

PROTEOME QUEST™

Correlogic Systems, Inc.

User's Guide

BY TOM HOLLON

Table of Contents

INTRODUCTION.....	3
ABOUT THE MANUAL.....	4
ABOUT THE INVENTORS.....	5
ABOUT THE AUTHOR.....	7
CHAPTER 1: UNDER THE HOOD	8
PREDICTIVE MODELS OF BIOLOGICAL STATES.....	8
SIDESTEPPING THE COMPUTATION BARRIER.....	11
A KNOWLEDGE DISCOVERY ENGINE EXAMPLE	12
CLUSTER ANALYSIS	14
CLUSTER CREATION STEP BY STEP	16
THE GENETIC ALGORITHM	19
PRESENTING THE MODEL.....	21
TESTING THE MODEL	22
THE SUCCESSFUL MODELER.....	26
CHAPTER 2: GETTING STARTED.....	27
SYSTEM REQUIREMENTS	27
INSTALLING PROTEOME QUEST	28
STARTING PROTEOME QUEST	33
EXITING PROTEOME QUEST	35
CHAPTER 3: HOW TO SEARCH FOR A MODEL.....	36
THE FORMAT OF INPUT DATA.....	36
ORGANIZE INPUT DATA BY BIOLOGICAL STATE	37
A DATA ORGANIZATION EXAMPLE.....	38
CREATING A NEW PROJECT	41
CHAPTER 4: HOW TO RECOGNIZE A GOOD MODEL.....	69
THE MODEL TEST AND VALIDATION RESULT TABLES.....	70
THE MODEL RESULT TABLE.....	79
THE MODEL SUMMARY TABLE.....	83
THE MODEL DESCRIPTION TABLE.....	84
THE MODEL TRAINING SET TABLE	85
OPENING AN EXISTING PROJECT	86

About the Manual

It's hard to resist taking a new program out for a spin before reading the manual, so you'll be glad to know it isn't necessary to read the entire manual before starting. However, it *is* necessary to read parts of it. Otherwise, the computer screens won't make sense.

Recommended Order of Reading. Read Chapters 1, 3 and 4 before spending much time at the computer. Then look at Chapter 8, the step-by-step summary of using Proteome Quest successfully. You won't completely understand Chapter 8 until you read the entire manual, but it will be useful as soon as you've read Chapters 3 and 4. Defer reading the other chapters until they are needed. Here is a summary of what they contain.

Chapter 1: Under the Hood

This overview of the Knowledge Discovery Engine—the Proteome Quest biomarker search algorithm—describes how cluster analysis and genetic algorithms find biomarkers for predictive models of biological states. You'll use Proteome Quest with more confidence if you understand how the Knowledge Discovery Engine works.

Chapter 2: Installing Proteome Quest

For most people a skim is all that's needed because installment is straightforward. But note the system requirements: Proteome Quest is computation intensive and a fairly up-to-date processor is recommended.

Chapter 3: How to Search for a Model

Chapter 3 is half of the heart of the manual—how to organize input data and set up the Knowledge Discovery Engine to search for a model.

Chapter 4: How to Recognize a Good Model

The other half. How to read model tables.

Chapter 5: Visualizing Models

This chapter describes charts that enhance understanding of predictive models and help you communicate your results.

Chapter 6: Advanced Modeling

Shortcut commands **Train** and **Retrain** are useful timesavers once you have experience searching for models. Commands **Validate**, **Recall** and **Update** help decide when one predictive model is better than another.

About the Author

Before Tom Hollon earned a Ph.D. in microbiology, he made use of his M.S. in mathematics as a bioinformatics programmer during the early days of gene sequencing. He was a researcher for 20 years before becoming a science writer and editor. He was the founding editor of *Modern Drug Discovery* magazine and his articles have appeared in *The Scientist*, *Signals*, *Nature Medicine*, *Nature Biotechnology*, and *The Lancet*. He was trained at the National Institutes of Health, the Pasteur Institute, the Fred Hutchinson Cancer Research Center, and the University of Washington. thollon@starpower.net

References

(1) Emanuel F Petricoin III, et al., "Use of proteomic patterns in serum to identify ovarian cancer." *The Lancet* 359: 572-577, 2002.

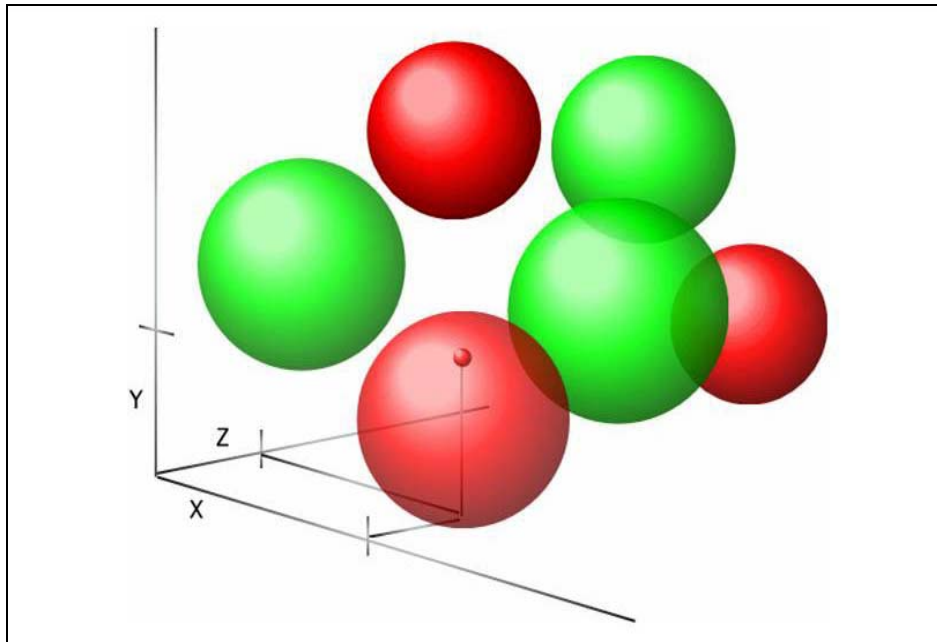


Figure 1-2: Cluster creation: A point representing a protein pattern becomes a new member of a cancer cluster. Cancer clusters are red, benign clusters are green. The point matches the cluster with the nearest centroid, provided that distance \leq decision boundary. This illustration is suggestive of the matching process in higher dimensions.

Distance calculations match all 100 cancer and benign points to clusters. For each point distances are calculated between the point and all existing centroids. The point either matches an existing cluster, or becomes the centroid of a new one-point cluster.

Cluster States and Misclassification Rates. Matching never considers what state a point represents. However, once all points are matched, biological states are considered in two ways. First, each cluster is assigned the biological state of its initial point. A cluster begun with a cancer point is a cancer cluster; a cluster initiated with a point from a benign sample is a benign cluster.

Second, biological states are used to tabulate the matching errors that prevent clusters from being homogeneous. An error is a point that has a different state than its cluster. Cancer points matching benign clusters are errors, and vice versa. The number of errors divided by the number of clusters equals the average number of errors per cluster. This is the misclassification rate.

Example: A model has ten clusters: Five are homogeneous (only cancer points and only benign). Four heterogeneous cancer clusters each have one (benign) error. One heterogeneous benign cluster has three (cancer) errors. Then

How to Search for a Model

Putting the Knowledge Discovery Engine to work

The heart of Proteome Quest is the Knowledge Discovery Engine, the algorithm for finding models and proteomic fingerprints. To use the Knowledge Discovery Engine, you must tell it where to find folders with input data, where to store the output, and set parameters that determine what kind of model to search for.

Projects. We speak of searching for a model in terms of projects: Every time you run the Knowledge Discovery Engine with different data folders or parameters, you create a new project. Each project has a project folder for storing output.

Projects are like stepping stones—one leads to another. Studying your output data often suggests something else to try—different input data, perhaps, or you wonder what will happen with different parameter settings. This is all part of the search for an optimal model. Recognizing the close relationship between projects, Proteome Quest lets you set up master folders that store related project folders together.

Note: This chapter describes how to run one project at a time. A later chapter, “*Batch Modeling*,” describes running many projects in succession.

The Format of Input Data

Proteome Quest can analyze data from a variety of instruments, liquid chromatographs, for example. This chapter, however, assumes that data comes from serum protein analysis by a SELDI-TOF (Surface-Enhanced Laser Desorption and Ionization-Time of Flight) mass spectrometer

Samples. An input data file, variously called a sample or case or record, contains one SELDI-TOF mass spectrum of one individual’s serum proteins. The format for a sample (Figure 3-1) is a two-column Comma-Separated Values file (file

extension .csv). Column 1 lists mass/charge (M/Z) lines. Column2 shows corresponding intensities. An intensity, or amplitude, is the relative amount of the protein represented by a mass/charge line.

	A	B	C	D	E	F	G
1	M/Z	Intensity					
2	-0.0000786	-0.4963137					
3	2.18E-07	-0.4806275					
4	0.000096	-0.2453333					
5	0.000366	-0.5433726					
6	0.0008102	-0.4649412					
7	0.0014286	0.4762353					
8	0.0022211	-2.3237647					
9	0.0031879	-2.739451					
10	0.0043288	-3.4374902					
11	0.0056439	-3.8139608					
12	0.0071332	-3.8218039					
13	0.0087967	-3.8218039					
14	0.0106344	-3.8218039					
15	0.0126463	-3.8184016					
16	0.0148324	-3.7990183					
17	0.0171926	-0.7440464					
18	0.0197271	4.9622004					
19	0.0224357	10.480507					
20	0.0253185	7.1677348					
21	0.0283755	4.2081987					

Figure 3-1: The two-column CSV format of a sample (input file). A sample may contain more than 15,000 SELDI-TOF mass/charge lines and intensities.

Note: Mass spectra from individuals differ by intensities, not mass/charge. Proteome Quest requires samples to have identical mass/charge lines and the same total number of lines. Models are worthless if mass/charge lines vary between samples.

Note 2: It is beyond the scope of this manual to discuss data preparation, but unprocessed data is recommended. Avoid processing negative numbers such as those in Figure 3-1 (included here as a deliberate bad example). Lines and intensities should be positive. SELDI-TOF miscalibration is the major cause of negative mass/charges. Negative intensities are instrument artifacts or result from baseline adjustment methods. If negative numbers occur only at the beginning of the sample, use the Set Data Range window (discussed below in “Creating a New Project”) so that only positive numbers are processed.

Organize Input Data by Biological State

The Knowledge Discovery Engine permits great flexibility in organizing samples into folders. That said, you will save time and make model interpretation easier by organizing samples by biological states. Good data organization also helps track biological substates and check data quality.

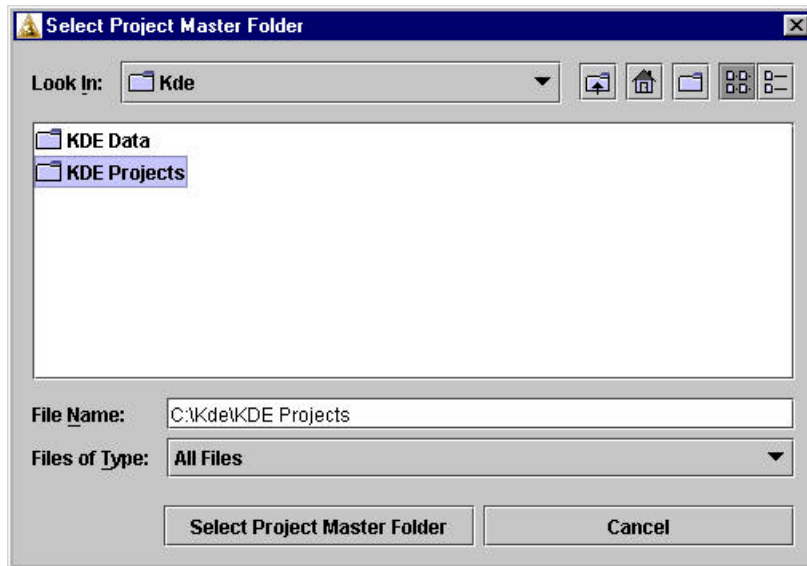





Figure 3-9: Use the Look-In box and the icons in the Select Project Master Folder box to select or create a project master folder.

The **Select Project Master Folder** box uses the **Look In** box and the icons at the top to navigate to folder and file locations, create new folders, and order folder appearance in the large white box. Here is how they work.

- **Short Cut.** The  button on **Look In** shows folder hierarchy—a short cut to changing folders.
- **Open Folder.** Double-click a folder name to move into that folder.
- **Up One Level.** Click the  button to move up one level in the folder hierarchy.
- **New Folder.** The  button creates a new folder inside the current folder; select the new folder, type a new name, and press **Enter**;
- The last two icons show or hide folder sizes, dates and other attributes.

In Figure 3-9, placing folder *Kde* within the **Look In** box displays the folder's contents, one level down, in the large white box. One level down are folders *KDE Data* (master input folder) and *KDE Projects* (master output folder). Clicking once on *KDE Projects* highlights the folder and shows its file path in the **File Name** box. Double clicking on *KDE Projects* moves into the folder (Figure 3-10).



Figure 3-29: A confirmation message signals completion of the Test and Train screen.

Frequently Asked Question

How many samples do these sets need?

Training, Test and Validation Sets require a minimum of three samples. There is no maximum. The number of Training Set samples needed to yield a useful model depends partly on the problem. The Training Set that yielded the ovarian cancer model reported in *The Lancet* had 50 ovarian cancer samples and 50 control samples. Modeling a different disease might require hundreds or thousands of training samples.

The number of samples also depends on data quality. Automation technologies, (robotic pipetting, for example) reduce noise in mass spectra and sharpen intensity pattern distinctions between biological states. Consequently, the Knowledge Discovery Engine finds superior models in less time and with fewer samples. Thanks in part to automation, we have found an ovarian cancer model that has so far been flawless—no false positives or negatives. The cleaner the data, the fewer training samples needed for outstanding results.

It is reasonable to place similar numbers of samples in Test and Validation sets, but consider making the Training Set larger in order to help the Knowledge Discovery Engine learn the patterns that best distinguish biological states.

Step 3: Set the Range of Mass Spectra Data

Clicking **OK** on the confirmation message in Figure 3-29 brings up the **Set Data Range** screen, which sets the range of Training Set mass/charge values within which the Knowledge Discovery Engine looks for a model. Figure 3-30 shows density plots of SELDI-TOF spectra from six samples randomly chosen from the Training Set. At the top are density plots from three samples for biological State 0 (Control, in this example). The lower plots represent State 1 (Diseased).

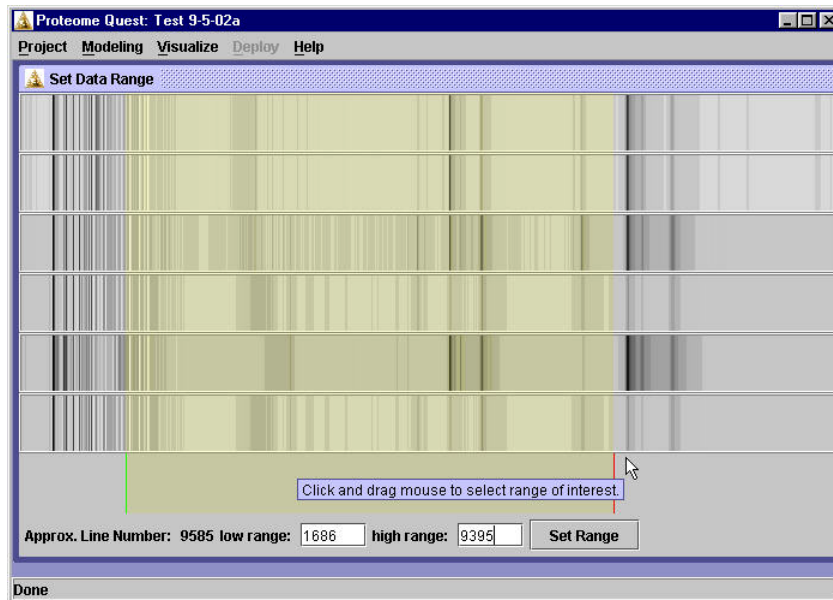


Figure 3-31: Drag the mouse from left to right to highlight and select the range of line numbers used for all Training Set samples. Range boundaries set by clicks appear in the low and high range boxes.

Tip: You may also set the range by typing into the range boxes and clicking Set Range. Typing overrides (but does not alter the appearance of) highlighting with the mouse.

Caution: Take a moment to notice how well major bands line up between density plots, especially upper range bands. Mass/charge lines are supposed to be identical between samples, so major bands usually align. Band shifts, such as the four shifted upper-range bands at the bottom of Figure 3-32, are a sign of mass spectrometer miscalibration and result in misleading models.



Figure 3-32. Watch out for unusual density plots. Notice how the higher range bands in the bottom plot are displaced from those above. Reexamine how displaced spectra were obtained.

Index

3

3D charts, 90–94
definition, 89

A

anchor points, 81

B

band shifts, density plot, 60
batch modeling, 127–54
 Accept Table adjustments, 139
 Accept Table reordering, 137
 Accept Tables, 135–38
 Accept Tables opened with
 Excel, 142–44
 advantages of, 127
 avoiding batches too large,
 152
 batch master folder, 128
 Batch Modeling box, 127
 Batch Modeling command,
 127
 batch name, 128
 by increasing Population,
 144–46
 by maximizing Population,
 149–50
 by reusing parameter settings,
 148–49
 by varying Feature, 150
 by varying Match, 146–48
 ending early, 143
 errors, 136
 estimating generation time,
 152
 filter for cluster properties,
 130
 filter for sensitivity and
 sensitivity, 131

filtering for underpopulated
 nodes, 147
filters compared, 146
finding model tables, 139–42
how-to summary, 159
investigate model details, 154
launch, 134
parameter ranges and
 increments, 129
preparing for, 151–53
project set-up, 127–34
Reject Tables, 138–39
sensitivity, 136
set data range, 133
specificity, 136
status boxes, 134
Test and Train box, 133
tips for better batch runs,
 153–54
Training Parameters box, 129
underpopulated nodes, 136
biological significance
 of clusters, 22
biological state
 and batch modeling, 129
 and data organization, 37
 default states, 42
 designation in Test and Train
 box, 53
 existence of numerous models,
 25
 how assigned to clusters, 18
 not limited to disease, 8
 predicted by cluster matching,
 14, 23
 principle for predicting state,
 10
 relationship to
 misclassification rate, 18
 sensitivity and specificity, 42
biological substate, 38
 and Test and Train box, 55
 control substates, 39
 folders, 39

 patterns associate with, 22
biomarker
 and clusters, 22
 multiple, 9, 21
 patterns in Model Result
 Table, 82
 patterns in Model Update
 Table, 123–26
 predictive potential, 11
 single vs multiple, 22
 single-protein, 9

C

centroid
 first point in cluster, 16
 in fixed position, 69
 pattern, 21
 relationship to decision
 boundary, 16
 relationship to Learning, 17
 relationship to proteomic
 fingerprint, 79
 shifts, 17
 shifts during updating, 123
charting tips, 91, 93
cluster analysis, 14–19, 16–19
 and biological significance, 22
 pattern recognition as
 geometry, 14
 transition to genetic algorithm,
 19
cluster creation
 assigning biological state, 18
 first point, 16
 subsequent points, 17
cluster matching
 according to the closest
 centroid, 23
 match failure. *See* match
 failure
 negative probability, 71
 used to predict biological
 state, 14, 23