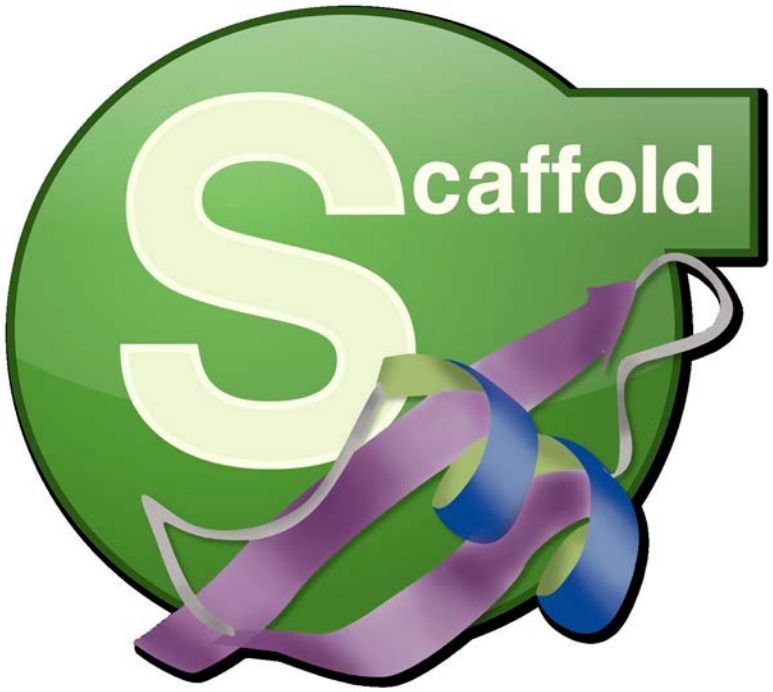


Getting Started with Scaffold 4

Proteome Software Users Group 2013

Tutorial Session

Module 1: Loading and Organizing Your Data

Getting Started with Scaffold 4

Module 1: Loading and Organizing Your Data

User's Tutorial - Proteome Software Users Group 2013

Many resources are available to help you get started using Scaffold. The Proteome Software website offers video tutorials, white papers and FAQ's on various topics at http://proteome-software.wikispaces.com/Resource+Library. The Scaffold User's Guide is found under the Help menu, and there is an online Help system built into the program which provides answers to many of your questions. You can find quick answers to many questions by clicking on the help icons scattered throughout the program. This tutorial is intended to supplement these resources. It does not give a complete introduction to Scaffold. Rather, it will focus on some of the features that we have found to cause the most confusion in new users.

The modules are designed to be independent, so you may choose to work on the units of greatest interest to you.

Locate the Scaffold tutorial data on our website at http://www.proteomesoftware.com/products/demo-data#scaffold. For the exercises in this module you will need to download the data from Scaffold Tutorial 3 – Sequest and Scaffold Tutorial 3- Mascot .

Contents:

Getting Started with Scaffold 4 ..... **0**

Module 1: Loading and Organizing Your Data **1**

Exercise 1: Organizing your samples 1

Exercise 2: To MuDPIT or not to MuDPIT 3

Exercise 3: Combining Search Engines 5

Exercise 4: Running X!Tandem 7

Exercise 9

Exercise 5: Selecting a Scoring System 10

Exercise 6: Protein Grouping and Clustering 13

Exercise 7: Applying a FASTA Database 17

Module 1: Loading and Organizing Your Data

Scaffold's Load Data Wizard makes it fairly simple to load your data, but there are still a few decisions you will need to make.

Exercise 1: Organizing your samples

- When loading your data into Scaffold you will be creating BioSamples. Each BioSample is made up of at least one MS Sample. MS Samples can represent individual MS/MS runs, technical replicates, or fractions of a BioSample.

- BioSamples are grouped by Sample Category for comparison.

For purposes of this exercise, suppose that we had three biological samples, one from a healthy animal and two from diseased animals. Each of these samples was run on a 2-D gel and spots were cut out and subjected to mass spectrometry. The spots are numbered as follows:

Healthy Cow: spots 06 – 10
Diseased Cow 1: spots 11-15
Diseased Cow 2: spots 12-16

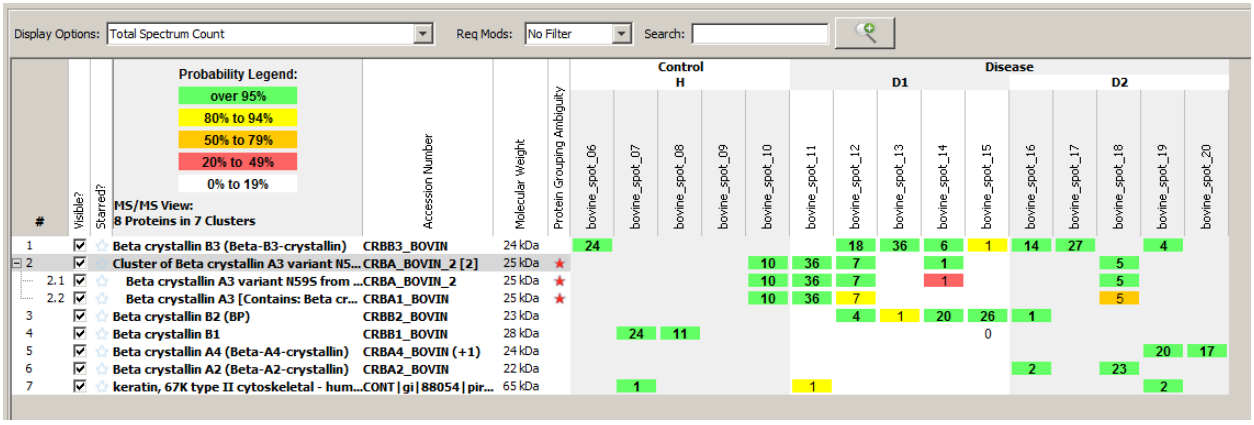
Start the Scaffold Load Data Wizard by clicking on the new experiment icon. Load the data found in the folder tutorial_3seq into three BioSamples, named H, D1 and D2. Place BioSample H into category "Control" and BioSamples D1 and D2 into category "Disease".

Accept defaults for all loading options.

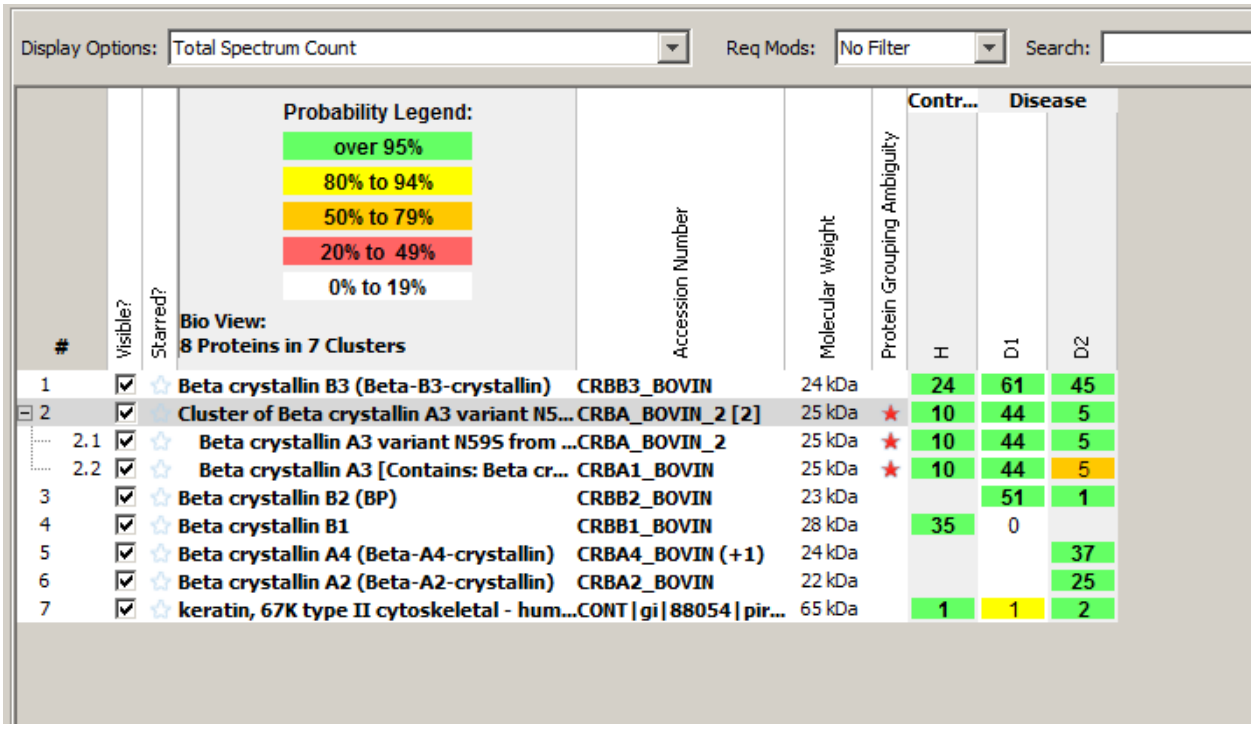
Apply database swissprot_bovine FASTA Database that came with your downloaded data files.

Exercise 1 Result:

Select “Total Spectrum Count” from the Display Options dropdown. Click on the MS View icon. The Samples View should show:



Click on the BioSample view icon. Counts will be combined and columns collapsed to the BioSample level.



Exercise 2: To MuDPIT or not to MuDPIT

MuDPIT**,** or Multidimensional Protein Identification Technology, is a well-known technique in Proteomics. In MuDPIT, 2D-LC is coupled with MS/MS to resolve and identify peptides from complex mixtures. The result is that peptides from the same protein are sometimes separated into different fractions. If each fraction were analyzed independently, there might not be sufficient evidence to identify the protein in any one sample. Scaffold deals with this by allowing you to combine the fractions treating them as a single MS Sample. In this way all of the peptide evidence will be considered together, giving a better chance of identifying all of the proteins that were actually present in the original complex sample.

MuDPIT should **only** be used in cases where samples have been separated at the peptide level and it is necessary to reassemble them for protein identification. It is important that the searches combined in this way were carried out against the same database and with the same search parameters and that the samples were digested with the same enzyme.

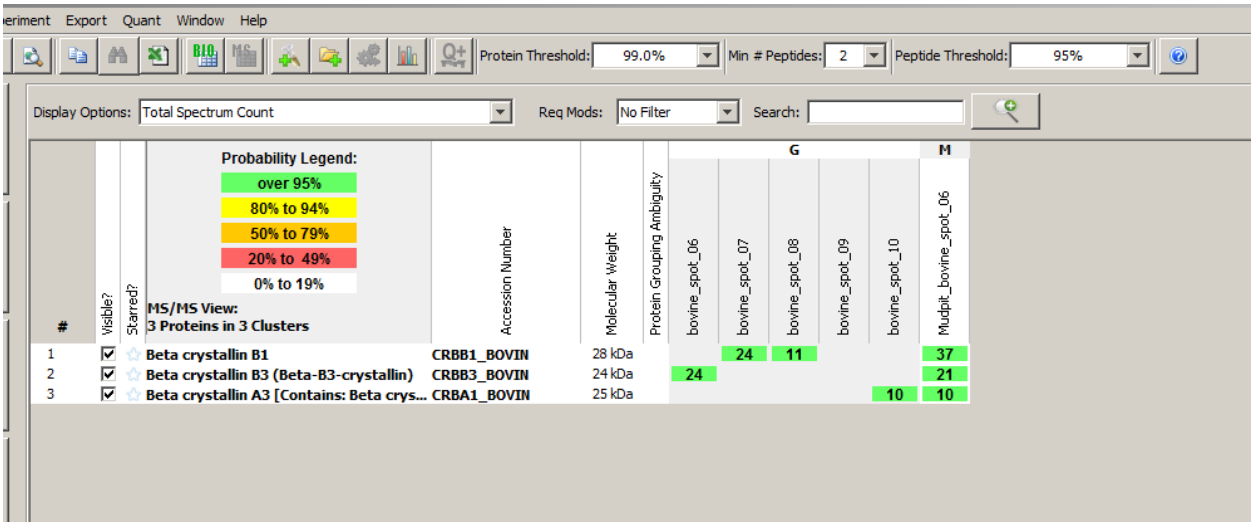
This exercise illustrates Scaffold's handling of data loaded with and without the MuDPIT option. Load the same set of data files into two different BioSamples, selecting the MuDPIT option for the first BioSample, and unchecking the option for the second BioSample.

Queue the files for spots 06-10 into a BioSample called "M" with the MuDPIT option checked, then add another BioSample called "G" and queue the same files but with the MuDPIT option unchecked.

Load the data without running X!Tandem and accepting all other default options.

When the Samples View appears, set the Display Option to "Total Spectrum Count" and click the MS view icon.

Exercise 2 Result:



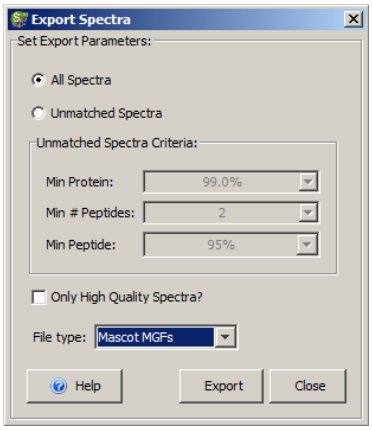
Exercise 3: Combining Search Engines

Different search engines apply different criteria in evaluating potential peptide-spectrum matches and so often identify different peptides. Combining different searches of the same spectra can result in more confident identifications or can provide better coverage.

Scaffold allows you to easily combine results from different search engines. The program aligns the corresponding spectra from the different searches and combines the scores to produce a more confident estimate of the probabilities of the peptide identifications.

There are a few important caveats:

- You cannot combine searches of the same spectra against different FASTA databases.
- Searches should be carried out with similar search parameters.
- You must use the same peak-picker for all searches to be combined. If you have search results and would like to perform a complementary search with another search engine but do not have access to the spectrum files used in the original search, Scaffold can help. Load the data into Scaffold, then export all spectra using Export>Spectra…



You can choose to export the spectra in a variety of formats. Select the best option for the search engine you wish to use for re-searching the spectra.

Start the Loading Wizard, and queue the files bovine_spot 06 through bovine_spot09 from the tutorial_3seq folder for loading. Select "Queue More Files For This BioSample" and select the corresponding files from tutorial_3mas. Continue with the wizard and load the queued files, using the swissprot_bovine FASTA Database and accepting all default options.

a) Go to the Load Data view.
b) Go to the Statistics View. Set the Peptide Threshold to 95%.

Exercise 3a Result:

In Load Data, you should see the loaded samples aligned as shown:

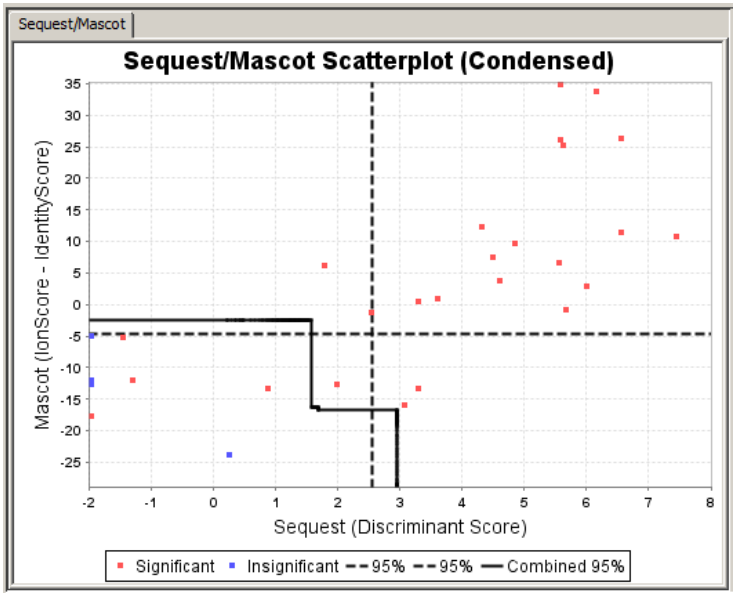
Files Currently Loaded

Mascot	Sequest
bovine_spot_06	bovine_spot_06
bovine_spot_07	bovine_spot_07
bovine_spot_08	bovine_spot_08
bovine_spot_09	bovine_spot_09

This indicates that Scaffold recognized and aligned the corresponding spectra in the Mascot and Sequest search results**. It is important to check for this alignment when loading data from multiple search engines.** If the corresponding files do not appear on a single line, Scaffold has been unable to align the spectra. Most often this is a result of using different peak selectors or of searching different databases.

Exercise 3b Result:

In the lower left quadrant of the Statistics View, you should see a graph that looks like this:



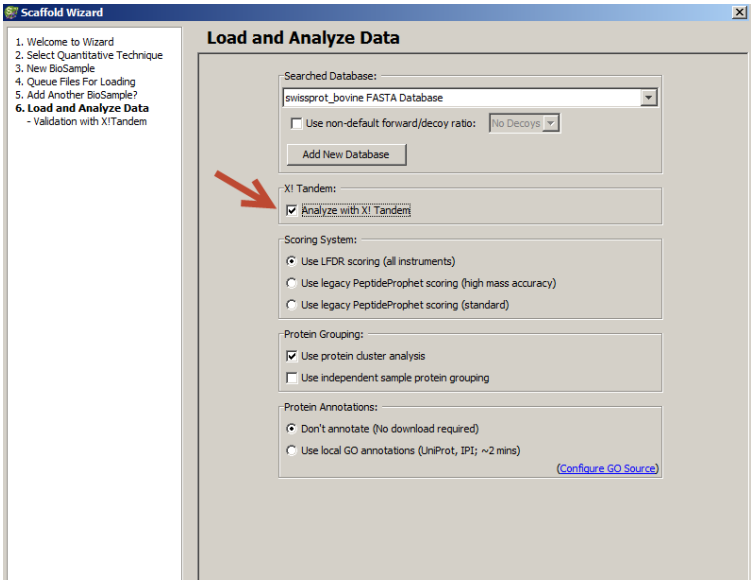
The blue dots indicate spectra that have been assigned as incorrect, while the red dots indicate spectra that have been assigned as correct. The Mascot scores are shown on the y-axis and the Sequest scores on the x-axis. The dotted lines show where the cutoff would be for each individual search engine, while the solid line shows the cutoff for the combined search. Combining the searches results in more peptides being confidently identified.

Exercise 4: Running X!Tandem

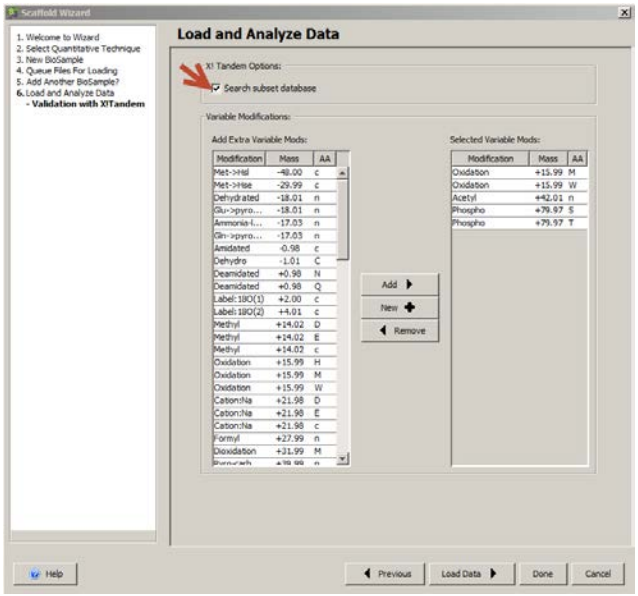
Scaffold includes the X!Tandem search engine as part of its installation. During loading, you can choose to have Scaffold run X!Tandem and use it as a complementary search engine.

If you plan to run X!Tandem, be sure to apply the same FASTA database that was used in the original search.

Queue spots 06-10 into a single BioSample without choosing MuDPIT. When you reach the Load and Analyze Data page, select the database swissprot_bovine FASTA, **check the box to "Analyze with X!Tandem" and then click "Next".**



With X!Tandem selected, the Wizard brings up an additional dialog to allow you to set the parameters for the X!Tandem run.



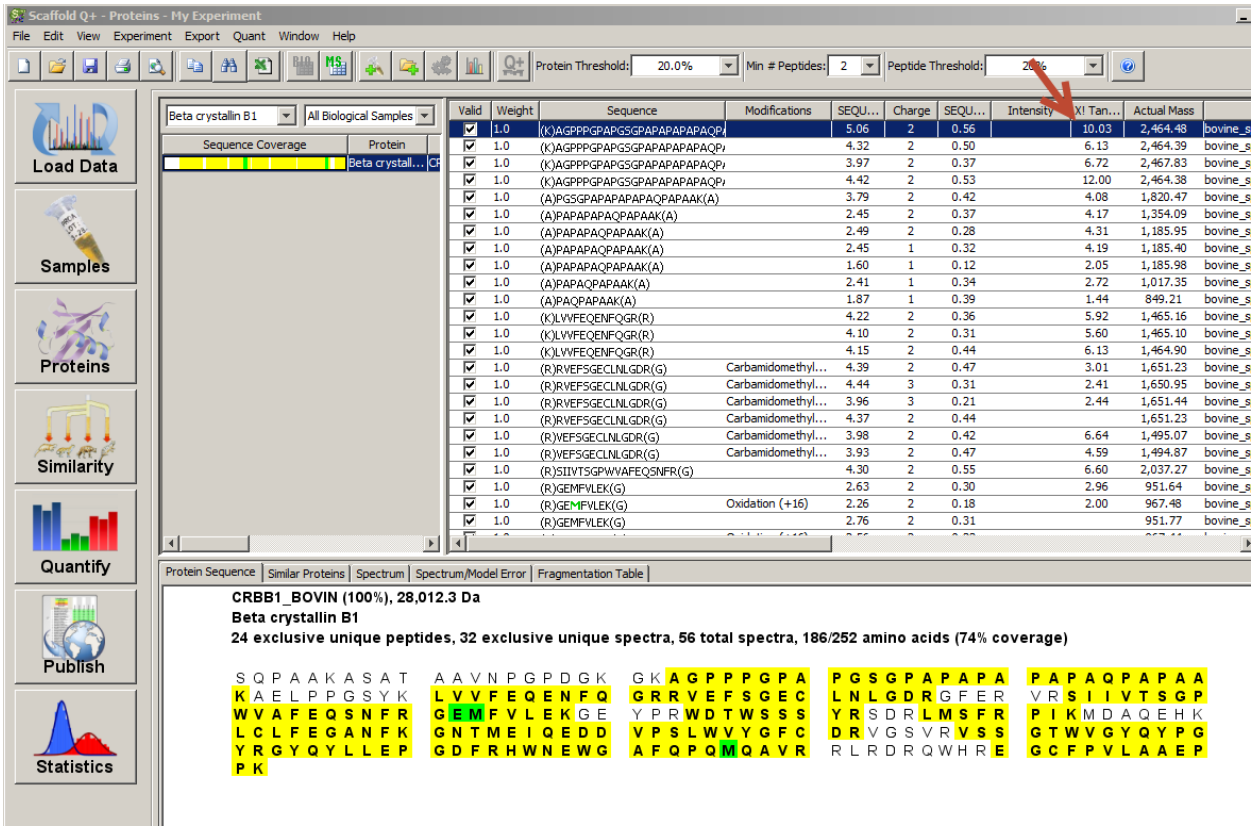
All of the modifications from the original search are automatically selected, and you can add other modifications.

If the “Search subset database” box is checked, Scaffold constructs a new database consisting only of the proteins identified in the original search. If the subset database is too small to give reasonable results, or if a decoy search was performed but the subset does not contain enough decoys for the X!Tandem search, Scaffold adds some additional proteins drawn from the FASTA database.

Select “Search Subset Database” then click “Load Data” to begin the X!Tandem search and then load the results.

When the data has loaded, click the “Proteins” button and look for the X!Tandem scores in the table at the upper left of the Proteins View.

Exercise 4 Result:



Exercise 5: Selecting a Scoring System

Scaffold 4 offers three alternative scoring systems. Previous versions of Scaffold depended on the PeptideProphet algorithm to assess the probabilities of peptide assignments. In Scaffold 3, an option to adjust the probabilities based on mass accuracy was added. Both the original PeptideProphet and the High Mass Accuracy-adjusted Prophet model are still offered in Scaffold 4, but recent technological developments and the increasing reliance on False Discovery Rate as a measure of reliability led us to develop an entirely new scoring algorithm.

The new method is known as LFDR, and is the default scoring system in Scaffold 4. It uses a Bayesian approach to estimate Local False Discovery Rate. Like other scoring methods, LFDR incorporates multiple scores when they are reported by a search engine. The algorithm builds a training set of "false" identifications drawn from the decoy hits and "true" identifications drawn from the highest scoring target hits. It constructs a discriminant score by using a naïve Bayes analysis to decide how much each score or metric reported by the search engine should count in the calculation in order to achieve the best fit in the training set. It then calculates peptide probabilities for the full set of peptides based on the resulting classifier.

Choose LFDR Scoring when:

- A decoy search has been performed
- There is no need for consistency with a previous Scaffold analysis
- You are analyzing QExactive

If you choose LFDR and there are no or very few decoy matches in your data, Scaffold will automatically switch to the PeptideProphet with HMA method.

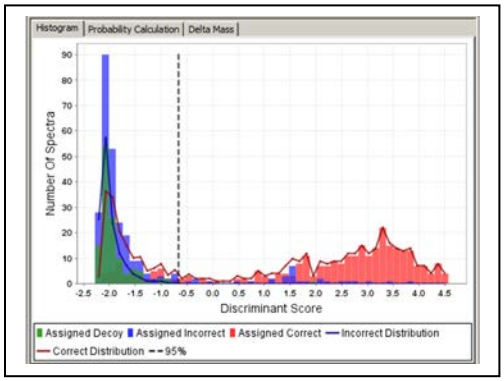
Choose the original PeptideProphet model when:

- You want compatibility with an earlier Scaffold analysis that used this model
- A decoy search was not performed and the spectra were not captured by a high mass accuracy device

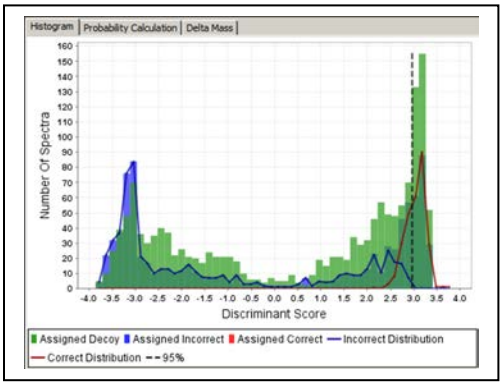
Choose PeptideProphet with High Mass Accuracy when:

- You want compatibility with an earlier Scaffold analysis that used this model
- A decoy search was not performed and the spectra were captured by a high mass accuracy device

It is always a good idea to check the scoring graphs in the Statistics View, however, to be sure that the selected scoring method has performed well with your particular data set. To judge whether LFDR has worked well, check whether there is separation of the correct distribution from the incorrect and decoy distributions, and whether there is a good fit in the Probability Calculation graph.

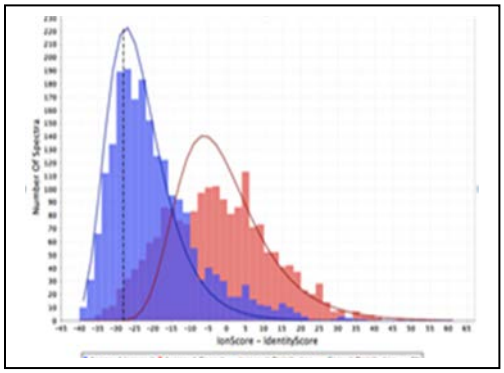


Good LFDR Fit

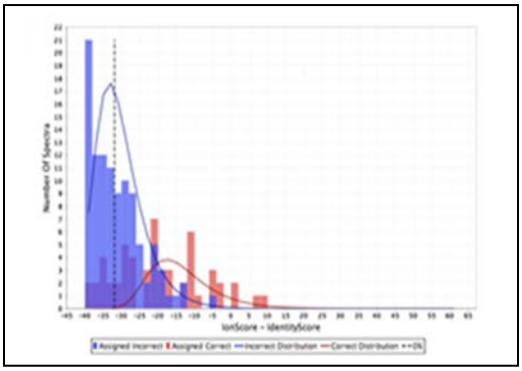


Bad LFDR Fit

If PeptideProphet has been effective on your data set, you should see two fairly well-resolved curves. You should always check this graph after loading your data to see whether the scoring model you selected has worked well with your data.



Good PeptideProphet Fit



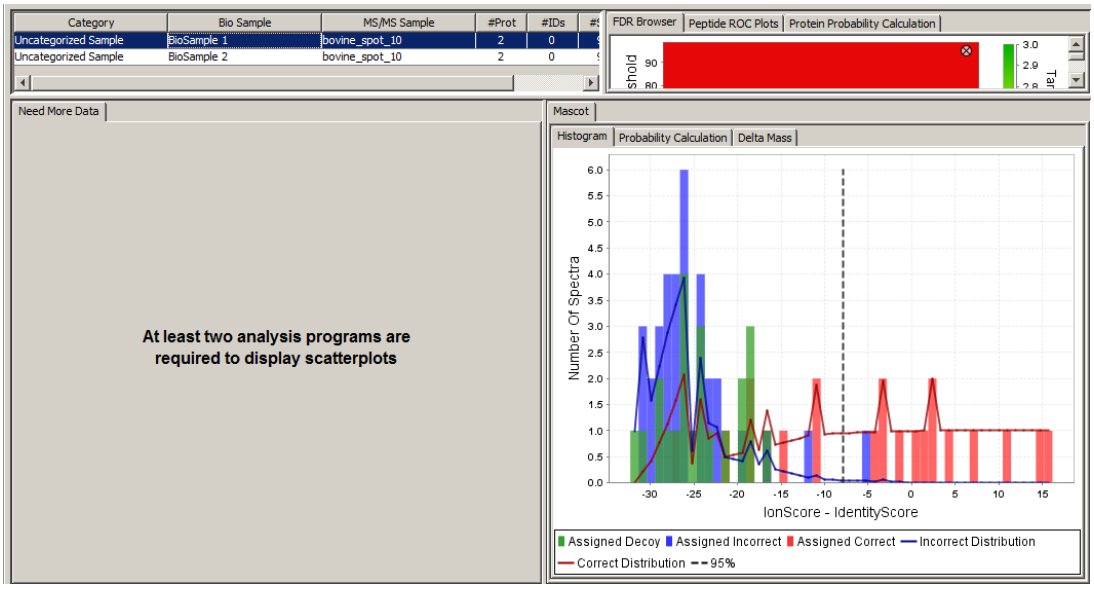
Bad PeptideProphet Fit

Load files bovine_spot_10 from tutorial_3mas into BioSample 1, then add another BioSample and load bovine_spot_10 from tutorial_3seq. Select LFDR as the scoring option. Accept the remaining default options, and click Load Data.

a) **Go to the Statistics View, and click on the line for BioSample 1 in the table at the upper left. Note the histograms in the lower right quadrant. Click through them.**
b) **Next select BioSample 2 in the upper left table. Compare the histograms at lower right.**

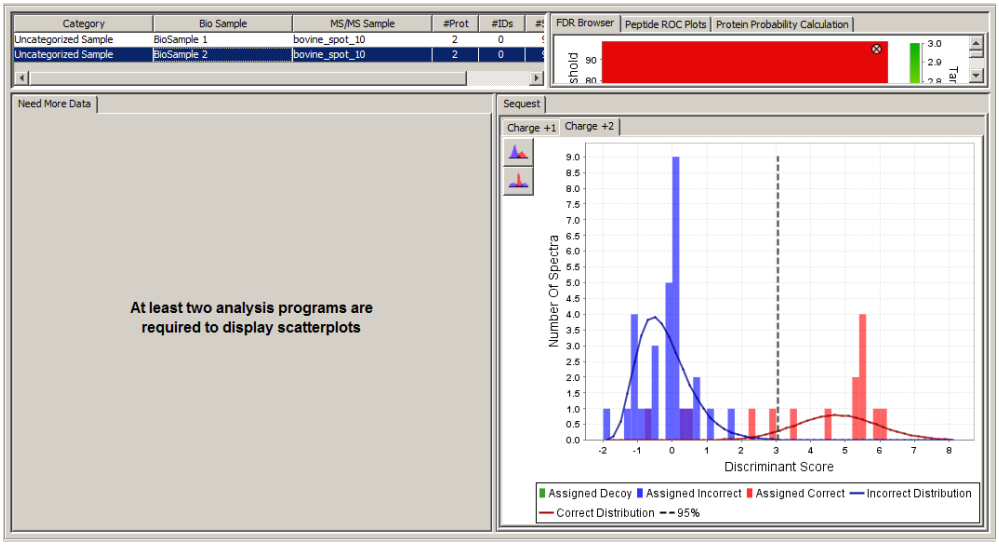
Exercise 5 Results:

a) BioSample 1



The Mascot searches were run with the decoy search option selected, so Scaffold was able to use the LFDR scoring method, and the graphs you see in BioSample 1 illustrate the workings of LFDR on this data.

b) Biosample 2



No decoys were used in the Sequest searches, so Scaffold was unable to run LFDR on this data, and defaulted to PeptideProphet with high mass accuracy instead.

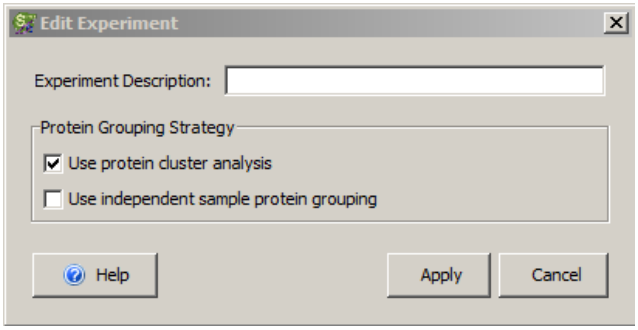
Exercise 6: Protein Grouping and Clustering

Scaffold groups proteins that share all of the same peptides and displays them as if they were a single protein, although it allows you to pick which protein name you would like to display. Grouped proteins appear as a single line in the Samples View, but the accession number is followed by an indication of the number of additional proteins in the group:

In Scaffold 4, we have also introduced the concept of protein clustering. If you choose to apply clustering, Scaffold groups together proteins that share some but not all of their peptides. This allows you to treat closely related proteins as a group. In the Samples View, you can choose whether to expand the cluster and see all of its proteins or collapse it into a single line.

When clustering is selected, Scaffold also changes the way it handles peptides that are shared between proteins. In the traditional non-clustered model, each peptide is assigned to a single protein. This often results in the loss of isoforms or closely related proteins that are actually present in the sample, so in this model Scaffold considers the peptide to be present in both proteins and applies a weighting function when computing the protein probabilities.

You can select the grouping/clustering model as you are loading your files, or you can change it at any time through the Experiment menu. Select Edit Experiment and you will see two check boxes under Protein Grouping Strategy.

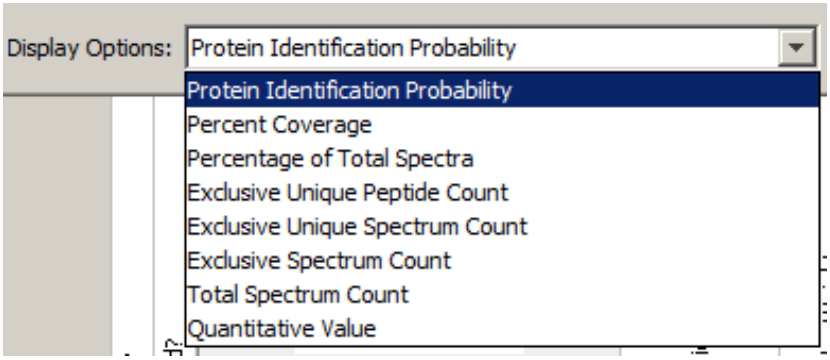


The first determines whether proteins will be grouped according to the traditional Scaffold method, or, if the box is check will be grouped using the new shared peptide grouping method and clustered as well.

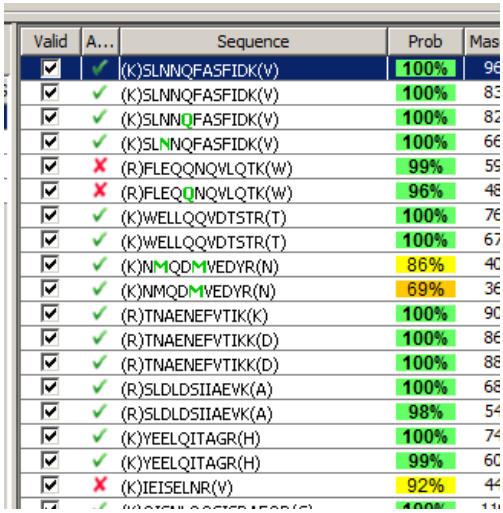
The second option tells Scaffold not to group proteins across MS Samples, but we do not recommend using this option with protein clustering.

Open tutorial_6 by selecting Help>Open Demo Files and selecting tutorial_6.sf3. Click "Open".

This file was loaded with the traditional grouping method. Expand the dropdown for the display options. You will see:



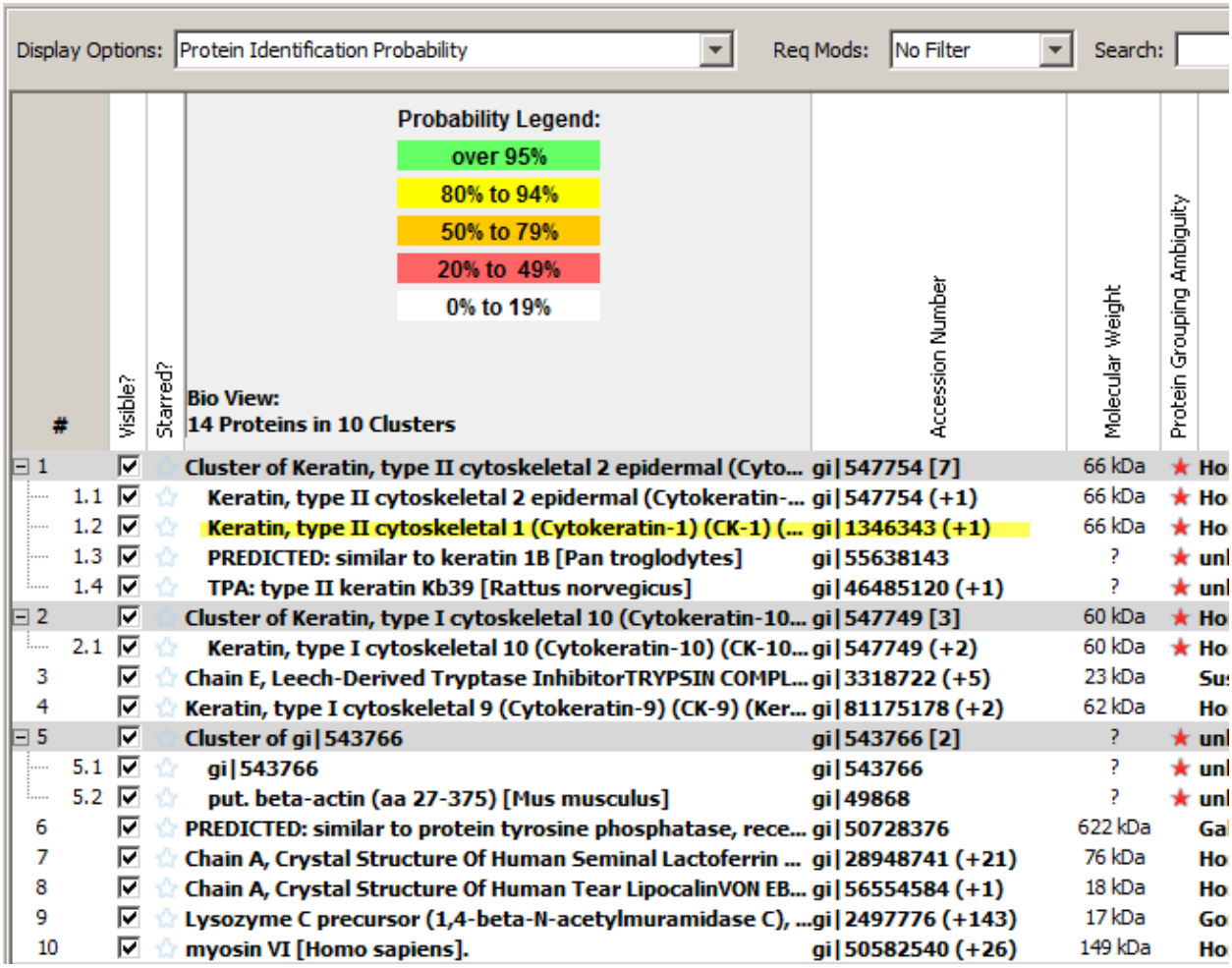
Lower the Peptide Threshold to 50% and select the second protein, gi|1346343, then go to the Proteins View. The peptides table looks like this:



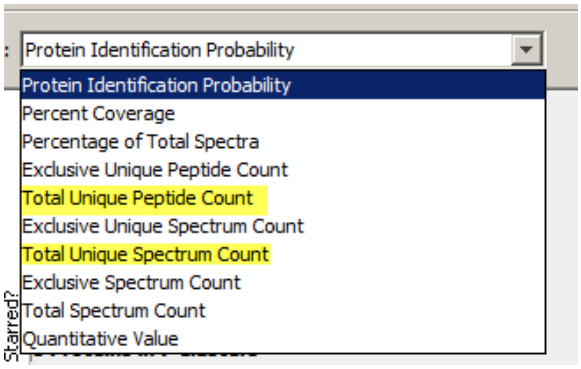
Each peptide has either a green check to indicate it has been assigned to this protein or a red X to indicate that it was identified with this protein but was assigned to another.

Now select Experiment>Edit Experiment, check the “Use protein cluster analysis” box and click “Apply”.

Notice that protein gi|1346343 is now part of a cluster.

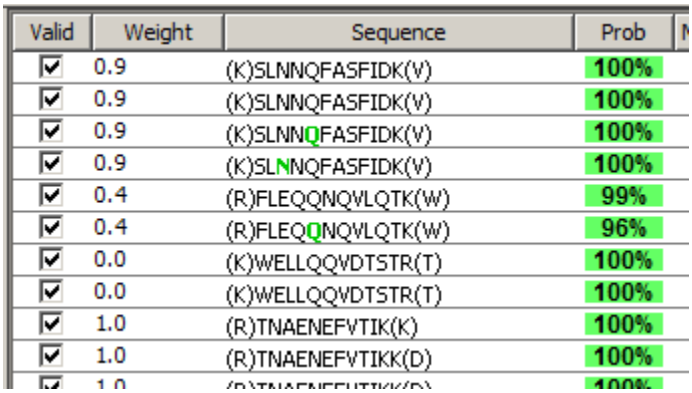


Expand the dropdown for the Display Options:



Notice that two additional counts are available. These total counts include peptides or spectra shared between proteins. Why?

Choose one of the proteins in the cluster and go to the Proteins View. In place of the "Assigned" column, you will now see a "Weight" column showing the apportionment of the peptide among the proteins in which it appears.



Valid	Weight	Sequence	Prob
☑	0.9	(K)SLNNQFASFIDK(V)	100%
☑	0.9	(K)SLNNQFASFIDK(V)	100%
☑	0.9	(K)SLNNQFASFIDK(V)	100%
☑	0.9	(K)SLNNQFASFIDK(V)	100%
☑	0.4	(R)FLEQQNQVLQTK(W)	99%
☑	0.4	(R)FLEQQNQVLQTK(W)	96%
☑	0.0	(K)WELLQQVDTSTR(T)	100%
☑	0.0	(K)WELLQQVDTSTR(T)	100%
☑	1.0	(R)TNAENEFVTIK(K)	100%
☑	1.0	(R)TNAENEFVTIKK(D)	100%

You should also look at the differences in the Similarity View.

Return to the Samples View and note the number of proteins and clusters at the top of the protein name column and in the FDR dashboard. In the View menu, uncheck Show Entire Protein Clusters and note the how the counts change.

Exercise 7: Applying a FASTA Database

Scaffold requires you to specify a FASTA database when loading data. The database supplies the protein sequences and molecular weights; all other data is read from the search engine results. It is important to specify the correct database if you wish to run X!Tandem. Otherwise, if you specify the wrong database, or if you have searched against more than one database, you can apply the correct database(s) after loading.

If you look at the Samples View and see non-decoy proteins with "?" for the molecular weight, it usually means that either you have specified the wrong database during loading or that you need to parse the database differently in order for Scaffold to be able to successfully look up the proteins in the FASTA file.

Before you can use a FASTA database in Scaffold, you must add it to the list of parsed databases. This can be done by choosing "Add New Database" during loading, or at any time by selecting Edit>Edit Fasta Databases>Add Database. When adding a new database, Scaffold gives you the choice to "Auto Parse" or "Use Regular Expressions". In most cases, you will want to choose "Auto Parse". If you have applied the correct database and still find "?" in the molecular weight column for non-decoy proteins, it may mean that you need to "Use Regular Expressions". In that case, our tech support staff will be happy to help you.

In this exercise, we will load some data with the wrong database, and then apply the correct database in order to fill in the molecular weights and protein sequences.

Start a new experiment and queue files bovine_spot_06-bovine_spot10 from the tutorial_3seq folder. In the "Load and Analyze Data" dialog, click "Add New Database", then click "Add Database…" in the window that appears. Browse to the Q+S folder, and select SILAC-demo.fasta. Click "Open" and then choose "Auto Parse". The database will be indexed. Select the new database in the list and click "OK".

Do not check the box labeled Use non-default forward/decoy ratio. This is only to be used in very rare circumstances in which the data has been searched against a database with a very different ratio of proteins to decoys than the one being used for loading. If your search engine has done an automatic decoy search and you are loading the non-decoy database you searched, Scaffold will assume a 1:1 ratio, so you do not need to check this box.

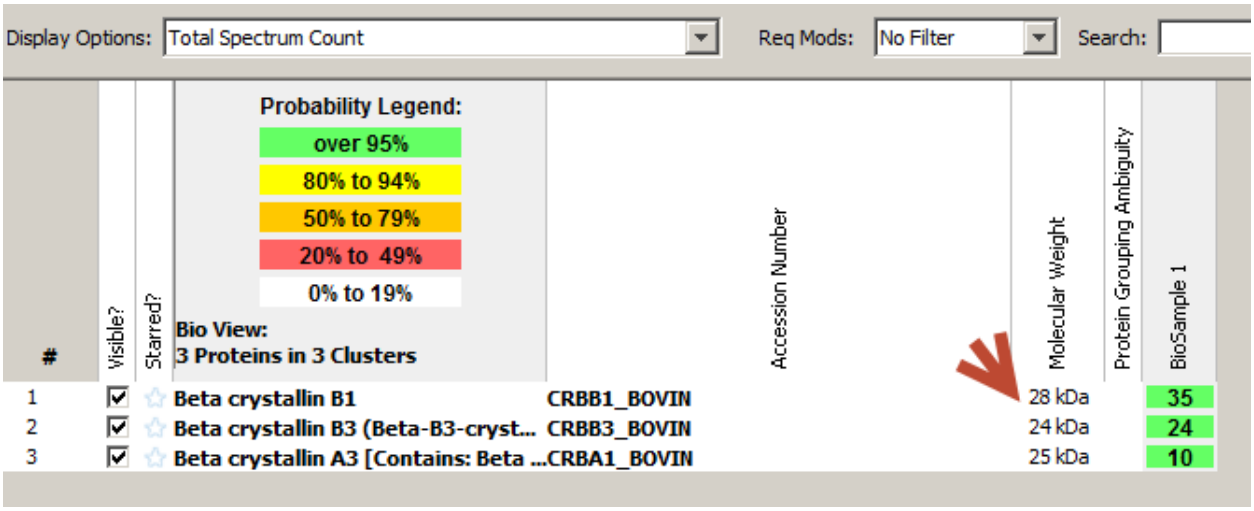
Now go ahead and load the data.

You should see that the proteins are missing their molecular weights, and if you select a protein and go to the Proteins View, you will see that they are also missing their sequences.

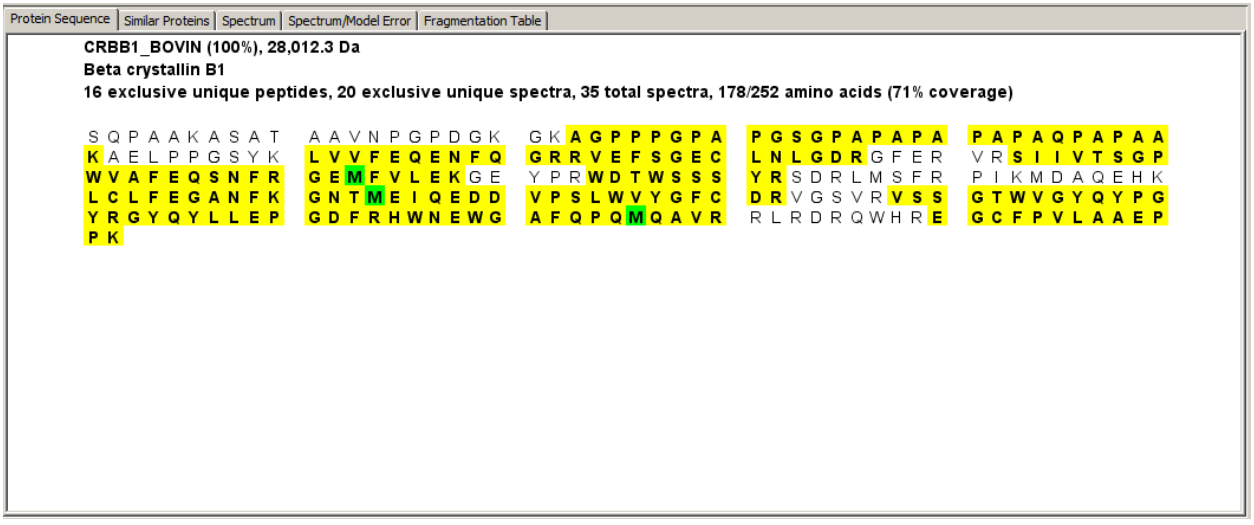
To apply the correct database, go to the Experiment menu and select "Apply New Database". Locate the database called "swissprot_bovine FASTA Database". If you do not find it, you will need to add it. Select the database and click "Apply".

Exercise 7 Result:

Samples View



Proteins View



Release Information

This document is applicable for Scaffold, Release 4.0 or greater, and is current until replaced.

Customer Support - Customer support is available to organizations that purchase *Scaffold, Scaffold Q+* or Scaffold Q+S 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

Document Version Number Scaffold 4.0-Tutorial-001
Document Release Date June 5, 2013