



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:

Display Options: Total Spectrum Count Req Mods: No Filter Search:

Probability Legend:
over 95%
80% to 94%
50% to 79%
20% to 49%
0% to 19%

MS/MS View:
8 Proteins in 7 Clusters

#	Visible?	Starred?	Accession Number	Molecular Weight	Protein Grouping Ambiguity	Control H				Disease D1				Disease D2			
						bovine_spot_06	bovine_spot_07	bovine_spot_08	bovine_spot_09	bovine_spot_10	bovine_spot_11	bovine_spot_12	bovine_spot_13	bovine_spot_14	bovine_spot_15	bovine_spot_16	bovine_spot_17
1	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Beta crystallin B3 (Beta-B3-crystallin) CRBB3_BOVIN	24 kDa		24				18	36	6	1	14	27		4
2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Cluster of Beta crystallin A3 variant N5... CRBA_BOVIN_2 [2]	25 kDa	★				10	36	7	1					5
2.1	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Beta crystallin A3 variant N595 from ... CRBA_BOVIN_2	25 kDa	★				10	36	7	1					5
2.2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Beta crystallin A3 [Contains: Beta cr... CRBA1_BOVIN	25 kDa	★				10	36	7						5
3	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Beta crystallin B2 (BP) CRBB2_BOVIN	23 kDa						4	1	20	26	1			
4	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Beta crystallin B1 CRBB1_BOVIN	28 kDa		24	11						0				
5	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Beta crystallin A4 (Beta-A4-crystallin) CRBA4_BOVIN (+1)	24 kDa													20
6	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Beta crystallin A2 (Beta-A2-crystallin) CRBA2_BOVIN	22 kDa									2			23	
7	<input checked="" type="checkbox"/>	<input type="checkbox"/>	keratin, 67K type II cytoskeletal - hum...CONT gi 88054 pir...	65 kDa		1				1							2

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

Display Options: Total Spectrum Count Req Mods: No Filter Search:

Probability Legend:
over 95%
80% to 94%
50% to 79%
20% to 49%
0% to 19%

Bio View:
8 Proteins in 7 Clusters

#	Visible?	Starred?	Accession Number	Molecular Weight	Protein Grouping Ambiguity	Contr... Disease		
						H	D1	D2
1	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Beta crystallin B3 (Beta-B3-crystallin) CRBB3_BOVIN	24 kDa		24	61	45
2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Cluster of Beta crystallin A3 variant N5... CRBA_BOVIN_2 [2]	25 kDa	★	10	44	5
2.1	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Beta crystallin A3 variant N595 from ... CRBA_BOVIN_2	25 kDa	★	10	44	5
2.2	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Beta crystallin A3 [Contains: Beta cr... CRBA1_BOVIN	25 kDa	★	10	44	5
3	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Beta crystallin B2 (BP) CRBB2_BOVIN	23 kDa			51	1
4	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Beta crystallin B1 CRBB1_BOVIN	28 kDa		35	0	
5	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Beta crystallin A4 (Beta-A4-crystallin) CRBA4_BOVIN (+1)	24 kDa				37
6	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Beta crystallin A2 (Beta-A2-crystallin) CRBA2_BOVIN	22 kDa				25
7	<input checked="" type="checkbox"/>	<input type="checkbox"/>	keratin, 67K type II cytoskeletal - hum...CONT gi 88054 pir...	65 kDa		1	1	2

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:

Experiment Export Quant Window Help

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

Display Options: Total Spectrum Count Req Mods: No Filter Search:

Probability Legend:

- over 95%
- 80% to 94%
- 50% to 79%
- 20% to 49%
- 0% to 19%

MS/MS View:
3 Proteins in 3 Clusters

#	Visible?	Starred?	Protein Name	Accession Number	Molecular Weight	Protein Grouping Ambiguity					M
						bovine_spot_06	bovine_spot_07	bovine_spot_08	bovine_spot_09	bovine_spot_10	Mudpit_bovine_spot_06
1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Beta crystallin B1	CRBB1_BOVIN	28 kDa		24	11			37
2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Beta crystallin B3 (Beta-B3-crystallin)	CRBB3_BOVIN	24 kDa	24					21
3	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Beta crystallin A3 [Contains: Beta crys...	CRBA1_BOVIN	25 kDa				10		10

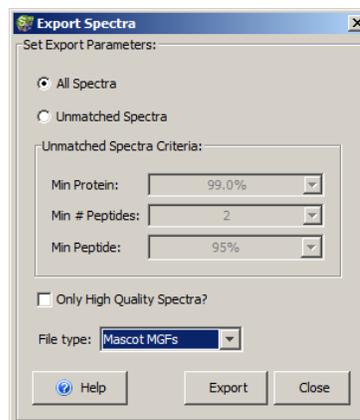
Exercise 3: Combining Search Engines

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:

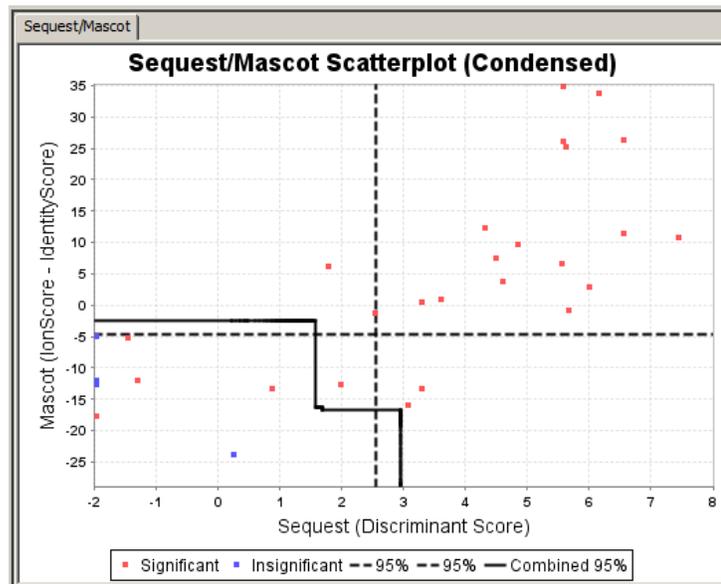
In Load Data, you should see the loaded samples aligned as shown:

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”.

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 X!Tandem

Load and Analyze Data

Searched Database:
swissprot_bovine FASTA Database
 Use non-default forward/decoy ratio: No Decoys
Add New Database

X! Tandem:
 Analyze with X! Tandem

Scoring System:
 Use LFDR scoring (all instruments)
 Use legacy PeptideProphet scoring (high mass accuracy)
 Use legacy PeptideProphet scoring (standard)

Protein Grouping:
 Use protein cluster analysis
 Use independent sample protein grouping

Protein Annotations:
 Don't annotate (No download required)
 Use local GO annotations (UniProt, IPI; ~2 mins) [Configure GO Source](#)

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

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 X!Tandem

Load and Analyze Data

X! Tandem Options:
 Search subset database

Variable Modifications:

Modification	Mass	AA
Met-Hist	-48.00	C
Met-Hist	-29.99	C
Dehydrated	-18.01	N
Glu-Spino...	-18.01	N
Asparagin...	-17.03	N
Glu-Spino...	-17.03	N
Amidated	-0.98	C
Dehydro	-1.01	C
Deamidated	+0.98	H
Deamidated	+0.98	Q
Label:18O(1)	+2.00	C
Label:18O(2)	+4.01	C
Methyl	+14.02	D
Methyl	+14.02	E
Methyl	+14.02	C
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	C
Formyl	+27.99	N
Dioxidation	+31.99	H
Dioxidation	+31.99	M
Dioxidation	+31.99	W

Selected Variable Mods:

Modification	Mass	AA
Oxidation	+15.99	M
Oxidation	+15.99	W
Acetyl	+42.01	N
Phospho	+79.97	S
Phospho	+79.97	T

Add
New
Remove

Help Previous Load Data Done Cancel

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:

The screenshot shows the Scaffold Q+ Proteins software interface. The top menu includes File, Edit, View, Experiment, Export, Quant, Window, and Help. The main window displays a list of peptides with columns for Valid, Weight, Sequence, Modifications, SEQU..., Charge, SEQU..., Intensity, XI Tan..., and Actual Mass. A red arrow points to the Intensity column. Below the list, the protein sequence for CRBB1_BOVIN (100%), 28,012.3 Da is shown, along with a summary of 24 exclusive unique peptides, 32 exclusive unique spectra, 56 total spectra, and 186/252 amino acids (74% coverage). The protein sequence is displayed in a grid format with highlighted residues.

Valid	Weight	Sequence	Modifications	SEQU...	Charge	SEQU...	Intensity	XI Tan...	Actual Mass	
✓	1.0	(K)AGPPPAPGSGPAPAPAPAPAPAQ		5.06	2	0.56	10.03	2,464.48	bovine_s	
✓	1.0	(K)AGPPPAPGSGPAPAPAPAPAPAQ		4.32	2	0.50	6.13	2,464.39	bovine_s	
✓	1.0	(K)AGPPPAPGSGPAPAPAPAPAPAQ		3.97	2	0.37	6.72	2,467.83	bovine_s	
✓	1.0	(K)AGPPPAPGSGPAPAPAPAPAPAQ		4.42	2	0.53	12.00	2,464.38	bovine_s	
✓	1.0	(A)PGSGPAPAPAPAPAPAPAAK(A)		3.79	2	0.42	4.08	1,820.47	bovine_s	
✓	1.0	(A)PAPAPAPAPAPAPAAK(A)		2.45	2	0.37	4.17	1,354.09	bovine_s	
✓	1.0	(A)PAPAPAPAPAPAPAAK(A)		2.49	2	0.28	4.31	1,185.95	bovine_s	
✓	1.0	(A)PAPAPAPAPAPAPAAK(A)		2.45	1	0.32	4.19	1,185.40	bovine_s	
✓	1.0	(A)PAPAPAPAPAPAPAAK(A)		1.60	1	0.12	2.05	1,185.98	bovine_s	
✓	1.0	(A)PAPAPAPAPAPAAK(A)		2.41	1	0.34	2.72	1,017.35	bovine_s	
✓	1.0	(A)PAPAPAPAPAAK(A)		1.87	1	0.39	1.44	849.21	bovine_s	
✓	1.0	(K)LVVFEQENFQGR(R)		4.22	2	0.36	5.92	1,465.16	bovine_s	
✓	1.0	(K)LVVFEQENFQGR(R)		4.10	2	0.31	5.60	1,465.10	bovine_s	
✓	1.0	(K)LVVFEQENFQGR(R)		4.15	2	0.44	6.13	1,464.90	bovine_s	
✓	1.0	(R)RVVFEQENFQGR(G)	Carbamidomethyl...	4.39	2	0.47	3.01	1,651.23	bovine_s	
✓	1.0	(R)RVVFEQENFQGR(G)	Carbamidomethyl...	4.44	3	0.31	2.41	1,650.95	bovine_s	
✓	1.0	(R)RVVFEQENFQGR(G)	Carbamidomethyl...	3.96	3	0.21	2.44	1,651.44	bovine_s	
✓	1.0	(R)RVVFEQENFQGR(G)	Carbamidomethyl...	4.37	2	0.44	6.64	1,651.23	bovine_s	
✓	1.0	(R)RVVFEQENFQGR(G)	Carbamidomethyl...	3.98	2	0.42	6.64	1,495.07	bovine_s	
✓	1.0	(R)RVVFEQENFQGR(G)	Carbamidomethyl...	3.93	2	0.47	4.59	1,494.87	bovine_s	
✓	1.0	(R)RVVFEQENFQGR(G)	Carbamidomethyl...	4.30	2	0.55	6.60	2,037.27	bovine_s	
✓	1.0	(R)GEMFVLEK(G)		2.63	2	0.30	2.96	951.64	bovine_s	
✓	1.0	(R)GEMFVLEK(G)	Oxidation (+16)	2.26	2	0.18	2.00	967.48	bovine_s	
✓	1.0	(R)GEMFVLEK(G)		2.76	2	0.31	951.77	bovine_s		

CRBB1_BOVIN (100%), 28,012.3 Da
Beta crystallin B1
24 exclusive unique peptides, 32 exclusive unique spectra, 56 total spectra, 186/252 amino acids (74% coverage)

```

S Q P A A K A S A T   A A V N P G P D G K   G K A G P P P G P A   P G S G P A P A P A   P A P A Q P A P A A
K A E L P P G S Y K   L V V F E Q E N F Q   G R R V E F S G E C   L N L G D R G F E R   V R S I I V T S G P
W V A F E Q S N F R   G E M F V L E K G E   Y P R W D T W S S S   Y R S D R L M S F R   P I K M D A Q E H K
L C L F E G A N F K   G N T M E I Q E D D   V P S L W V Y G F C   D R V G S V R V S S   G T W V G Y Q Y P G
Y R G Y Q Y L L E P   G D F R H W N E W G   A F Q P Q M Q A V R   R L R D R Q W H R E   G C F P V L A A E P
P K
    
```

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.

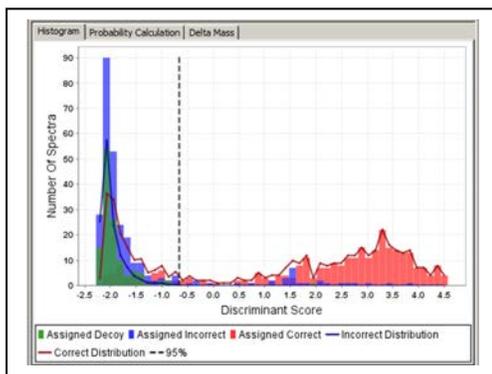
Choose the original PeptideProphet model when:

- 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

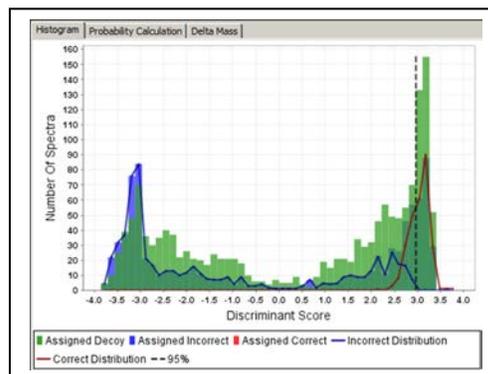
Choose PeptideProphet with High Mass Accuracy when:

- 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.

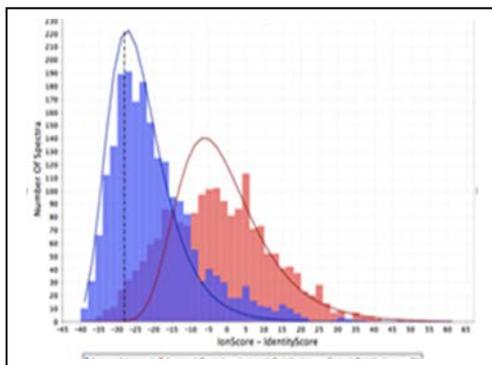


Good LFDR Fit

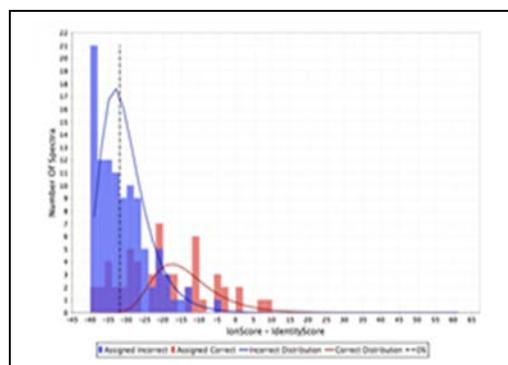


Bad LFDR Fit

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

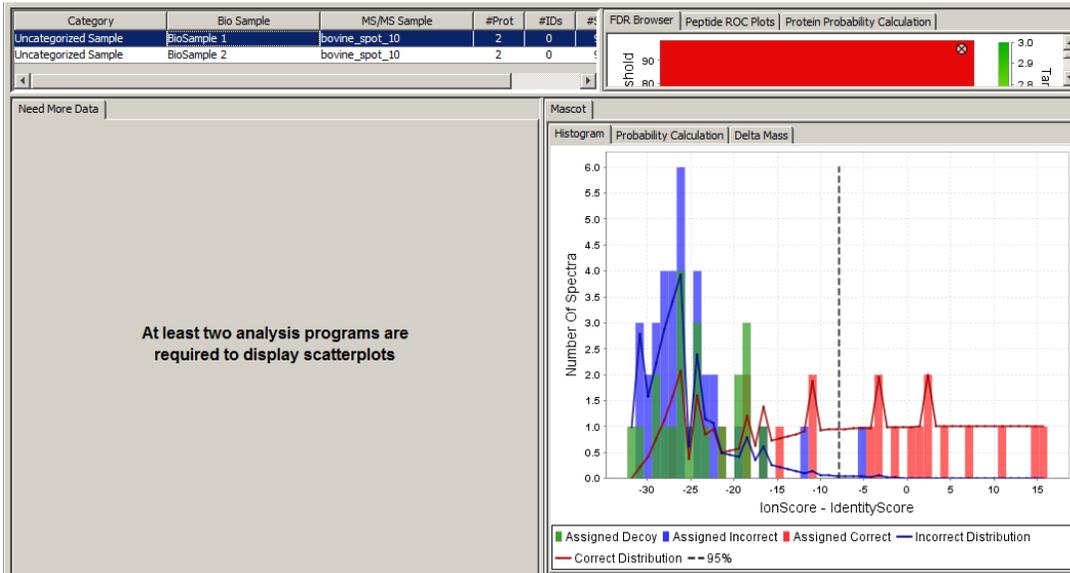


Load files bovine_spot_10 from tutorial_3mas into BioSample 1, then add another BioSample and load bovine_spot_10 from tutorial_3seq. Select LFDR as the scoring option. Accept the remaining default options, and click Load Data.

- 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.**

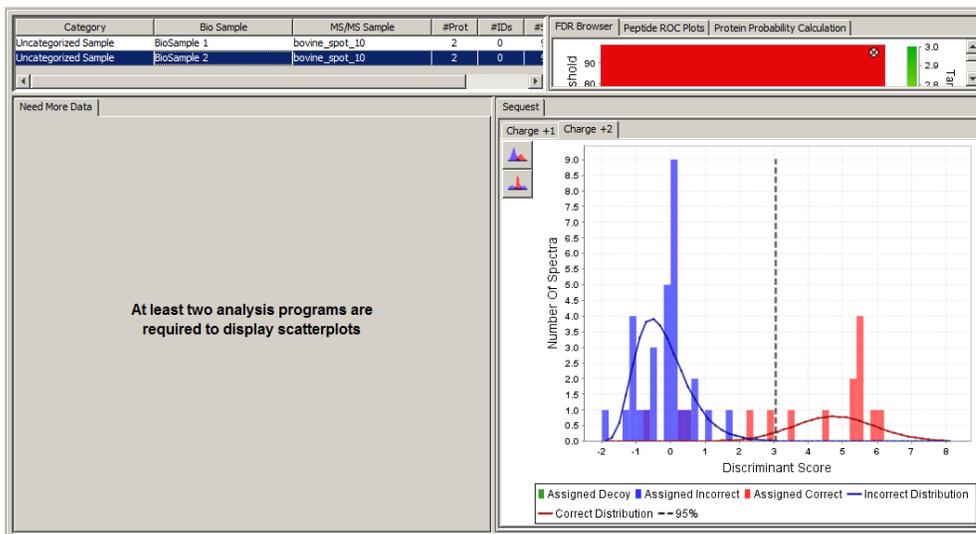
Exercise 5 Results:

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.

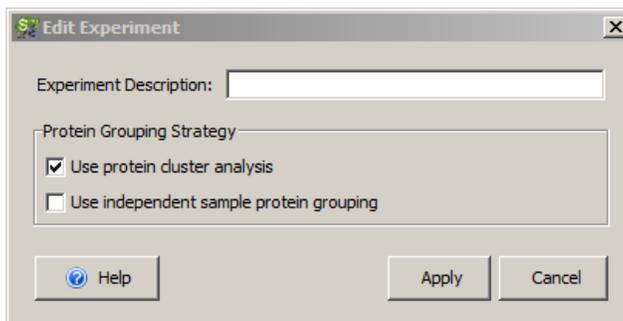
Grouped proteins appear as a single line in the Samples View, but the accession number is followed by an indication of the number of additional proteins in the group:

1 ☆ Keratin, type I cytoskeletal 10 (Cytokeratin-10) (CK-1... gi|547749 (+2)

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.



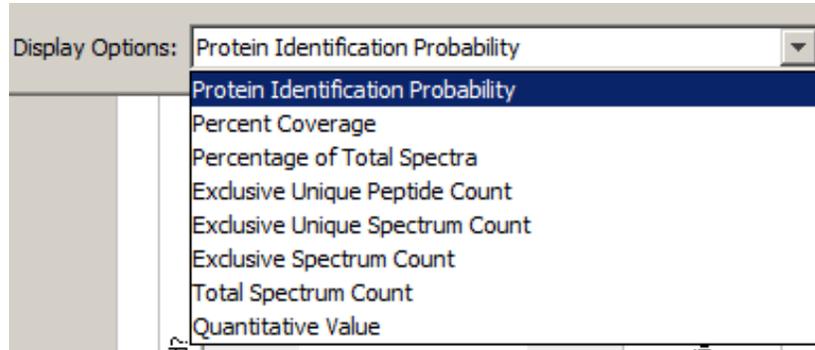
The first determines whether proteins will be grouped according to the traditional Scaffold method, or, if the box is checked will be grouped using the new shared peptide grouping method and clustered as well.

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:

Valid	A...	Sequence	Prob	Mas
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)SLNNQFASFDK(V)	100%	96
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)SLNNQFASFDK(V)	100%	83
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)SLNNQFASFDK(V)	100%	82
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)SLNNQFASFDK(V)	100%	66
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(R)FLEQQNQVLQTK(W)	99%	55
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(R)FLEQQNQVLQTK(W)	96%	48
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)WELLQQVDTSTR(T)	100%	76
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)WELLQQVDTSTR(T)	100%	67
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)NMQDMVEDYR(N)	86%	40
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)NMQDMVEDYR(N)	69%	36
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(R)TNAENEFVTIK(K)	100%	90
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(R)TNAENEFVTIK(D)	100%	86
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(R)TNAENEFVTIK(D)	100%	88
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(R)SLDLDSIIAEVK(A)	100%	68
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(R)SLDLDSIIAEVK(A)	98%	54
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)YEELQITAGR(H)	100%	74
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)YEELQITAGR(H)	99%	60
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)IEISELNR(V)	92%	44
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)IEISELNR(V)	100%	11

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.

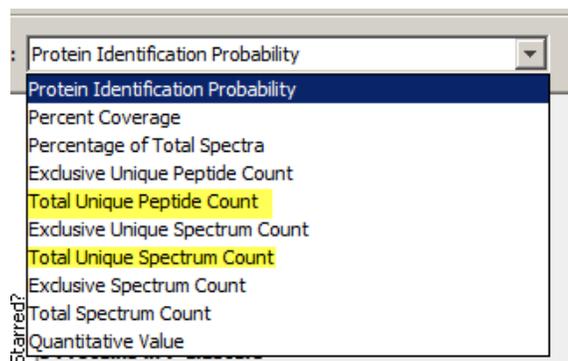
Now select Experiment>Edit Experiment, check the “Use protein cluster analysis” box and click “Apply”.

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

Display Options: Protein Identification Probability Req Mods: No Filter Search:

#	Visible?	Starred?	Accession Number	Molecular Weight	Protein Grouping Ambiguity
Probability Legend: over 95% 80% to 94% 50% to 79% 20% to 49% 0% to 19%					
Bio View: 14 Proteins in 10 Clusters					
1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Cluster of Keratin, type II cytoskeletal 2 epidermal (Cyto... gi 547754 [7]	66 kDa	★ Ho
1.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Keratin, type II cytoskeletal 2 epidermal (Cytokeratin-... gi 547754 (+1)	66 kDa	★ Ho
1.2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Keratin, type II cytoskeletal 1 (Cytokeratin-1) (CK-1) (... gi 1346343 (+1)	66 kDa	★ Ho
1.3	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	PREDICTED: similar to keratin 1B [Pan troglodytes] gi 55638143	?	★ unl
1.4	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	TPA: type II keratin Kb39 [Rattus norvegicus] gi 46485120 (+1)	?	★ unl
2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Cluster of Keratin, type I cytoskeletal 10 (Cytokeratin-10... gi 547749 [3]	60 kDa	★ Ho
2.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Keratin, type I cytoskeletal 10 (Cytokeratin-10) (CK-10... gi 547749 (+2)	60 kDa	★ Ho
3	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Chain E, Leech-Derived Trypsin InhibitorTRYPSIN COMPL... gi 3318722 (+5)	23 kDa	Su
4	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Keratin, type I cytoskeletal 9 (Cytokeratin-9) (CK-9) (Ker... gi 81175178 (+2)	62 kDa	Ho
5	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Cluster of gi 543766 gi 543766 [2]	?	★ unl
5.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	gi 543766 gi 543766	?	★ unl
5.2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	put. beta-actin (aa 27-375) [Mus musculus] gi 49868	?	★ unl
6	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	PREDICTED: similar to protein tyrosine phosphatase, rece... gi 50728376	622 kDa	Ga
7	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Chain A, Crystal Structure Of Human Seminal Lactoferrin ... gi 28948741 (+21)	76 kDa	Ho
8	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Chain A, Crystal Structure Of Human Tear LipocalinVON EB... gi 56554584 (+1)	18 kDa	Ho
9	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Lysozyme C precursor (1,4-beta-N-acetylmuramidase C), ...gi 2497776 (+143)	17 kDa	Go
10	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	myosin VI [Homo sapiens]. gi 50582540 (+26)	149 kDa	Ho

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
<input checked="" type="checkbox"/>	0.9	(K)SLNNQFASFIDK(V)	100%
<input checked="" type="checkbox"/>	0.9	(K)SLNNQFASFIDK(V)	100%
<input checked="" type="checkbox"/>	0.9	(K)SLNNQFASFIDK(V)	100%
<input checked="" type="checkbox"/>	0.9	(K)SLNNQFASFIDK(V)	100%
<input checked="" type="checkbox"/>	0.4	(R)FLEQQNQVLQTK(W)	99%
<input checked="" type="checkbox"/>	0.4	(R)FLEQQNQVLQTK(W)	96%
<input checked="" type="checkbox"/>	0.0	(K)WELLQQVDTSTR(T)	100%
<input checked="" type="checkbox"/>	0.0	(K)WELLQQVDTSTR(T)	100%
<input checked="" type="checkbox"/>	1.0	(R)TNAENEFVTIK(K)	100%
<input checked="" type="checkbox"/>	1.0	(R)TNAENEFVIKK(D)	100%
<input checked="" type="checkbox"/>	1.0	(R)TNAENEFVTIK(K)	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

Display Options: Total Spectrum Count Req Mods: No Filter Search:

#	Visible?	Starred?	Accession Number	Molecular Weight	Protein Grouping Ambiguity	BioSample 1
Probability Legend: <div style="display: flex; flex-direction: column; gap: 5px;"> <div style="background-color: #90EE90; padding: 2px;">over 95%</div> <div style="background-color: #FFFF00; padding: 2px;">80% to 94%</div> <div style="background-color: #FFD700; padding: 2px;">50% to 79%</div> <div style="background-color: #FF6347; padding: 2px;">20% to 49%</div> <div style="background-color: #FFFFFF; padding: 2px;">0% to 19%</div> </div>						
Bio View: 3 Proteins in 3 Clusters						
1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Beta crystallin B1 CRBB1_BOVIN	28 kDa		35
2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Beta crystallin B3 (Beta-B3-cryst... CRBB3_BOVIN	24 kDa		24
3	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Beta crystallin A3 [Contains: Beta ...CRBA1_BOVIN	25 kDa		10

Proteins View

Protein Sequence | Similar Proteins | Spectrum | Spectrum/Model Error | Fragmentation Table

CRBB1_BOVIN (100%), 28,012.3 Da
Beta crystallin B1
 16 exclusive unique peptides, 20 exclusive unique spectra, 35 total spectra, 178/252 amino acids (71% coverage)

S Q P A A K A S A T	A A V N P G P D G K	G K A G P P P G P A	P G S G P A P A P A	P A P A Q P A P A A
K A E L P P G S Y K	L V V F E Q E N F Q	G R R V E F S G E C	L N L G D R G F E R	V R S I I V T S G P
W V A F E Q S N F R	G E M F V L E K G E	Y P R W D T W S S S	Y R S D R L M S F R	P I K M D A Q E H K
L C L F E G A N F K	G N T M E I Q E D D	V P S L W V Y G F C	D R V G S V R V S S	G T W V G Y Q Y P G
Y R G Y Q Y L L E P	G D F R H W N E W G	A F Q P Q M Q A V R	R L R D R Q W H R E	G C F P V L A A E P
P K				

Release Information

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

Copyright © 2013. 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 used their 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.

Trademark - The name *Proteome Software*, the Proteome Software logo, *Scaffold*, *Scaffold Q+*, *Scaffold Q+S*, and the Scaffold, Scaffold Q+, and Scaffold Q+S 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+* 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