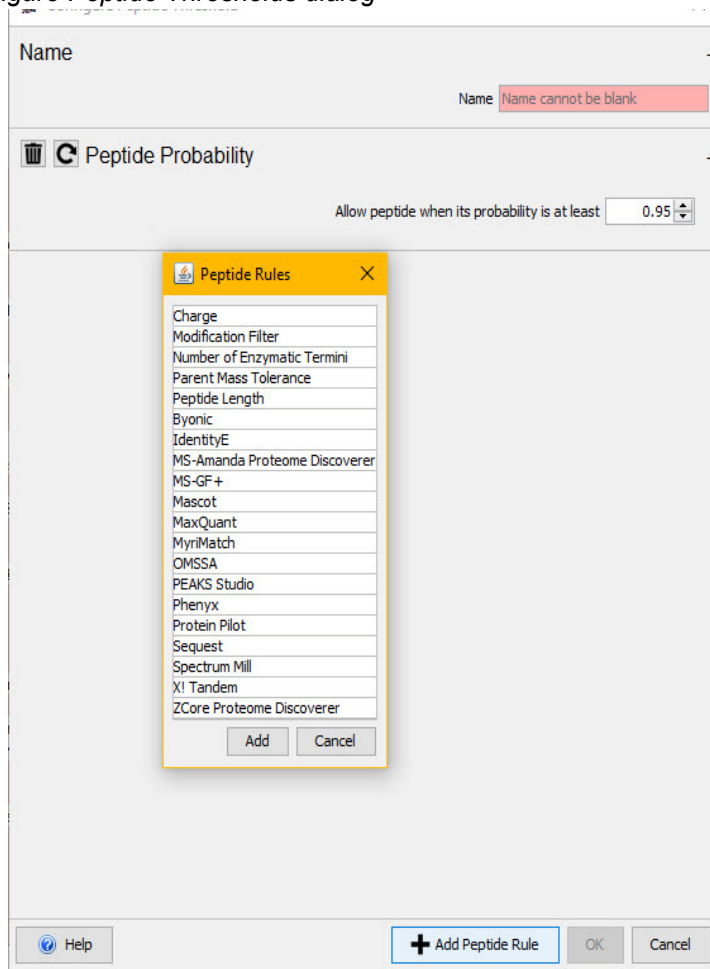
2

Custom peptide thresholds provide a way to threshold peptides based on the underlying search engine scores, modification states, charges, NTT, mass tolerances, etc. When the user chooses custom thresholds, Scaffold ignores the protein probabilities and filters the proteins exclusively on the number of peptides that pass the selected custom peptide threshold.

- *The threshold criteria in effect can be reviewed in the Publish View and in the Publications Report.*
- *Scaffold may be slower when applying custom filters.*
- *The ability to define and apply custom filters can also be controlled by a password, which can make sharing correctly displayed datasets between colleagues easier, see Preferences, [Password](#).*

Through the Configure Peptide Thresholds dialog the user can define custom peptide thresholds to augment the available standard ones.

Figure 6-13: Configure Peptide Thresholds dialog



The following options are available for configuring peptide thresholds:

- **Name Peptide Threshold** - Assigns a name to the custom Peptide threshold built in this dialog box.
- **Peptide Probability** - By default the opportunity to set probability thresholds is offered. These can be combined with other threshold rules. To create a custom threshold with no probability component, click on the trash can icon, then choose other rules.
- **Add Peptide Rule** - brings up a dialog that allows the user to select additional peptide thresholding rules.

Note: If data is prefiltered, unless probabilities can be computed from Percolator results, no peptide probability threshold may be used, but other types of peptide rules are allowed.

Types of Peptide Rules which may be selected:

- *Charges* - Use to define which charges Scaffold displays.
- *Parent Mass Tolerance* - The Parent Mass Tolerance is an after-the-probability-calculation filter on the mass accuracy.
- *Min # Enzymatic Termini (NTT)*

- [Min Peptide Length](#)
- *Program Scores*: Determine which scores from each search engine will be accepted for peptide validation.



Some Scaffold filtering operations are faster using the standard peptide thresholds than using the custom peptide thresholds

Min # Enzymatic Termini (NTT)

When peak lists are searched with a search engine such as Sequest, Mascot, OMSSA, Phenyx, Spectrum Mill or X! Tandem, two of the parameters set are the digestion enzyme and the number of missed cleavages. The search engine only matches spectra to peptides which conform to these parameters.

One approach to increasing the likelihood that the peptides found are correct is to specify that there is no enzyme when running the search engine, and then restricting the search to peptides conforming to the digestion enzyme. Since trypsin is the most common digestion enzyme, the filter in Scaffold is called NTT (Number of Tryptic Termini).

By excluding the peptides with good scores which are non-tryptic, the number of false positives can be decreased but this also decreases the sensitivity. Filtering on NTT is similar to searching with a loose mass tolerance and then restricting the results to only peptides within a tight mass tolerance. Both approaches are ways to filter the data which are independent of the peptide and protein probabilities calculated by Scaffold.

Min Peptide Length

Eliminates peptides with less than the minimum peptide length. This threshold can be used to exclude short peptides, which are seldom exclusive to a single protein. These short peptides may cause a very large number of similar proteins to be displayed in the Similarity View.



-
- *Most search engines (Mascot, Sequest, X! Tandem, etc.) have a minimum peptide length filter option. The Scaffold minimum peptide length threshold is only useful if this filtering was not done in the search engine.*
 - *Using the minimum peptide filter option in the search engine will reduce the processing and file sizes in Scaffold.*

FDR Thresholding

Scaffold allows the user to filter the displayed data based on peptide and/or protein FDR rates when analyzing results of a decoy search. When search results that include decoy matches are loaded into Scaffold, the Peptide and Protein Threshold pull-down list includes **%FDR** values in the list of selectable values. In addition, it is possible to type a custom FDR threshold directly into the box by adding “FDR” to the end of the string, e.g. “10.3% FDR”.

FDR thresholding in Scaffold works by finding the combination of peptide and protein probability thresholds that maximizes the number of proteins identified without exceeding the FDR thresholds and using the selected minimum number of peptides as a lower bound. An FDR landscape, a matrix with all

possible combinations of protein and peptide thresholds, is created and the exact point which maximizes number of proteins while obeying the desired FDR limitations is found. When different threshold combinations would result in the same number of target proteins identified, points at which the protein probability is highest are considered, and of these the point with the highest possible peptide probability is selected.

The actual thresholding is then done using the resulting probability threshold settings. The Minimum Peptide Probability and Minimum Protein Probability thresholds selected by the program are shown in the [FDR Dashboard](#), in the lower left corner of the Scaffold Window. The actual peptide and protein FDR levels are calculated and displayed in the FDR Dashboard as well.



- *Note that FDR Thresholding may select a minimum peptide count higher than the minimum specified by the user. This means that the program found a combination of probability thresholds with that higher minimum peptide count that produced more identified proteins.*

How FDR values are calculated in Scaffold

- Peptide FDR is calculated as the sum of the [Exclusive Spectrum Counts](#) of decoy proteins divided by the sum of the Exclusive Spectrum Counts of target proteins, converted to a percentage.
- Protein FDR is the number of decoy proteins D divided by the number of target proteins T:

$$FDR = \frac{D}{T}$$

expressed as a percentage. This approach assumes that decoy proteins can be filtered out of the considered protein list and that the user is interested in the FDR value of the remaining proteins in the list or target proteins T. This can also be formalized by defining FDR as:

$$FDR = \frac{I}{(C + I)}$$

Where *I* is the count of incorrect proteins and *C* the count of the correct ones and *C+I* is implicitly considered as the total number *T* of target proteins. Scaffold makes the assumption that *D=I*, or that there is an equal number of decoy proteins *D* as incorrect proteins *I* from the original search, providing *FDR=D/T*.

The Display pane

Through the Display pane, the user can specify the type of value that is displayed for each protein in each BioSample or MS Sample in the Samples Table. The pane also contains filtering options for limiting the display to only those proteins that meet specific criteria.

Figure 6-14: Scaffold Display pane



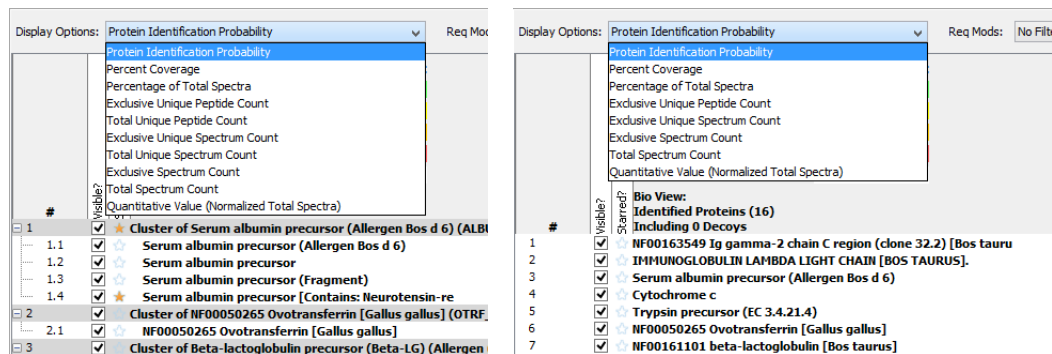
The Display pane contains the following features:

- [Display Options](#)
- [ReqMods](#)
- [Search - Text Box](#)
- [Advanced Search](#)

Display Options

Scaffold reports statistics other than the identification probability. The Display Options pull-down list offers a range of statistics values that once selected are then displayed for each protein under each BioSample or MS Sample in the Samples Table. Depending on whether or not the clustering algorithm option is selected, a slightly different list of options is available.

Figure 6-15: List of Display Options with and without clustering option selected



- **Protein Identification Probability**-- Scaffold's calculated probability that the protein identification for any of the MS Samples is correct. Results are color-coded to indicate significant differences in protein ID confidence.
- **Percentage Coverage** --The percentage of all the amino acids in the protein sequence that were detected in the sample.
- **Percentage of all Spectra**-- The number of spectra matched to a protein, summed over all MS Samples, as a percentage of the total number of spectra in the sample.

- **Exclusive Unique Peptide Count** - (corresponding to Number of Unique Peptides in Scaffold3) -- The number of different amino acid sequences (unique peptides), regardless of any modification that are attributed to a single protein/protein group.
- **Total Unique Peptide Count** - *only available with clustering algorithm selected* -- Number of different amino acid sequences (unique peptide) that are attributed to a single protein including those shared with other proteins
- **Exclusive Unique Spectrum Count** - (corresponding to Number of Unique Spectra in Scaffold3) -- Number of unique spectra attributed to a single protein/protein group. A spectrum is considered unique when it identifies a different charged state and/or a different modified form of an amino acid sequence.
- **Total Unique Spectrum Count** - *only available with clustering algorithm selected* -- Number of unique spectra attributed to a single protein/protein group including those shared with other proteins
- **Exclusive Spectrum Count** - (corresponding to Number of Assigned Spectra in Scaffold3) -- The number of spectra, associated only with a single protein/protein group.
- **Total Spectrum Count** - (corresponding to Unweighted Spectrum count in Scaffold3) - The total number of spectra associated to a single protein/protein group, including those shared with other proteins.
- **Quantitative Value (Selected quantitative method)** -- Scaffold will display the results of the Quantitative Method selected from the dialog.



When a display option different from Protein Identification Probability is selected, the colored highlights don't change. The colors continue to represent the probability ranges specified by the legend. This is true no matter which statistic is chosen for viewing, so that a feel for how probable a given identification is, is always available.

ReqMods

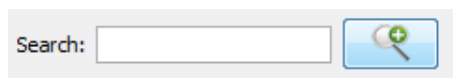
The Required Modifications filter lists all the post-translational modifications (PTMs) selected during the search phase of data processing. Choose a modification on the drop-down list to filter the display to only those proteins, peptides, and spectra that contain the selected modification.

- **No Filter** - No filtering is applied. All proteins, peptides, and spectra that meet all other display and filtering options are displayed.
- **Unmodified Only** - Display only those proteins, peptides, and spectra that do not have any associated PTMs.
- **Variable Modifications** - Display only those proteins, peptides, and spectra that were identified as having the selected variable modification

Search - Text Box

The Scaffold search box allows the user to type in search terms to quickly identify specific proteins by protein names or accession numbers, but it can also filter on peptide sequences and/or spectra information.

Figure 6-16: Search text box



The Search field accepts regular expressions and filters the results based on accession number or protein name. Only those proteins that meet all the search criteria are displayed. Your search is limited to the exact order of the characters in the string, but the string is not case-sensitive and it can appear anywhere in the search results. For example, a search string of **ATP** returns both **ATP** synthase and calcium-transporting **ATPase**. In another example, a search string of **sodium|transport** returns all values that have sodium and/or transport in the protein name - **Sodium/potassium transporting...** and **Calcium-transporting ATPase** sar..., and so on.

Click on the magnifier glass button to the right side of the search text box for more advanced search features. See [Advanced Search](#).

Advanced Search

When the magnifying glass button on the right side of the search text box in the Samples View is clicked, the **Configure Advanced Protein Filter** dialog opens.

Figure 6-17: Samples View - Configure Advanced Protein Filter

Configure Advanced Protein Filter

Search for proteins based on accession number, protein name, peptide sequences (or sub sequences) and spectrum ID names. All searches support regular expressions. For example, you can find possible CaMKII phosphorylation sites by searching for peptides with the "R...[ST]" motif. Please visit [this site](#) for more help with forming regular expressions.

Warning:
Searching by spectrum name can be extremely slow. It works best on small data sets.

Matching **All** of the text filters:

Accession/Protein Name:

Protein Sequence Motif:

Identified Peptide Sequence:

Spectrum Name (SLOW!):

Category Filtering Options

Present in **All** of the Categories:

Absent from the Categories:

Quantitative Profile Filtering Options ☒ Filter by: Significant

Compared Categories:

☒ Control

☒ Treatment

High Categories:

Low Categories:

Starred as... ☐ ☐ ☐ ☐ ☐ ☐

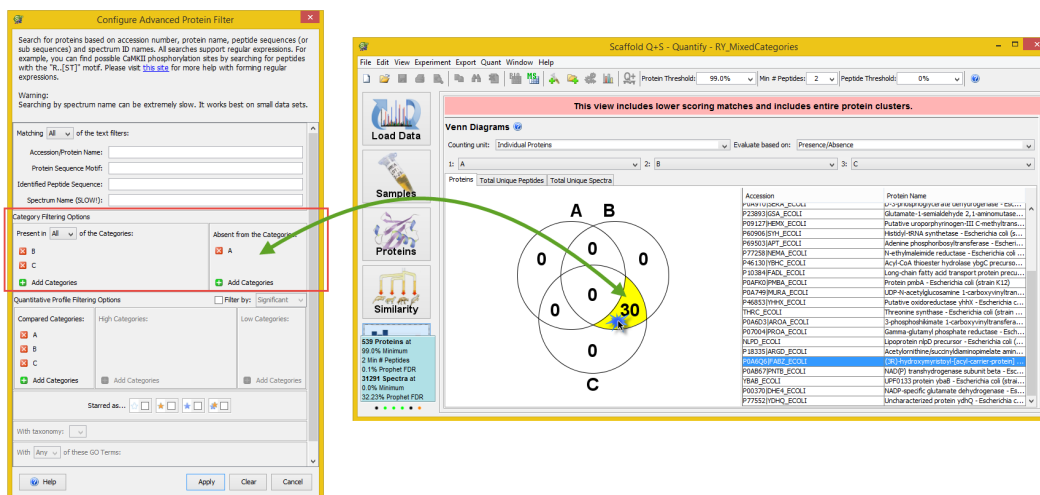
With taxonomy:

With **Any** of these GO Terms:

The dialog contains a number of search tools organized in different panes.

- **Pane A** - Useful to search for specific proteins, peptides, spectra (handy in peptidomic studies whenever questions arise about a peptide assignment) or peptide motifs (useful to investigate potential modification sites). All searches support regular expressions.
- **Pane Category Filtering Options** - Allows for searches based on the intersection of categories. It displays only proteins found in a category, or proteins found in one category and not in another category. This is the filter engaged when double clicking on a section of the diagram located in [The Venn Diagrams pane](#) when the evaluation is based on presence/absence, see [Figure 6-18](#).

Figure 6-18: Filtering options when double clicking on Venn Diagram if evaluation is based on Presence/Absence.



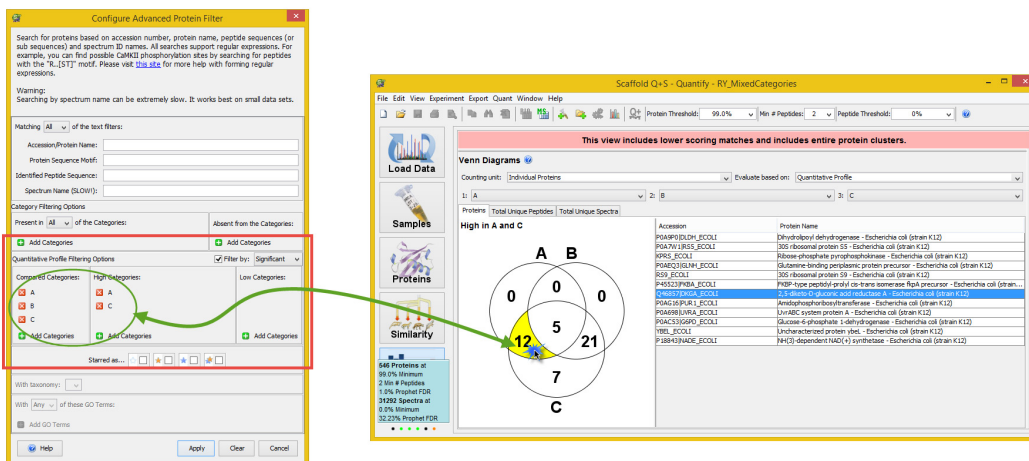
- Quantitative Profile Filtering Options** - This pane is active only when a statistical test has been selected from the setup dialog and filters only over proteins that are significantly quantitatively different among categories. After a statistical test is applied both the p-value column and the **Quantitative profile** column appear in the Samples Table. Clicking the magnifier lens close to the search text box, opens the **Configure Advanced Protein** filter dialog where the Quantitative Profile filtering pane is now active. The pane shows all the categories selected for the test in the Compared Categories sub-pane. On the left side of each category there is a red cross that can be used to deselect the category, allowing the possibility of choosing a subset of the categories available. When a subset of categories is selected it is possible to check whether they are up-regulated or down-regulated with respect to each other. The mean of the averages is then recalculated, since it depends on the number of categories selected. Within the pane it is also possible to add other categories using the **Add Categories** green plus button.

Note: Deselecting all categories or just leaving one category clears the **Quantitative profile** column since comparing one or no category has no meaning.

Selecting the **Filter by:** check box activates the High Categories: and Low Categories: sub-panes. The user can define here which high category or low category he wants to filter in. The pull-down menu gives the option to filter over significant or not.

A combination of these filters is engaged when double clicking on a section of the diagram located in **The Venn Diagrams** pane when the evaluation is based on Quantitative Profile, see [Figure 6-19](#).

Figure 6-19: Filtering options when double clicking on Venn Diagram if evaluation is based on Quantitative Profile.



Searches can also be performed over the full list of identified proteins or within different groups of proteins like categories or starred proteins.

- **Pane starred** - Filters over the presence or absence of proteins tagged with stars.
- **Pane Taxonomy** - Filters over type of taxonomies listed in a pull-down menu. This sub-pane is active only when GO terms have been searched.
- **Pane GO Terms** - Filters over GO Terms that can be added to the sub-pane. This pane is active only when GO terms have been searched.

Information Panes

The bottom section of the Samples View contains three information panes:

- Protein Information pane
- Annotations pane
- Sample Information Pane

Each pane provides further diversified information related to each row in the Samples Table

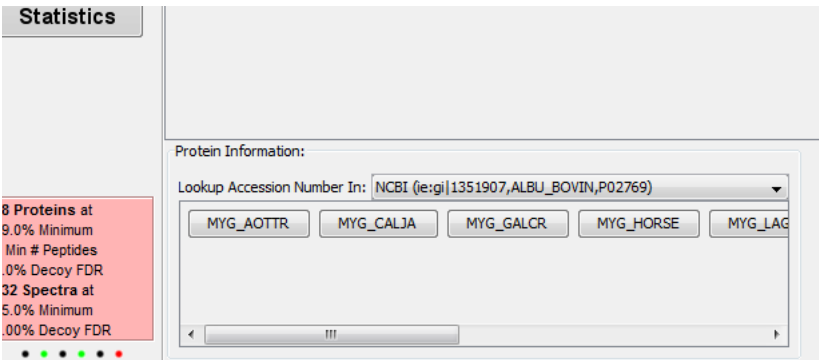
Figure 6-20: Samples View - Information Panes



Protein Information pane

The Protein Information pane is displayed in the lower left section of the Samples View. It includes a pull-down list of Online protein databases such as SwissProt, NCBI or WikiPathways in which the user may look up the selected protein. For each selected row in the Samples Table, the pane shows the set of proteins included in the corresponding protein group as click-able buttons. A button is displayed for each selected accession number, and if Alternate ID's have been added via FASTA parsing or NCBI annotation, additional buttons are provided for looking up selected proteins via their alternate ID's. Clicking one of the buttons opens an Internet browser to the address selected from the pull-down list and searches for the selected protein identifier. If the identifier is found, additional information for the selected proteins is then easily available to the user.

Figure 6-21: Protein Information pane

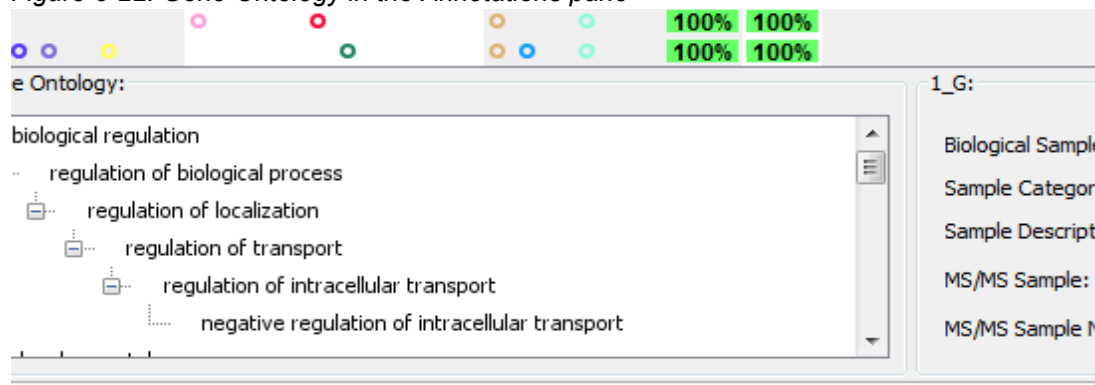


Annotations pane

The Gene Ontology pane, when displayed, is located in the lower center section of the Samples View. The pane is displayed only when GO terms or pathways have been applied. The annotations that apply to the selected protein appear in this pane. Annotations may be added to the Samples Table during the loading phase or after the data has been loaded by going to the **Experiment> Add or Edit Annotations**. For more information see [GO Annotations and Pathways in Scaffold](#).

GO terms are displayed as an ontology tree, with the high-level GO annotations showing as colored dots (which match the colors shown in the Samples Table) and its subsequent children. Pathways

Figure 6-22: Gene Ontology in the Annotations pane



Pathways associated with the selected protein are listed in the same space. Clicking on a pathway in this pane opens the listing for that pathway in the website from which it was obtained, either Reactome or WikiPathways.

It is possible to display both GO terms and pathways at the same time.



- Double-clicking a GO term or pathway in this pane opens a page in a browser with detailed information about the annotation.
- GO terms may be hidden by un-checking the menu in **View > Show GO Annotations**. See [Add or Edit Annotations](#)

Sample Information Pane

The Samples Information Pane displays the Biological and MS Sample names and descriptions for the selected MS Samples. Biological Sample name and notes, and the MS Sample name and notes can be edited here. To populate this pane, click on a Bio/MS Sample column.

Figure 6-23: Sample Information pane

50% 100%
100% 100%
100% 100%

1_G:

Biological Sample: 1_G

Sample Category: Bovine lens

Sample Description:

MS/MS Sample: bovine_mudpit_10

MS/MS Sample Notes:

sport



To change a category name for a BioSample, go to **Experiment > Edit BioSample**.
See [Edit BioSample](#).

Chapter 7

Proteins View

The information in the view is organized in three different interconnected panes. This type of framework simplifies the selection of the protein of interest, the manual inspection of its spectra and identified peptides; the viewing of proteins coverage and other characteristics of the MS experiment.

The different panes in the view are:

- [Proteins pane](#), which provides two perspectives of locating the protein of interest and shows characteristics for the selected protein, in each BioSample or MS sample, depending on the selected summarization level, containing that protein
- [Peptides pane](#), which shows characteristics of each peptide spectrum associated with the selected protein
- [Protein Sequence pane](#), which through different tabs provides detailed information about the spectrum for a selected identified peptide, its model error and fragmentation table.

The user can select the protein of interest in the [Samples View](#) Proteins table and then click the Proteins View button to open the corresponding Proteins View, or simply double-click the selected protein. Once in the Proteins View, the user can select another protein from the [Proteins pane](#) pull-down list.

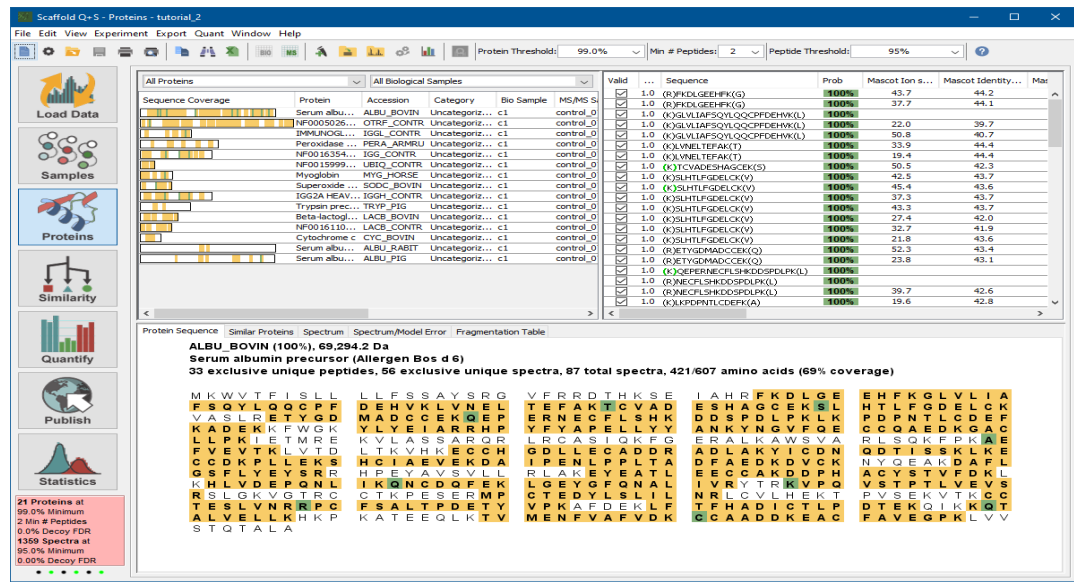
Note: All tables included in the view behave as described in the [Display pane](#) section. Several options for filtering and sorting allow for adjusting the content and organization of the tables.

- **Filtering proteins and peptides in the Proteins View** - What appears in the Proteins view is affected by the settings of proteins and peptide thresholds available in the main Scaffold's window. Adjusting these thresholds determines which proteins are visible in the [Proteins pane](#) and which peptides are listed in the [Peptides pane](#).

A modification of the Peptide Probability changes not only the list of peptides shown in the peptides pane but it is also reflected in the Sequence Coverage shown in the [Similar Proteins tab](#) and in the [Protein Sequence tab](#).

The [Proteins pane](#) also contains two additional filter options, which work in conjunction with the general protein and peptide thresholds.

Figure 7-1: Proteins View



Proteins pane

The Proteins pane is located in the upper left corner of the Proteins View. It includes two pull-down lists and a table.

Figure 7-2: Proteins Pane

Serum albumin precursor (Allergen Bos d 6)												All Biological Samples
Sequence Coverage	Protein	Accession	Category	Bio Sample	MS/MS Sample	Prob	%Spec	#Pep	#Uni...	#Spec	%Cov	m.w.
	Serum albu...	ALBU_BOVIN	Uncategoriz...	c1		100%	12%	33	56	87	69%	69 kDa
	Serum albu...	ALBU_BOVIN	Uncategoriz...	c2		100%	12%	35	54	82	70%	69 kDa
	Serum albu...	ALBU_BOVIN	Uncategoriz...	c3		100%	13%	37	57	93	70%	69 kDa
	Serum albu...	ALBU_BOVIN	Uncategoriz...	c4		100%	12%	37	59	90	68%	69 kDa
	Serum albu...	ALBU_BOVIN	Uncategoriz...	c5		100%	12%	34	53	88	65%	69 kDa

- **Sequence coverage table** - For the selected protein each row shows its coverage in a BioSample/MS sample and category and displays along its columns a number of protein characteristics and quantitative values. The list of coverages reported in the table depends on the parameters selected from the pull-down list located above the table.

Some columns in the table provide a description of the protein and list which of the BioSamples and MS samples include it:

- *Sequence Coverage* - a string that shows highlighted in yellow the sequence of identified peptides and in green the modifications. It provides a visual of the identified peptides and modifications in a specific BioSample/MS Sample. The list of sequence coverages along the column provide a way to compare the protein sequence coverage among different BioSamples or MS samples.
- *Protein* - protein description
- *Accession* - protein accession number
- *Category* - name of the category the protein belongs to
- *BioSample* - [Appears when summarization is set to Biological Sample View] - name of the BioSample the protein belongs to
- *MS/MS Sample* - [Appears when summarization is set to MS/MS Sample View] - name of the MS sample the protein belongs to
- *m.w.* - protein molecular weight

The remaining columns report quantitative values for the protein as computed within a specific BioSample/MS Sample:

- *Prob* - protein probability as calculated using the selected peptide and protein validation algorithm. No probabilities are shown in Prefiltered Mode.
- *%Spec* - Protein [Percentage of all Spectra](#)
- *#Pep* - Exclusive unique peptide count

- *#Unique* - Exclusive unique spectrum count
- *#Spec* - This variable reports the count of rows in the Peptides table. Depending on the type of grouping used to analyze the data, the value corresponds to different Display Options. (1) When clustering is involved the *#Spec* provides the **Total Spectrum Count**, which reports the number of spectra or rows in the table or (2) with no clustering, it provides the **Exclusive Spectrum Count** which reports the number of green checks appearing under the “Assigned” column.
- *%Cov* - Percentage of amino acids identified or sequence coverage.

Selecting a row in the Proteins table updates the list of peptides appearing in the Peptides pane Spectrum table and the related information is updated in the Spectrum pane as well.

- **Upper Left Pull-Down list** - It can either filter the table so that the rows show information related to only one selected protein within each of the BioSamples and MS samples in the experiment, or show all the proteins. The pull-down includes the option “All proteins” and the list of protein groups and clusters appearing in the [The Samples Table](#). This means that the proteins found in the list can change depending on the filters and thresholds applied to the Samples table, see [Thresholding pane](#).

Note: When a protein group is selected, the Upper Right Pull-Down List defaults to “All biological samples” or “All MS/MS Samples” depending on the selected summarization.

- **Upper Right Pull-Down list** - Depending on the summarization level currently selected, it filters the list of proteins according to the BioSamples/MS Samples included in the experiments. The pull-down shows the list of available BioSamples or MS Samples.

Note: When a specific BioSample or MS Sample is selected the Upper Left Pull Down list defaults to “All Proteins”

A context menu is available when the user right clicks the mouse over the Sequence Coverage table, see [Context Menus Right Click Commands](#)

Peptides pane

The Peptides pane is located in the upper right corner of the Proteins view and it contains the [Spectrum table](#). This table shows the list of spectra that identify peptides belonging to the protein selected in the [Proteins pane](#).

Figure 7-3: Proteins View: Spectrum table in Peptide Pane

Valid	...	Sequence	Prob	Mascot...	Mascot Identity...	Mascot Delta Ion Sc...	XI Ta...	NTT	Modifications	Observed	Actual Mass	Charge	Delta ...	Delta ...	R...	Intensity	TIC	Start	S
<input checked="" type="checkbox"/>	1.0	(K)GLVLIAPSQYLOQCFDEHW(L)	100%	22.0	39.7	21.0	7.27	2	Carbamidomethyl...	831.83	2,492.46	3	1.2	62			3828000	45	
<input checked="" type="checkbox"/>	1.0	(K)GLVLIAPSQYLOQCFDEHW(L)	100%	50.8	40.7	49.4	5.33	2	Carbamidomethyl...	1,246.26	2,490.50	2	-0.75	100			4885000	45	
<input checked="" type="checkbox"/>	1.0	(K)GLVLIAPSQYLOQCFDEHW(L)	100%	33.9	44.4	11.7	3.37	2	Carbamidomethyl...	1,246.69	2,491.36	2	0.11	43			5,223E7	45	
<input checked="" type="checkbox"/>	1.0	(K)LVMLTFPAK(T)	100%	19.4	44.4	15.6	2.44	2		582.28	1,162.55	2	-0.07				39E7	66	
<input checked="" type="checkbox"/>	1.0	(K)LVMLTFPAK(T)	100%	19.4	44.4	15.6	2.44	2		1,163.70	1,162.69	1	0.06				953E7	66	
<input checked="" type="checkbox"/>	1.0	(K)YCAVDSHAGDCK(S)	100%	50.5	42.3	37.5		2	Carbamyl (+43), ...	753.76	1,505.51	2	-0.07				377E7	76	
<input checked="" type="checkbox"/>	1.0	(K)SLHTLFGDELQ(V)	100%	42.5	43.7	36.0	4.80	2	Carbamidomethyl...	710.32	1,418.62	2	-0.063	-45			1,561E7	89	
<input checked="" type="checkbox"/>	1.0	(K)SLHTLFGDELQ(V)	100%	45.4	43.6	32.9		2	Carbamyl (+43), ...	731.92	1,461.83	2	0.14	96			2,641E7	89	
<input checked="" type="checkbox"/>	1.0	(K)SLHTLFGDELQ(V)	100%	37.3	43.7	31.9	3.44	2	Carbamidomethyl...	710.31	1,418.61	2	-0.073	-52			1,763E7	89	
<input checked="" type="checkbox"/>	1.0	(K)SLHTLFGDELQ(V)	100%	43.3	43.7	25.6	7.38	2	Carbamidomethyl...	1,419.69	1,418.68	1	-0.0039	-2.8			8,955000	89	
<input checked="" type="checkbox"/>	1.0	(K)SLHTLFGDELQ(V)	100%	27.4	42.0	22.9	4.26	2	Carbamidomethyl...	710.08	1,418.15	2	-0.53	330			7951000	89	
<input checked="" type="checkbox"/>	1.0	(K)SLHTLFGDELQ(V)	100%	32.7	41.9	18.9	3.68	2	Carbamidomethyl...	474.14	1,419.38	3	0.70	-220			2,386E7	89	
<input checked="" type="checkbox"/>	1.0	(K)SLHTLFGDELQ(V)	100%	21.8	43.6	10.8	2.72	2	Carbamidomethyl...	474.26	1,419.76	3	1.1	52			1,497E7	89	
<input checked="" type="checkbox"/>	1.0	(R)ETYSRMADCCCK(Q)	100%	52.3	43.4	40.8	4.27	2	Carbamidomethyl...	1,478.63	1,477.62	1	0.11	72			7465000	106	
<input checked="" type="checkbox"/>	1.0	(R)ETYSRMADCCCK(Q)	100%	23.8	43.1	17.5	2.92	2	Carbamidomethyl...	739.29	1,476.56	2	-0.95	34			3,203E7	106	
<input checked="" type="checkbox"/>	1.0	(K)QEPERNECHLSKDGDPK(L)	100%				4.24	2	Ammonia-loss (-1...	842.73	2,525.15	3	2.0	5.6			3,516E7	118	
<input checked="" type="checkbox"/>	1.0	(R)NECFLSHKDSQPK(L)	100%				3.09	2	Carbamidomethyl...	634.98	1,901.90	3	1.0	20			1,383E7	123	
<input checked="" type="checkbox"/>	1.0	(R)NECFLSHKDSQPK(L)	100%	39.7	42.6	27.7	3.89	2	Carbamidomethyl...	951.52	1,901.02	2	0.16	84			4936000	123	
<input checked="" type="checkbox"/>	1.0	(K)KPPPTLCOEPIA	100%	19.6	42.8	5.3	5.00	2	Carbamidomethyl...	788.99	1,575.97	2	0.21	130			9029000	139	
<input checked="" type="checkbox"/>	1.0	(K)KPPPTLCOEPIA	100%	47.0	40.6	41.4	5.68	2	Carbamidomethyl...	1,010.76	2,019.50	2	0.54	230			1,678E7	139	
<input checked="" type="checkbox"/>	1.0	(K)PPPTLCOEPIA	100%				6.40	1	Carbamidomethyl...	889.95	1,777.88	2	0.10	56			7911000	141	
<input checked="" type="checkbox"/>	1.0	(K)LYEIAR(R)	100%	33.0	43.7	7.9	2.32	2		464.17	926.33	2	-0.15	-160			7,408E7	161	

Spectrum table

Each row in the table represents a spectrum while the columns provide information like the peptide sequence it identifies, its search score information and other parameters that are meant to help the user manually validate the spectrum.

Spectrum table Columns:

- Valid** - When checked the spectrum contributes to the calculation of the protein probability while unchecked spectra do not. When loading data, identified peptides are assigned a probability using the selected scoring algorithm. Scaffold then automatically defines a spectrum as valid when its probability is greater than the selected peptide probability threshold. While manually inspecting a spectrum the user can decide if a peptide identification is valid or not and accordingly manually update the **Valid** check box in the table. It is also possible to globally define a minimum acceptance probability or reset manually validated peptides through the command [Reset Peptide Validation](#).
- Weight/Assigned** - This column shows **Weights** when the clustering algorithm is applied or **Assigned** when the legacy grouping algorithm is instead selected.
 - Assigned** - Under this heading a green check box is visible when a particular peptide has been assigned to the current protein or a red cross if the peptide has been assigned to another protein. The column *Other Protein* shows the accession number of the protein that won the assignment, for further information see [Standard Protein grouping](#).
 - Weight** - When clustering is selected the grouping algorithm used considers all peptides appearing in a protein and assigns a weight to those peptides that appear in more than one protein. This column reports the weight assigned to an identified peptide according to a [Weighting Function](#).

- **Sequence** - Peptide sequence identified by the spectrum. The string of amino acids in the peptide is prefaced and followed by one residue symbol in parentheses. For example: (K)TGQAPGFSYTDANK(N). The symbols in parentheses represent the residue just before and just after the peptide, within the entire protein sequence. This gives the user additional information such as confirming whether the peptide is tryptic or semi-tryptic, based on whether it begins just after a K or R. See also NTT.
- **Prob** - Probability assigned by Scaffold to the peptide identification according to the selected validation algorithm at the time of data load. The peptide probabilities are shaded according to the legend shown in the Samples Table. No probabilities are shown in Prefiltered Mode.
- **Search engine scores** - A series of columns reports the search engine scores used to analyze the loaded data. Scaffold combines and aligns results of the same data analyzed with different search engines and reports here the different assigned scores. As an example find below the different score listed for some of the most commonly used search engines.
 - *SEQUEST* - Xcorr and DeltaCn
 - *Mascot* - Ion score, Identity score and Delta Ion Score.
 - *X! Tandem* - log of expectation score
- **NTT** - Number of termini consistent with the enzymatic cleavage or tryptic termini
- **Modifications** - List of modifications identified by the spectrum
- **Mass measurements** - See also [Computation for Peptide mass values](#):
 - *Observed* - Mass over charge (M/Z) of the parent or precursor ion measured by the mass spectrometer.
 - *Actual Mass* - Peptide mass in Dalton obtained by multiplying the charge to the subtraction of one proton from the observed M/Z.
 - *Charge* - Peptide charge
 - *Delta Da* - (Actual Mass - Theoretical Peptide Mass) in Dalton, where the Theoretical Peptide Mass or Calculated peptide mass, is given by the sum of amino acid residue masses included in the peptide plus a water molecule.
 - *Delta PPM* - (Actual Mass - Theoretical Peptide Mass) in PPM also referred to in the spectrum as the Parent error. It is calculated by dividing the delta mass expressed in Dalton by the Actual Mass and then multiplied by one million (see [Adjusting the Delta PPM Value](#)).
- **Retention Time** - Measured in seconds, it is included in the table only if the information is listed in the peak list of the loaded data.
- **Intensity** - Peptide precursor intensity area under the curve. This column is populated only when the data imported include this type of computation.

- **TIC** - MS/MS Total Ion Current, for more information see [Total Ion Current \(TIC\)](#).
- **Start** - Peptide start index.
- **Stop** - Peptide stop index.
- **# Other Proteins** - Number of other proteins where the peptide is found in.
- **Other Proteins** - Accession number of other proteins where the peptide is found in. When the data is analyzed with the legacy protein grouping algorithm, if a red cross is assigned to a peptide in the current protein it lists the protein to which it has been assigned to.
- **Spectrum ID** - Name of the spectrum from which the information shown has been extracted from.

A context menu is available when right clicking the mouse over the Spectrum table, see [“Context Menus Right Click Commands”](#)



- Go to www.proteomesoftware.com/pdf/file_compatibility_matrix.pdf to check if the search engine used to search the data imported in Scaffold is supported for Retention Time and Label Free Quantitation-Precursor Intensity.
- Go to <https://www.proteomesoftware.com/data-formats> for information on how to run the different supported search engines to generate Retention Time and Label Free Quantitation data to be imported in Scaffold.

Computation for Peptide mass values

The Spectrum Report provides for each peptide four different mass values, which are:

- **Observed m/z**, experimentally measured precursor ion mass expressed in M/Z provided by the mass spectrometer.
- **Actual Peptide Mass (AMU)**.
- **Calculated +1H Peptide Mass (AMU)**.
- **Actual minus Calculated peptide mass (AMU)**, defined as Delta Mass in Dalton
- **Actual minus calculated peptide mass (PPM)**.

Although it is not directly a mass, the Spectrum charge is a very important part of Scaffold's mass calculations.

The Actual Peptide Mass expresses the weight of the peptide as measured by the mass spectrometer and it is calculated using the Observed precursor ion mass as follows:

$$\text{Actual Peptide Mass (Da)} = \text{Observed (M/Z)} \times \text{Spectrum Charge} - 1 \text{ proton} \times \text{Spectrum Charge}$$

The Theoretical Peptide Mass is given by the sum of the amino acid residue masses included in the peptide plus the mass of a water molecule.

$$\text{Theoretical Peptide Mass (Da)} = \sum \text{Amino Acid Residue Mass} + 1 \text{ water}$$

The spectrum report provides the theoretical peptide mass plus one proton mass under the column “Calculated + 1H Peptide Mass”. The delta mass in Dalton is then calculated as follows:

$$\begin{aligned} \text{Delta Mass (Da)} = \\ \text{Actual Peptide Mass} - \text{Theoretical Peptide Mass (Da)} = \\ (\text{Actual Peptide Mass}) - (\text{Calculated} + 1\text{H Peptide Mass} - 1\text{Proton}) \end{aligned}$$

The Delta Mass in PPM is calculated as follows:

$$\text{Delta Mass(PPM)} = \frac{\text{Delta Mass(Da)} \times 10^6}{\text{Actual Peptide Mass(Da)}}$$



Adjusting the Delta PPM Value

If the Delta Mass (Da) value is approximately +/-1, Scaffold assumes that it is the result of a triggering error caused by misidentification of the monoisotopic peak. It therefore adjusts the mass by the mass of a neutron when calculating the delta PPM. This adjustment causes a discrepancy between the Delta Mass (Da) value and the Delta Mass (PPM) value, but is expected to give a more accurate assessment of Delta Mass (PPM).

Here is an example calculation:

$$\text{Actual Mass} - \text{Calculated Mass} = \text{Delta Mass (Da)}$$

$$3344.65 \quad - \quad 3345.62 \quad = \quad -0.9653$$

Since this value is approximately -1 amu, add the mass of a neutron:

$$\text{Delta Mass (Da)} + \text{mass of neutron} = \text{adjusted Delta Mass (Da)}$$

$$-0.9653 \quad + \quad 1.00273 \quad = \quad 0.0374$$

Then using the adjusted value to obtain the PPM:

$$\text{adjusted Delta Mass (Da)} / \text{Actual Mass} = \text{Delta Mass (PPM)}$$

$$0.0374 \quad / \quad 3344.65 = \quad 11.18$$

Protein Sequence pane

The Protein Sequence pane provides detailed information about any peptide selected from the Peptides pane; it is organized in 5 different tabs. Each tab contains pieces of evidence meant to help the user inspect and assess whether the selected peptide identification appears meaningful or not.

- [Protein Sequence tab](#)
- [Similar Proteins tab](#)
- [Spectrum tab](#)
- [Spectrum/Model Error tab](#)
- [Fragmentation Table tab](#)

Figure 7-4: Proteins View: Protein sequence pane

Protein Sequence	Similar Proteins	Spectrum	Spectrum/Model Error	Fragmentation Table
ALBU_BOVIN (100%), 69,294.2 Da Serum albumin precursor (Allergen Bos d 6) 33 exclusive unique peptides, 56 exclusive unique spectra, 87 total spectra, 421/607 amino acids (69% coverage)				
MKQVTFISLL FSQYLQCCFF VASLRLETYGD KADEKFWGK LLPKIETMRE FVEVTKLVTD QCDKPLLEKS GSFLYEYSRR KHLVDEPQNL RSLGKVGTRC TESLVNRRPC ALVELLKHHP STOTALA	LLFSSAYSRL DEHVKLVLNEL MADCCCKQEP YLVEIARRHP KVLASSARQR LTKVHKECCH HCLAEVEKDA HPEYAVSVLL IKQNCDDQFEK CTKPESESRMP FSALTPDETY KATEEQCLKTY	VFRRDTHKSE TEFAKTCVAD ERNECFLLSHK YFYAPELLYY LRCASIQKFG GDLLECADDR IPENLPPLTA RLAKEYEATL LGEYGFQNAL CTEDYLSLIL VPKAFDEKLF MENFVAFVDK	IAHRFKDGLGE ESHAGCEKSL DDSPDLPKLK ANKYNGVFQE ERALKAWSVA ADLAKYICDN DFAEDKDVCK EECCAKDDPH IVRYTRKVPQ NRLCIVLHEKT TFHADICTLP GCAADDKEAC	EHFKGLVLIA HTLFGDELCK PDPNTLCDEF CCQAEDKGAC RLSQKFPPKE QDTISSKLLKE NYQEAQDAFL ACYSTVLVEVS VSTPTLVEVS PVSEKVTKCC DTEKQIKKQT FAVEGPKLVV

Protein Sequence tab

It displays the entire amino acid sequence for the protein selected in the view or a protein selected from the [Similar Proteins tab](#). Identified peptides meeting the minimum peptide threshold are highlighted in yellow while modifications are shown in green, see [Figure7-4](#).

Above the sequence the protein accession number, molecular weight and protein name are shown together with the number of exclusive unique peptides, exclusive unique spectra, total spectra and % coverage. The sequence is derived from the fasta database that was loaded into Scaffold with the data set.



If the fasta database is not identical to the external protein database, including the version, used for searching experimental data, then the protein sequence and molecular weight might not be available for display. When this happens a question mark also appears in the Samples table Molecular weight column.

A context menu is available when the user right clicks the mouse over the protein sequence, see [Context Menus Right Click Commands](#).

Similar Proteins tab

The Similar Proteins tab lists, for the selected protein, all the members of its protein group. It

includes a table very similar to the sequence coverage table shown in the [Proteins pane](#). It visually depicts the fraction of amino acids identified in each protein in the group through a sequence coverage string. Yellow indicates an identified amino acid. Green indicates a modification. The protein name, accession number, and descriptor are also shown.

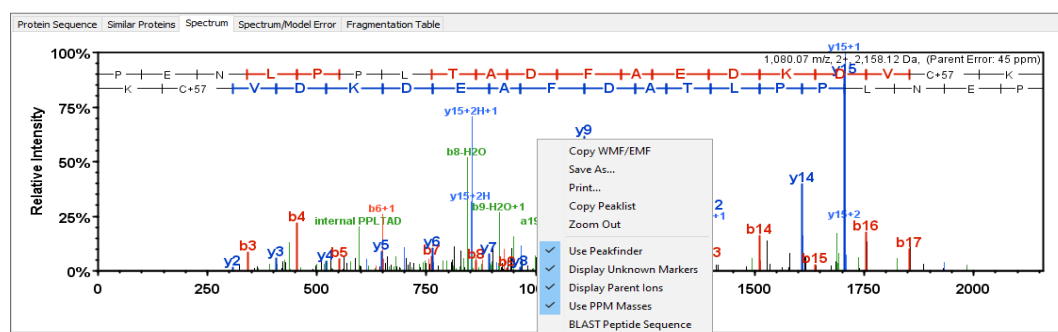
The molecular weight and percent sequence coverage displayed correspond to the selected protein. If there are several similar proteins, one of these proteins is chosen as the representative protein. The m.w. and %Coverage for the representative protein are displayed in the sequence coverage table shown in the [Proteins pane](#) and its sequence is displayed by default in the [Protein Sequence tab](#). To see the sequence of a different protein present in the list of similar proteins, the user can click on it and then switch to the Protein Sequence tab.

A context menu is available when the user right clicks the mouse over the protein sequence, see [Context Menus Right Click Commands](#).

Spectrum tab

The Spectrum tab displays the peptide/spectrum selected in the [Peptides pane](#).

Figure 7-5: Proteins View: Spectrum tab



The spectrum graph is interactive:

- The user can click anywhere on the spectrum to display the M/Z value for the position.
- The user can click and hold the left mouse button anywhere on the spectrum and then drag the mouse pointer to any position in the spectrum of his/her choosing. As the user drags the mouse pointer, the Start and Stop M/Z values for the segment are displayed as well as the length for the segment. The user can release the button to zoom in on the selected region. A single click of the mouse returns the zoom out magnification to 100%.
- The user can right-click anywhere on the spectrum to open the context menu which provides a number of menu options as described in the [Context Menus Right Click Commands](#) section.

Note: For Waters-PLGS, the Spectrum tab will be split into the following two tabs:

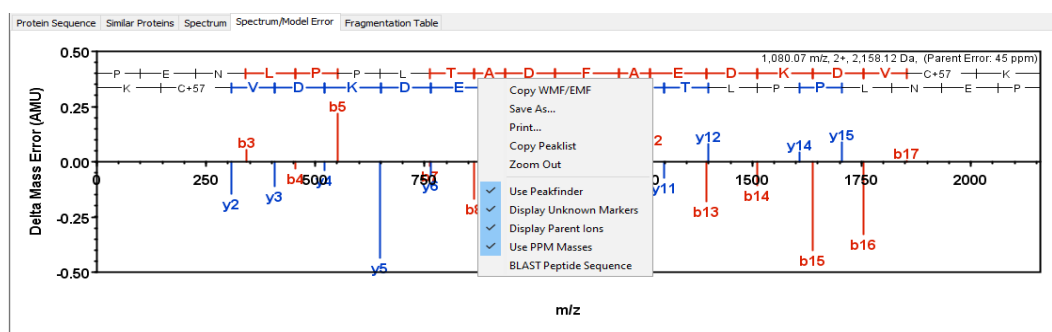
- **RT Summary Plot tab** - Shows the collected data versus the retention time. The various measured ion fragments are color coded depending whether they are unidentified or identified peptides or non peptides fragments. The size of the points shown in the plot can be toggled to be either scaled or not to their intensity values.

- **Spectrum tab** - The second tab shows the regular spectrum for the selected peptide.

Spectrum/Model Error tab

The Spectrum/Model Error tab displays a graph of the errors, defined as Actual peptide fragment peak - measured fragment peak, for the peptide/spectrum selected in the [Peptides pane](#). Differences are shown in AMUs. The maximum difference displayed equals the fragment error level set under “mass accuracy” in the parameter set chosen when the sample was loaded.

Figure 7-6: Proteins View: Spectrum/Model Error tab



The spectrum graph is interactive:

- The user can click anywhere on the spectrum to display the M/Z value for the position.
- The user can click and hold the left mouse button anywhere on the spectrum and then drag the mouse pointer to any position in the spectrum of his/her choosing. As the user drags the mouse pointer, the Start and Stop M/Z values for the segment are displayed as well as the length for the segment. The user can release the button to zoom in on the selected region. A single click of the mouse returns the zoom out magnification to 100%.
- The user can right-click anywhere on the spectrum to open the context menu which provides a number of menu options as described in the [Context Menus Right Click Commands](#) section.

Fragmentation Table tab

The tab includes the Fragmentation Table which shows what peptide fragmentation ions match the spectrum shown in the [Spectrum tab](#). The table and the peak labels on the spectrum are complementary ways of judging how well the peptide explains the spectrum. The fragmentation table is particularly good at showing ladders of fragment ions while the spectrum shows ion intensity, non-matching peaks and it also labels internal ions.

The Fragmentation table shows the following ions:

Table 7-1: Fragment ions

Fragment Ion	Note
B ions	

Table 7-1: Fragment ions

Fragment Ion	Note
B+2H	Doubly charged B ion
B-NH ₃	B ion with loss of ammonia
B-H ₂ O	B ion with loss of water
Y ions	
Y+2H	Doubly charged Y ion
Y-NH ₃	Y ion with loss of ammonia
Y-H ₂ O	Y ion with loss of water

The cells in the Fragmentation table are color coded according to [Table 7-1](#)

- Colored cells indicate that the peptide fragment ion is in the spectrum.
- White cells indicate the predicted peptide ion is missing from the spectrum.
- Blank cells indicate the ion is not chemically possible.

A context menu is available when the user right clicks the mouse over the Fragmentation table, see [Context Menus Right Click Commands](#).

Chapter 8

Similarity View

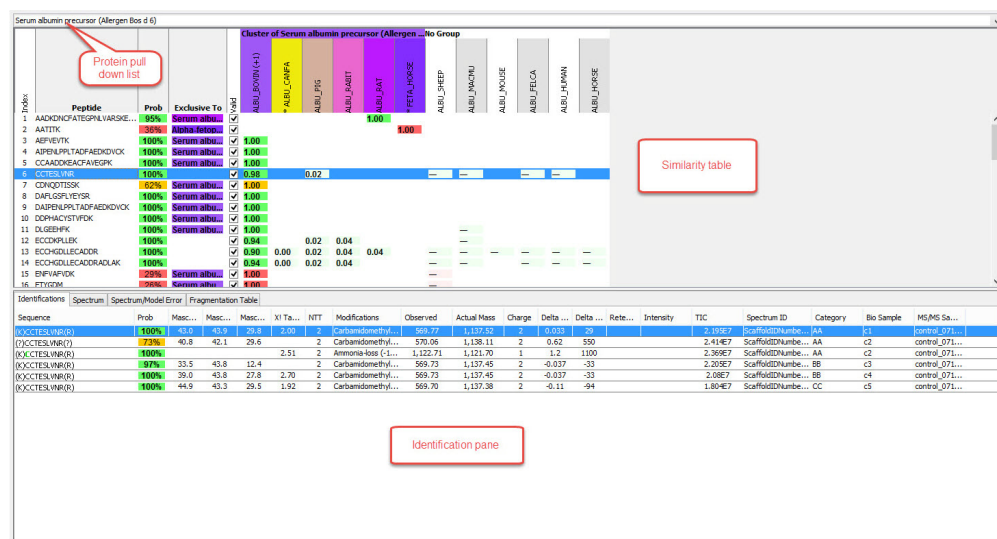
The Similarity View is designed to help the user view and understand which peptides support the identification of the proteins listed in [The Samples Table](#).

The user can select the protein of interest from the [The Samples Table](#) and then click the Similarity View button located in the [Navigation pane](#) or click the menu option **Window > Similarity** or double-click on the Protein Grouping Ambiguity star to switch to this view. Once in the Similarity View, the user can choose another protein from a pull down list.

The Similarity View, see [Figure8-1](#), includes:

- **Protein pull down list** - It provides a selectable list of the protein names present in the [The Samples Table](#) and it is located above the Similarity pane.
- **Similarity table** - a table where the peptide assignment is elucidated in detail and where the user can decide whether or not a peptide is valid. In experiments with traditional protein grouping (no clusters), the probabilities of peptides that are associated with a protein but not “assigned” to that protein are enclosed in parentheses. These peptides are also marked with a red X in the Proteins View.
- **Identifications pane** - It includes a number of tabs that helps the user look through the experimental evidence and judge whether the peptide protein mapping was properly done.

Figure 8-1: Similarity View



In [The Samples Table](#), proteins which have peptides shared with other proteins are marked with a star in the Protein Grouping Ambiguity column. These proteins are the most

interesting to investigate within the Similarities View. Double clicking on the Protein Grouping Ambiguity stars will also switch to the Similarities View. The Protein Grouping Ambiguity star is colored in red when the user has not yet viewed the Similarity table for the protein, otherwise it is shown as colored in green.

Since peptides belong to proteins independent of the sample in which the peptides were detected, the Similarity view shows the user the combined evidence across all of samples.

All protein quantification methods based on measuring peptide quantities in MS/MS data depend upon a satisfactory treatment of the mapping of peptides to proteins. The Similarity view let's the user review and control the peptide/protein mapping.

Similarity table

The Similarity table is located in the upper half of the Similarity View. It shows how Scaffold groups peptides and proteins based on evidence included in the loaded data. It does this by showing how peptides are associated with multiple proteins and how those proteins are grouped together based upon the peptide evidence.

Scaffold uses three different grouping algorithms that treat shared peptides in quite distinct ways. The grouping algorithms, described in [Chapter 12, “Protein Grouping and Clustering,” on page 203](#) affect the way the Similarity table is organized. The user can select how Scaffold treats shared peptides from the dialog in its protein grouping pane.

Figure 8-2: Similarity table; Shared peptide grouping

Myoglobin					Cluster of Myoglobin (MYG_HORSE)								Cluste... No Group
Index	Peptide	Prob	Exclusive To	Valid	MYG_HORSE (+1)	MYG_CASFI	MYG_GALCR (+2)	MYG_OCHPR	MYG_ORYAF	MYG_RABIT	MYG_GLOME (+2)	MYG_ELEMA (+2)	
1	ADIAGHGQEVLR	100%		<input checked="" type="checkbox"/>	0.76	0.02		0.11		0.11			
2	ALELFR	33%		<input checked="" type="checkbox"/>	0.63	0.01	0.09	0.09	0.09		0.09	—	
3	ETLEKFDKFNLIKSEDEMGKS...	100%	Myoglobin	<input checked="" type="checkbox"/>				1.00					
4	GDFGADAQGAMTK	100%	Myoglobin	<input checked="" type="checkbox"/>	1.00								
5	GLSDGEWQQVLNWWGK	100%	Myoglobin	<input checked="" type="checkbox"/>	1.00								
6	HGTVLTALGGILK	100%	Myoglobin	<input checked="" type="checkbox"/>	1.00								

Each row in the Similarity table shows a peptide and which proteins this peptide is part of. After the first descriptive columns, each column shows one of the similar proteins and which peptides are in the proteins.

The number and type of descriptive columns depends on the selected grouping algorithm:

- **Index** - numbers the list of peptides
- **Peptide** - shows the peptide sequence
- **Prob** - [appears only when Shared peptide grouping is selected] - Probability assigned to the peptide. All probabilities are shown as 99% in Prefiltered Mode.
- **Exclusive to** - shows to which protein the peptide is exclusive to
- **Valid** - it contains a check box. Based on the evaluation of the validity of the peptide assignments, the user can check or un-check the peptides from the table. Un-checking a peptide means that the user doesn't believe that the mass spectral evidence supports this peptide being in the experiment. By allowing or disallowing some peptide assignments, the user can regroup the proteins to exclude shared peptides or merge proteins into larger protein groups.

When the Shared peptide grouping algorithm is selected, at the intersection of a peptide row and a protein column, the weight which represents the apportionment of the peptide among the proteins where it is found is reported, see [Figure8-2](#).

When the Legacy protein grouping algorithm is selected at the intersection of a peptide row

and a protein column is a peptide probability, see [Figure8-3](#). This probability is the highest peptide probability throughout all the samples for the peptide. The probability is shown in parentheses if the peptide is not assigned to this protein (see [Standard Protein grouping](#)). In Prefiltered Mode, the primary search engine score for a peptide replaces the probability.

Figure 8-3: Similarity table; Legacy protein grouping

Serum albumin precursor (Allergen Bos d 6)																
Index	Peptide	Exclusive To	Valid	Serum albumi... Seru... Seru... Seru... No Group												
				ALBU_BOVIN	ALBU_CONTR	* ALBU_PIG	ALBU_RABIT	* ALBU_RAT	ALBU_SHEEP	ALBU_MACMU	ALBU_CANIFA	ALBU_FELCA	ALBU_HUMAN	ALBU_HORSE	ALBU_MOUSE	FETA_HORSE
1	AADKDNCFATEGPNILVARSK...	Serum albu...	<input checked="" type="checkbox"/>													
2	AATITK		<input checked="" type="checkbox"/>													(36%)
3	AEFVEVTK	Serum albu...	<input checked="" type="checkbox"/>	100%	100%											
4	AIPENLPPLTADFAEDKDVCK	Serum albu...	<input checked="" type="checkbox"/>	100%	100%											
5	CCAADDKEACFAVEGPK	Serum albu...	<input checked="" type="checkbox"/>	100%	100%											
6	CCTESLVNR		<input checked="" type="checkbox"/>	100%	100%	(100%)				(100%)	(100%)		(100%)	(100%)		
7	CDNQDTISSK	Serum albu...	<input checked="" type="checkbox"/>	62%	62%											
8	DAFLGSFLYEYSR	Serum albu...	<input checked="" type="checkbox"/>	100%	100%											
9	DAIPENLPPLTADFAEDKDVCK	Serum albu...	<input checked="" type="checkbox"/>	100%	100%											
10	DDPHACYSTVFVK	Serum albu...	<input checked="" type="checkbox"/>	100%	100%											
11	DLGEEHFK	Serum albu...	<input checked="" type="checkbox"/>	100%	100%						(100%)					
12	ECCDKPLEK		<input checked="" type="checkbox"/>	100%	100%	(100%)	(100%)				(100%)					
13	ECCHGDILLECADDR		<input checked="" type="checkbox"/>	100%	100%	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)
14	ECCHGDILLECADDRADLAK		<input checked="" type="checkbox"/>	100%	100%	(100%)	(100%)		(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	
15	ENFVAFVK	Serum albu...	<input checked="" type="checkbox"/>	29%	29%					(29%)						
16	ETYGDM	Serum albu...	<input checked="" type="checkbox"/>	26%	26%					(26%)						
17	ETYGDMADCCCK	Serum albu...	<input checked="" type="checkbox"/>	100%	100%					(100%)						
18	EYEATLEECCA	Serum albu...	<input checked="" type="checkbox"/>	100%	100%											
19	EYEATLEECCA/DDPHACYST...	Serum albu...	<input checked="" type="checkbox"/>	100%	100%											
20	FKDLGEEHFK	Serum albu...	<input checked="" type="checkbox"/>	100%	100%						(100%)					

Each protein column is identified by the protein's accession number. The accession numbers are written vertically. Above the protein's accession number is the name of the protein group that the protein is assigned to.

Proteins that have exactly the same peptides are grouped together. The protein groups are color coded. There is no experimental evidence to decide between the proteins in a protein group, so Scaffold includes them all. The name given to the protein group is by default that of one of the proteins, but this protein group name can be changed either specifically in the [Protein Information pane](#) located on the lower left side of the Samples View or globally using the command [Apply Protein Annotation Preferences](#).

Proteins whose peptides are a subset of the peptides in one of the named protein groups are identified by the column header “No Group”. While these proteins may be in the experiment's samples, there is no independent proof that they are. All the experimental evidence, that is all the MS/MS spectra, can be explained without these “No Group proteins”. For more information see [Protein Paring](#) and [Standard Protein grouping](#).

Scaffold's parsimony principle uses the simplest set of proteins which will explain all the data. Scaffold does not include these “No Group” proteins as part of the experiment.

Almost all protein quantitation methods calculate protein abundance from peptide abundance. Each peptide abundance may be due to the abundance of any of the proteins it is a part of. So it is necessary to resolve the peptide to protein assignments before calculating the protein abundance.

This resolving can be done by excluding ambiguous peptides from the experiment by unchecking them. By doing so, the several proteins can be merged into a protein group, or peptides which are shared between groups can be removed.

- Note:** The similarity table shows all the identified peptides independently of the filters applied in the Samples view. Those proteins that do not pass the filter, and so do not appear in the samples view protein list, will be marked by a star close to their accession number in the protein column headings, as shown in [Figure8-3](#).
- Note:** The column “Other Proteins”, shown in the [Peptides pane](#) found in the Proteins View, will list only proteins containing the peptide and passing the Samples view filters.

Identifications pane

The Identifications pane is located in the lower half of the Similarity View and it includes four tabs from where the user may review evidence supporting the assignment of the peptides.

- [Identifications tab](#)
- [Spectrum tab](#)
- [Spectrum/Model Error tab](#)
- [Fragmentation Table tab](#)

Figure 8-4: Identification pane

Sequence	Prob	Masc...	Masc...	Masc...	Xi Ta...	NTT	Modifications	Observed	Actual Mass	Charge	Delta ...	Delta ...	Rate...	Intensity	TIC	Spectrum ID	Category	Bio Sample	MS/MS Sa...
(K)ASLPLK(E)(K)	100%	28.3	37.7	13.5	1.64	1	Acetyl (+42)	513.81	1,025.61	2	-0.0062	-6.0	2680		82430	8126: Scan 1319...	Control	BioSample 1	qe2_101220...
(K)ASLPLK(E)(K)	100%	29.0	37.8	13.6	1.14	1	Acetyl (+42)	513.81	1,025.61	2	-0.0057	-5.5	2740		351300	5985: Scan 1350...	Control	BioSample 1	qe2_101220...
(K)ASLPLK(E)(K)	99%	16.6	37.8	6.8	1.64	1	Acetyl (+42)	513.81	1,025.61	2	-0.0052	-5.1	3160		774050	6147: Scan 1369...	Treatment	BioSample 2	qe2_101220...
(K)ASLPLK(E)(K)	100%	30.4	37.8	0.0	0.68	1	Acetyl (+42)	513.81	1,025.61	2	-0.0050	-4.8	3220		13090	4209: Scan 1394...	Treatment	BioSample 2	qe2_101220...

Identifications tab

The Identifications tab lists all the spectra found in all samples in the experiment that have been matched to the peptide highlighted in the [Similarity table](#). The peptide that shows the best probability is the one reported in the [Similarity table](#).

Identification table Columns:

- **Sequence** - Peptide sequence identified by the spectrum. The string of amino acids in the peptide is prefaced and followed by one residue symbol in parentheses. For example: (K)TGQAPGFSYTDANK(N). The symbols in parentheses represent the residue just before and just after the peptide, within the entire protein sequence. This gives the user additional information such as confirming whether the peptide is tryptic or semi-tryptic, based on whether it begins just after a K or R. See also NTT.
- **Prob** - Probability assigned by Scaffold to the peptide identification according to the selected validation algorithm at the time of load. The peptide probabilities are shaded according to the legend shown in the Samples Table.
- **Search engine scores** - A number of columns report the search engine scores used to analyze the loaded data set. Scaffold combines and aligns results of the same data analyzed with different search engines and reports here the different assigned scores. As an example find below the different score listed for some of the most commonly used search engines.
 - *SEQUEST* - Xcorr and DeltaCn
 - *Mascot* - Ion score, Identity score and Delta Ion Score.

- *X! Tandem* - log of expectation score
- **NTT** - Number of termini consistent with the enzymatic cleavage or tryptic termini
- **Modifications** - List of modifications identified by the spectrum
- **Mass measurements** - See also [Computation for Peptide mass values](#):
 - *Observed* - Mass over charge (M/Z) of the parent or precursor ion measured by the mass spectrometer.
 - *Actual Mass* - Actual peptide mass in Dalton
 - *Charge* - Peptide charge
 - *Delta Da* - (Actual Mass - Calculated peptide mass) in Dalton
 - *Delta PPM* - (Actual Mass - Calculated peptide mass) in PPM also referred to in the spectrum as the Parent error.
- **Retention Time** - Measured in seconds, it is included in the table only if the information is listed in the peak list of the loaded data.
- **Intensity** - Peptide precursor intensity area under the curve. This column is populated only when the data imported include this type of computation.
- **TIC** - MS/MS Total Ion Current, for more information see [Total Ion Current \(TIC\)](#).
- **Spectrum ID** - Name of the spectrum from which the information shown has been extracted from.
- **Category** - see [Category or Sample Category](#).
- **BioSample** - see [BioSample](#).
- **MS/MS sample** - [MS or MS/MS Sample](#).

Spectrum tab

The Spectrum tab displays one of the spectra that identify the peptide selected in the [Similarity table](#). It is typically the first spectrum appearing in the [Identifications tab](#) where all the identification spectra in the experiment for the selected peptide are listed.

The spectrum graph is interactive:

- The user can click anywhere on the spectrum to display the M/Z value for the position.
- The user can click and hold the left mouse button anywhere on the spectrum and then drag the mouse pointer to any position in the spectrum of his/her choosing. As the user drags the mouse pointer, the Start and Stop M/Z values for the segment are displayed as well as the length for the segment. The user can release the button to zoom in on the selected region. A single click of the mouse returns the zoom out magnification to 100%.

- The user can right-click anywhere on the spectrum to open the context menu which provides a number of menu options as described in the [Context Menus Right Click Commands](#) section.

Note: For Waters-PLGS, the Spectrum tab will be split into two tabs:

- **RT Summary Plot** - Shows the collected data versus the retention time. The various measured ion fragments are color coded depending whether they are unidentified or identified peptides or non peptides fragments. The size of the points shown in the plot can be toggled to be either scaled or not to their intensity values.
- The second tab shows the regular spectrum for the selected peptide.

Spectrum/Model Error tab

Works as described in the [Spectrum/Model Error tab](#) in the Proteins View.

Fragmentation Table tab

Works as described in the [Fragmentation Table tab](#) in the Proteins View.

Chapter 9

Quantify View

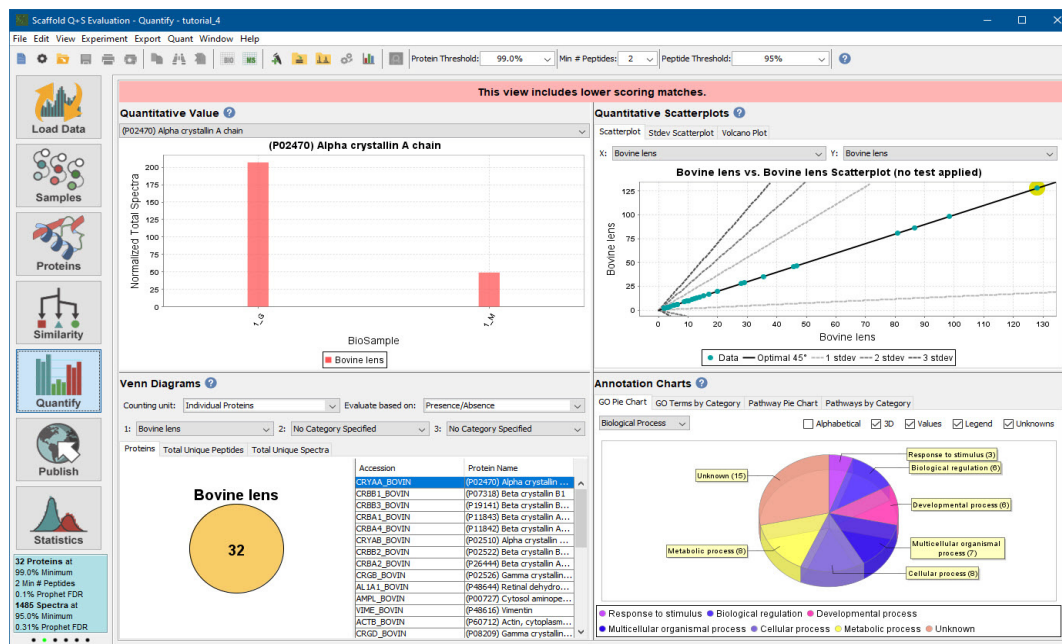
Scaffold's Quantify View provides graphical tools to help the user visualize experiments and draw conclusions about the quantitative relationships demonstrated in the data. From the Quantify View, the user can compare quantitative values between samples and categories, analyze the biological functions of the proteins identified in the experiment, and assess the reliability of the statistical analysis of the data.

The Quantify View, see [Figure 9-1](#), can be reached through the Quantify button located in the [Navigation pane](#) or through the menu **Window > Quantify**.

The Quantify view includes the following panes:

- [The Quantitative Value pane](#), located in the upper left of the Quantify View, provides information about quantitative values of a specific protein and allows comparisons among biosamples and categories.
- [The Quantitative Scatterplots pane](#), located in the upper right of the Quantify View, includes tools to analyze differences and correlations among protein quantitative values in the different categories.
- [The Venn Diagrams pane](#), located in the lower right of the Quantify View, shows the relationship between proteins, total unique peptides and total unique spectra in various categories, and allows the user to easily identify proteins, peptides or spectra of interest.
- The [Annotations pane](#), located in the lower right of the Quantify View, helps identify which proteins may be biologically significant.

Figure 9-1: Scaffold Quantify View



Graphical User Interface Actions in the Quantify View

The different panes included in the Quantify view share the following graphical user interface (GUI) features:

- The [Quantitative Value pane](#), the [Quantitative Scatterplots pane](#) and the [Venn Diagrams pane](#) are coordinated in such a way that selecting a point in one of them updates the other two panes accordingly. For example:
 - A single click on a point in the [Quantitative Scatterplots pane](#) highlights the point with a yellow circle in all applicable scatterplots, updates the bar chart or histograms in the [Quantitative Value pane](#) and updates the pull-down list associated with the bar chart.
 - Selecting a protein from the pull-down list in the [Quantitative Value pane](#) marks points on scatter charts and brings up appropriate bar chart.
- Zooming into a plot retains the zoomed view allowing the user to click on any point in the graph.
- Double clicking on a point in any scatter plot selects the particular protein in all views, and in the [Stdev Scatterplot tab](#) it shifts to the [Proteins View](#).

- Initially, no protein groups are selected, and it is only when the user either (A) single-clicks on a point in a scatter plot or (B) selects the protein from the drop-down menu does the protein become selected, and then stays selected.
- Right-clicking in any of the panes brings up a context menu as described in [Context Menus Right Click Commands](#).

The Quantitative Value pane

The Quantitative Value pane contains a bar chart or histogram called Quantitative bar chart. The chart provides a view of the relative abundance of a specific protein (selected through a pull-down list) across different BioSamples and categories.

For a protein, selected among the ones appearing in the Samples Table, the chart plots the quantitative values for all biosamples in the experiment. It highlights the categories by coloring the bars according to which category a biosample belongs to. Through a pull-down list, located above the chart, the user can select a different protein.

- The Y-axis displays types of quantitative values according to the Quantitative Method selected in the dialog from the related pull-down list. The plotted values are the same as the ones appearing in the Samples table when Quantitative Values is selected from of the Display Options pull-down list. These values typically depend upon the protein, peptide, required mods search filters and thresholds set on the [The Samples Table](#).
- X-axis displays bars for each BioSample in the Scaffold experiment. The bars are color coded according to the defined categories.

If the loaded dataset contains replicates, from this pane the user can assess the consistency of quantitative values across replicates within each category while comparing expression levels of the protein among categories. This allows visual inspection of the data and provides insight into the meaning of statistical comparisons such as the T-test or ANOVA.

The Quantitative Scatterplots pane

The Quantitative Scatterplots pane is located in the top right corner of the [Quantify View](#). It has three option tabs:

- The [Scatterplot tab](#)
- The [Stdev Scatterplot tab](#)
- The [Volcano Plot tab](#)

Scatterplot tab

This tab contains a scatter plot that helps visualize differences in the proteins quantitative values belonging to two different categories. The data points in the chart represent the mean of the quantitative values within a category versus the mean of the quantitative values in another category for all proteins appearing in [The Samples Table](#).

- **X-axis:** Mean of protein quantitative values for all samples in one category.
- **Y-axis:** Mean of protein quantitative values for all samples in a different category.

The type of quantitative value used in the plot depends on the Quantitative method selected from the pull-down list found in the dialog. The values will also depend on whether the **Use Normalization** option is selected and on the minimum value selected from the pull-down menu **Minimum Value**. Both options are available in the same dialog.

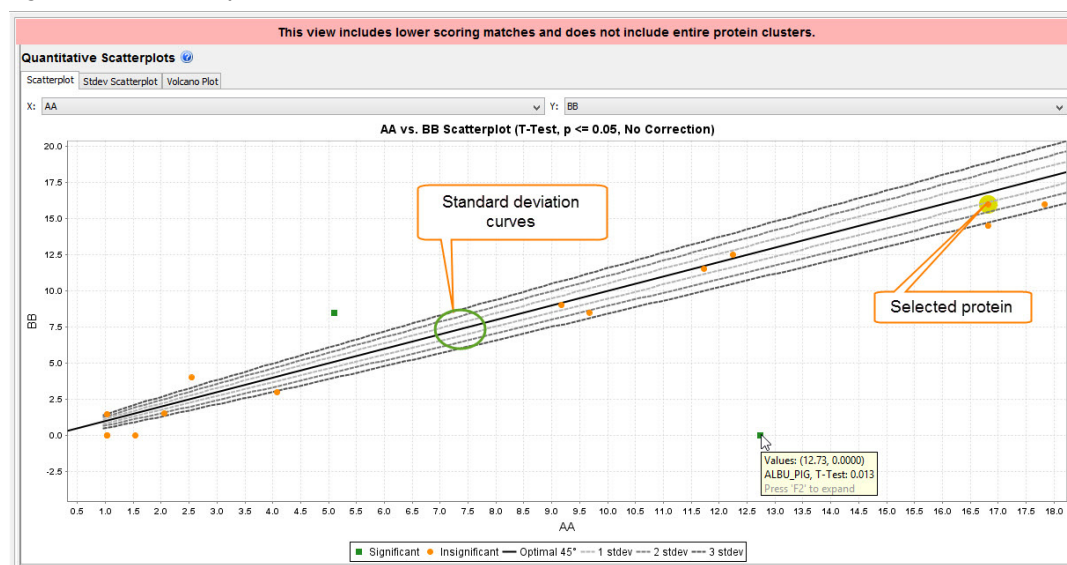
The categories assigned to each axis can be changed through the two pull-down menus located above the scatter plot.

- [Information tools in the plot area](#)
- [Tools to assess differences](#)

Information tools in the plot area

- Hovering over any point in the graph activates a tool tip that shows the accession number of the protein the point represents and the related plotted values. If two or more proteins have the same coordinates the tool tip will list them all.
- Clicking on a data point in the graph automatically selects (1) the protein it represents in the Quantitative Value pane and (2) in the Venn Diagram pane highlights the section where the protein is included in the diagram.

Figure 9-2: Scatterplot tab: tools



Tools to assess differences

To help assess and verify differences Scaffold draws in the chart a 45 degree line and 1, 2 and 3 standard deviations curves which show the value of the mean displaced by 1, 2 and 3 standard deviations above and below the 45 degree line. These error curves are a function of the mean so their traces form exponential curves.

The standard deviation at a given mean value is taken from the curve fitting determined in the $\text{Log}_{10}(\text{Mean})$ vs. $\text{Log}_{10}(\text{Stdev})$ scatter plot, see the [Stdev Scatterplot tab](#). Then, the adjusted mean (Mean+1 Stdev, Mean+ 2 Stdev etc.) is computed and the collection of all data points is curve fitted to determine a trace of 1 Stdev, 2 Stdev etc. The Stdev curves are always drawn from the standpoint of the X-Axis, to draw from the Y-Axis the user can simply switch the dataset around.

When the average quantitative values for a protein in the two different categories are the same, the data point will fall on the 45 degree line. The standard deviation lines help assess the degree of differences between the mean quantitative values in the two categories. When proteins plot outside the error curves they can be considered as differentially expressed.

When a Quantitative Analysis Test is applied through the dialog, all data points in the graph are colored according to the statistical significance determined based on the test. Note that depending on the type of test the data used might be a subset of the whole data included in the experiment and this is the data visualized in the Scatterplot.

Significance is defined depending on the test applied as follows:

- **Tests with p-values (T-test, ANOVA and Fisher's Exact test)**- Significance is defined according to the p-value threshold selected from the **Significance Level** pull-down list that becomes available when a p-value test is selected in the dialog. Multiple Testing Correction is also available in the same dialog.
- **Coefficient of Variance** - a value is defined "significant" when it is equal or larger than 1.

- **Fold Change** - a value is defined “significant” when it is either greater than or equal to 2.0 or it is less than or equal to 0.5.

In general, reasonable agreement should be expected between the distance of the data points from the 45 degree line and the coloring of the points according to the statistical test applied.

When no statistical test is applied, all data points are rendered in a single color.

Stdev Scatterplot tab

The Stdev Scatterplot provides a method of estimating the coefficient of variation or variance (CV) of the estimates of protein abundance. The tab includes a graph that plots the Log10(Mean) and Log10(Standard Deviation) for each protein appearing in the protein list for the whole data set and categories when no quantitative test is applied. When a statistical test is applied the Stdev scatter plot will include only the subset of data selected through the dialog.

- **X-axis:** Log10 of the mean value of the estimated protein abundance across all samples
- **Y-axis:** Log10 of standard deviation of the estimated protein abundance computed across all samples.
- [Information tools in the plot area](#)
- [Assessing quality of the data](#)

Information tools in the plot area

- All data points are colored according to the category from which they are derived.
- Hovering over any point in the graph activates a tool tip that shows the accession number of the protein the point represents, its category and the related plotted values.
- Double clicking on a data point takes the user to the Proteins View where the protein identification for all samples and categories is provided.
- Selecting a data point highlights all points that refer to the same protein in all categories defined in the experiment.

Assessing quality of the data

A regression line is calculated to provide a model that defines the two standard deviation lines shown in the plot included in the [Scatterplot tab](#).

$$\text{Log}_{10}(\text{Stdev}) = m \cdot \text{Log}_{10}(\text{Mean}) + b$$

Where m and b are the regression parameters. This theoretical estimation is represented as a dashed line in the plot and shows that in general the larger the estimated protein abundance, the larger is the absolute uncertainty in the estimate.

Another way of using this graph is to evaluate if the percent uncertainty, the CV, is roughly

constant, see reference [Pavelka \(2004\)](#) and [Pavelka \(2008\)](#). This method of estimating the CV uses all the available data. In most instances using all the data for an estimate gives the best estimate. However like any time a line is fit to data, it is possible for outliers to cause inaccuracies. Outliers in this data that will introduce the most inaccuracies are proteins with a high estimated abundance in the samples of one category and low abundance in another sample category. The implicit assumption is that if some abundant proteins are greatly suppressed in one category of samples, they will be balanced by roughly the same number of abundant proteins with elevated levels.

Note that this curve-fitting technique has only been explored in the literature using NSAF values, and using this quantitative metric leads to the best fit.

Note: If a visual inspection of this graph suggests that outliers have distorted the estimate of the CV, care should be taken when interpreting the [Scatterplot tab](#).

Caution: At least one category must contain multiple samples, or no standard deviation values can be calculated. When this is the case this plot will not be generated, and the standard deviation traces will not be visualized on the Q-Q [Scatterplot tab](#).

Volcano Plot tab

A volcano plot is a type of scatter plot used to quickly identify changes in large datasets composed of multiple replicate data. It plots significance versus fold-change on the y- and x-axes, respectively, combining in this way a measure of statistical significance from a statistical test (e.g. a p-value from a T-test) with the magnitude of the change between categories. It enables a quick visual identification of those data points that display large magnitude changes that are also statistically significant.

A volcano plot is constructed by plotting:

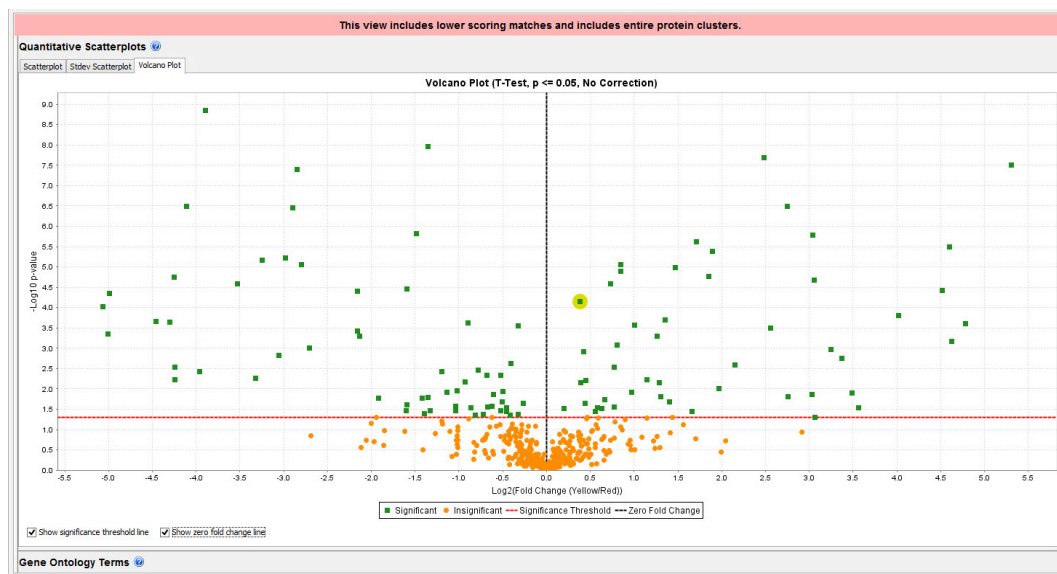
- X-Axis: the log of the fold change between the two conditions. The log of the fold-change is used so that changes in both directions (up and down) appear equidistant from the center.
- Y-Axis: the negative log of the p-value (usually base 10). This results in data points with low p-values (highly significant) appearing towards the top of the plot.

Plotting points in this way result in two regions of interest in the plot: those points that are found towards the top of the plot that are far to either the left- or the right-hand side. These represent values that display large magnitude fold changes (hence being left- or right- of center) as well as high statistical significance (hence being towards the top).



The categories assigned to each axis in the Volcano Plot can be changed through the two pull-down menus located above the scatter plot in the Scatterplot tab.

Figure 9-3: Volcano Plot



In Scaffold the Volcano plot is generated when statistical tests like [T-Test](#) and [Fisher's Exact Test](#) are selected. These tests compute a p-value and compare only two categories.

- [Information tools in the plot area](#)

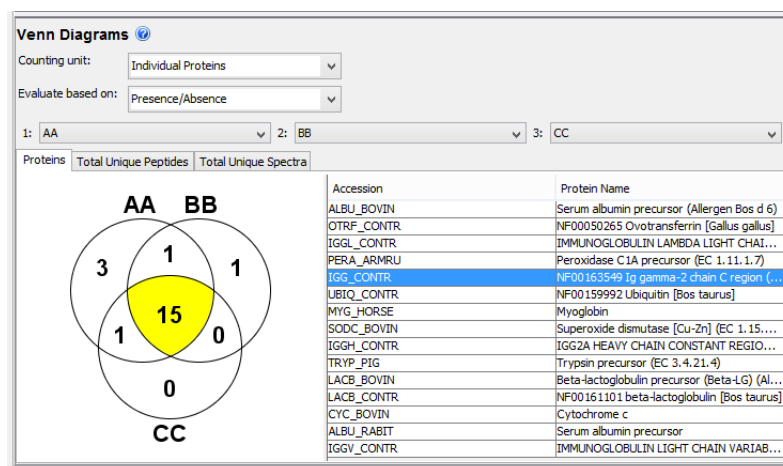
Information tools in the plot area

- Hovering over any point in the graph activates a tool tip that shows the accession number of the protein the point represents, the related plotted values and a p-value. If two or more proteins have the same coordinates the tool tip will list them all.
- Clicking on a data point in the graph automatically selects (1) the protein it represents in the Quantitative Value pane and (2) in the Venn Diagram pane highlights the section where the protein is included in the diagram.
- Under the graph area there are check boxes to display the statistical significance threshold, as well as the 0-fold change line, see [Figure 9-3](#).

The Venn Diagrams pane

The Venn Diagram pane includes three tabs each containing Venn diagrams with different types of quantitative values and a table placed on the right side of the diagram. The table becomes visible whenever the user selects a section of the diagram. When a section of the diagram is selected it appears highlighted in yellow. Over the tabs there are three pull-down lists showing the categories available in the experiment and pull-down lists to define the counting units and counting methods applied to the diagrams.

Figure 9-4: Venn Diagram pane



Through the Venn Diagram pane, the user can take a look at the relationship among proteins, total unique peptides, or total unique spectra identified in the various categories. Each of the tabs display a Venn diagram showing the overlap of up to three categories and reflect the current filters and thresholds applied in the Samples View. The user can decide which category is visible in the diagram by selecting them through the three drop-down lists located above the Venn Diagram.

- **Proteins tab** - The diagrams show the number of proteins groups identified in each category and in the overlap between two or among up to three categories. When a section of the diagram is selected, the table shows the list of proteins included in the highlighted section. Clicking on a protein group in the list will highlight the row and update accordingly [The Quantitative Value pane](#) and [The Quantitative Scatterplots pane](#).

Note: For this tab only and when the **Use protein Clustering Analysis** is selected, the **Counting unit** pull-down list allows the user to select among protein or cluster counts as counting units.

Note: For this tab only and when a Quantitative test is selected, the option **Evaluate based on:** *Quantitative Profile* becomes available, see [Quantitative profile](#).

- **Total Unique Peptides tab** - The diagrams show the sum of [Total Unique Peptide Count](#) for each protein in a category for up to three categories. When a section of the diagram is selected, the table shows the list of peptides included in the highlighted section.
- **Total Unique Spectra** - The diagrams show the sum of the [Total Unique Spectrum Count](#) for each protein in a category for up to three categories. When a section of the diagram is

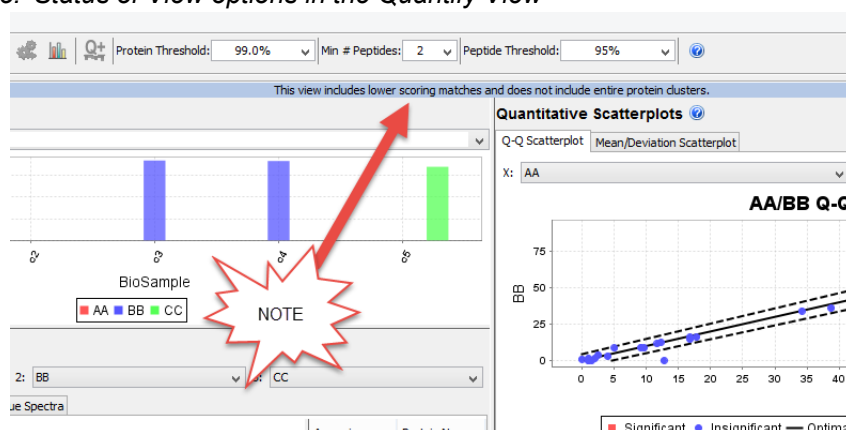
selected, the table shows the list of identified peptides, with their charge and modifications, included in the highlighted section.

The numbers shown in the Venn diagrams include all proteins and peptides displayed in the Samples view.

If the View option Show Entire Protein Clusters is selected, the counts will include the lower-scoring proteins that are part of displayed clusters, see [Clusters in the Samples Table](#).

If the View option [Show Lower Scoring Matches](#) is selected, the numbers include peptides not meeting the current thresholds but currently displayed because they do meet the thresholds in other biosamples for the same protein. The status of these options is shown at the top of the window. To count only proteins and peptides that meet thresholds, the user can go to the View menu and turn off these options.

Figure 9-5: Status of View options in the Quantify View



- The Venn Diagram is interactive. When a region of the diagram is selected, the protein accession numbers, peptide sequences, or spectral peptide sequence and nominal charge (spectra display) display in the table next to the Venn diagram.
- When a region of the Venn diagram is double clicked on, Scaffold switches to the [Samples View](#) and applies an Advanced Filter so that only those proteins in the selected region are displayed. The **Search** input box becomes highlighted in yellow and displays **Advanced...**
- To remove the Venn Diagram's applied advanced filter the user needs to clear the contents of the highlighted yellow input box or simply double-click outside the Venn Diagram in the diagram pane.

Annotations pane

The Annotations pane is populated only when GO terms or pathways have been searched and found, see [Add or Edit Annotations](#) and [GO Annotations and Pathways in Scaffold](#). Each protein in the Samples Table may show one or more Gene Ontology Terms or pathways which are associated with that protein. These terms are very useful to attach biological

significance to the results.

The detailed GO terms describing each protein are summarized in the pane in broader categories called ontologies. Each one of these ontologies has its own pie and bar charts. The user can select which ontology to display using a drop-down above the chart. The three ontologies available are:

- Biological process
- Cellular component
- Molecular function

Separate tabs provide pie and bar charts showing all displayed pathways.

Pie Charts

Each slice of the pie chart corresponds to one column of the GO term or pathway annotations in the Samples view. The annotation represented by a slice is shown in a box linked to the slice. If Show Values is checked, the number of proteins annotated with that GO term is also shown. Since a single GO term may be associated with more than one protein, these numbers may sum to a value greater than the number of proteins.

Double-clicking on a section of the pie chart filters the proteins in the experiment to show only the proteins annotated with that GO term or pathway, and opens the Samples view. A filtered set can be further filtered by returning to the pie chart and double-clicking again. More sophisticated GO filtering can be done through the Advanced Filters dialog.

Bar Charts

The bar charts are organized by category. Each bar displays the number of proteins annotated with a certain GO term or pathway in a certain category. This allows you to compare if proteins associated with a certain biological structure or function are over-represented in one category as compared to another.

Note that the bar chart only indicates presence or absence of proteins in the samples in a certain category. It does not reflect quantitative values.

Chapter 10

Publish View

The Scaffold Publish View displays information about the MS data loaded and analyzed in the program that is typically required for publication in a number of Proteomics journals, such as:

- Molecular & Cellular Proteomics,
- Proteomics
- Journal of Proteomics Research

Experimental Methods tab

This tab includes a table placed on the right side of the view and on the left side a page with a description of the experiment and three buttons that generate excel exports of the peptide and protein data reports and of the publication report itself.

- [Parameter table](#)
- [Description of the experiment](#)

Parameter table

This table contains two columns: the column **Parameter** where the parameters are described and the column **Value** where the values are reported when known. Rows are gathered in the following groups:

- **Peak List Generator** - which includes a number of parameters used to generate peaks from the raw data files.
- **Database Set** - which includes information about the databases searched to identify the proteins, with the parameters used by the search engine (or engines), and the criteria used for protein identification.
- **Search Engine Set** - which lists the search engines used for protein identification and the parameters used in the searches.
- **Scaffold** - which lists all the versions of Scaffold used to analyze the data along with the threshold settings and validation methods applied to the loaded data.

Note: Much of this data has been filled in by Scaffold from the information in the data files. Red highlights missing information that is required for publication in Proteomics journals. This information was not available in the loaded data files. It can be filled in manually.

Completing this table is the first step toward uploading Scaffold sf3 files and spectra for

publication purposes. To describe the experimental methods please ensure that ALL blanks in the Experiment Methods section in Scaffold's Publish view are filled in. It is very easy to miss some of the blanks. The table can be exported to excel by using the Export Publication report button located at the bottom of the Description of the experiment section.

Description of the experiment

The right side of the page has a narrative description of the same information. This can be used as a rough draft of the methods section of a journal article. Although the user will undoubtedly want to refine this computer generated text to improve its readability, it provides the user with a starting point.

Under the text window there are buttons which produce the Protein and Peptide reports. These reports may be useful as supplemental data supporting a publication in a Proteomics journal.

Figure 10-1: Publish View

Experiment Methods

Parameter	Value
Experiment:	tutorial_4
Peak List Generator:	
Version:	
Charge States Calculated:	
Desiotope:	
Textual Annotation:	
Database Set:	5 Databases
Database Name:	a subset of the swissprot_bovine database
Version:	
Taxonomy:	All Entries
Number of Proteins:	159
Database Name:	a subset of the swissprot_bovine database
Version:	
Taxonomy:	All Entries
Number of Proteins:	162
Database Name:	a subset of the swissprot_bovine database
Version:	
Taxonomy:	All Entries
Number of Proteins:	171
Database Name:	a subset of the swissprot_bovine database
Version:	
Taxonomy:	All Entries
Number of Proteins:	209
Database Name:	a subset of the swissprot_bovine database
Version:	
Taxonomy:	All Entries
Number of Proteins:	257
Database Name:	the swissprot_bovine database
Version:	
Taxonomy:	All Entries
Number of Proteins:	1705
Explain Database w/ < 1000 entries:	
Does database contain common contaminants:	
Search Engine Set:	2 Search Engines
Search Engine:	Sequest
Version:	27, rev. 17

DESCRIPTION OF THE EXPERIMENT

DATABASE SEARCHING-- Tandem mass spectra were extracted by (unknown) version (unknown). Charge state deconvolution and desiotope were not performed. All MS/MS samples were analyzed using Sequest (Thermo Fisher Scientific, San Jose, CA, USA, version 27, rev. 12) and XI Tandem (The GPM, ThermoFinnigan, version 2007.01.01.1). Sequest was set up to search the swissprot_bovine database (unknown version, 1705 entries) assuming the digestion enzyme non-specific. XI Tandem was set up to search a subset of the swissprot_bovine database (unknown version, 159 entries) (only "bovine_mudpit_11") also assuming non-specific, a subset of the swissprot_bovine database (unknown version, 162 entries) (only "bovine_mudpit_10") also assuming non-specific, a subset of the swissprot_bovine database (unknown version, 171 entries) (only "bovine_mudpit_12") also assuming non-specific, a subset of the swissprot_bovine database (unknown version, 209 entries) (only "bovine_mudpit_13") also assuming non-specific and a subset of the swissprot_bovine database (unknown version, 257 entries) (only "Mudpit_10") also assuming non-specific. Sequest was searched with a fragment ion mass tolerance of 0.00 Da and a parent ion tolerance of 2.5 Da. XI Tandem was searched with a fragment ion mass tolerance of 0.50 Da and a parent ion tolerance of 2.5 Da. Iodoacetamide derivative of cysteine was specified in Sequest and XI Tandem as a fixed modification. Oxidation of methionine and tryptophan, acetylation of the N-terminus and phosphorylation of serine and threonine were specified in Sequest and XI Tandem as variable modifications.

CRITERIA FOR PROTEIN IDENTIFICATION-- Scaffold (version Scaffold_5.0.0beta, Proteome Software Inc., Portland, OR) was used to validate MS/MS based peptide and protein identifications. Peptide identifications were accepted if they could be established at greater than 95.0% probability by the Peptide Prophet algorithm (Keller, A et al Anal. Chem. 2002;74(20):5383-90). Protein identifications were accepted if they could be established at greater than 99.0% probability and contained at least 2 identified peptides. Protein probabilities were assigned by the Protein Prophet algorithm (Nesvizhskii, A et al Anal. Chem. 2003;75(17):4646-58). Proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principles of parsimony.

Export Protein Report Export Peptide Report Export Publication Report

Chapter 11

Statistics View

The Statistics View provides tools to check how the protein and peptide validation algorithms, which assign probabilities to peptides and proteins, have been applied to the data in the MS Samples loaded in Scaffold.

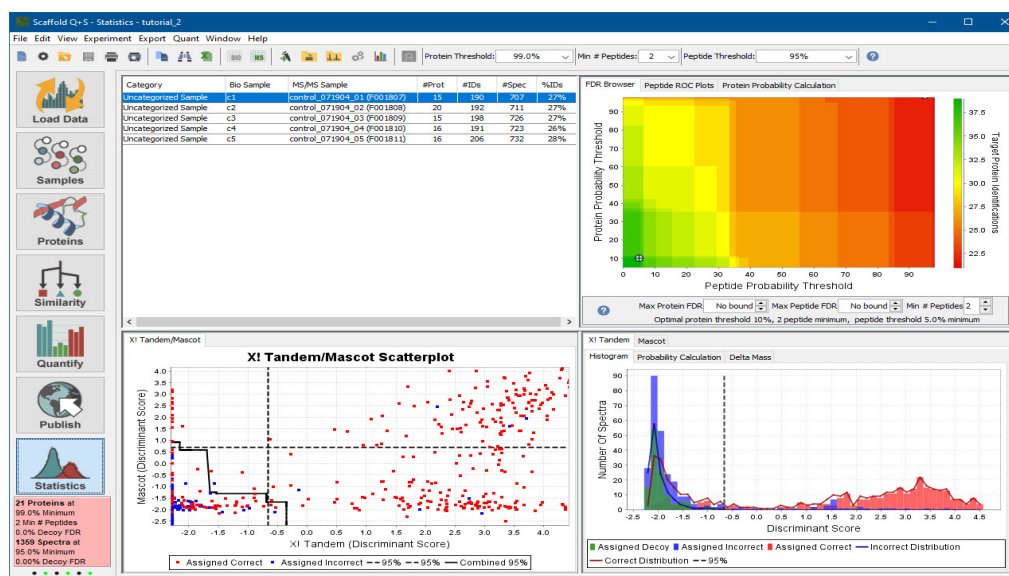
The Statistics View is not shown when data is loaded in Prefiltered Mode, unless probabilities can be calculated from Percolator results.



If you don't see the Statistics view, and the data was not loaded in prefiltered mode, it might be hidden.

Go to [Preferences](#), Display Settings tab and select Show Statistics View.

Figure 11-1: Statistics View



The available tools, which are tables and plots of statistical distributions, are arranged in four different panes. The MS/MS Samples table is used to select the specific sample to be analyzed in the remaining panes.

- [MS Samples Table](#) - Table that lists all the MS Samples loaded in the experiments. The charts in the other panes represent the sample selected here.

- [Statistics View Upper Right pane](#) - Includes three different panes containing graphs:
 - FDR Browser - plotting false discovery rates, (FDRs) versus peptide and protein probabilities (shown only for decoy searches)
 - Peptide ROC Plots - false positive rate versus number of identified peptides
 - Protein Probability Calculation - protein probabilities versus peptide probabilities.
- [Multiple Search Engine Scatter Plot pane](#) - Compares peptide scores assigned to spectra by two different search engines (shown only if two or more search engines were used).
- [Peptide Validation pane](#) - Shows the histograms built by the selected validation algorithms from which the peptide probability assignments are inferred.

Graphical User Interface Actions in the Statistics View

The different panes and tables included in the Statistics View share the following graphical user interface (GUI) features:

- Selecting a row from the [MS Samples Table](#) re-plots the graphs appearing in all the other panes included in the view
- Adjusting thresholds from the [Thresholding pane](#) updates the [Statistics View Upper Right pane](#), the [Multiple Search Engine Scatter Plot pane](#) and the [Peptide Validation pane](#).
- Hovering on a point in any scatter plot shows a tool tip reporting the values plotted in the graph.
- Right-clicking in any of the panes brings up a context menu as described in [Context Menus Right Click Commands](#).

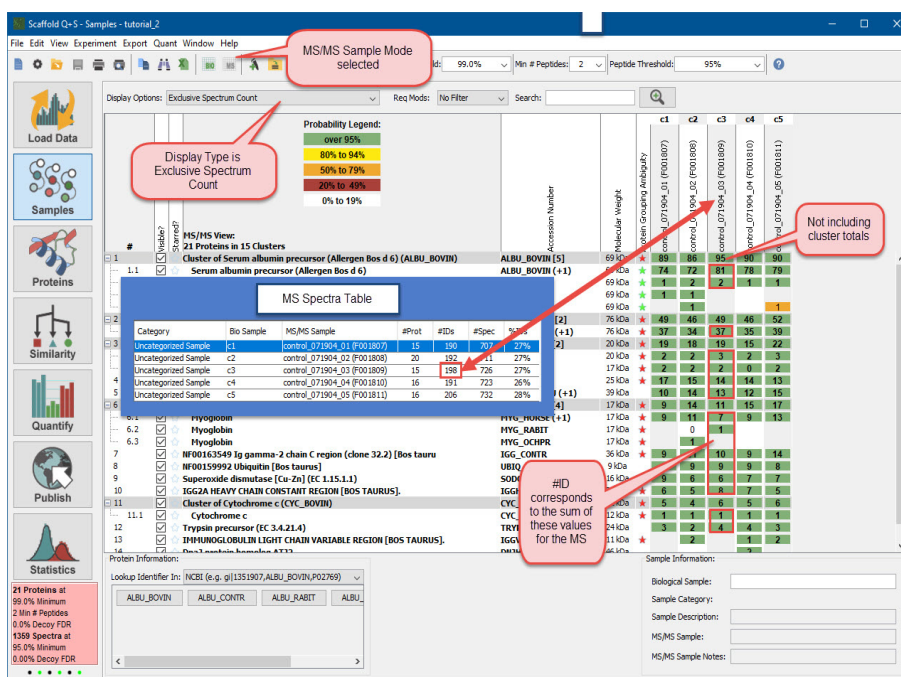
MS Samples Table

The MS Samples table is located in the upper left of the Statistical View. All MS samples loaded in the experiment, one for each row, are listed in the table.

MS Samples table columns

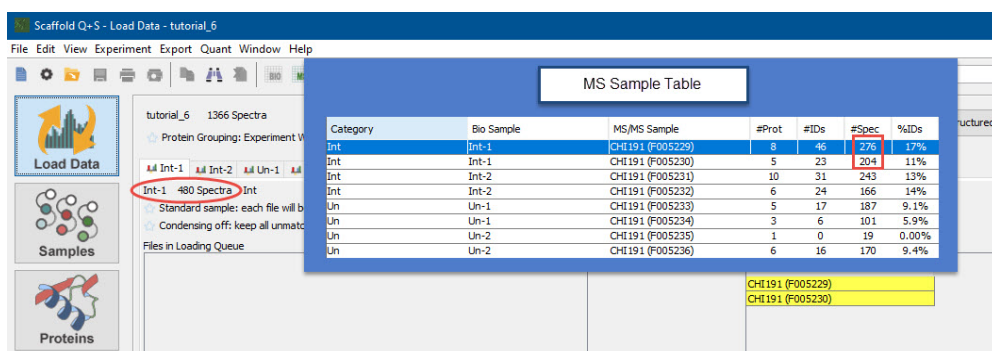
- **Category** - see [Category or Sample Category](#).
- **BioSample** - see [BioSample](#).
- **MS/MS sample** - see [MS or MS/MS Sample](#).
- **#Prot** - Number of identified proteins in the MS Sample (This number depends on the filters and thresholds applied to [The Samples Table](#)).
- **#IDs** - Number of Identified Spectra. (This number depends on the filters and thresholds applied to). The value shown corresponds to the sum over all proteins of the Exclusive Spectrum Counts within an MS sample. To check this value, set the summarization mode to [The MS/MS Sample Mode](#); select **Display Options: Exclusive Spectrum Count** and then sum the values appearing in the table over all proteins for each MS Sample. When clustering is selected, sum only the values for each protein in the list and not those corresponding to the overall cluster, see [Figure 11-2](#).

Figure 11-2: .Statistics View: MS Samples table: Computing #ID from Samples table.



- **#Spec** - Total number of spectra in the MS Sample. If the BioSample considered includes only one MS Sample, #Spec will be the same as the number of spectra reported in the [Load Data View](#) under the specific BioSample tab. Otherwise the number of spectra under the BioSample tab is the sum of the numbers of spectra in each MS Sample in the BioSample, see [Figure 11-3](#).

Figure 11-3: Statistics View: MS Samples table: Computing #Spec from Load data View



Category	Bio Sample	MS/MS Sample	#Prot	#IDs	#Spec	%IDs
Int	Int-1	CHI191 (F005229)	5	45	276	17%
Int	Int-1	CHI191 (F005230)	5	23	204	11%
Int	Int-2	CHI191 (F005231)	10	31	243	13%
Int	Int-2	CHI191 (F005232)	6	24	166	14%
Un	Un-1	CHI191 (F005233)	5	17	187	9.1%
Un	Un-1	CHI191 (F005234)	3	6	101	5.9%
Un	Un-2	CHI191 (F005235)	1	0	19	0.00%
Un	Un-2	CHI191 (F005236)	6	16	170	9.4%

- **%IDs** - Percentage of the total spectra in a MS Sample that has been identified as a peptide, which means for each peptide spectrum match (PSM). This is given by the ratio between #IDs and #Spec.

When a row in the table is selected, which selects an MS Sample, the [Statistics View Upper Right pane](#), the [Multiple Search Engine Scatter Plot pane](#) and the [Peptide Validation pane](#) are updated. They will display statistical information about the PSMs for the selected MS Sample.

Statistics View Upper Right pane

The Statistics View upper right pane provides information about probability assignments to peptides and proteins and how those assignments are translated into protein and peptide FDR values. The information is contained in three tabs each including graphical representations of the peptide and protein probability assignments calculated using the different validation algorithms available in the program.

- [FDR Browser pane](#)
- [Peptide ROC Plots pane](#)
- [Protein Probability Calculation tab](#)

FDR Browser pane

The FDR Browser tab appears only when decoy searches are loaded into Scaffold. It includes the following tools:

- **A Heat Map** - which shows how many target protein identifications are listed in the Samples table for any combination of peptide and protein thresholds. In the graph the x-axis represents the peptide probability threshold, the y-axis the protein probability threshold and the coloring the number of proteins identified as a function of the combination of peptide and protein probability thresholds.

Scaffold makes a series of these heat maps, one for each possible value of the minimum number of peptides. Changing the **Min # Peptides** pull down list located under the chart allows the user to step through these maps. Hovering over any point in the Heat Map activates a tool-tip which displays the number of proteins and other information about a selected point.

- **A Threshold pane** - Which includes three pull down menus used to adjust the desired FDR thresholds and select which heat map to explore through the **Min# Peptides pull** down list. In any of the FDR threshold pull down lists the user can select from a series of percentage FDR values or type in the desired one. At the bottom of the pane a suggestion for the optimal thresholds to provide the largest number of proteins is also shown.

The FDR Browser can help the user reach a better understanding of the loaded data and make informed choices about threshold settings. A user who is doing a peptidomics experiment, for instance, may be willing to settle for fewer protein identifications in order to get very high quality peptide assignments would explore the areas toward the right side of the graph. On the other hand, if the user is primarily interested in identifying a large number of candidate proteins, but peptide quality is not as important, it might be better to look along the top of the graph.

Often there are many possible threshold combinations that give nearly equivalent numbers of protein identifications, by default, Scaffold selects a point that maximizes the number of target proteins, then the protein probability and then the peptide probability, but depending on the purpose of the experiment, it may be desirable to adjust these priorities.

When looking at the browser, there are two spots marked with cross-hairs.

- The blue dot indicates the threshold settings that give the most proteins and meets the criteria specified in the controls under the browser.
- The pink dot represents the current settings in the program thresholds at the top of the screen. It is only seen if the Min # Peptides setting below the Browser matches the current Min # Peptides setting.

If it is impossible to meet the FDR cutoffs specified below the Browser with a pair of probability thresholds, a gray background is shown in the Browser.

Through the FDR Browser, the user may explore various settings and identify the best threshold settings for the specific experiment. These settings may then be applied by selecting the appropriate probability-based settings and minimum number of peptides in the experiment threshold settings at the top of the screen.

Peptide ROC Plots pane

The Peptide ROC Plots pane includes a graph that plots Receiver Operator Characteristic (ROC) curves, see [ROC curves](#), of the distribution of good and bad peptide spectrum matches (PSMs) or peptide assignments that appear in the Histogram tab of the [Peptide Validation pane](#). The validity of an assignment, defined as good or bad, and highlighted in red or blue in the histogram, is assessed by the selected peptide validation algorithm applied by Scaffold when the data is loaded. Note that one or more ROC curves might appear in the plot depending on whether the data was analyzed using one or more search engines.

As shown in the histogram, there may be two or three distributions of data, one defined as good and the other or others as bad. Some parts of the distributions may overlap making it difficult to distinguish whether PSMs are good or bad with 100% accuracy. This type of uncertainty is typically resolved by defining a cut off level beyond which the PSMs are considered good. In Scaffold this level is determined by the values assigned to the protein and peptide thresholds and by the min. number of peptides. The position of the cut off point determines the number of true positives, true negatives, false positives and false negatives assignments, see [Figure11-4](#), with a further imposition that the true peptide assignments belong to proteins listed in the Samples table.

Generally a good cut off level includes a small fraction of false negatives and a small fraction of false positives. Using a ROC plot it is possible to estimate a proper cut off level for assessing good PSMs. The plot typically compares the sensitivity, or true Positive Rate (PR), of a test with its False Positive Rate (FPR) as determined by varying the cut off level. Scaffold instead of using the PR compares the number of good PSMs or true peptide identifications, against the FPR, thus illustrating the trade-offs between the number of identified spectra and the FPR of the peptide assignments.

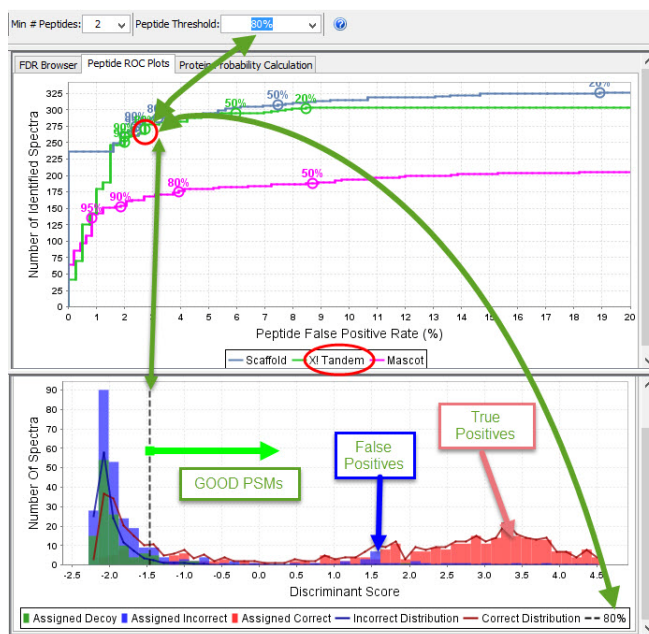
The FPR is defined as a function of the number of False Positives (FP) and the number of True Negatives (TN) as follows:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

The corresponding number of identified spectra are then calculated from the histogram by using the Peptide probability as the cutoff point. By varying the cutoff point the ROC plot is then constructed. The calculation includes all the information related to the different charge states.

An ideal ROC plot hugs the upper left corner of the graph, indicating that multiple identified spectra with a low false positive rate are present in the experiment.

Figure 11-4: *ROC plot: cut off level*



Protein Probability Calculation tab

The Protein Probability Calculation tab for each MS sample shows through a graph how the protein probability is correlated to the peptide probabilities.

Scaffold assigns probabilities to peptides using one of the validation algorithms described in [Increased Confidence Using Peptide and Protein Validation Algorithms](#). If the loaded data has been analyzed using more than one search engine, Scaffold assigns to a peptide a joint probability computed from several search engines.

The protein probability is computed separately for each MS sample based on the probabilities of the peptide or peptides of that protein. If the sample has proteins with more than one peptide, the one-peptide curve shows that the calculation is biased against “one hit wonders”.

This is Scaffold's way of adjusting the probability of finding a protein to the characteristics of the sample, the mass spec parameters, and the size of the FASTA database searched.

If the sample has too few spectra to make the statistics valid, default statistical distributions are used and the graphs are omitted.

This graph also shows that the protein probability can be high if there are two, three or more

peptides, even if each of those peptide hits are of relatively modest. This is somewhat in the spirit of PMF (Protein Mass Fingerprinting).

If after examining the Protein Probability Calculation graph, you feel that this calculation is not applicable for your sample, we recommend that you reduce the protein probability filter to its lowest value and filter exclusively by the number of peptides meeting the peptide probability filter.

Multiple Search Engine Scatter Plot pane

The Multiple Search Engine pane includes a tab containing a scatter plot that compares peptide scores assigned to spectra by two different search engines.

The tab is active only when the loaded spectra have been analyzed using more than one search engine.

Each plotted point corresponds to one Peptide Spectrum Match (PSM). The x and y axes of the scatter plot are the search engine discriminant scores. These scores tell how well the search engine rated the PSM. Bigger numbers are better.

The vertical and horizontal dashed lines depend on the value the user chooses for the “Peptide Threshold”. Changing this threshold changes where the lines are drawn. Roughly speaking, dots above the horizontal or to the right of the vertical lines represent PSMs that are above the filter, that is “good” matches. Scaffold uses a somewhat more complex algorithm to combine the scores from multiple search engines to separate the good from the bad.

From these PSM scores Scaffold calculates first peptide probabilities (see the Histograms) and protein probabilities (see the Protein Probability Calculation). The “assigned correct” proteins are determined from the minimum probability and number of peptide filters.

The scatter plot dots which correspond to PSMs assigned to “correct” proteins are colored red. The “incorrect” peptide dots are colored blue. As the user changes the filters to accept more or fewer proteins, some of the dots may change color.

Note: When the data is loaded in the condensed mode, the scatter plot will not include those points that correspond to spectra discarded in the loading phase. As a reminder, the heading of the Scatter Plot will report how the data has been loaded.

Peptide Validation pane

The Peptide Validation pane shows how Scaffold uses the validation algorithms described in [Increased Confidence Using Peptide and Protein Validation Algorithms](#) to turn peptide scores into probabilities for each MS sample loaded in the program.

The pane includes tabs for each search engine used to analyze the mass spec data loaded into Scaffold. Each of the search engine tabs includes three sub-tabs that vary in type and content depending on the validation algorithm selected when loading data into the program:

- [Percolator](#) - When Percolator is used to assess peptide identifications, this pane displays a histogram showing the correct and incorrect distributions of spectral matches as determined by Percolator. The bins are based on discriminant score.
- [PeptideProphet](#) - It includes a sub-tab for each charge state. The sub-tab shows a histogram which bins PSMs according to a discriminant score customized for each search engine used to search the experimental data. Two distribution curves are fit in every histogram. The lower curve (blue) is the distribution of incorrect matches. The higher curve (red) is the distribution of correct matches. For each MS Sample the user can judge from looking at the graph how well these distributions match the histogram. Bayesian statistics says that the peptide probability for a given discriminant score is the ratio of $(\text{Correct} / (\text{Correct} + \text{Incorrect}))$ distributions for any discriminant score.
- [PeptideProphet](#) and High Mass accuracy - When this option is selected at the time of loading, the Peptide Prophet histogram is augmented with a second histogram showing the distribution of the delta masses for all the spectra included in the experiment. To switch between the two histograms two icon appear under each of the charge tabs.

In the delta mass histogram the peptide identifications assumed to be correct by PeptideProphet are labeled in red, while incorrect identifications are labeled in blue. The ratio between red and blue is calculated within each bin and plotted using a smoothing function. The resulting curve is called the Mass Accuracy Adjustment Factor, shown as a dotted line in the graph.

Adjustment factors greater than 0.5 increase the peptide probability, while factors less than 0.5 decrease the probability.

- [LFDR-based scoring system](#) - This algorithm works specifically with searches performed against databases that include decoys. It is designed to improve upon the Peptide Prophet and Peptide Prophet and High Mass Accuracy approaches in three different ways. First, the curve-fitting in PeptideProphet is changed to an alternative approach that uses local false discovery rate (LFDR). Second, a new classifier system for combining search engine scores, such as XCorr and DeltaCN, has been implemented. Third, the delta mass error modeling has been adjusted for better performance. While this system requires target/decoy searches, any ratio of targets to decoys will work. For example, if the user has a QExactive instrument and he/she wants to search with tight parent and fragment ion tolerances, the user might want to consider searching a 1:10 ratio target to decoy database.

To ensure that LFDR is producing a good probability model, for each score Scaffold forms its distribution of forwards and decoys, and calculates the Mann-Whitney U Test rho-statistic, $\rho = U/(n1*n2)$. This rho is a measure of how much the two distributions overlap. Always $0 \leq \rho \leq 1$ and the closer rho is to 0.5 the more the distributions overlap. Each score is excluded from use in creating the classifier if $|\rho - 0.5| < 0.05$. If no scores remain that satisfy this, Scaffold will fall back to peptide prophet with high mass accuracy correction.

The LFDR tab includes the following sub-tabs:

- **Histogram sub-tab** - When dealing with a target / decoy search the user knows something about the data set that PeptideProphet doesn't. Half of the "incorrect" identifications should be assigned to decoy peptides, labeled in the histogram in green. In each of those bins the user can look at the ratio of the green bar to the non-green bars and estimate the likelihood a peptide in that bin should correspond to a decoy, which corresponds to the probability that a peptide in that bin is "incorrect". This ratio is essentially a local false discovery rate for each score bin, hence the acronym LFDR. The histogram allows to calculate probabilities for peptides without making any assumption about possible underlying distributions.
- **Probability Calculation sub-tab** - Shows the distribution of the probability a peptide is correct
- **Delta mass sub tab** - the delta mass analysis is performed as described in the Peptide Prophet and High Mass accuracy approach. This graph shows that when getting closer to 0 delta mass the probability a peptide is correct increases noticeably and this can be factored in directly into the probability model.

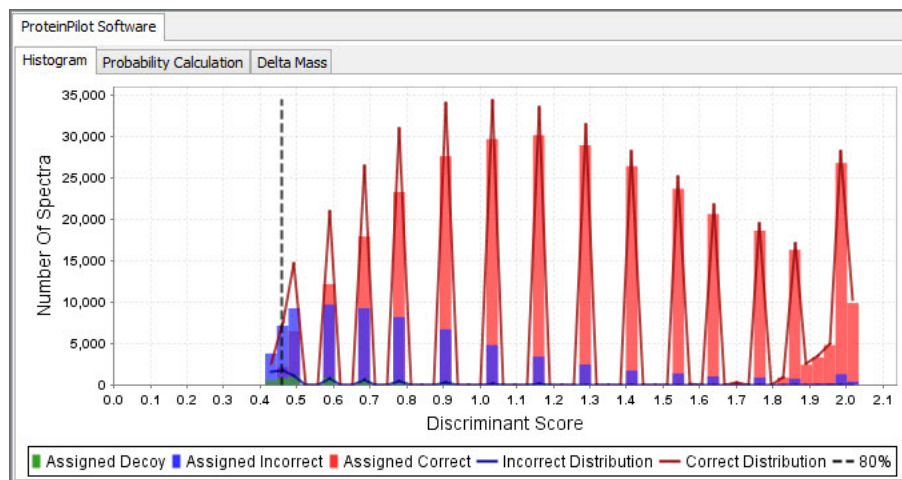
The black vertical dashed line appearing in the histograms indicates the point at which the ratio between correct and incorrect matches the Min Peptide probability threshold. By adjusting the Min Peptide threshold the user can move the line.

Note

When loading mzIdentML files created by AB Sciex ProteinPilot, Scaffold does not apply any validation algorithm and uses the scores computed by Paragon directly. The histogram appearing in the peptide validation pane will look quite different, see [Figure11-5](#).

When MSF files created by Proteome Discoverer using the Percolator node are loaded into Scaffold in Prefiltered Mode, Scaffold calculates probabilities directly from the PEP scores assigned by Percolator. In this case, the histogram shows the distribution of search engine scores, color-coded to indicate whether they are considered correct or incorrect. No probability function is applied by Scaffold, so no Probability Calculation or Delta Mass charts are displayed.

Figure 11-5: Scaffold's PSM histogram of a ProteinPilot AB Sciex MS sample



Chapter 12

Protein Grouping and Clustering

The different protein grouping and clustering algorithms offered in Scaffold are:

- [“Shared Peptide Grouping and Protein Cluster Analysis” on page 204](#), which provides a description of the Shared Peptide Grouping algorithm used for grouping, of the paring and clustering of proteins appearing in the Samples Table for version 4 and higher.
- [“Standard Protein grouping” on page 211](#), which provides a description of the grouping algorithm used in versions 3 and older and still applied in Scaffold version 4 and higher when the clustering option is not selected.

Note that in Prefiltered Mode” protein grouping and clustering proceed as usual, but all peptides are considered to have equal probabilities, unless their probabilities have been calculated based on ProteinPilot or Percolator scores.

Shared Peptide Grouping and Protein Cluster Analysis

Scaffold, in versions 4 and above, offers the option of applying a protein grouping method called Shared Peptide Grouping. Scaffold versions 3 and lower only provided a different grouping algorithm referred to as Standard Experiment Wide Protein Grouping.

Shared Peptide Grouping is designed to lessen the probability of discarding a valid protein identification when the protein happens to share many peptides with another identified protein. Scaffold versions 4 and higher also include the option to assemble proteins into clusters based on shared peptide evidence using Protein Cluster Analysis.

These two options are selected during file loading by checking **Use protein cluster analysis** in the **Load and Analyze** Wizard page within the **Protein Grouping** pane. Choosing this option enables the application of both Shared Peptide Grouping and Protein Cluster Analysis.

Figure 12-1: Load and Analyze Window

Load and Analyze Data

Searched Database(s):

1. unprot_sprot FASTA Database

Edit Databases...

MSFrager

☐ Analyze with MSFrager

Scoring System

☒ Use Percolator [help me choose](#)

☐ Use PeptideProphet scoring (high mass accuracy)

☐ Use prefiltered mode (accept search engine probabilities, e.g. from Percolator, when available)

☐ Other ...

Protein Grouping:

☒ Use protein cluster analysis

☐ Use standard experiment wide protein grouping

☐ Use legacy independent sample protein grouping

Optional Loading Steps

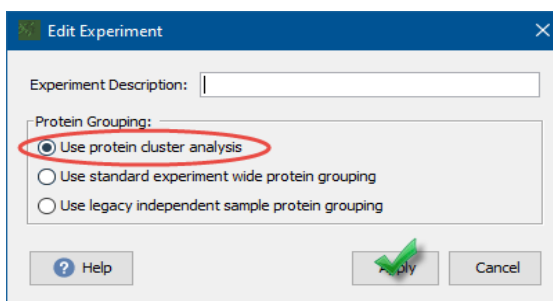
☒ Precompute FDR thresholds

☐ Annotate proteins [Choose source](#) [Configure Annotation Sources...](#)

If Protein Cluster Analysis is not selected during file loading, it can be applied to the already

loaded data by going to the menu **Experiment> Edit Experiment**. The **Edit Experiment Window** opens and, in the fully licensed version of Scaffold, the check box **Use Protein Cluster Analysis** is available. Selecting the check box and clicking **Apply** rearranges the protein groups and creates clusters using Shared Peptide Grouping and Protein Cluster Analysis.

Figure 12-2: Edit Experiment Window



For explanation purposes the grouping and clustering processes can be broken down into the following three phases:

- Protein Grouping
- Protein Paring
- Protein Clustering

Protein Grouping

The way Shared Peptide Grouping assigns peptides to proteins is quite different from the way it is done in the Legacy Protein Grouping algorithm. Rather than assigning each peptide to a single protein, Shared Peptide Grouping includes a peptide in each of its matching proteins. It then proceeds to form “Protein Groups” and assign weights to each shared peptide, see “Weighting Function”

Protein Groups

Scaffold considers proteins that share peptide evidence. In cases where two or more proteins share all of their peptides, there is no basis for discrimination amongst them and the proteins are grouped and treated as a unit called a *Protein Group*. These proteins appear in the Samples Table as a single line with the accession number of one of them followed by a plus and the number of other proteins in the group. The “preferred” or named protein is arbitrarily selected from the group and may be changed by the user.

Figure 12-3: Samples Table, Protein grouping

Cytochrome b-c1 complex subunit 8 OS=Mus musculus GN=U...	QCR8_MOUSE	10 kDa	LK3 transgenic mice	
40S ribosomal protein S29 OS=Bos taurus GN=RPS29 PE=3 SV...	RS29_BOVIN (+6)	7 kDa	Bos Taurus	
Homerin OS=Homo sapiens GN=HRNR PE=1 SV=2	HRNR_HUMAN	282 kDa	Homo sapiens	

Weighting Function

For the purpose of calculating the protein probabilities, shared peptides are apportioned among proteins according to a weighting function.

The weights are assigned by using the following formula:

$$W(p, A) = \frac{PE_{excl}(A)}{\sum_{All(B \supseteq p)} PE_{excl}(B)}$$

Where $W(p, A)$ is the weight assigned to shared peptide p contained in protein A and in other proteins. $PE_{excl}(A)$, the exclusive peptide evidence, is defined as the sum of the probabilities of each exclusive unique valid peptide X belonging to protein A .

$$PE_{excl}(A) = \sum_{X \subset A} P_X$$

This value is then normalized by the sum of the exclusive peptide evidence for each of the proteins that contain peptide p .



A peptide can be set “valid” either manually, by un-checking peptides in the Proteins View Peptide Table or globally by using the Experiment menu option **Reset Peptide Validation**. The Scaffold default cut off is 0%.

As an example [Figure 12-4](#) shows peptides shared by multiple proteins with their related weights listed in the Similarity view.

Figure 12-4: Similarity View- Peptide weights

Myoglobin									
Cluster of Myoglobin (MYG_HORSE)									
Index	Peptide	Prob	Exclusive To	Valid	MYG_HORSE (+1)	MYG_CASFI	MYG_GALCR (+2)	MYG_OCPR	MYG_ORYAF
1	ADIAGHGQEVLR	100%		<input checked="" type="checkbox"/>	0.76	0.02		0.11	
2	ALELFR	33%		<input checked="" type="checkbox"/>	0.63	0.01	0.09	0.09	0.09
3	ETLEKFDKFKNLKSEDEMKG...	100%	Myoglobin	<input checked="" type="checkbox"/>				1.00	
4	GDFGADAQGAMTK	100%	Myoglobin	<input checked="" type="checkbox"/>	1.00				
5	GLSDGEWQQVLNVWGK	100%	Myoglobin	<input checked="" type="checkbox"/>	1.00				
6	HGTVLTALGGILK	100%	Myoglobin	<input checked="" type="checkbox"/>	1.00				

These weights are also displayed in the Proteins View Peptide Table in the column titled “Weight”, which replaces the traditional “Assigned” column when the cluster grouping model is used.

Protein Paring

Next, the protein list is pared down according to the principle of parsimony. As in the case of the Legacy Protein Grouping, the Shared Peptide Grouping algorithm thins down the list of proteins by eliminating any for which there is no independent evidence. However, independent evidence is defined differently in the two grouping algorithms. In the Shared Peptide Grouping a protein is considered having independent evidence when it contains at least one exclusive unique peptide.

Proteins for which there is no exclusive evidence are then eliminated from the protein identification list. This process can best be seen in Scaffold's Similarity View. Here all proteins sharing peptide evidence are assembled into a table. Proteins with exclusive peptides are placed to the left and included in the experiment. Proteins for which all of the associated peptides are subsumed by these identified proteins are eliminated from further consideration, as there is no independent evidence of their presence in the experiment. These proteins appear in the "No Group" columns in the Similarity View but are otherwise invisible in Scaffold.

Figure 12-5: Similarity View - No group columns

Histone H2A type 1-B/E OS=Homo sapiens GN=HIST1H2AB PE=1 SV=2

Index	Peptide	Prob	Exclusive To	Valid	Cluste... H2A1B_HUMAN (+14)	Cluste... H2A1D_HUMAN (+18)	Cluste... H2AV_BOVIN (+21)	Cluste... H2AZ_CANAL (+1)	No Group H2A1A_HUMAN (+25)	No Group H2A1_CANAL (+30)	No Group H2A1_A5HGO (+40)
1	ATIAGGGVIPHIK	52%	Histone H2...	<input checked="" type="checkbox"/>			1.00				
2	HLQLAIR	80%		<input checked="" type="checkbox"/>	0.39	0.39	0.21	0.01			
3	HLQLAIRNDEELNK	100%		<input checked="" type="checkbox"/>	0.50	0.50					
4	NDEELNKLLGK	100%	Histone H2...	<input checked="" type="checkbox"/>		1.00					
5	NDEELNKLLGR	100%	Histone H2...	<input checked="" type="checkbox"/>	1.00						
6	VTIAQGGVLPNIQAVLLPK	100%		<input checked="" type="checkbox"/>	0.50	0.50					

Protein Clustering

Assembling proteins into clusters is based on shared peptide evidence. While akin to Mascot's hierarchical family clustering, Scaffold's Protein Cluster Analysis is more stringent in its requirement for two proteins to appear in the same cluster. This added stringency often succeeds in separating proteins into sets of biologically meaningful isoforms.

In essence, a cluster is a set of proteins with overlapping peptide evidence, and may be treated as a proxy for a single identification. This view allows interpretation of identification probability, spectral counts, and normalized quantitative values calculated on the level of clusters.

Cluster formation begins with the creation of protein groups as described above, see "[Protein Groups](#)". Next, these protein groups are grouped into clusters of similar proteins. Two proteins (or protein groups) are considered similar if their joint weighted peptide evidence is

at least half of the weighted peptide evidence of either protein. A protein is iteratively added to a cluster if it is similar to at least one other protein in the cluster. This information can be translated into the following rules of thumb for cluster formation:

1. For two proteins to be clustered, the sum of the probabilities of their shared peptides must be at least 95%.
2. The proteins must share at least 50% of their evidence. This is determined by summing the probabilities of the shared peptides and comparing this value with the summed probabilities of all of the peptides for each individual protein. If the sum of the probabilities of the shared peptides is greater than or equal to half of the sum of the peptide probabilities for either of the individual proteins, a cluster is formed.
3. A protein may be included in an existing cluster if it meets the above criteria with a member protein of the cluster.

For a detailed example of how a cluster is formed see an extended version of this document published on our website: [scaffold_protein_grouping_clustering.pdf](#).

Clusters in the Samples Table

Thresholds and filters do not affect the formation of clusters, but they do determine which clusters and proteins or protein groups are displayed in the Samples Table. Scaffold builds the Samples Table applying thresholds and filters to the formed clusters, proteins and proteins groups in the following order:

1. Select all clusters that pass thresholds.
2. Include all proteins and protein groups belonging to the selected clusters.
3. Prune proteins or protein groups based on selected filters
4. Remove clusters that do not include proteins
5. Prune proteins and proteins groups based on thresholds.

This order of applying thresholds and filters keeps clusters in the Samples Table that might not include proteins or protein groups that pass thresholds and filters.



Filters apply only to proteins or protein groups.

Clusters are shown in the Samples View as a line with protein name “Cluster of ...” and the name of one of the constituent proteins. This protein is designated as the “primary protein” of the cluster, but the primary protein may be changed by clicking on the accession number field of a cluster and selecting a different accession number from the drop-down list when it appears. A cluster may be expanded in the Samples View by clicking on the “+” at the left of the cluster’s row. When a cluster expands, it displays all of its constituent proteins or protein groups, including the primary protein. The right click menu provides bulk operations to expand or collapse all clusters simultaneously.

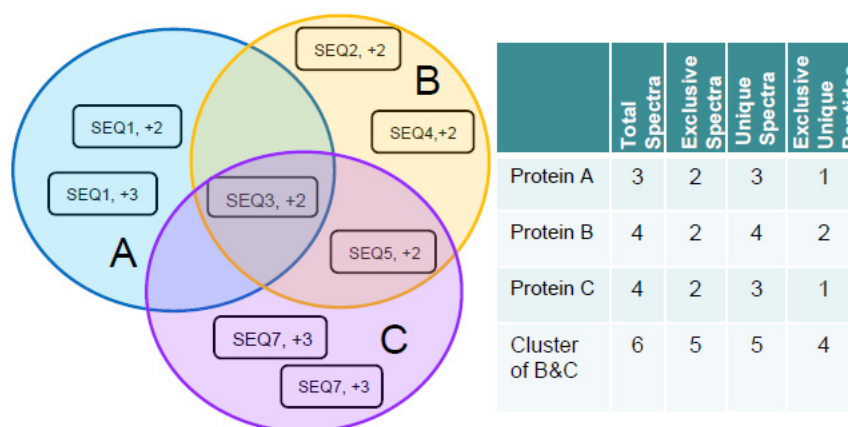
The menu option **View > Show Entire Protein Cluster** will show in a gray font proteins or protein groups in clusters that do not pass thresholds

Cluster Display Values

Display values are calculated for a cluster as a whole based on the set of peptides that make up the cluster. Note that these values are different from the values of any individual protein, including the primary protein of the cluster. Selecting a cluster and going to one of the other views displays all of the information for the entire cluster.

The diagram in [Figure 12-6](#) illustrates Scaffold's method of spectra and peptide counting in clustered proteins. The circles A, B and C represent three proteins of which B and C form a cluster. The little squares in the circles represent the spectra included in the proteins. Their charge is also indicated. The table on the side shows how the different quantitative values are counted for each protein and for the cluster. Note that the total spectra of the cluster does not correspond to the sum of the total spectra of the proteins included in the cluster because some of the peptides are shared.

Figure 12-6: Spectral counting in clustered proteins.



Standard Protein grouping

Scaffold groups proteins with the Legacy Protein Grouping algorithm if either the Use standard experiment wide protein grouping or the Use legacy independent sample protein grouping option is selected during loading or when data is already loaded in Scaffold and one of these options is selected in **Edit Experiment Window**.

Generally, the Legacy Protein Grouping algorithm groups proteins using a table very similar to the one shown in the Similarity View when the Protein Cluster Analysis option is not selected, see [Figure 12-7](#).

Figure 12-7: Scaffold Legacy Similarity View

Serum albumin precursor (Allergen Bos d 6)										
Index	Peptide	Exclusive To	Valid	Serum albumi...	Seru...	Seru...	Seru...	No Group	ALBU_BOVIN	ALBU_CONTR
1	AADKDNCFATEGPNLVARSK...	Serum albu...	<input checked="" type="checkbox"/>				95%			
2	AATITK		<input checked="" type="checkbox"/>							
3	AEFVEVTK	Serum albu...	<input checked="" type="checkbox"/>	100%	100%					
4	AIPTNLPLLTADFAEDKDVCK	Serum albu...	<input checked="" type="checkbox"/>	100%	100%					
5	CCAADDKEACFAVEGPK	Serum albu...	<input checked="" type="checkbox"/>	100%	100%					
6	CCTESLVNR		<input checked="" type="checkbox"/>	100%	100%	(100%)		(100%)	(100%)	(100%)
7	CDNQDTISSK	Serum albu...	<input checked="" type="checkbox"/>	62%	62%					
8	DAFLGSFLYEYSR	Serum albu...	<input checked="" type="checkbox"/>	100%	100%					
9	DAIPENLPLLTADFAEDKDVCK	Serum albu...	<input checked="" type="checkbox"/>	100%	100%					
10	DDPHACYSTVFDK	Serum albu...	<input checked="" type="checkbox"/>	100%	100%					
11	DLGEEHFK	Serum albu...	<input checked="" type="checkbox"/>	100%	100%			(100%)		
12	ECCDKPLLEK		<input checked="" type="checkbox"/>	100%	100%	(100%)	(100%)	(100%)		
13	ECCHGDILLECADDR		<input checked="" type="checkbox"/>	100%	100%	(100%)	(100%)	(100%)	(100%)	(100%)
14	ECCHGDILLECADDRDLAK		<input checked="" type="checkbox"/>	100%	100%	(100%)	(100%)	(100%)	(100%)	(100%)
15	ENFVAFVDK	Serum albu...	<input checked="" type="checkbox"/>	29%	29%			(29%)		
16	ETYGDM	Serum albu...	<input checked="" type="checkbox"/>	26%	26%			(26%)		

Assigning peptides to proteins

Initially the table of peptides and proteins has a column for every protein to which a peptide could potentially be assigned, and a row for every valid peptide that can be found in the listed proteins. When a peptide is found in a protein the peptide probability is shown in the appropriate cell. The sum of the probabilities is then calculated for each protein, see [Figure 12-8](#).

Figure 12-8: Initial Similarity Table

Scaffold Table Export.xls															
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Index	Peptide	Exclusive To	Valid	#1157831969	#113279011	#130585149	#1109935949	#111513661	#12136563	#173989799	#191092998	#1146747386	#1237572	#1169158861
2	1	AKWYPEVR		FALSE	9%	9%	9%	9%	9%			9%	9%	9%	
3	2	CVVVGDAVGK		FALSE	28%	28%	28%			28%		28%			
4	3	DDKDTIEK		TRUE	73%	73%	73%	73%	73%				73%	73%	73%
5	4	GSPQAIK	Chain A, Small G-Protein	TRUE	75%										
6	5	IISAMQTIKCVVVGDAVGK		TRUE											
7	6	KLTPITYPQGLAMAK	Chain A, Small G-Protein	TRUE	95%			95%	95%	95%	95%	95%	95%	95%	
8	7	LIPITYPQGLAMAK	Ras-related C3 botulinum tox	TRUE		95%	95%								
9	8	LTPITYPQGLAMAK	Chain A, Small G-Protein	TRUE	95%			95%	95%	95%	95%	95%	95%	95%	
10	9	LVPITYPQGLAMAK		TRUE											
11	10	TVFDEAIR		TRUE	95%	95%	95%	95%	95%	95%	95%	95%			95%
12	11	VDSKPVNLGLWDAGQEDYDR		TRUE											92%
13			Sum of probabilities		433%	263%	263%	358%	358%	285%	285%	285%	263%	263%	260%
14															
15															

Each peptide is then assigned to the protein that has the highest total probability among all those where the peptide is found, see Figure 12-9. If two or more proteins have equal total probabilities and that is the highest for that peptide, it is assigned to all of them.

Figure 12-9: Assigned peptides are shown in green, unassigned in gray

Scaffold Table Export.xls															
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Index	Peptide	Exclusive To	Valid	#1157831969	#1109935949	#111513661	#12136563	#173989799	#191092998	#113279011	#130585149	#1146747386	#1237572	#1169158861
2	1	AKWYPEVR		FALSE	9%	9%	9%			9%	9%	9%	9%	9%	
3	2	CVVVGDAVGK		FALSE	28%		28%	28%		28%	28%	28%	73%	73%	73%
4	3	DDKDTIEK		TRUE	73%	73%	73%								
5	4	GSPQAIK	Chain A, Small G-Protein	TRUE	75%										
6	5	IISAMQTIKCVVVGDAVGK		TRUE											
7	6	KLTPITYPQGLAMAK	Chain A, Small G-Protein	TRUE	95%	95%	95%	95%	95%	95%	95%	95%	95%	95%	
8	7	LIPITYPQGLAMAK	Ras-related C3 botulinum tox	TRUE		95%	95%				95%	95%			
9	8	LTPITYPQGLAMAK	Chain A, Small G-Protein	TRUE	95%	95%	95%	95%	95%	95%	95%		95%	95%	
10	9	LVPITYPQGLAMAK		TRUE											
11	10	TVFDEAIR		TRUE	95%	95%	95%	95%	95%	95%	95%	95%			95%
12	11	VDSKPVNLGLWDAGQEDYDR		TRUE											92%
13					433%	358%	358%	285%	285%	285%	263%	263%	263%	263%	260%
14															

Defining protein Groups

Now the grouping begins. Proteins with no peptides assigned are eliminated from consideration, the evidence for those proteins has already been accounted for in proteins which are more likely to be present in the analyzed sample. Proteins with the same peptides assigned to them are combined into a group, see Figure 12-10.

Figure 12-10: Protein groups formation

Scaffold Table Export.xls															
	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1				Group 1	Group 2		Group 3								
2	Peptide	Exclusive To	Valid	#115781969	#113275011	#130685149	#1169158861								
3	AKWYPEVR		FALSE	9%	9%	9%									
4	CVVVGDAVGK		FALSE	28%	28%	28%									
5	DDKDTIEK		TRUE	73%	73%	73%	73%								
6	GSPQAIK	Chain A, Small G-Protein	TRUE	75%											
7	IISAMQTIKCVVVGDAVGK		TRUE												
8	KLTPITYPQGLAMAK	Chain A, Small G-Protein	TRUE	95%											
9	LIPITYPQGLAMAK	Ras-related C3 botulinum tox	TRUE		95%	95%									
10	LTPITYPQGLAMAK	Chain A, Small G-Protein	TRUE	95%											
11	LVPITYPQGLAMAK		TRUE												
12	TVFDEAIR		TRUE	95%	95%	95%	95%								
13	VDSKPVNLGLWDTAGGEDYDR		TRUE				92%								
14				433%	263%	263%	260%								

There is one further complication, however. If the only evidence for a group is a single protein with probability less than 95%, Scaffold disregards this group. This is based on a heuristic rule built into the algorithm which cuts down on the number of false protein matches displayed. In this case it would eliminate Group 3, see [Figure 12-11](#).

Figure 12-11: Formed protein groups

Scaffold Table Export.xls															
	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1				Group 1	Group 2										
2	Index Peptide	Exclusive To	Valid	#115781969	#113275011	#130685149									
3	1 AKWYPEVR		FALSE	9%	9%	9%									
4	2 CVVVGDAVGK		FALSE	28%	28%	28%									
5	3 DDKDTIEK		TRUE	73%	73%	73%									
6	4 GSPQAIK	Chain A, Small G-Protein	TRUE	75%											
7	5 IISAMQTIKCVVVGDAVGK		TRUE												
8	6 KLTPITYPQGLAMAK	Chain A, Small G-Protein	TRUE	95%											
9	7 LIPITYPQGLAMAK	Ras-related C3 botulinum tox	TRUE		95%	95%									
10	8 LTPITYPQGLAMAK	Chain A, Small G-Protein	TRUE	95%											
11	9 LVPITYPQGLAMAK		TRUE												
12	10 TVFDEAIR		TRUE	95%	95%	95%									
13	11 VDSKPVNLGLWDTAGGEDYDR		TRUE												
14				433%	263%	263%									

Generally, this approach works well to eliminate false assignments, however in certain instances, it can result in a protein that may actually be found in the sample being eliminated from consideration, and thus not seen in Scaffold's other views. Unfortunately, changing the filter settings has no effect upon this type of grouping algorithm. A different approach can now be tried by using the clustering option available in Scaffold 4. The new grouping algorithm does not forcefully assign peptides uniquely to a protein but considers shared peptides among different proteins.

Chapter 13

Quantitative Methods and Tests

Scaffold supports label free quantitative methods. Some of them are based on spectrum counting and others are based on MS1 intensity measurements. For the purpose of establishing differential expressions among the categories present in a Scaffold experiment it is important to normalize the values and accommodate for systematic differences and experimental errors. Scaffold provides option for normalizing values and taking care of missing values.

- [Label Free Quantitative Methods](#)
- [Normalization among BioSamples in Scaffold](#)
- [Quantitative Tests](#)

Label Free Quantitative Methods

There are two widely used label free quantification strategies which are quite different in their approach and methods of accounting for the presence of proteins in a sample. There is also a third method that is sort of an in-between method. The methods are the following:

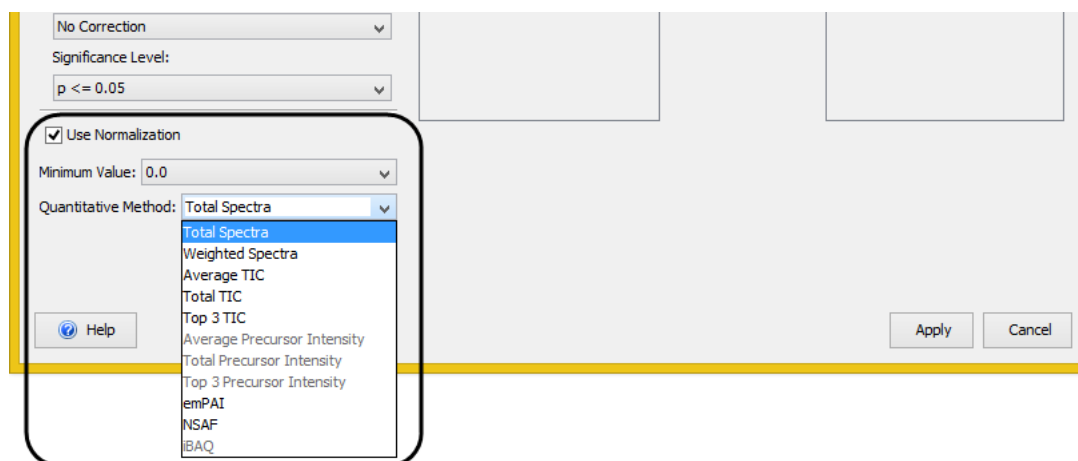
- **Spectrum Counting**, which counts and compares the number of fragment spectra identifying peptides of a given protein.
- **Total Ion Current (TIC)**, which considers peak intensities from MS/MS spectra combined with counting of the spectra
- **Precursor Ion Intensity**, which measures and compares the mass spectrometric signal intensity of peptide precursor ions belonging to a particular protein

For each of these main general methods Scaffold provides a number of variations that are commonly proposed in the standard literature:

- **Spectrum Counting**
 - Total Spectra - (default)
 - Weighted Spectra
 - emPAI
 - NSAF
- **Total Ion Current (TIC)**
 - Average TIC
 - Total TIC
 - Top Three TIC
- **Precursor Ion Intensity quantitation**
 - Average Precursor Intensity
 - Total Precursor Intensity
 - Top Three Precursor Intensities
 - iBAQ

In Scaffold a quantitative method is selected through a pull down list located in the dialog, see [Figure 13-1](#). The quantitative values can be viewed in the Samples Table when the option **Quantitative Value** is selected from the **Display Options** list.

Figure 13-1: Quantitative Method pull down list



How to choose the proper quantitative method

When setting up an experiment researchers typically have a question in mind. The question determines the way the experiment is organized and conducted and also which quantitative method needs to be used to find answers to the question asked.

Typical questions asked in a mass spectrometry Proteomics experiment are:

1. Is anything changing?
2. How much is the amount of change I am dealing with?

Of the three main label free quantitative methods available in Scaffold, [Spectrum Counting](#) methods are the most reliable in answering question [number 1](#). The [Total Ion Current \(TIC\)](#) methods can answer both questions but they provide only estimated quantites based on of the peak intensities from identified MS/MS spectra. [Precursor Ion Intensity quantitation](#) methods are very reliable in answering question [number 2](#).

Spectrum Counting

Scaffold includes the following Spectrum Counting methods:

- [Total Spectra - \(default\)](#)
- [Weighted Spectra](#)
- [emPAI](#)
- [NSAF](#)

Total Spectra - (default)

This method uses the sum of all the spectra associated with a specific protein within a sample which includes also those spectra that are shared with other proteins and is referred to as the Total Spectrum Count.

Weighted Spectra

This method uses the sum of all weighted spectra associated with a specific protein and within a sample, where the weight is a measure of how much a spectrum is shared by other proteins. For more details on how the weight is calculated see [Weighting Function](#).

emPAI

Spectrum Counting methods can also be used in the determination of absolute abundance of proteins. Initially the parameter used to measure this absolute abundance was the Protein Abundance Index (PAI) defined as the number of observed peptides divided by the number of all possible tryptic peptides from a particular protein, that are within the mass range of the employed mass spectrometer.

$$PAI = \frac{N_{observed}}{N_{observable}}$$

Where $N_{observed}$ is the number of experimentally observed peptides and $N_{observable}$ is the calculated number of observable peptides for each protein. In a subsequent refinement PAI was transformed into an exponential form called emPAI and defined as follows, see [Ishihama \(2005\)](#):

$$emPAI = 10^{PAI} - 1$$

In Scaffold the algorithm developed to calculate *emPAI* is modeled after the one used by Mascot as described in the following document: www.matrixscience.com/help/quant_empai_help.html.

While from the definition of *emPAI* one would expect values ranging from 0 to 9, for some proteins higher values can be found as shown in [Ishihama \(2005\)](#).

On the other hand we also have observed that extremely high *emPAI* values can be reached if there are too many modifications considered for a search and the search space is noticeably larger than the apparent number of peptides (without considering mods). Extremely high values can appear for other conditions like non-tryptic searches, etc

When Scaffold encounters these extreme cases it adjusts *PAI* as follows:

- If $PAI > 3 \rightarrow emPAI \sim 10^3$ the final *emPAI* value is truncated to $10^3 - 1 + (PAI - 3)$.

This will only be seen in very rare cases.

NSAF

The Normalized Spectral Abundance Factor ([NSAF](#)) quantitative method is used when comparing the abundance of individual proteins in multiple independent samples and is typically applied to quantify the expression changes in various complexes. It is generally calculated using the number of spectra (SpC) identifying a protein divided by the protein length (L), referred to as Spectral Abundant Factor (SAF) and then normalized over the total

sum of spectral counts/length in a given analysis. This means that SAF is then divided by the sum of SpC/L for all proteins in the experiment.

The NSAF values shown in [The Samples Table](#), when NSAF is selected as a quantitative value, are calculated using the NSAF strategy 2-a listed in Table 1 of [Zhang \(2010\)](#).

NSAF calculations in Scaffold:

The calculation used in Scaffold translates to the following expression:

SAF = number of exclusive spectra/ length of proteins (expressed in number of amino acids)

The SAF value is then normalized using the regular Scaffold quantitative value normalization scheme, see [Normalization among BioSamples in Scaffold](#). NSAF values are derived by normalizing the SAF values so that they sum to one across all proteins in a sample resulting in the values shown in [The Samples Table](#).

Total Ion Current (TIC)

Scaffold includes the following [TIC- Total Ion Current](#) methods:

- [Average TIC](#)
- [Total TIC](#)
- [Top Three TIC](#)

Average TIC

Average of all the TIC values of the spectra assigned to a protein. When selected the user needs to adjust the **Minimum Value**: accordingly by selecting **Other...** in the pull down list.

Total TIC

Sum of all the TIC values of all spectra assigned to a protein. When selected the user needs to adjust the **Minimum Value**: accordingly by selecting **Other...** in the pull down list.

Top Three TIC

Sum of the top three TIC values among the spectra assigned to a protein. When selected the user needs to adjust the **Minimum Value**: accordingly by selecting **Other...** in the pull down list.

Precursor Ion Intensity quantitation

Scaffold supports label-free quantitation based on precursor ion intensity for data analyzed with Proteome Discoverer, Mascot Distiller, MaxQuant and Spectrum Mill. For more updated information on supported data check the following document on Proteome Software website:

www.proteomesoftware.com/pdf/loading_search_engine_results_into_scaffold.pdf.

Precursor ion intensity refers to the area under an MS1 spectrum peak corresponding to a specific peptide, whereas spectral counting counts the number of spectra identified for a given peptide. Scaffold provides four options for considering a protein's precursor intensity in its Quantitative Methods drop-down list in the dialog. For more information on how the values shown in the Samples Table when one of these methods is selected, are computed see [Calculation of Precursor Intensities](#).

Scaffold includes the following Precursor Intensity Quantitation methods:

- [Average Precursor Intensity](#)
- [Total Precursor Intensity](#)
- [Top Three Precursor Intensities](#)
- [iBAQ](#)

Average Precursor Intensity

This method takes the geometric mean of the peptide intensity values for a given protein. When selected the user needs to adjust the Minimum Value accordingly by choosing **Other...** in the pull down list.

Total Precursor Intensity

The sum of all distinct intensity values for a protein. When selected the user needs to adjust the **Minimum Value**: accordingly by selecting **Other...** in the pull down list.

Top Three Precursor Intensities

The sum of the three highest peptide intensity values for a protein. If fewer than three peptides have intensity values, the intensities that are present are summed. When selected the user needs to adjust the **Minimum Value**: accordingly by selecting **Other...** in the pull down list.

iBAQ

iBAQ (Intensity-Based Absolute Quantification) is a popular approach for absolute quantification of proteins.

It is similar in its approach to the [emPAI](#) method, which belongs to the spectral count methods based on counting the number of identified unique parent ions per protein¹. In contrast, iBAQ and similar algorithms are called intensity-based because they calculate the sum of parent or precursor ion intensities of identified peptides per protein. In both types of methods, the numbers of theoretically possible peptides per protein for the protease used in sample preparation enter the equation to account for different protein lengths and distribution and frequency of cleavage sites.

In iBAQ, the sum of intensities of all tryptic peptides for each protein is divided by the number of theoretically observable peptides (fully tryptic, 6-30 amino acids, no missed

1. Mann, K and Edsinger, E, Proteome Science 2014, 12:28 doi:10.1186/1477-5956-12-28

cleavages). The resulting iBAQ intensities provide an accurate determination of the relative abundance of all proteins identified in a sample.

To calculate iBAQ Scaffold performs an in-silico digest of a considered protein based on a specified enzyme and gets the count of possible peptides. It counts only the resulting peptides of length 6 - 30 amino acids ignoring the possibility of missed cleavages. The precursor ion intensity of the peptide is then divided by the number of possible peptides. To check the values we suggest the following tool for performing in-silico digests http://web.expasy.org/peptide_cutter/

A comparison between the spectral count based versus the intensity based methods show a higher accuracy of the intensity-based methods, including iBAQ². It has also been noted that the emPAI method in its original form³ has become somewhat obsolete because of the recent progress in technology. For instance, modern mass spectrometers and the associated software provide high-confidence identifications of much longer peptides than previously possible. Consequently these long peptides are not included into emPAI calculations but are included in iBAQ calculation.

Note: This option is available only when precursor intensity data is loaded in Scaffold, see [Preparing Data for Precursor Intensity Quantitation in Scaffold](#).

-
2. Schwanhäusser B., Busse D., Li N., Dittmar G., Schuchhardt J., Wolf J., Chen W. and Selbach M. Nature 473 (2011), 337–342 doi:10.1038/nature10098
 3. Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J, Mann M, Mol Cell Proteomics 2005, 4:1265-1272. OpenURL

Normalization among BioSamples in Scaffold

To allow comparisons, Scaffold normalizes the MS/MS data. The user can then compare abundances of a protein between samples. The normalization scheme used works for the common experimental situation where individual proteins may be up-regulated or down-regulated, but the total amount of all proteins in each sample is about the same. It is not appropriate if the total amount of protein varies widely from one sample to the next.

In Scaffold there are two levels of summarization: the MS level which shows the samples run through the mass spectrometer; and the BioSample level, where BioSamples can contain one or more MS samples. Frequently the biological sample, or BioSample in Scaffold, is fractionated into multiple MS samples. Scaffold allows the user to view the MS samples within a BioSample or to combine all the MS samples into a single sample using the “MuDPIT” option. Normalization is performed at the MS sample level.

The normalization scheme in Scaffold adjusts the sum of the selected quantitative value for all proteins in the list within each MS sample to a common value: the average of the sums of all MS samples present in the experiment. This is achieved by applying a scaling factor for each sample to each protein or protein group adjusting in this way the selected value to a normalized “Quantitative Value”.

Note: For Precursor Intensities, which operate at the peptide level, the precursor intensity might be associated with more than one spectrum derived from the same peptide. These spectra all exhibit the same Intensity value. In the normalization scheme this value will only be considered once for calculation purposes. For more information see [Precursor Intensity Quantitation in Scaffold](#) and [Performing Quantitation in Scaffold](#).

Note on Low abundance peptides

The normalization method used in Scaffold, as mentioned above, distorts the data if the total protein loaded varies considerably from sample to sample. This is because low abundance peptides may be on the edge of detectability. If for example, sample A has a lot of protein loaded, the low abundance peptides may be detected. If sample B has much less protein loaded, these low abundance peptides might not be detected; that is, their spectral count is zero. Scaling cannot compensate for the failure to detect these peptides.

The user can view the normalized data by selecting the **Quantitative Values** option from the Display Options pull down list in the Samples View. When viewing Quantitative Value (Normalized Total Spectra), the default quantitative method in Scaffold, if the user switches from this value to Total Spectrum Count and notices that all of the values in one column change much more than the values in the other columns, this is evidence of uneven protein loading. When this happens it is important to be careful about how the data is used.

Normalization schemes not supported in Scaffold

There are more sophisticated normalization schemes that attempt to normalize the data in a way that allows the user to compare in a semi-quantitative way the abundance of one protein with another protein in the same sample. Scaffold does not support these schemes. This means that the user should exercise caution about trying to draw conclusions about the stoichiometry of the proteins from quantitative values as presented in Scaffold.

In particular, the user should be cautious about drawing conclusions about differential abundances for proteins where the spectral counts are small numbers. Scaffold tries to mitigate this problem by its treatment of [Missing values](#).

Missing values

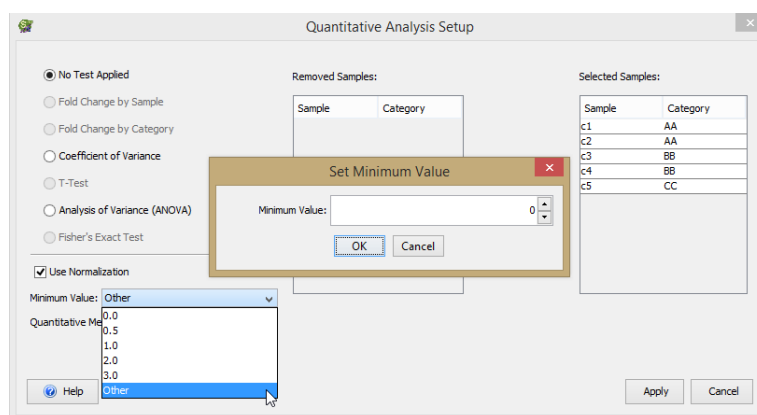
For differential protein expression tests, Scaffold replaces missing values with a specified [Minimum Value](#). Whenever a sample has no assigned spectra for a specific protein and that protein is found in a different sample, the specified minimum value is used to calculate the normalized values.

Minimum Value

The minimum value option allows the user to set a floor when calculating Label-Free quantitative values. Higher values will output shorter lists of highly confident changes; lower values will output longer lists that may contain less confident changes.

The minimum value is set by the user in the dialog through the Minimum Value pull down list and it defaults to 0. When selecting the option Other from the list the Set Minimum Value dialog opens allowing the user to record and use a minimum value different from the ones shown in the pull down.

Figure 13-2: Setting the minimum value



Whenever a sample has no assigned spectra for a specific protein and that protein is found in a different sample, the specified minimum value is used instead of zero for the sample with no assigned spectra. When Normalization is selected, the Missing Values are replaced by the set minimum value when quantitative values are calculated.

All quantitative values that are lower than the selected minimum value will also be replaced by the minimum value. This is true even if no statistical test is selected and the dialog controlling this value is grayed out.



- *Scaffold does not display intensities lower than the minimum value.*
- *The default minimum value is set at zero.*
- *Select “Other” in the “Minimum Value” drop-down list to specify a custom value.*
- *In the MS/MS Sample View parentheses indicate that the value shown in the cell was substituted with the minimum value.*
- *In the BioSample View parentheses indicate that the subsumed value shown in the cell was derived from a set of values that contained values substituted with the minimum value.*
- *For Fold change, if a zero appears in the denominator INF is shown in the Fold Change column.*

Quantitative Tests

Scaffold provides several tests to identify proteins which show different quantitative abundances in different categories. Which test to use depends upon the experimental design, particularly the number of replicates available.

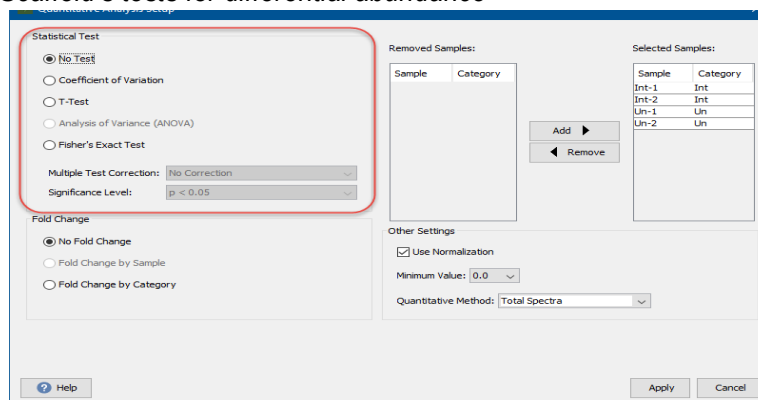
The tests are available for selection through the Quantitative Analysis dialog, accessed by selecting **Experiment>Quantitative Analysis...** or clicking the Quantitative Analysis icon in the toolbar.:



Through this dialog the user chooses the categories to be tested, the quantitative value to be used, whether the data should be normalized and the statistical tests to be performed.

Tests are based upon the data that is displayed in the [The Samples Table](#). Adjusting the peptide and protein filters and thresholds or the ReqMod filter or toggling the [Show Lower Scoring Matches](#) changes the number of proteins shown in the table and as a result the tests will analyze different sets of data and may identify different proteins as exhibiting statistically significant changes in abundance level.

Figure 13-3: Scaffold's tests for differential abundance



Up to seven tests are potentially applicable, depending on the number of categories and replicate samples included in the categories:

- [Fold Change by Sample](#)
- [Fold Change by Category](#)
- [Coefficient of Variation](#)
- [T-Test](#)
- [ANOVA](#)
- [Fisher's Exact Test](#)

When a quantitative test is selected and applied, two columns appear in [The Samples Table](#): one shows the test results and the other the [Quantitative profile](#). Sorting one of these columns brings the differentially expressed proteins together. These proteins can then be:

- Checked for the quality of the spectra supporting the identification in the [Proteins View](#).
- Checked for peptides shared between proteins in [Similarity View](#).
- Checked for differential expression using the [The Quantitative Value pane](#) in the Quantify View.

Fold Change by Sample

The simplest [Quantitative Tests](#) is the Fold Change, which reports by how much two variables differ. It is defined as the ratio of the quantitative value in one BioSample over the quantitative value in a second BioSample. The Fold Change by Sample can be used when only two BioSamples are selected in the quantitative Analysis setup dialog. Because the specified Minimum Value replaces any Missing Values, if a zero appears in the denominator an INF will appear in the Fold Change column.

Notes:

- Scaffold currently only shows the ratio, not the log base 2 of the ratio.
- Fold Change values need to be interpreted cautiously. A fold change of 2 is much more likely to be significant if the ratio is between 48 and 24 than if it is between 2 and 1. Scaffold's Q-Q scatter plot may help in this matter.
- If you sort data based on the fold change, it is important to check both the top and the bottom of the sorted table. A 4 to 1 ratio will display as 4, but a 1 to 4 ratio will display as 0.25.
- If the fold change is less than 0.5 or 2.0, Scaffold colors the box green to help highlight possible proteins of interest. This does not necessarily signify any important statistical difference.
- Scaffold versions older than 3.5 report the ratio as 1 if there are any missing values.

Fold Change by Category

Fold Change by Category is available for selection only when samples belonging to exactly two different categories are selected in the Quantitative Analysis Setup dialog.

It is defined as the ratio between the average among the quantitative values of each BioSample included in one category versus the average of the quantitative values in the other category.

Figure 13-4: Fold Change by category

Statistical Test

☒ No Test

☐ Coefficient of Variation

☐ T-Test

☐ Analysis of Variance (ANOVA)

☐ Fisher's Exact Test

Multiple Test Correction: No Correction

Significance Level: $p < 0.05$

Fold Change

☐ No Fold Change

☐ Fold Change by Sample

☒ Fold Change by Category

Reference Category: Int

Removed Samples:

Sample	Category
--------	----------

Add

Remove

Selected Samples:

Sample	Category
Int-1	Int
Int-2	Int
Un-1	Un
Un-2	Un

Other Settings

☒ Use Normalization

Minimum Value: 0.0

Quantitative Method: Total Spectra

Help Apply Cancel

The Reference Category pull down list allows the selection of which category is used as denominator when calculating the Fold Change.

Figure 13-5: Fold change by category in the Samples Table

Display Options: Quantitative Value (Normalized Total Spectra)																	Req Mods:		No Filter	Search:				
#	Visible?	Starred?	Protein Grouping Ambiguity	bility Le	wer 95	% to 94	% to 71	% to 4	% to 19	MS/MS	View:	Identif Protein (11)	Accession Number	Molecular Weight	Fold Change by Category	Int				Un				
																Int-1		Int-2		Un-1		Un-2		
																	Denominator				Numerator			
																CH191 (F005229)	CH191 (F005230)	CH191 (F005231)	CH191 (F005232)	CH191 (F005233)	CH191 (F005234)	CH191 (F005235)	CH191 (F005236)	
1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	★	Kera...	gi 54...	60 kDa	0.8	4	6	5	7	6	11	(0)	1									
2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	★	Kera...	gi 13...	66 kDa	0.7	6	4	4	6	4	4	(0)	7									
3	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		Chai...	gi 33...	23 kDa	1.3	3	5	4	2	7	7	0	5									
4	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		Kera...	gi 54...	66 kDa	0.10	4	5	1	3	0	(0)	(0)	1									
5	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		Kera...	gi 81...	62 kDa	0.3	3	(0)	1	2	(0)	(0)	(0)	1									
6	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		RecN...	gi 11...	42 kDa	3.5	1	0	1	(0)	(0)	(0)	(0)	5									
7	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		PRED...	gi 50...	622 kDa	3.2	0	(0)	1	1	5	(0)	(0)	(0)									
8	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		Chai...	gi 28...	76 kDa	0.0	(0)	(0)	2	(0)	(0)	(0)	(0)	(0)									
9	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		Chai...	gi 56...	18 kDa	0.0	(0)	(0)	2	(0)	(0)	(0)	(0)	(0)									
10	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		Lyso...	gi 24...	17 kDa	0.0	(0)	(0)	1	(0)	(0)	(0)	(0)	(0)									
11	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		myo...	gi 50...	149 kDa	0.0	1	(0)	(0)	(0)	(0)	(0)	(0)	(0)									

The column headers of the samples included in the category considered as the numerator are highlighted in purple, the headers of the category considered as the denominator are highlighted in orange.

Coefficient of Variation

A coefficient of variation (CV) can be calculated and interpreted in two different settings: analyzing a single variable or interpreting a model. The standard formulation of the CV, the ratio of the standard deviation to the mean, applies in the single variable setting. In the modeling setting, the CV is calculated as the ratio of the root mean squared error (RMSE) to the mean of the dependent variable. In both settings, the CV is often presented as the given ratio multiplied by 100.

The CV for a single variable aims to describe the dispersion of the variable in a way that does not depend on the variable's measurement unit. The higher the CV, the greater the dispersion in the variable.

The CV for a model aims to describe the model fit in terms of the relative sizes of the squared residuals and outcome values. The lower the CV, the smaller the residuals relative

to the predicted value. This is suggestive of a good model fit.

$$C_v = \frac{\sigma}{|\mu|}$$

In Scaffold the user can select the Coefficient of Variation test only when he/she adds two or more BioSamples to the Select Samples Table in the Quantitative Analysis set up dialog.

- The CV is only defined for a nonzero mean. Because the CV is expressed as a percentage, Scaffold multiplies the ratio of the standard deviation to the mean by 100
- The CV is typically used to describe the dispersion of the variable independently of its measurement unit. The higher the CV, the greater the dispersion in the variable. For example, when analyzing four samples - A, B, C, and D - the coefficient of variation outputs how dispersed are the values in respect to their mean. A small coefficient of variation means that the four samples have values close together compared to their average value. If the coefficient of variation is big, then at least one of the four samples is different, but it doesn't specify A, B, C or D. Examining Scaffold's Quantitative Value Bar Graph helps determine which it is.
- Coefficient of Variation is typically used in place of an ANOVA test when not enough replicates are available to give sufficient statistical power to apply ANOVA.

T-Test

The T-test is a measure of the distance between the mean of the replicate samples in one category from the mean of the replicate samples in another category. This distance is scaled by the standard deviation of the replicates. Scaffold performs a two-tailed T-test. The result of a T-test is reported as the probability (p-value) that this distance between means could occur by chance.

To be able to apply the T-test in Scaffold, the BioSamples in the experiment need to be organized at least in two different categories, see [Organize Samples Into Categories](#). Among the various samples in the experiment only samples belonging to two different categories need to be included in the Selected Sample table in the Quantitative analysis set up dialog to have access to the T-test option.

Each of the two categories should include three or more replicate BioSamples. Examples of typical categories are “treated/untreated”, “disease/control” or “cell line1/cell line2”. Since the test is computed using quantitative values most of the time normalized, the user should keep in mind potential issues surrounding [Missing values](#) and Normalization as described in [Normalization among BioSamples in Scaffold](#).

A small p-value means that the BioSamples in one category are most likely different from those in the other category. A threshold or alpha level or significance level of 0.05 is commonly used to assess how statistically significant the result of the T-test is. This value should be appropriately adjusted in Proteomics experiments since differences are evaluated among many proteins at once, see [Multiple test significance levels and corrections](#).

The T-test is generally considered a fairly robust test. This means that even if its basic assumptions are violated somewhat, it still tends to be fairly reliable at separating the

categories which are the same from those that differ. Some researchers believe that spectral data should be transformed in some way, for example by taking its log, before doing a T-test. Other researchers may think that 3 replicates is not enough to apply the T-test. Still others believe that more advanced non-parametric tests would work better. So if the T-test gives a borderline result, the user may want to check it carefully. But if the T-test has a very small p-value, the robustness of the T-test means that it is unlikely that a more sophisticated statistical analysis will give a different result.

If the user computes the T-test with less than 3 replicates, it is unlikely to give informative and trustworthy results. The [Fisher's Exact Test](#) may be more appropriate for samples with few replicates and low abundance proteins with few spectral counts.

Scaffold does not make a distinction between technical replicates and biological replicates. This is a more advanced statistical analysis concept and is not supported by Scaffold.

ANOVA

The ANOVA (Analysis of Variance) is an analysis method for testing equality of means across treatment groups or categories. It tells if there are differences among categories. The result of the test is a p-value which when low indicates a large probability for variation among the different categories considered for the test.

The ANOVA test in Scaffold requires three or more replicates in the categories. Like the T-test, having fewer than 3 replicates produces untrustworthy results and more replicates are better.

Like the Coefficient of Variation test, the ANOVA test shows that something is different, but it does not tell which categories are different from each other. Checking the Bar Chart in [The Quantitative Value pane](#) helps in understanding which category is different.

The ANOVA test in Scaffold is a simple ANOVA. More sophisticated ANOVA tests are beyond Scaffold's capability.

Before applying the ANOVA test the user should understand the issues regarding [Protein Grouping Ambiguity](#), [Missing values](#), and Normalization as described in [Normalization among BioSamples in Scaffold](#).

Fisher's Exact Test

The **Fisher's Exact Test**, like the T-test, compares the relative abundance between two sample categories. It is used in the analysis of contingency tables where sample sizes are small. It is called exact since it calculates the significance of the deviation from the null hypothesis with an exact method not using an approximation dependent on the size of the sample statistics. This means that the Fisher's exact test is more appropriate than the T-test if there are fewer replicates. Like the T-test, the Fisher's Exact Test produces a p-value.

Scaffold calculates the Fisher's exact test p-value according to a model discussed in [Zhang \(2010\)](#). The paper performs a systematic analysis of the various approaches to quantify differential expressions among different experiments. Particularly, it describes how and when it is reasonable to apply a Fisher's Exact test in a pair wise experiment.

As described in the paper, to calculate the Fisher exact test for a target protein, Scaffold

arranges the spectral counts for a pair of categories into a two-way contingency table where the first row contains the counts for the target protein in each category and the second row contains the rest of the counts for the rest of proteins listed in the Samples table. The test is based on the assumption that the row and column totals are fixed, which means that any entry in the table completely determines the others. The probability assigned to a particular arrangement of spectral counts in the table is calculated using a hyper-geometric distribution.

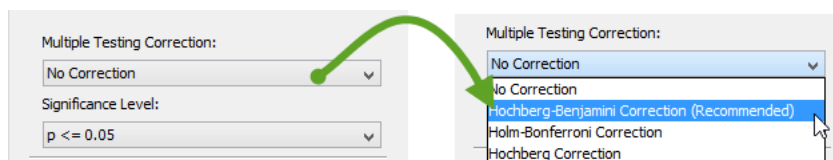
The p-value assigned by the Fisher's Exact test to the target protein is the sum of all p-values over all the possible configurations of the two-way table that have p-values less or equal to the initial target protein two way table (which means integrating over the tail of the distribution).

For proper calculation of the test, it is important to turn off normalization when using the Fisher's exact test. This is due to the fact that the test relates a value to the sum of its column and compares it with the corresponding value in another column and its related sum. When Normalization is turned on, see [Normalization among BioSamples in Scaffold](#), the sums are forced into equality and the test results would be skewed. For convenience when the Fisher's test is selected, Scaffold grays out the option for selecting normalization.

Multiple test significance levels and corrections

In the dialog under the list of available statistical tests, there are two pull down menus: the Multiples testing correction and the Significance Level that allow the user to set a significance level for the selected inference test and choose methods to control the familywise error rate.

Figure 13-6: Significance Level tab



When considering a set of statistical inferences simultaneously and doing multiple comparisons the risk of making one or more false discoveries or a Type I error grows quite quickly. In these cases it is common to adjust p-values for the number of hypothesis tests performed. There are many different methods that provide a way to perform this adjustment. A common one is to control the familywise error rate, which is defined as the probability of making Type I errors.

One of the initial and still quite common methods used to control this error is provided by the Bonferroni correction where the significance level α for an individual test is found by dividing the familywise error rate (usually 0.05) by the number of performed tests. Thus when doing 100 statistical tests, the α level for an individual test would be $0.05/100=0.0005$, and only individual tests with $P<0.0005$ would be considered significant.

The Bonferroni approach is a fairly conservative one and for a very large number of independent comparisons it may lead to a high rate of false negatives.

To address this issue Scaffold provides three different types of corrections:

- [Benjamini-Hochberg correction \(recommended\)](#)
- Holm-Bonferroni correction - a step down method see [Control FWER with Hochberg's step-up and Holm's step-down](#)
- Hochberg correction - a step-up method [Control FWER with Hochberg's step-up and Holm's step-down](#)

Control FWER with Hochberg's step-up and Holm's step-down

There are various methods described in the literature that control the Familywise error rate (FWER) using less conservative corrections than the Bonferroni one but are still based on the Bonferroni inequality. These methods are usually quite appropriate to control the FWER in control trial experiments in which a limited number of comparisons are of interest and where the use of the False Discovery Rate is inappropriate. In these cases the corrections guard against any false positive occurring.

For more information on how the two methods are developed in Scaffold please see the [Holm's and Hochberg's Techniques to Control the Familywise Error Rate](#).

Benjamini-Hochberg correction (recommended)

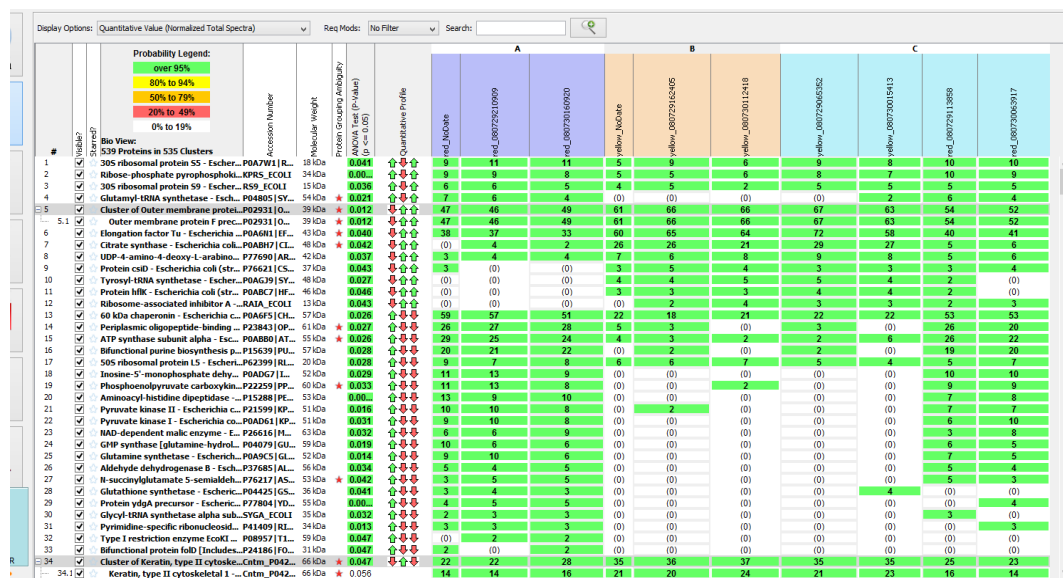
This method of controlling the error rate in multiple experiments is particularly useful in high-dimensional type of experiments where a more common goal is to identify as many true positive findings as possible, while incurring a relatively low number of false positives. The false discovery rate (FDR) is designed to quantify this type of trade-off, making it particularly useful for performing many hypothesis tests on high-dimensional data sets.

Scaffold computes the FDR using the Benjamini-Hochberg procedure as developed in the original paper, see [Benjamini \(1995\)](#).

Quantitative profile

When a statistical test is applied Scaffold creates a profile associated with statistically significant category-specific averages and shows graphically the up and down regulations in respect to this average under the Quantitative Profile column that appears in the [The Samples Table](#).

Figure 13-7: Quantitative Profile



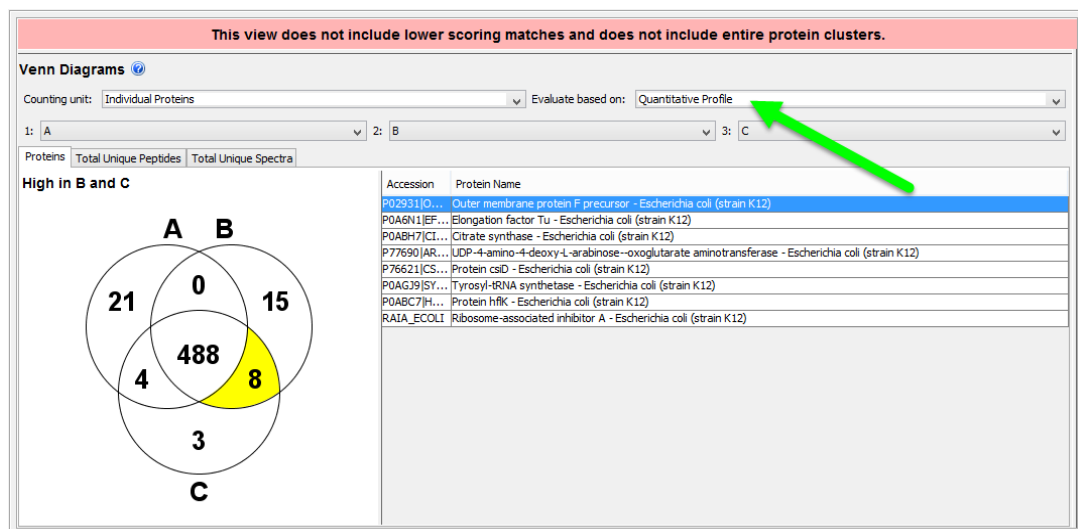
To compute a quantitative profile Scaffold considers the categories selected for the statistical test and calculates, within each category, the average of the values given by the quantitative method selected in the Quantitative analysis setup dialog. Then Scaffold calculates the mean of these averages, or averages of averages, and compares to it every category-specific average.

The averages that are greater or equal to the mean are marked high and graphically represented in the Samples Table by a green arrow pointing upward. The averages that are smaller than the mean are marked low and graphically represented by a red arrow pointing downward.

Quantitative Profile Venn Diagram

In [The Venn Diagrams pane](#) it is possible to view a Venn diagram that uses the information shown in the Quantitative Profile column by selecting the appropriate option from the **Evaluate based on:** pull down list. This option is available only when a Quantitative Test is active.

Figure 13-8: Quantitative Profile in the Venn Diagram



In the diagram, see [Figure 13-8](#), the central region reports the number of proteins that show statistically insignificant changes across categories, and those proteins that have identical cross-category averages. All the other sections report the number of proteins that are marked as high. Labels will appear on the chart when one of the section is highlighted to help clarify confusion.

The categories available for selection from the pull down lists above the diagram include only those categories that are selected to perform the statistical test or selected through the advanced filter and are those used to compute the category averages.

Chapter 14

Precursor Intensity Quantitation

Introduction

The Scaffold Suite of programs supports a variety of methods for performing quantitative mass spectrometry. Scaffold offers full support for spectral counting techniques, including emPAI and NSAF, and also calculates Total Ion Current (TIC), the sum of the intensities of the peaks in the MS/MS spectra. Scaffold Q+ supports isobaric labeling techniques, such as iTRAQ and TMT, and Scaffold Q+S adds support for SILAC and dimethyl labeling.

An increasingly popular option for quantitative Proteomics, however, is Precursor Intensity Quantitation, which offers a good compromise between the accuracy of labeled techniques and the simplicity and lower cost of label-free quantitation. This method relies on measuring the signal intensity of the peptide precursors representing a specific protein at the MS level and comparing these intensities across samples. Scaffold and Scaffold Q+/Q+S each support this method in different ways.

Scaffold is designed to provide easy and confident validation, visualization and quantitation of search results. It does not read raw files and does not have direct access to precursor information; instead it reads intensity data already computed by the identification software. Currently, Scaffold is able to obtain precursor intensity information from Thermo Proteome Discoverer, Mascot Distiller, Agilent Spectrum Mill, and MaxQuant files. Scaffold normalizes precursor intensity values across samples and calculates fold changes at the BioSample or Category level. Statistical tests of differences in the calculated intensities are also offered, including the T-Test, ANOVA and Coefficient of Variation as appropriate to the experimental design.

Because of its dependence on search results, Scaffold's approach to Precursor Intensity Quantitation is to work backward from the peptides that have been identified through their MS/MS spectra and compare the intensities of the MS peaks from which they were derived. By contrast, some programs align the MS peaks of all samples and calculate their intensities. The peaks that appear to be biologically important are subsequently identified. This method has the advantage of providing quantitative information for low abundance proteins, but it relies on complicated peak alignment algorithms and may in some cases incorrectly identify the corresponding peaks.

Working from identifications is simpler, since it obviates the need for retention time warping and peak alignment, and it has the advantage of depending on more reliable data. On the other hand, in this approach missing values become an issue. Often a peptide is identified in one sample and not in another, producing a missing value even if there may have been a detectable MS peak in the corresponding position in the second sample. Generally, however, higher abundance peptides are more likely to be identified and MS peaks that do not result in identified spectra are relatively weak signals, minimizing the effect of treating them as

missing values. Scaffold further reduces the effect by choosing algorithms that are less sensitive to missing values: for example it uses the geometric mean rather than the average in calculating protein level fold-changes.

Calculation of Precursor Intensities

Precursor Intensity Quantitation is based on the principle that the area of the peak in the MS1 chromatogram provides a measure of the relative abundance of the corresponding peptide in the sample. Peptides are identified based on their MS/MS spectra, and then the corresponding MS1 peaks are identified in each LC-MS/MS run. The areas under these peaks are calculated and normalized and their ratios are used as a measure of the relative abundance of the peptides in different samples. Relative quantities of proteins are estimated by combining the precursor intensities of the constituent peptides in various ways.

The following illustration of the typical LC-MS/MS analysis of a peptide is reproduced from [Lai \(2013\)](#).

Figure 14-1: Identification of a peptide through LC-MS/MS analysis

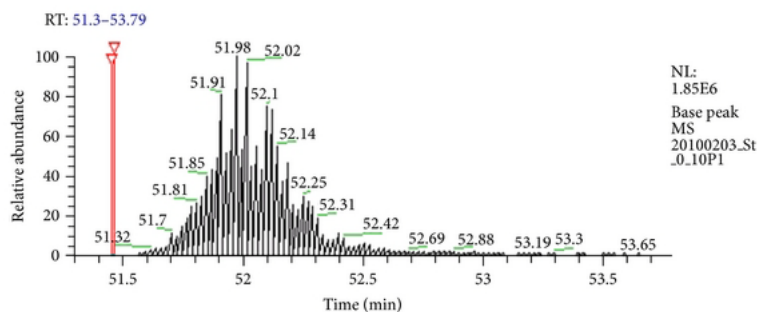


Figure 14-1-a: The peptide is eluted from the LC column and its ion intensity is plotted as a function of the retention time.

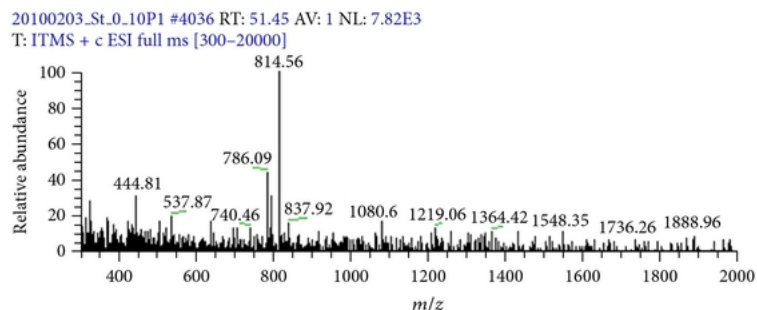


Figure 14-1-b: At the first scan time shown in red in (a), a full MS scan is performed. The ion with m/z 786.09 is selected as a precursor ion for MS/MS analysis.

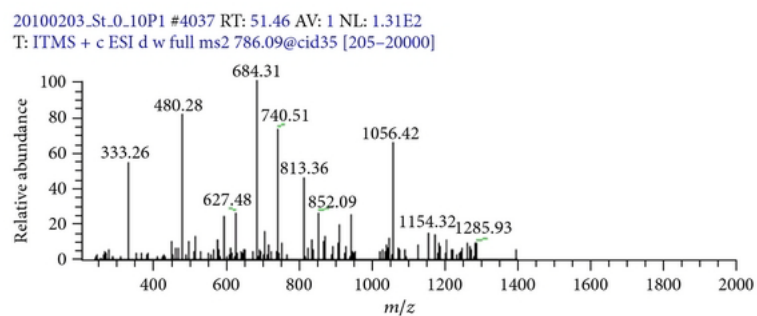


Figure 14-1-c: At the next scan (also shown in red in (a)), an MS/MS scan is performed, providing peptide fragmentation information for peptide identification.

Once a peptide has been identified, a program can go back to the MS1 scans and find a series of spectra which contain peaks corresponding to the same peptide as it continues to elute from the column. These spectra are then aligned and the intensities of the peaks for the specific m/z value which represents the parent ion of this peptide are plotted against the retention time, giving an extracted ion chromatogram.

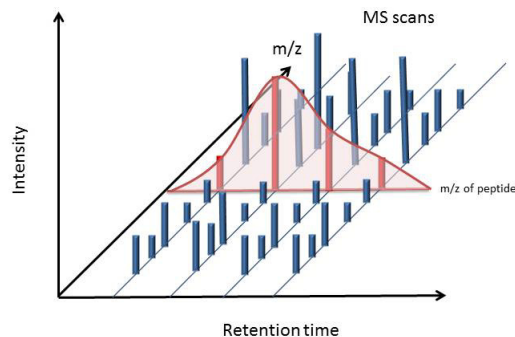


Figure 14-2: The intensities of the MS peaks at the same m/z value are plotted as a function of the retention time. The area under this curve (enclosed in red) is the precursor ion intensity.

In the extracted ion chromatogram, a curve is fit to the intensities at a specific m/z . The area under this curve represents the total amount of the specified peptide that eluted. Scaffold reads these values from its input files and uses them to do quantitative analysis.

Preparing Data for Precursor Intensity Quantitation in Scaffold

Scaffold reads precursor intensity information from various identification programs provided that the user has requested this type of quantitation during the search. Following are instructions for preparing input files for quantitation in Scaffold:

Proteome Discoverer

Proteome Discoverer (PD) provides a workflow template for computing precursor intensity values. The template **WF_LTQ_Orbitrap_Sequest_Precursor_ions_Area_Detector** can be used as a starting point, and the search engine choice or instrument settings may be changed. Scaffold reads the precursor intensities from the MSF file.

Similar instructions work for the PD version 2.0 and higher. More information on how to load newer PD search results in Scaffold is provided in the document: <https://support.proteomesoftware.com/hc/en-us/articles/360004385172-Configuring-Proteome-Discoverer-2-x>.

Mascot Distiller

When setting up the Mascot search, select **Average[MD]** as the quantitation method. When the search is complete, in Distiller select **Analysis>Calculate XIC**, and then **Analysis>Quantitate**. Export the results as an XML file using **Analysis>Quantitative Report>Save as XML**. Also create an ROV file by saving the project with **File>Save Project As....** Place the ROV file and the XML file in the same directory, and if the DAT file

is not accessible directly from the Mascot Server, also place that file in the same location.
Load only the XML file into Scaffold.

Spectrum Mill

No special settings are required. Load the entire Spectrum Mill results directory into Scaffold.

MaxQuant

MaxQuant 1.3 will only compute precursor intensity when two or more raw files are processed together. Each of the samples to be compared must be labeled with a different experiment name in the experiment.txt file.

Generally, all MaxQuant results in a single directory load into Scaffold as a single sample. For precursor quantitation, however, the samples to be compared must be loaded into different BioSamples. Accordingly, Scaffold has a special dialog that opens when the program recognizes the presence of an experiment file. To place each experiment into its own BioSample, from the loading wizard select the MaxQuant output directory and click **Add to Import Queue** then when the dialog appears, select the first experiment. Click Next, then **Add another BioSample** and select the same directory, but choose a different experiment from the dialog box.

In **MaxQuant 1.4**, precursor intensity may be computed even when analyzing a single raw file if the user selects the Label Free Quantitation option. Individual results may then be loaded into separate BioSamples in the usual way and used for Precursor Intensity Quantitation in either Scaffold or Scaffold Q+.

If two or more raw files are analyzed together in MaxQuant 1.4 with the Label Free Quantitation option selected, and no **Experiment.txt** file is provided, they form a single combined folder which loads into Scaffold as a single sample. In this case, Scaffold and Scaffold Q+ are unable to perform Precursor Intensity Quantitation. It is possible, although not required, in MaxQuant 1.4 to create an experiment file. The experiments can be named through the MaxQuant 1.4 GUI, and then an experiment file can be exported by right-clicking and choosing **Export**. The user should name the file **Experiment.txt** and then Scaffold will recognize it and loading can proceed as for MaxQuant 1.3 results.

Performing Quantitation in Scaffold

Scaffold reads the precursor intensity values already computed by the search engine software from the input files. For each peptide-spectrum match, it reports the intensity value in the Peptides Table in the upper right of the Proteins View ([Figure 14-3](#)).

Charge	Delta ...	Delta ...	Reten...	Intensity	TIC	Start	Stop	# Ot...	Other Pr
2	-0.011	-6.5	2720	1.88E7	728700	2	15	0	
2	-0.011	-6.5	2780	1.88E7	190600	2	15	0	
3	-0.0086	-5.2	2780	8.33E7	191400	2	15	0	
3	-0.0091	-5.5	2720	8.33E7	559200	2	15	0	
3	-0.0096	-5.8	2880	403000	63620	2	15	0	
3	-0.0095	-5.9	1700	20800	301700	2	15	0	
2	-0.018	-7.9	5230	30000	547400	16	33	0	
2	-0.013	-6.0	3990	195000	2466000	16	33	0	
4	0.0063	2.8	3990		28220	16	33	0	
3	-0.013	-6.0	5250	5920000	256800	16	33	0	
4	0.0048	2.1	5250		24070	16	33	0	
3	-0.015	-6.6	5190	5920000	18650	16	33	0	
3	-0.015	-6.9	3990	4260000	522900	16	33	0	
4	-0.013	-5.7	5190	5530000	10210	16	33	0	
3	-0.012	-6.0	5600	2.33E7	519400	18	33	0	
3	-0.013	-6.4	5540	2.33E7	93110	18	33	0	

Figure 14-3: Precursor Intensities in the Peptides Table

Scaffold provides three methods of using these values to perform relative quantitation at the protein level. These methods are available through **Experiment > Quantitative Analysis** or by clicking on the bar-graph icon at the top of the screen. Either of these methods brings up the **Quantitative Analysis Setup** dialog (Figure 6). Because it is a relative quantitative method, when using label-free quantitation, it is necessary to select at least two samples. It is also important to adjust the **Minimum Value** setting to a value that is appropriate for intensities. Values other than zero require the use of the **Other** option in the dropdown. A checkbox allows the user to choose whether or not to normalize between samples.

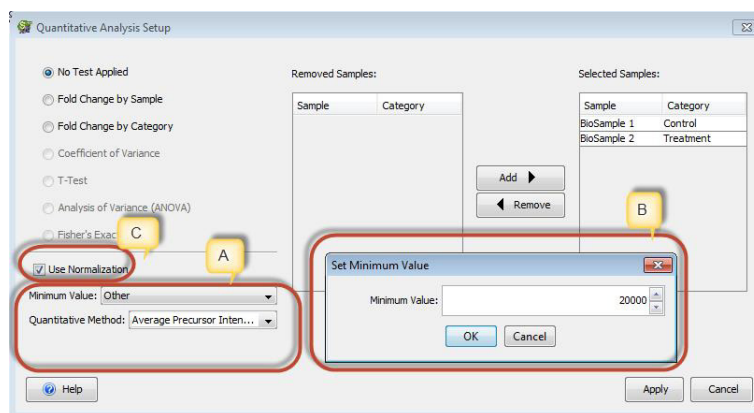


Figure 14-4: The Quantitative Analysis Setup Dialog – A. Selecting the Quantitative Value, B. Specifying the Minimum Value, C. Selecting Normalization option

There are a number of options for combining peptide precursor intensities to provide an estimate of relative quantities at the protein level. Scaffold provides three methods: **Average Precursor Intensity**, **Total Precursor Intensity** and **Top 3 Precursor Intensity**. These are calculated from the **Intensity** values shown in the Peptides Table as follows:

Valid	...	Sequence	Modifications	Charge	Intensity	Observed
<input checked="" type="checkbox"/>	1.0	(R)NYDSMKDFFEMRK(A)	Oxidation (+16), Oxidation (+16)	4	2.61E7	131
<input checked="" type="checkbox"/>	1.0	(R)NYDSMKDFFEMRK(A)	Oxidation (+16), Oxidation (+16)	3	2.61E7	171
<input checked="" type="checkbox"/>	1.0	(R)NYDSMKDFFEMRK(K)	Oxidation (+16), Oxidation (+16)	3	2650000	331
<input checked="" type="checkbox"/>	1.0	(M)SSGALLPKPQMR(G)	Oxidation (+16)	3	5610000	431
<input checked="" type="checkbox"/>	1.0	(M)SSGALLPKPQMR(G)	Oxidation (+16)	2	6100000	651
<input checked="" type="checkbox"/>	1.0	(K)AGIFQSAK(-)		2	5.13E7	411
<input checked="" type="checkbox"/>	1.0	(R)NYDSMKDFFEMRK(A)	Oxidation (+16)	2	1390000	851
<input checked="" type="checkbox"/>	1.0	(R)NYDSMKDFFEMRK(A)	Oxidation (+16)	3	6060000	571
<input checked="" type="checkbox"/>	1.0	(K)KAYAEFYR(N)		2	2430000	521
<input checked="" type="checkbox"/>	1.0	(R)NYDSMKDFFEMRK(A)	Deamidated (+1), Oxidation (+16)	2		851
<input checked="" type="checkbox"/>	1.0	(R)NYDSMKDFFEMRK(A)	Oxidation (+16)	3		571
<input checked="" type="checkbox"/>	1.0	(R)NYDSMKDFFEMRK(K)	Oxidation (+16)	3	980000	521
<input checked="" type="checkbox"/>	1.0	(R)NYDSMKDFFEMRK(K)	Oxidation (+16)	2	2000000	791
<input checked="" type="checkbox"/>	1.0	(M)SSGALLPKPQMR(G)	Acetyl (+42), Oxidation (+16)	2	6.65E7	671
<input checked="" type="checkbox"/>	1.0	(M)SSGALLPKPQMR(G)	Acetyl (+42), Oxidation (+16)	2	6.65E7	671
<input checked="" type="checkbox"/>	1.0	(R)NYDSMKDFFEMRK(A)		2	846000	841
<input checked="" type="checkbox"/>	1.0	(M)SSGALLPKPQMR(G)	Acetyl (+42)	2	3.81E7	661
<input checked="" type="checkbox"/>	1.0	(M)SSGALLPKPQMR(G)	Acetyl (+42)	2	3.81E7	661

Figure 14-5: Intensity values for a singleBioSample for a specific protein - A. Note that often multiple MS2 spectra are collected from a single MS1 spectrum. This results in duplicate reports of the same Intensity value. Scaffold counts each value only once. **B.** Intensities for different charge states of the same peptide are summed to give the total intensity for that peptide.

First, peptide intensity values are calculated. As shown in Figure 14-5, duplicate intensity values for the same peptide are discarded. If there are multiple peptide-spectrum matches with the same peptide sequence and modifications but with different intensity values, their intensities are summed and the sum is used as the intensity value for that peptide. The peptide intensity values are then used in the following calculations:

- **Average Precursor Intensity:** The geometric mean of the peptide intensity values for a given protein.
- **Total Precursor Intensity:** The sum of all distinct intensity values for a protein.
- **Top 3 Peptides Precursor Intensity:** The sum of the three highest peptide intensity values for a protein. If fewer than three peptides have intensity values, the intensities that are present are summed.

When one of these methods is selected through the **Quantitative Method** drop-down, it becomes available for display in the Samples View. Choosing **Quantitative Value** from the **Display Options** drop-down causes the Samples View to show precursor intensity values calculated according to the selected method in the Samples Table. The name of the method is displayed in Display Options and if the values have been normalized, that is also indicated (Figure 14-6).

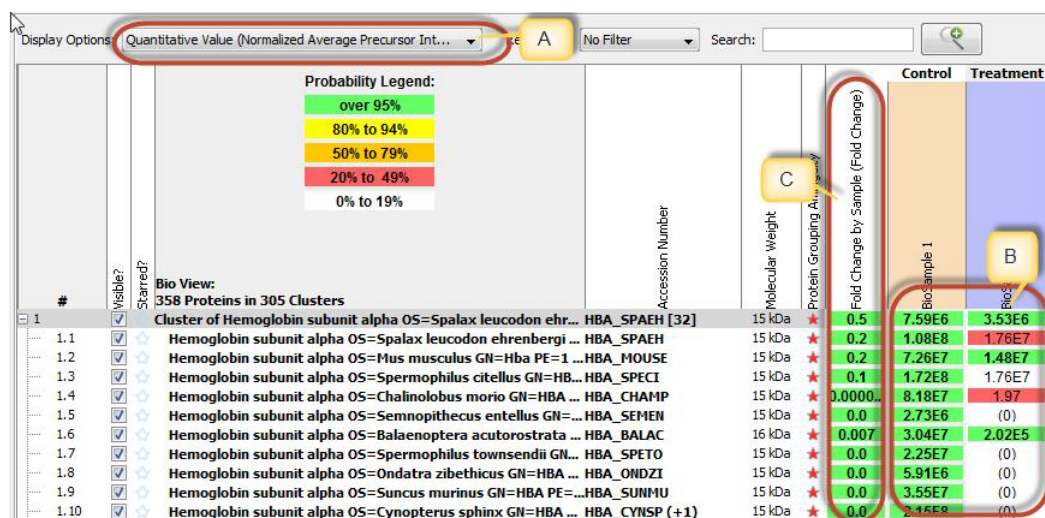


Figure 14-6: The Samples View – A. Quantitative Method selected B. Protein-level intensity values C. Fold Change calculated from the precursor intensity values.

Using the Quantitative Values based on precursor intensity, Scaffold can also calculate fold change at either the BioSample or the Category level. The desired fold change option is specified in the Quantitative Analysis Setup, which also allows selection of which BioSample or Category should serve as the reference (Figure 14-7).

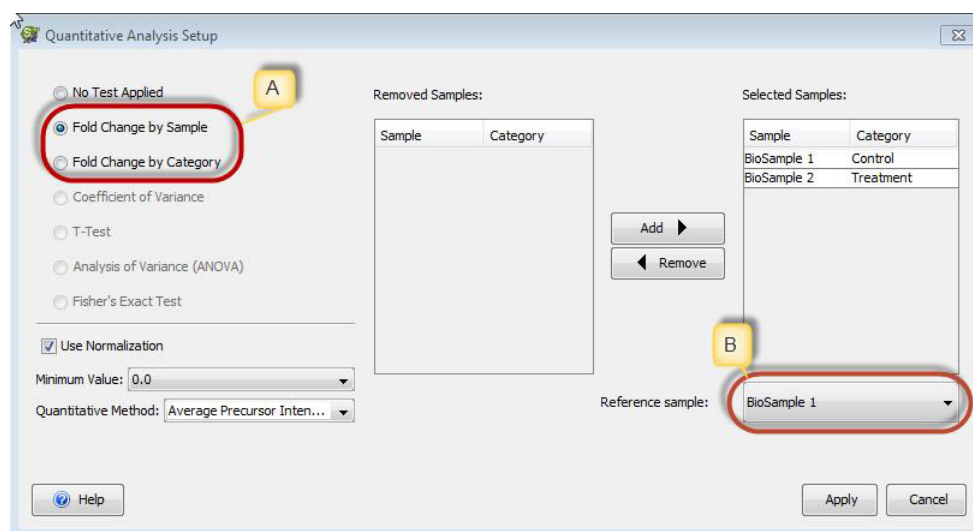


Figure 14-7: Requesting display of Fold Change – A. Choice of Fold Change by Sample or by Category, B. Specification of Reference Sample or Category.

When a Fold Change option is selected, an additional column is displayed in the Samples View. Fold Change is based on the Quantitative Method selected in the Quantitative Analysis Setup dialog even if a different display type (such as Total Spectrum Count) is displayed in the Samples View.

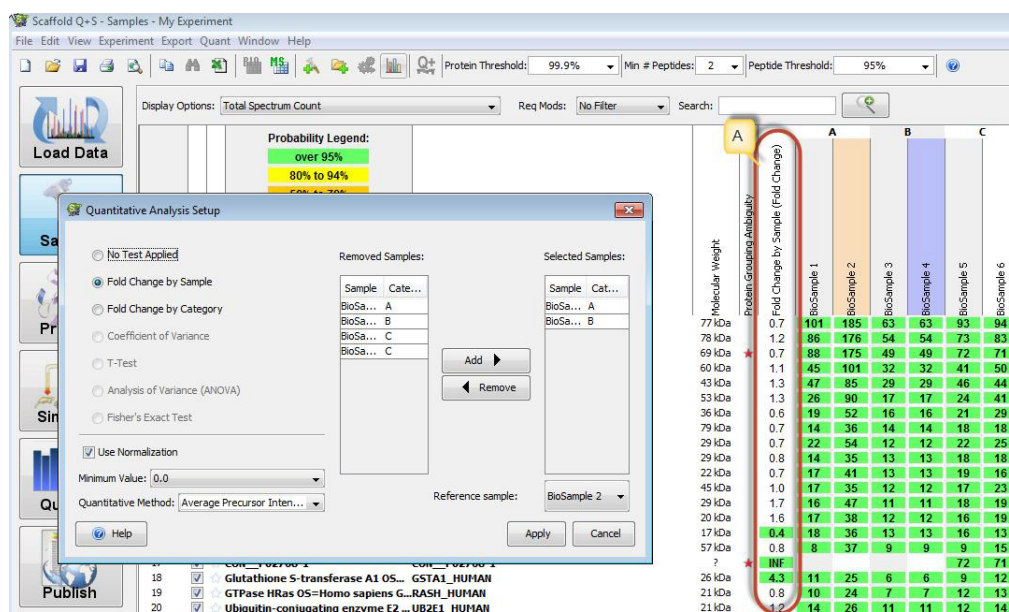


Figure 14-8: Fold Change by BioSample – A. The Fold Change column, showing the ratio of the Average Precursor Intensity of BioSample 4 to the Average Precursor Intensity of BioSample 2.

Fold Change by Sample is only available if exactly two BioSamples are selected for quantitation. It displays the ratio of the quantitative value of the non-reference BioSample to the quantitative value of the reference BioSample for each protein. The reference Sample is indicated by peach coloring in the column header, and the sample being compared is indicated by a purple header.

If samples from exactly two categories are selected, Fold Change by Category is available for display.

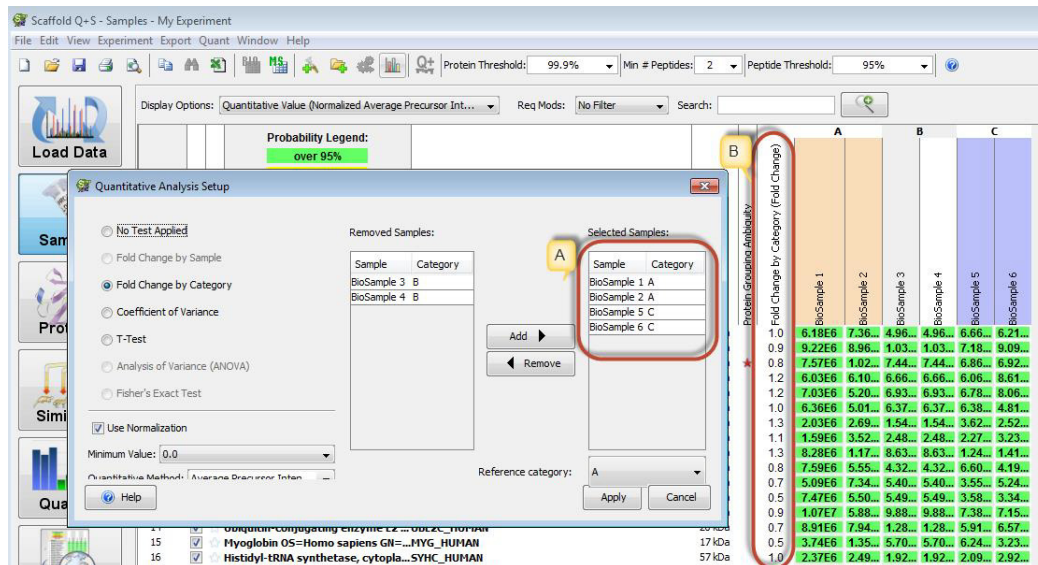


Figure 14-9: Fold Change by Category – A. Selected samples must belong to exactly two categories. B. The Fold Change values are displayed in a column in the Samples View.

The fold change values represent the ratio of the average of the quantitative values of the selected samples in the comparison category to the average of the quantitative values of the samples in the reference category.

Chapter 15

Reports

All of the reports are listed in the Scaffold main menu option **Export**. When a report is selected, a file browser opens to allow the user to save the resulting file. A default name in the format <Report Name><Scaffold File name> is provided for the report, but its value can be modified. The user cannot change the format of a report.

When saving a Scaffold ProtXML report, the user must provide a name for the report.

Excel-style reports are saved as CSV files, but are given the .XLS file extension so that they will automatically open in Excel.

Peptide and protein probabilities are omitted from all reports when data has been loaded in Prefiltered Mode unless probabilities were calculated from PEP values.

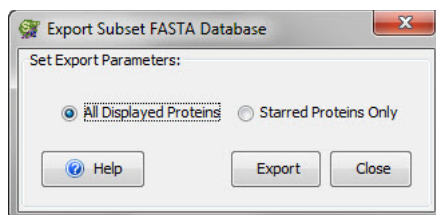
The following reports are available in Scaffold:

- [Subset Database](#) 246
- [Spectra](#) 246
- [BLIB export](#) 248
- [ProtXML report](#) 248
- [mzIdentML](#) 248
- [SFDB...](#) 252
- [Scaffold Batch...](#) 253
- [Scaffold Batch Archive...](#) 253
- [Exports compatible with Excel](#) 254

Subset Database

The command **Export > Subset Database** exports a FASTA database subset of the original sequence database used for searching the imported data.

Figure 15-1: Subset Database dialog



When selected the command opens a dialog containing the following export parameters:

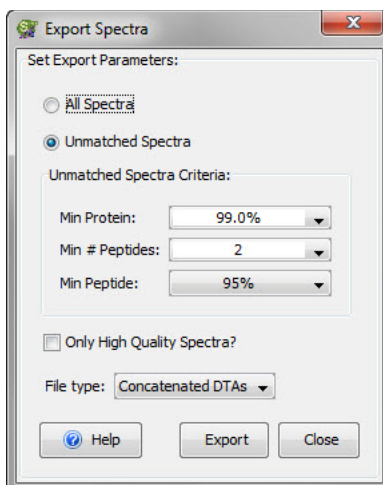
- **All Displayed Proteins** - when chosen the created subset database contains only proteins appearing in the proteins list. Adjusting the protein and peptide thresholds determines which identified proteins are included in the list and subsequently in the subset database. (More restrictive parameters result in fewer proteins in the exported database.)
- **Starred Proteins Only** - When selected the subset database will include only proteins labeled using stars.

The exported subset can facilitate a more thorough search for protein modifications with the original search engine.

Spectra

The menu command **Export > Spectra** exports spectra loaded in the current Scaffold experiment as peak lists. A list of different formats is available for the user to choose how to save the exported peak list.

Figure 15-2: Export Spectra



When selected the command opens a dialog containing the following parameters options:

- **All Spectra** - This option exports all the spectra loaded in the current Scaffold experiment.
- **Unmatched spectra** - This option exports only spectra that do not meet the filters criteria set by the user to allow further targeted searches on these types of spectra. The criteria are based on the probabilities assigned by Scaffold through its scoring algorithms.

Unmatched Spectra Criteria:

- Min Protein
- Min # Peptides
- Min Peptide
- **Only High Quality Spectra?** - When selected Scaffold chooses for export only those spectra that identify peptides with probabilities higher than 50% or if the peptide probability happens to be lower, the spectra has to be assigned to proteins that have a probability of at least 95%.
- **Types of peak list files:**
 - Concatenated DTAs
 - Individual DTAs
 - Mascot MFGs
 - Micromass PKLs
 - SEQUEST MS2s

BLIB export

This option exports Scaffold results in the BiblioSpec format¹ which is commonly used by Skyline. BLIB files may be used as reference libraries in **Scaffold DIA**, or transition lists for subsequent analysis in Skyline. The BLIB export respects threshold and filter settings.

ProtXML report

Exports all quantitative data in the protXML format, which is an open XML file format for the storage of data at the raw spectral data, peptide, and protein levels. This format enables uniform analysis and exchange of MS/MS data generated from a variety of different instruments, and assigned peptides using a variety of different database search programs.



See Molecular Systems Biology, 1:2005.0017 for more information.

mzIdentML

Scaffold fully supports the mzIdentML standard format for Proteomics data developed by the HUPO Proteomics Standards Initiative (Proteomics Informatics Standards group). A description of the standard specifications is available at www.psidev.info/mzidentml and a Java desktop software for validating mzIdentML can be downloaded at code.google.com/p/psi-pi/downloads/list.



Scaffold supports both mzIdentML 1.0.0 and the latest version 1.1.0.

Exports are compatible with:

- PRIDE. Now PRIDE supports uploading mzIdentML files created in Scaffold.
- SKYLINE for building spectral libraries. See [Creating a spectral library in Skyline](#)

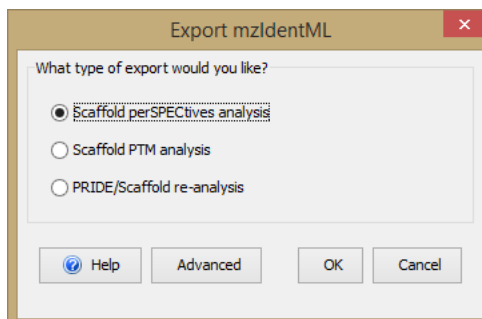
Selecting the menu command **Export > mzIdentML** opens the **Export mzIdentML** dialog where the user can easily customize his/her mzIdentML exports.

The **Export mzIdentML** dialog shows three basic options for creating mxIdentML exports optimized for the following uses:

- **Scaffold PTM analysis**
- **Scaffold DIA analysis**
- **PRIDE/Scaffold re-analysis** -- This export is suggested both for loading data in PRIDE or reloading data in Scaffold using mzIdentML instead of the regular search engines files.

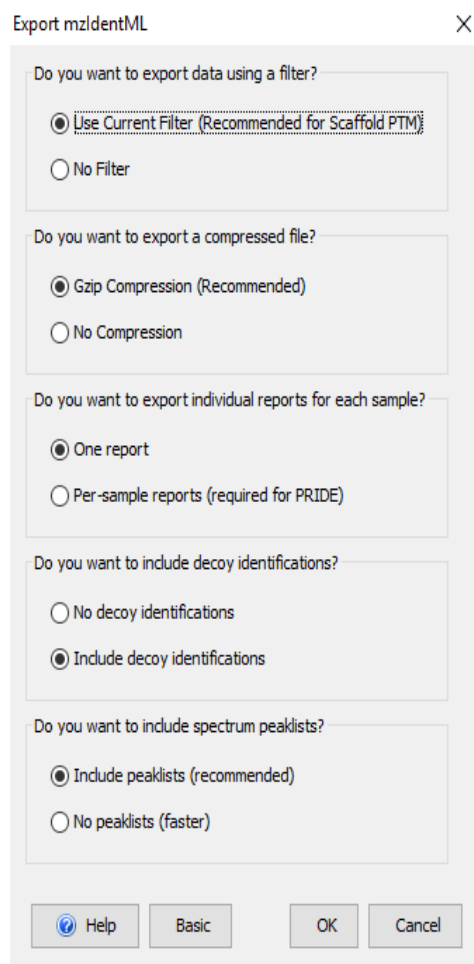
1. MacCoss Lab, University of Washington - <https://skyline.ms/wiki/home/software/BiblioSpec/page.view?name=BiblioSpec%20input%20and%20output%20file%20formats>

Figure 15-3: Export mzIdentML short dialog



Clicking the **Advanced** button expands the dialog to show the full list of options available to further customize the mzIdentML export.

Figure 15-4: Export mzIdentML expanded dialog



The available options are the following:

- Selection of the list of proteins to include in the MzIdentML through the set filters. Use Current Filter exports the set of proteins and peptides currently displayed in Scaffold, respecting current threshold and filter settings. No Filter exports all peptides and proteins.
- Selection of the type of compression
- Selection of the number of reports exported- With multiple BioSamples it is possible to create mzIdentML exports for each BioSample included in the experiment
- Inclusion of decoys
- Inclusion of peak lists - The peak list is saved using the MGF format.



*The mzIdentML export creates one or more *.MZID files and a series of *.MGF files, if the inclusion of peaks option is selected, saved in a newly created meaningfully named directory.*

Creating a spectral library in Skyline

Skyline is a popular application used to create and iteratively refine targeted methods for proteomics studies, see [Skyline](#). It also provides tools to build spectral libraries from validated peptide spectrum matches.

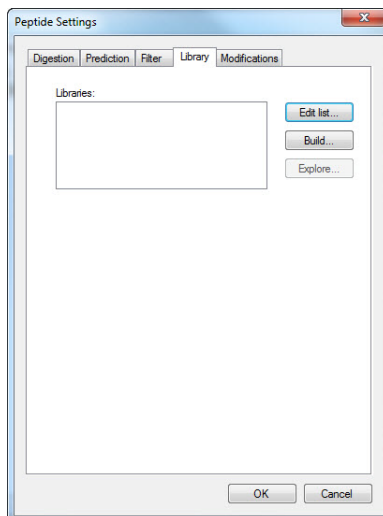
Scaffold experiments can include a variety of validated spectrum matches coming from different sources and analyzed using multiple search engines. This means that Scaffold experiments can be a good source for validated spectrum matches. The way Scaffold exports spectral identification results is through mzIdentML reports and associated peak lists in MGF format. Within these files Scaffold embeds precursor intensity, retention time, and a reference to the original RAW file, which are requirements for creating transition libraries in Skyline. The mzIdentML exports are now compatible with Skyline and can be used to create spectral libraries within that application.

Furthermore, Scaffold supports a large variety of search engine reports, some of which are not currently compatible with Skyline. In particular Proteome Discoverer is a common platform for MS/MS based proteomics which is now compatible with Skyline using Scaffold as an interface.

The user can create a spectral library in Skyline following these instructions:

1. From a Scaffold experiment select the menu option **Export > mzIdentML...**, the Export [mzIdentML](#) dialog opens. In the dialog select the option Scaffold PTM analysis and then click **Advanced** see [Figure 15-4](#).
2. To the question: **Do you want to export a compressed file?** select the answer **No compression**. Click **OK** to save the mzIdentML files.
3. To create a spectral library in Skyline go to the menu option **Settings > Peptide Settings....** The **Peptide Settings** dialog opens onto the tab **Library**.

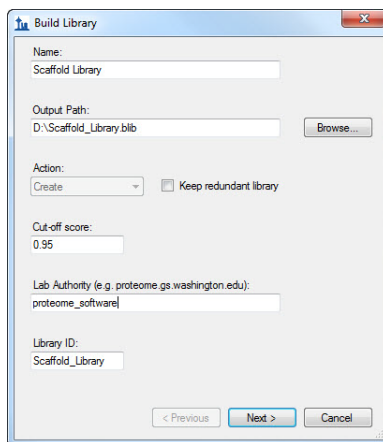
Figure 15-5: Skyline: Peptide Settings, Library tab



4. Click **Build** to add a new library. The Build Library wizard opens.

Figure 15-6: Skyline: Build library wizard initial page

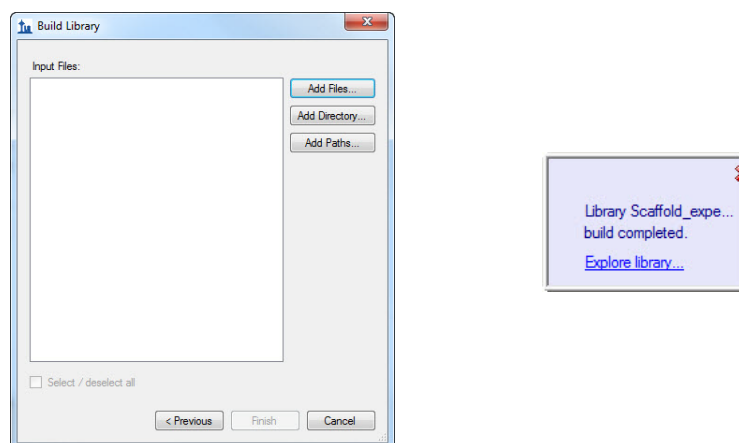
5. Assign a name to the library and if needed adjust the output path where the library is



saved, click **Next**.

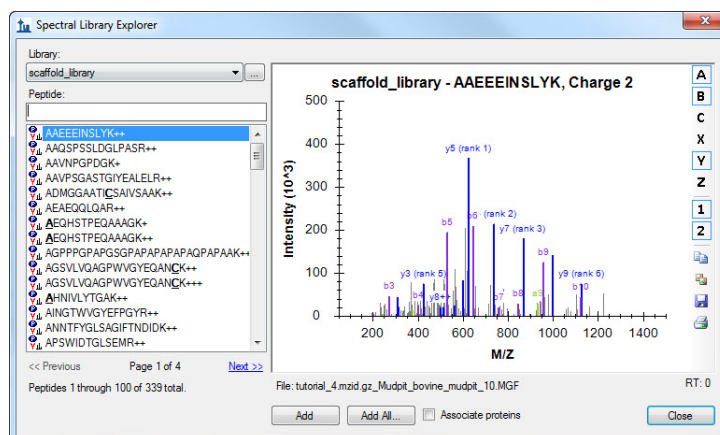
6. In the new page of this wizard click **Add Files** and point Skyline to the location where the Scaffold *.MZID is saved.
7. Once the file is selected it appears listed in the Input file text area. Click **Finish**.

Figure 15-7: Skyline: wizard and message when the library is loaded



8. A message appears once the library is built up and ready for use.
9. The newly built library is listed in the text area Libraries in the **Peptide Settings** dialog, **Library** tab.
10. To view the newly loaded library first select it in the **Peptide Settings** dialog and then choose **View > Spectral Libraries**. The Spectral Library Explorer dialog opens.

Figure 15-8: Skyline: Spectral Library Explorer



SFDB...

This option exports the experiment in the SFDB format, which is a Scaffold LFQ experiment file. This makes it easy to perform further analysis of the data with Scaffold LFQ. When this export option is clicked, a dialog appears, asking whether the export should respect the current thresholds and filters or should export the complete unfiltered data set. Exporting filtered allows creation of a smaller Scaffold LFQ file, which may lead to faster, more responsive behavior in that program. If, however, the user may wish to adjust the thresholds

or look at data currently hidden by filters, the unfiltered option is better.

Scaffold Batch...

Scaffold Batch is a command line version of Scaffold designed to run in the background of a system where it is installed. Scaffold Batch reads its commands from an XML type driver file (*.SCAFML) rather than from the graphical user interface (GUI) and creates an *.SF3 Scaffold experiment.

When selected the command option **Export > Scaffold Batch** creates a SCAFML driver file that if run in Scaffold Elements reproduces the Scaffold experiment from which it is exported.

SCAFML files are often used as an interface between Scaffold and another program. A number of labs have created custom software that uses SCAFML driver files as an interface between a Laboratory Information Lab System (LIMS) and Scaffold.

Commercial versions of Scaffold Batch are available bundled into Sage N Research's Sorcerer and Genologics' Proteus Analytics. Matrix Science has an interface between their Integra LIMS system and Scaffold that uses SCAFML files.

Scaffold Batch Archive...

When selected the command option **Export > Scaffold Batch Archive** bundles into one package a SCAFML driver file that, if run in Scaffold Elements, reproduces the Scaffold experiment from which it is exported, along with all the files that are referenced in the SCAFML driver. This package typically contains the input files and the FASTA database and it is saved in a compressed format like zip or tar. Before running it on a computer that has Scaffold Elements installed the user needs to unzip or untar the package.

Exports compatible with Excel

Scaffold provides a number of tab delimited reports containing different types of information related to the analysis performed in the current Scaffold experiment:

- [Publication report](#)
- [Samples report](#)
- [Spectrum report](#)
- [Peptide report](#)
- [Protein report](#)
- [Current View report](#)
- [Complete report](#)

How to open Scaffold reports in Excel:

The exported reports can be viewed in Microsoft Excel for further analysis of the data they contain. When importing any of these reports into the spreadsheet, it may be necessary to specify that the report is TAB delimited. Excel may also show its Text Import Wizard the first time the user opens an exported Scaffold report. Selecting delimited file, then tab-delimiters completes the conversion to the Excel format. Saving as an XLS file avoids repeating the conversion in the future



To create an export that includes the GO annotations see [Samples report](#)

Publication report

The Publication report lists the data analysis information required for publication in a number of the Proteomics journals. This report is a copy of the information reported on the Publish view.

The top of the report lists in a structured manner:

- How the peak lists were generated
- What databases were searched to identify the proteins
- What parameters were used by the search engine (or engines)
- What criteria were used for protein identification.

Following this is a narrative description of the same information. This can be used as a rough draft for the methods section of any journal article. Although the user will undoubtedly want to clean up this computer generated text to improve its readability, it gives a place to start.

Exporting MCP supplemental table, the second step in the MCP Submission checklist,

exports the [Publication report](#). In the MCP Submission procedure the user must finish step 1, describing the experimental methods, before he/she can export the Publication Report in order to ensure that the report is complete.

For publication in Proteomics journals, the user might also use the [Protein report](#) and [Peptide report](#) as supplemental supporting data.

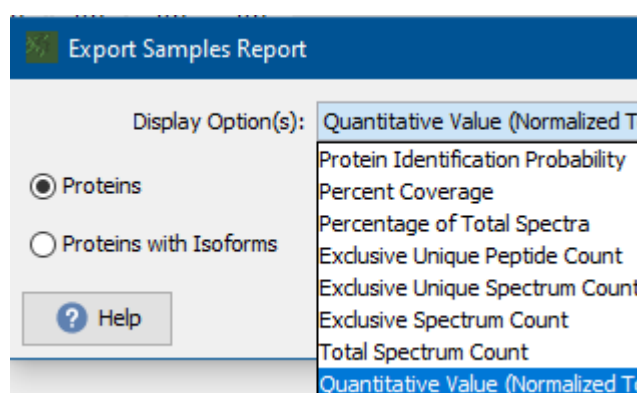
Figure 15-9: Publication report example

```
tutorial_2_multiple_cat, Publication report created on 01/09/2014
Experiment: tutorial_2_multiple_cat
  Peak List Generator: unknown
    Version: unknown
    Charge States Calculated: unknown
    Deisotoped: unknown
    Textual Annotation: unknown
  Database Set: 2 Databases
    Database Name: a subset of the control_sprot database
      Version: unknown
      Taxonomy: All Entries
      Number of Proteins: 766
    Database Name: the control_sprot_1 database
      Version: unknown
      Taxonomy: All Entries
      Number of Proteins: 127876
    Explain Database w/ < 1000 entries:
    Does database contain common contaminants?: unknown
  Search Engine Set: 2 Search Engines
    Search Engine: Mascot
      Version: 2.4.0
      Samples: All Samples
      Fragment Tolerance: 0.50 Da (Monoisotopic)
      Parent Tolerance: 1.2 Da (Monoisotopic)
      Fixed Modifications: +57 on C (Carbamidomethyl)
      Variable Modifications: +16 on M (Oxidation), +43 on n (Carbamyl)
      Database: the control_sprot_1 database (unknown version, 127876 entries)
      Digestion Enzyme: Trypsin
      Max Missed Cleavages: 1
      Probability Model:
        control_071904_01 (F001807): LFDR Model, Classifier data: Bayes, Good (50%) m:49.9/s
        control_071904_02 (F001808): LFDR Model, Classifier data: Bayes, Good (50%) m:74.5/s
        control_071904_03 (F001809): LFDR Model, Classifier data: Bayes, Good (50%) m:53.5/s
        control_071904_04 (F001810): LFDR Model, Classifier data: Bayes, Good (50%) m:49.9/s
        control_071904_05 (F001811): LFDR Model, Classifier data: Bayes, Good (50%) m:52.5/s
    Search Engine: X! Tandem
      Version: CYCLONE (2010.12.01.1)
      Samples: All Samples
      Fragment Tolerance: 0.50 Da (Monoisotopic)
      Parent Tolerance: 1.2 Da (Monoisotopic)
      Fixed Modifications: +57 on C (Carbamidomethyl)
      Variable Modifications: -18 on n (Glu->pyro-Glu), -17 on n (Ammonia-loss), -17 on n (Gln->pyro-Glu)
      Database: a subset of the control_sprot database
      Digestion Enzyme: Trypsin
      Max Missed Cleavages: 2
      Probability Model:
        control_071904_01 (F001807): LFDR Model, Classifier data: Bayes, Good (68%) m:51.1/s
        control_071904_02 (F001808): LFDR Model, Classifier data: Bayes, Good (79%) m:49.1/s
        control_071904_03 (F001809): LFDR Model, Classifier data: Bayes, Good (73%) m:51.5/s
        control_071904_04 (F001810): LFDR Model, No Classifier [all charge states]
        control_071904_05 (F001811): LFDR Model, No Classifier [all charge states]
  Scaffold: Version: Scaffold_4.2.1
    Modification Metadata Set: 1541 modifications
```

Samples report

The Samples report mimics the Samples View. The report header rows identify the data and how it was created, which is the same information that is contained in the [Publication report](#). Subsequently each row in the report represents a protein in the samples list. The number of proteins displayed depends on the current filter and threshold settings.

The Export>Samples Report menu option opens a dialog which allows selection of the Display Option(s) to be reported, and whether to include isoforms and/or clusters. The All display options produces a report that is very useful for publishing results. It can be opened in Excel and all values can easily be summed to the appropriate level of categorization.



If **Edit > Show GO Annotations** is selected, the Go annotation information appearing in the Samples View will also be included in the Samples Report.

There are three slightly different version of this report:

- *Samples report (regular)*- See [Samples report](#).
- *Samples report with clusters* - Available when protein cluster analysis is selected. It adds clusters to the regular report
- *Samples report with Isoforms* - Includes expanded protein groups to the regular report.
- *Spectrum Counting report* - It is like the regular Samples report but reports the samples quantitative values and the proteins identification probability.

Figure 15-10: Samples report columns

#	Visible?
	Starred?
	Identified Proteins (32)
	Accession Number
	Molecular Weight
	Protein Grouping Ambiguity
	Quantitative Variance
	Taxonomy
	biological adhesion
	biological regulation
	cell killing
	cellular process
	developmental process
	establishment of localization
	growth
	immune system process
	localization
	locomotion
	metabolic process
	multi-organism process
	multicellular organismal process
	pigmentation
	reproduction
	reproductive process
	response to stimulus
	rhythmic process
	viral reproduction
	Golgi apparatus
	cytosol
	cytoskeleton
	endoplasmic reticulum
	endosome
	extracellular region
	intracellular organelle
	membrane
	mitochondrion
	nucleus
	organelle membrane
	organelle part
	plasma membrane
	ribosome
	antioxidant activity
	auxiliary transport protein activity
	binding
	catalytic activity
	chaperone regulator activity
	chemoattractant activity
	chemorepellent activity
	electron carrier activity
	enzyme regulator activity
	metalochaperone activity
	molecular function
	molecular transducer activity
	motor activity
	nutrient reservoir activity
	protein tag
	structural molecule activity
	transcription regulator activity
	translation regulator activity
	transporter activity
	1_M
	1_G

Spectrum report

The Spectrum report details all the spectra passing the current filter and threshold settings. The report header rows identify the data and how it was created, which is the same information that is contained in the [Publication report](#). Afterwards, each entry represents a spectrum matching a peptide.

Figure 15-11: Spectrum report columns

Experiment name
Biological sample category
Biological sample name
MS/MS sample name
Protein name
Protein accession numbers
Database sources
Protein molecular weight (Da)
Protein identification probability
Exclusive unique peptide count
Exclusive unique spectrum count
Total spectrum count
Percentage of total spectra
Percentage sequence coverage
Manual validation
Assigned
Spectrum name
Peptide sequence
Previous amino acid
Next amino acid
Peptide identification probability
SEQUEST XCorr score
SEQUEST DCn score
X! Tandem -log(e) score
Number of enzymatic termini
Fixed modifications identified by spectrum
Variable modifications identified by spectrum
Observed m/z
Actual peptide mass (AMU)
Calculated +1H Peptide Mass (AMU)
Spectrum charge
Actual minus calculated peptide mass (AMU)
Actual minus calculated peptide mass (PPM)
Total Ion Current
Peptide start index
Peptide stop index
Exclusive
Other Proteins

Column quick notes:

- The first 14 columns of the table provide information available in the Protein Report identifying the sample and the protein.
- The Manual validation column reports if the user manually validated a spectrum. This is done by selecting or deselecting the check-box in the “Valid” column shown in the spectrum table displayed in the peptides pane. One of the following possible statuses of the Valid check box can appear in the report:
 - *Possibly Correct* - The user accepted the status of the box resulting from Scaffold analysis and did not touch it.
 - *Correct* - The user deselected and then selected the box again.
 - *Unchecked box* - If the box remains unchecked the spectrum does not appear in this report.
- Number of enzymatic termini (NTT). When the digestion enzyme is trypsin, this tells if the peptide is tryptic (2) semi-tryptic (1) or non-tryptic (0).

Peptide report

The Peptide report details all the peptides that pass the current filter and thresholds settings. The report header rows identify the data and how it was created, which is the same information that is contained in the [Publication report](#).

Afterwards, every row represents a peptide in each of the samples present in the Scaffold experiment. For example, if there are 3 samples each with 100 peptides, there will be 300 rows in the report. Even if several spectra match a peptide, the peptide only gets one line in

this report.

Figure 15-12: Peptide report columns

Experiment name
Biological sample category
Biological sample name
MS/MS sample name
Protein name
Protein accession numbers
Database sources
Protein molecular weight (Da)
Protein identification probability
Exclusive unique peptide count
Exclusive unique spectrum count
Total spectrum count
Percentage of total spectra
Percentage sequence coverage
Peptide sequence
Previous amino acid
Next amino acid
Best Peptide identification probability
Best SEQUEST XCorr score
Best SEQUEST DCn score
Best X! Tandem -log(e) score
Number of identified +1H spectra
Number of identified +2H spectra
Number of identified +3H spectra
Number of identified +4H spectra
Number of enzymatic termini
Calculated +1H Peptide Mass (AMU)
Median Retention Time
Total Precursor Intensity
Total TIC
Peptide start index
Peptide stop index
Star Category
Assigned
Other Proteins

Columns quick notes:

- The first 14 columns repeat the information available in the Protein Report which identify the sample and the protein.
- Next comes the peptide sequence followed by the best scores for the spectrum matching it. Then there are columns showing how many spectra matched the peptides in each charge state and a column for the number of tryptic termini (NTT)

There are two different versions of this report:

- *Peptide Report (regular)* - See [Figure 15-12](#)
- *Peptide Quantitative report* - which exports similar information as the regular report does but organizes it emphasizing the various quantitative values available in the experiment for each peptide in every sample, see [Figure 15-13](#)

Figure 15-13: Peptide quantitative report columns

	Total Spectrum Count					Weighted Spectrum Count					Total Precursor Intensity					Total TIC					Median Retention Time				
	1	M	1	G	1	G	1	M	1	G	1	G	1	M	1	G	1	G	1	M	1	G	1	G	
Protein Peptide	Mudpit_bovine_mudpit_10							Mudpit_bovine_mudpit_10												Mudpit_bovine_mudpit_10					
	bovine_mudpit_10							bovine_mudpit_10												bovine_mudpit_10					
	bovine_mudpit_11							bovine_mudpit_11												bovine_mudpit_11					
	bovine_mudpit_12							bovine_mudpit_12												bovine_mudpit_12					
	bovine_mudpit_13							bovine_mudpit_13												bovine_mudpit_13					
	Mudpit_bovine_mudpit_10							Mudpit_bovine_mudpit_10												Mudpit_bovine_mudpit_10					
	bovine_mudpit_10							bovine_mudpit_10												bovine_mudpit_10					
	bovine_mudpit_11							bovine_mudpit_11												bovine_mudpit_11					
	bovine_mudpit_12							bovine_mudpit_12												bovine_mudpit_12					
	bovine_mudpit_13							bovine_mudpit_13												bovine_mudpit_13					
	Mudpit_bovine_mudpit_10							Mudpit_bovine_mudpit_10												Mudpit_bovine_mudpit_10					
	bovine_mudpit_10							bovine_mudpit_10												bovine_mudpit_10					
	bovine_mudpit_11							bovine_mudpit_11												bovine_mudpit_11					
bovine_mudpit_12							bovine_mudpit_12												bovine_mudpit_12						
bovine_mudpit_13							bovine_mudpit_13												bovine_mudpit_13						

Protein report

The protein report details all the proteins passing the current filter and threshold settings. This report is designed to be used as part of the supplemental information supporting a journal article. The report header rows identify the data and how it was created, which is the same information that is contained in the [Publication report](#). Then there is a single report entry for each protein in every sample. For example, if there are 3 samples each with the same 12 proteins, there will be 36 rows in this report.

Figure 15-14: Protein report columns

Experiment name
Biological sample category
Biological sample name
MS/MS sample name
Protein name
Protein accession numbers
Database sources
Protein molecular weight (Da)
Protein identification probability
Exclusive unique peptide count
Exclusive unique spectrum count
Total spectrum count
Percentage of total spectra
Percentage sequence coverage
Peptide sequence
Previous amino acid
Next amino acid
Best Peptide identification probability
Best SEQUEST XCorr score
Best SEQUEST DCn score
Best X! Tandem -log(e) score
Number of identified +1H spectra
Number of identified +2H spectra
Number of identified +3H spectra
Number of identified +4H spectra
Number of enzymatic termini
Calculated +1H Peptide Mass (AMU)
Median Retention Time
Total Precursor Intensity
Total TIC
Peptide start index
Peptide stop index
Star Category
Assigned
Other Proteins

Columns quick notes:

- The first 4 columns identify the experiment, the biological sample, its category and the MS/MS sample (or run).
- This is followed by columns that identify the protein by name, accession number, database where the accession resides, and the protein's mass.
- The remainder of the columns provide the results of the analysis.

There are three different versions of this report:

- *Protein report (regular)* - See [Figure 15-14](#)
- *Protein cluster report* - Available when protein cluster analysis is selected. It adds clusters to the regular report
- *Protein Accession Number Report* - Similar to the regular report, its purpose is to provide an easy way to look up the protein description for each accession number. The report also provides the name of the database that was used for searching the data.

PSEA-Quant Report

PSEA-Quant² is a protein set enrichment analysis algorithm for mass spectrometry-based quantitative proteomics. It identifies protein sets that are statistically significantly enriched with abundant proteins, which are measured with high reproducibility across a set of replicates.

Scaffold offers PSEA-Quant as a method to identify groups of proteins based on shared GO terms, common pathways or user-specified grouping implemented using the star function. It can be applied while adding GO or Pathway Annotations (see [Select Displayed Annotations with PSEA-Quant](#)) and used to select meaningful annotations to be displayed in the Samples View. It can also be run independently of adding annotations in order to produce the PSEA-Quant report.

When run from the export menu, the PSEA-Quant Analysis dialog does not look exactly the same as when it is run for selection of annotation columns.

2. Lavallée-Adam, M., Rauniyar, N., McClatchy, D. B., & Yates, J. R. (2014). PSEA-Quant: A protein set enrichment analysis on label-free and label-based protein quantification data. *Journal of proteome research*

Figure 15-15: PSEA-Quant Analysis dialog as it appears when run from the Export menu

Protein Set Source - determines whether proteins are to be grouped on the basis of:

- GO Terms - proteins are grouped on the basis of shared GO terms.
- Star Color - allows creation of user-defined protein groups for PSEA-Quant analysis. This option is only available through **Export>PSEA-Quant Report**, and not when adding annotations.
- Pathways - proteins are grouped on the basis of inclusion in common pathways.

The experiment must have been previously annotated with GO Terms, Pathways or Stars, or an empty report will be produced.

The Star Color PSEA-Quant analysis is only available in report form.

The PSEA-Quant report lists all protein groups as defined by the protein set source, whether or not the result is significant for the protein groups. The columns in the report are:

Identifier - GO Term, Pathway name, star color

Size - number of proteins in the group

PES - the sum of all of the protein scores calculated by the algorithm for proteins in the group

Avg. PES - the average protein score for a protein in the group

p-Value - probability estimate for enrichment in the group

q-Value - estimated FDR based on the entire experiment

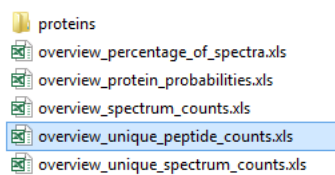
Proteins - list of proteins in the protein group

Current View report

The Current View report contains the information that is displayed in the current view. This report is applicable for the Samples View, the Proteins View, and the Publish View.

Complete report

This export is meant to provide the full results of the current analysis in a series of *.XML files saved in a separate directory. The directory created contains the following list of files:



Appendix

- Appendix A. [Algorithm References](#)
- Appendix B. [Terminology](#)
- Appendix C. [Terminology comparison between Scaffold 4 and Scaffold 3](#)
- Appendix D. [Context Menus Right Click Commands](#)
- Appendix E. [Holm's and Hochberg's Techniques to Control the Familywise Error Rate](#)

Appendix A. Algorithm References

The algorithm for calculating the peptide probabilities from the search engine scores is described in:

Keller (2002)

Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold R., *Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search.*

Anal. Chem., 2002, 74 (20), pp 5383–5392 DOI: 10.1021/ac025747h

The algorithm for calculating the protein probabilities from the peptide probabilities is described in:

Nesvizhskii (2003)

Nesvizhskii, A. I., Keller, A., Kolker, E., Aebersold, R., *A statistical model for identifying proteins by tandem mass spectrometry.*

Anal. Chem., 2003, 75 (17), pp 4646–4658 DOI: 10.1021/ac0341261

The algorithm for combining results from multiple searches is described in:

Searle (2008)

Searle, B.C., Turner, M., Nesvizhskii, A.I., *Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies.*

J Proteome Res., 2008, 7(1), pp 245-53 DOI: 10.1021/pr070540w

The algorithm for grouping proteins across samples is described in

Searle (2010)

Searle, B.C., *Scaffold: A bioinformatic tool for validating MS/MS-based proteomics studies*

Proteomics, 2010, 10(6), pp 1265-9. DOI: 10.1002/pmic.200900437.

The algorithm for X! Tandem is described in:

Craig (2003)

Craig, R., Beavis, R.C., *A method for reducing the time required to match protein sequences with tandem mass spectra.*

Rapid Communications in Mass Spectrometry, 2003, 17(20), pp 2310-6. DOI:10.1002/rcm.1198

Purity corrections calculations for iTRAQ data:

Shadforth (2005)

Shadforth, I.P., Dunkley T.P., Lilley K.S. and Bessant C., *i-Tracker: For quantitative proteomics using iTRAQ™*

BMC Genomics, 2005, 6:145. DOI:10.1186/1471-2164-6-145

Reference for estimation of CV in the [Stdev Scatterplot tab](#).

Pavelka (2004)

Pavelka, N., Pelizzola, M., Vizzardelli, C., Capozzoli, M., Splendiani, A., Granucci, F. and Ricciardi-Castagnoli, P., *A power law global error model for the identification of differentially expressed genes in microarray data*

BMC Bioinformatics 2004, 5:203 doi:10.1186/1471-2105-5-203

Pavelka (2008)

Pavelka, N., Fournier, M., L., Swanson, S., K., Pelizzola, M., Ricciardi-Castagnoli, P., Florens, L. and Washburn M., P., *Statistical Similarities between Transcriptomics and Quantitative Shotgun Proteomics Data*

Mol Cell Proteomics April, 2008 7: 631-644. DOI:10.1074/mcp.M700240-MCP200

Reference for calculating emPAI:

Ishihama (2005)

Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., Mann, M., *Exponentially Modified Protein Abundance Index (emPAI) for Estimation of Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein*

Molecular & Cellular Proteomics, 2005, 4, 1265-1272. DOI: 10.1074/mcp.M500061-MCP200

Reference for calculating SAF:

Zhang (2010)

Zhang, Y., Wen, Z., Washburn, M. P., & Florens, L., *Refinements to Label Free Proteome Quantitation: How to Deal with Peptides Shared by Multiple Proteins*

Anal. Chem., 2010, 82(6):2272-81. DOI: 10.1021/ac9023999

References for Precursor intensity quantitation:

Bantscheff (2007).

Bantscheff M., Schirle M., Sweetman G., Rick J. and Kuster. B. *Quantitative mass spectrometry in proteomics: a critical view*

Anal Bioanal Chem, 2007, 389:1017–1031. DOI:10.1007/s00216-007-1486-6

Lai (2013)

Xianyin Lai, Lianshui Wang, and Frank A. Witzmann, *Issues and Applications in Label-Free Quantitative Mass Spectrometry*

Appendix

International Journal of Proteomics, 2013, vol. 2013, Article ID 756039, 13 pages. DOI:10.1155/2013/756039

Raubenheimer (1992)

Raubenheimer, D. and Simpson, S. L., *Analysis of covariance: an alternative to nutritional indices*.

Entomologia Experimentalis et Applicata, 1992, 62: 221–231. DOI: 10.1111/j.1570-7458.1992.tb00662.x

References for Fisher's Exact test:

Zhang (2006)

Zhang, B., VerBerkmoes, N.C., Langston, M. A., Uberbacher, E., Hettich, R. L., Samatova, N. F. *Detecting differential and correlated protein expression in label-free shotgun proteomics*

J. Proteome Res., 2006, 5 (11), pp 2909–2918. DOI: 10.1021/pr0600273

References for techniques to control the Family Wise error rate:

Huang (2007)

Y. Huang and Hsu J.C. *Hochberg's Step-Up Method: Cutting Corners Off Holm's Step-Down Method* Biometrika, 2007, 94,4,pp.965–975. DOI: 10.1093/biomet/asm067

Benjamini (1995)

Benjamini Y. and Hochberg Y. *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing* Journal of the Royal Statistical Society, Series B (Methodological), 1995, Vol.57, No. 1: 289-300

Appendix B.Terminology

BioSample

Scaffold calls BioSample a physical sample, such as a drop of blood or biopsy from a patient, or a tissue sample from a model organism or cell line. The proteins or peptides in a BioSample are typically separated by 2D gels or liquid chromatography into several spots, bands, or fractions, each of which becomes one mass spectrometry sample or MS Sample. One BioSample is therefore typically made up of several MS samples. Both BioSamples and MS Samples are often referred to by practitioners just as “samples”.

- When running Scaffold Q+ or Scaffold Q+S, quantitative multiplexed samples are initially loaded in Scaffold and referred to as BioSamples.

Category or Sample Category

A group of BioSamples defined within a Scaffold experiment. Categories are useful to organize samples to find which protein is differentially expressed.

Contingency table

In statistics, a contingency table (also referred to as cross tabulation or cross tab) is a type of table in a matrix format that displays the (multivariate) [Frequency Table](#) or distribution of the variables.

Exclusive Spectrum Count

Total number of spectra associated with only one protein

Frequency Table

In statistics, a frequency table is a table that displays the frequency of various outcomes in a sample. Each entry in the table contains the frequency or count of the occurrences of values within a particular group or interval, and in this way, the table summarizes the distribution of values in the sample. Bivariate joint frequency tables or distributions are often presented as (two-way) [Contingency tables](#).

Mascot File Names

If a file containing Mascot data is named something like F*.DAT, for example: F123456.dat or F987654.dat, Scaffold looks inside the Mascot file to search for the name of the original file that Mascot searched. Since that name is more likely to be meaningful, Scaffold uses it to name the MS sample. On the Load Data page, Scaffold displays the “original” name followed by the F*.DAT name assigned by Mascot in parentheses. On the Samples and Proteins pages, Scaffold uses only the original name. The MS Sample name can be changed on the Samples View, see [“Sample Information Pane” on page 155](#).

Mascot Output File

Mascot is a search engine distributed by Matrix Science. Scaffold loads Mascot result files created with the following format: *.DAT. When Mascot is used through Proteome Discoverer the results are stored in *.MSF type of file.

MS or MS/MS Sample

Each individual band, spot, or LC fraction processed by a mass spectrometer in one run. These samples are a result of separation techniques such as 2D gels or liquid chromatography, which separate the

Appendix

proteins or peptides within a biological sample such as a drop of blood or tissue sample. In the case of MuDPIT experiments, Scaffold considers all the MS data in one BioSample to be one MS Sample.

NSAF

The Normalized Spectral Abundance Factor (NSAF) is a modified version of spectral counting. It was introduced and defined by the Washburn Lab group at the Stowers Institute to account for the effects of protein length, as large proteins tend to contribute to the spectral counts a greater amount of peptide/spectra than smaller ones do. For more detailed information visit the Washburn Lab website: <http://research.stowers.org/proteomics/Quant.html>

Percentage of all Spectra

It is defined as the number of spectra matched to a protein, summed over all MS Samples, as a percentage of the total number of spectra in the sample.

Showing the Percentage of Total Spectra in the Samples View gives an idea of how many of the overall spectra loaded in the experiment participate to the protein identification. Lower percentages point to the fact that very few recorded spectra are involved in the identification of a specific protein.

The value is calculated by dividing the total spectrum count appearing for a specific set of filters by the number of spectra loaded for that sample. Here Sample refers to a BioSample if the summarization level selected corresponds to the Biological Sample View or an MS sample if the summarization corresponds to the MS/MS sample View. When in the Biological Sample View the numbers appearing in the Samples table can be checked by selecting the display option Total Spectrum Count and dividing those values by the number of all the spectra loaded in that particular BioSample which is recorded in the [Load Data View](#) under each of the [BioSample](#) tabs.

It is important to note that not all the loaded spectra are assigned to proteins, this is the reason why even when lowering filters and thresholds to the minimum the sum of the Percentage of all Spectra over all the proteins in the list does not amount to 100%.

In the Statistics View, the Table appearing in the Samples pane shows the number of Exclusive spectra in each MS sample, named #IDs, the number of spectra loaded for the specific MS Sample, named #Spec and the percentage of loaded spectra assigned to proteins, named #%IDs, which is typically quite small.

ROC curves

Receiver operating characteristic (ROC) curves were developed to assess the quality of radar. In medicine, ROC curves are a way to analyze the accuracy of diagnostic tests and to determine the best threshold or “cutoff” value for distinguishing between positive and negative test results.

Diagnostic testing is almost always a trade off between sensitivity and specificity. ROC curves provide a graphic representation of this trade off. Setting a cutoff value too low may yield a very high sensitivity (i.e., no disease would be missed) but at the expense of specificity (i.e., a lot of false-positive results). Setting a cutoff too high would yield high specificity at the expense of sensitivity. The best cutoff has the highest sensitivity and lowest 1–specificity, and is therefore located as high up on the vertical axis and as far left on the horizontal axis as possible (upper left corner).

The area under an ROC curve is a measure of the usefulness or “discriminative” value of a test in general. The greater the area, the more useful the test. The maximum possible area under the curve is

simply a perfect square and has an area of 1.0. The diagonal 45° line represents a test that has no discriminative value—i.e., it's completely useless.

Sequest Output Files

Sequest is one of the search engines distributed by Thermo Scientific. Depending on the platform used to run it Sequest creates files with the following format: *.DTA and *.OUT, *.MS2 and *.SQT, and *.SRF types of file. The proteomics platform developed by Thermo Scientific called Proteome Discoverer creates output files with extension:*.MFS

TIC- Total Ion Current

The total ion current (TIC) is the sum of the areas under all the peaks contained in a MS/MS spectrum. Scaffold assumes that the area under a peak is proportional to the height of the peak and approximates the TIC value by summing the intensity of the peaks contained in the peak list associated to a MS/MS sample.

Total Unique Peptide Count

Number of different amino acid sequences that are associated with a protein including those shared with other proteins

Total Unique Spectrum Count

Number of unique spectra associated with a protein including those shared with other proteins

Appendix C. Terminology comparison between Scaffold 4 and Scaffold 3

Beginning with Scaffold version 4, Proteome Software added new terms to capture different types of evidence used in protein clustering. These new terms affect the display options in the Samples View. The tables below indicate the correspondence with Scaffold 3 display option terms (The user might want to print these tables).

Table 1: New Terminology

Unique peptides	Peptides with different amino acid sequences, regardless of any modifications
Unique spectra	Spectra that differ in amino acid sequence, charge state or modifications
Exclusive	Associated with a single protein group
Total	Associated with a protein group, whether or not it is shared with other protein groups

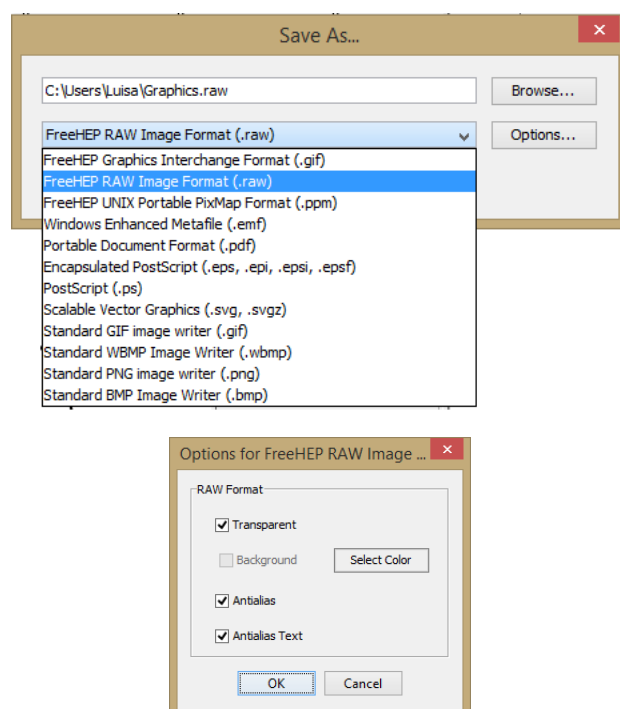
Table 2: Terminology comparison between Scaffold 4 and Scaffold 3

Scaffold 4 term	Scaffold 3 term	Description
Exclusive Unique Peptide Count	Number of Unique Peptides	Number of different amino acid sequences that are associated only with this protein
Exclusive Unique Spectrum Count	Number of Unique Spectra	Number of unique spectra associated only with this protein
Exclusive Spectrum Count	Number of Assigned Spectra	Number of spectra associated only with this protein
Total Spectrum Count	Unweighted Spectrum Count	Total number of spectra associated with this protein including those shared with other proteins
Total Unique Peptide Count	N/A	Number of different amino acid sequences that are associated with this protein including those shared with other proteins
Total Unique Spectrum Count	N/A	Number of unique spectra associated with this protein including those shared with other proteins

Appendix D.Context Menus Right Click Commands

- **BLAST Peptide Sequence** - opens <http://www.ncbi.nlm.nih.gov/blast/> site for current peptide
- **BLAST protein sequence** - opens <http://www.ncbi.nlm.nih.gov/blast/> site for current protein.
- **Copy All Data** - copies all the data listed in the table shown in the current pane to the clipboard.
- **Copy Image** - copies the image of the current view and current pane to the clipboard.
- **Copy Peak List** - copies the pick list of the current spectrum to the clipboard.
- **Copy Protein Sequence** - copies the sequence to the clipboard.
- **Copy publication Sized JPEG** - for publication purposes.
- **Copy Selected Cell** - from the table copies selected cell to the clipboard.
- **Copy Selected row** - from the table copies selected row to the clipboard.
- **Copy WMF/EMF** - copies picture using Windows Meta-file formats which are portable between applications. They contain both vector graphics and bitmap components. Images can be edited and scaled without compromising their resolution.
- **Delete Biological Sample** - a window pops up asking to confirm deletion.
- **Display parent Ions** - toggle function.
- **Display unknown markers** - toggle function
- **Export Peptides of starred to Excel...** - exports Peptide Report only for proteins tagged with a star. If there are no proteins starred, then a warning message pops up notifying the user that the export cannot be created.
- **Export to Excel...** - export information in current tab table
- **Edit BioSample** - See [Edit BioSample](#).
- **Print** - print image of current view and pane.
- **Queue Files for Loading** - See [Queue Search Engine Results for Loading](#), [Queue Raw Data for Loading](#), [Queue Structured Directory for Loading](#), [Queue Files From Mascot Server for Loading...](#) .
- **Save as** - provides the option of saving pictures in a large variety of graphical formats.

Appendix



- **Save JPEG Image** - saves image of current view and pane JPEG format.
- **Show Fixed Modifications** - it toggles the function of highlighting fixed modifications along the sequence.
- **Use Amino Acid Finder** - it toggles the activation of the tool tip that shows the peptides along the sequence.
- **Use Peak-finder** - displays the tool tip for the different peaks if checked.
- **Use PPM Masses** - toggle function.
- **Zoom Out** - zoom function.

Appendix E. Holm's and Hochberg's Techniques to Control the Familywise Error Rate

Scaffold supports different methods to control the familywise error rate, FWER. Among them it supports the Holm's step-down procedure and the Hochberg's step-up procedure, which are developed in the program as described in [Huang \(2007\)](#).

The two methods make the same type of comparisons, but Holm starts at the smallest p-value and works down the list until one fails the bound, while Hochberg starts at the largest p-value and works up the list until one passes the bound (and then declares that everything below that passes. Hence the Holm bound is in general more conservative than the Hochberg.

For example, let's suppose we have ($m=5$) proteins A, B, C, D, E with p-values 0.030, 0.014, 0.013, 0.060, and 0.009 respectively, and want to reject the null hypothesis at $\alpha = 0.05$. Let's sort the p-values and make the following table:

k	p-value	$\alpha / (m + 1 - k)$	p-value < $\alpha / (m + 1 - k)$
1	0.009	0.01	yes
2	0.013	0.0125	no
3	0.014	0.0167	yes
4	0.030	0.025	no
5	0.060	0.05	no

- The Holm step-down procedure would start at $k=1$ and reject $H_0(1)$ but it would stop at $k=2$ since the p-value is larger than the bound.
- The Hochberg step-up procedure would start at $k=5$, go to $k=4$, go to $k=3$, see that the bound passes and stop, accepting $H_0(1)$, $H_0(2)$, and $H_0(3)$.

The user should be aware of the fact that technically the Hochberg procedure should only be used if the hypothesis tests are independent (which they are certainly not for Fisher's Exact Test, and not usually really for the other tests as well).

Index

A

Acetylation 57
 Adjusting the Delta PPM value .. 164
 Advanced Preferences
 Proteome Discoverer settings... 90
 Algorithms references 264
 Alternate ID 80, 82
 ANOVA 229
 Appendix 263
 References algorithms 264
 Applying Thresholds 142
 Assumptions for the manual 8

B

BLIB export 248

C

Coefficient of variance 227
 Coefficient of variation 227
 Configure advanced protein filter 150
 Configure peptide threshold dialog box 144
 Conventions used in the manual.... 7
 Copyright 3
 Current Set 98

D

Default View Options (for new files). 87
 Display pane in the Samples View .. 148
 display options 148
 Req Mods..... 149

search feature 150

E

Emphasize windows options 110
 Export
 mzIdentML 248
 Pride..... 248
 ProtXML 248
 Scaffold perSPECTives..... 248
 ScaffoldPTM..... 248
 Spectra 246
 Subset database 245, 246
 Export reports 245

F

FASTA databases 76
 FASTA databases in Scaffold 52
 FDR
 How Scaffold calculates 147
 FDR filtering
 Filtering Samples 146
 Filtering samples
 Custom peptide filters 144
 FDR filtering 146
 Minimum number of peptides.. 143
 Peptide threshold 143
 Protein thresholds 142
 Fisher exact test..... 229
 Fold change
 by category..... 226
 by sample..... 226
 FragPipe decoys 78

G

Gene Ontology terms pane 187
 GO bar charts 188
 GO Pie charts 188

H

he categories assigned to each axis in the Volcano Plot can be changed through the two pull down menus located above the scatter plot in the Scatterplot tab. 184

I

Identification pane in the Similarity View 174

L

Label Free Quantitative Methods 216
 precursor intensity 219
 precursor intensity quantitation
 average 220
 top three 220
 total 220
 selecting quantitative method. 217
 spectrum counting..... 217
 empai 218
 NSAF..... 218
 total spectra..... 217
 weighted spectra 218
 total ion count (TIC)..... 219
 average TIC 219
 top three TIC 219
 total TIC..... 219

LFDR-based scoring system..... 30

Licensing the Program 11

M

Manual registration..... 13
 Mascot Scoring Function..... 91
 Minimum value 223
 Missing values..... 223
 Mouse right click commands..... 271

Moving the Program to a different computer.....	18
MSFragger.....	60
Searching Raw Data.....	60
Supplementary Searches	61
MSFragger decoy search	78
mzIdentML	
export.....	248

N

Normalization.....	222
Normalization among Biosamples in Scaffold.....	222
minimum value.....	223
missing values	223

O

Offline activation	13
--------------------------	----

P

parentheses in Similarity View...	169
Peptide report.....	257
Peptide Validation pane in the Statistics View	200
PeptideProphet.....	29
Peptides pane in the Proteins View .	161
Percolator	29
Precursor Intensity.....	219
Precursor Intensity quantitation	
average.....	220
top Three	220
total.....	220
Precursor intensity quantitation .	235
Calculations	236
Mascot Distiller	238
MaxQuant	239
Performing quantitation.....	239

Preparing data for	238
Proteome Discoverer	238
Spectrum Mill	239
Prefiltered Mode	48
Pride	
export.....	248
Protein and PeptideProphet.....	28
Protein annotation preferences..	139
Protein Information pane in the Samples View	154
Protein list.....	137
Hidden proteins.....	141
Protein cluster	138
Protein group	137
Proteins of interest.....	140
Sorting Feature	140
Protein N-Terminal Acetylation	57
Protein Sequence tab in the Proteins View	165
Protein Set Source.....	99
ProteinProphet.....	30
proteins	
hiding in the Samples View.....	141
proteins of interest	
identifying in the Samples View	
starring proteins	140
Proteins View	
Peptides pane	161
Protein Sequence tab	165
Spectra pane.....	165
Spectrum tab.....	166
ProtXML export.....	248
PSEA-Quant	98, 99
PSEA-Quant Report	259
Publication report.....	254

Q

Quantify View	
Gene Ontology terms pane.....	187

GO bar charts.....	188
GO pie charts	188
Quantitation in Scaffold	
Precursor intensity	239
Quantitative Analysis Tests	
ANOVA.....	229
coefficient of variation	227
Fisher exact test.....	229
fold change by category	226
fold change by sample	226
T-test.....	228
Quantitative methods in Scaffold	216
Quantitative Tests	225
Queue Raw Data for Loading	125

R

Release Information.....	3
Reports	245
reports	
Peptide.....	257
Publication.....	254
Samples	256
Reset Don't Show Messages button	87

S

Sample Information Pane	155
Sample Information pane in the Samples View	155
Samples report.....	256
Samples Table	135
Samples View	
advanced Search	150
Display pane	148
hiding proteins from.....	141
Protein Information pane.....	154
proteins of interest.....	140

Sample Information pane	155
sorting feature	140
Scaffold 3	
terminology comparison Scaffold 4	270
Scaffold DIA	248
Scaffold perSPECTives	
export	248
ScaffoldPTM	
export	248
Scoring Algorithm	
Prefiltered Mode	140
Scoring algorithm	
LFDR-based scoring	140
PeptideProphet scoring	140
Scoring Function	
Mascot	91
Sequest	90
Scoring system	
LFDR-based	30
PeptideProphet	29
ProteinProphet	30
Sequest Scoring function	90
show hidden proteins	141
Similarity pane in the Similarity View	171
Similarity View	
Identification pane	174
Similarity pane	171
Similarity View	
Spectrum tab	175
Skyline	250
sorting feature	
Samples View	140
Special information about the manual	7
Spectra pane in the Proteins View ..	165
Spectrum Counting	217
Spectrum counting	217

emPAI	218
NSAF	218
total spectra	217
weighted spectra	218
Spectrum tab in the Proteins View ..	166
Spectrum tab in the Similarity View .	175
Statistical Tests	225
Statistics View	
MS/MS Samples Table	193
Multiple Search Engine Scatter Plot pane	199
Peptides Validation pane	200
T	
Terminology	267
BioSample	267
Terminology comparison between Scaffold 3 and Scaffold 4	270
The	184
The Scaffold loading Wizard	41
TIC	219
Average	219
top three	219
total	219
TMTpro	108
Total ion count (TIC)	219
average	219
top three	219
total	219
Total Spectra	217
T-test	228

U

Using the manual	7
------------------------	---

V

Volcano Plot category selection	184
---------------------------------	-----

W

Web Link	85
Web lookup	137

