

Biomarker Discovery Using SELDI Technology

A Guide to Data Processing and Analysis
Using ProteinChip® Data Manager Software

BIO-RAD

Data Processing and Analysis Using ProteinChip® Data Manager Software

About This Guide

This guide describes the concepts and workflow involved in the processing and analysis of data generated with the ProteinChip SELDI system. A content map and summary of the general workflow steps are provided at right. The workflow begins with the organization and processing of the raw spectra that are obtained during data collection and ends with the multivariate statistical analysis methods available within ProteinChip data manager software.

The recommendations presented are based on the ProteinChip SELDI protein biomarker discovery process. They are intended as general workflows and settings for a majority of sample types and experiments. In some cases, modifications may be necessary. Included are recommendations for each stage of data processing and analysis, discussions of the different parameters and settings that may be used, and examples of how each step may affect final data analysis.

A more detailed summary of the workflow steps is provided in the Appendix. Details for performing each step in ProteinChip data manager software, however, are not provided; for this information, refer to the ProteinChip Data Manager Software Operation Manual.

Data Organization

Data Organization

- Ensure proper annotation of spectra
- Group spectra into folders

Mass Calibration and Data Processing

Mass Calibration

- Create the calibration equation using calibration standards
- Improve mass accuracy through external calibration

Processing Spectral Data

- Adjust peak intensity calculations using baseline subtraction
- Reduce noise by filter adjustment
- Improve the signal-to-noise ratio by setting the noise range

Processing Conditions

- Standardize intensities by normalization
- Improve mass precision by spectrum alignment

Generation and Analysis of Peak Clusters

Generation of Peak Clusters

- Associate peaks from multiple spectra into peak clusters
- Adjust peaks within clusters using cluster editing

Selection of Candidate Biomarkers

- Select biomarker candidates using P values
- Measure biomarker sensitivity and specificity using ROC curves

Assessment of Sample Relationships

- Examine sample hierarchical clustering patterns
- Visualize sample similarity using PCA

Notices

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage or retrieval system, without permission in writing from Bio-Rad Laboratories, Inc.

Bio-Rad reserves the right to modify its products and services at any time. This user guide is subject to change without notice.

Although prepared to ensure accuracy, Bio-Rad assumes no liability for errors, or for any damages resulting from the application or use of this information.

ProteinChip and ProteoMiner are trademarks of Bio-Rad Laboratories, Inc. Windows is a trademark of Microsoft Corporation.

The SELDI process is covered by U.S. patents 5,719,060, 6,225,047, 6,579,719, and 6,818,411 and other issued patents and pending applications in the U.S. and other jurisdictions.

Copyright © 2008 by Bio-Rad Laboratories, Inc. All rights reserved.

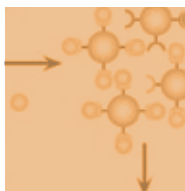
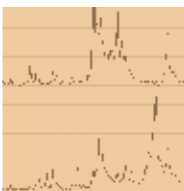


Contents

■ Part I: The Data Analysis Process	1
The Data Analysis Process	1
Overview	1
ProteinChip® SELDI Data Types.....	2
■ Part II: Data Organization	7
Data Organization	7
Ensure Proper Annotation of Spectra.....	7
Group Spectra Into Folders.....	8
■ Part III: Mass Calibration and Data Processing	11
Mass Calibration	11
Create the Calibration Equation Using Calibration Standards	11
Improve Mass Accuracy Through External Calibration.....	13
Processing Spectral Data	14
Adjust Peak Intensity Calculations Using Baseline Subtraction.....	14
Reduce Noise by Filter Adjustment	17
Improve the Signal-to-Noise Ratio by Setting the Noise Range	18
Processing Conditions	19
Standardize Intensities by Normalization.....	19
Improve Mass Precision by Spectrum Alignment.....	21
■ Part IV: Generation and Analysis of Peak Clusters	25
Generation of Peak Clusters	25
Associate Peaks From Multiple Spectra Into Peak Clusters	25
Adjust Peaks Within Clusters Using Cluster Editing	28
Selection of Candidate Biomarkers	30
Select Biomarker Candidates Using P Values.....	30
Measure Biomarker Sensitivity and Specificity Using ROC Curves.....	32
Assessment of Sample Relationships	34
Examine Sample Hierarchical Clustering Patterns	34
Visualize Sample Similarity Using PCA.....	36
■ Part V: Going Beyond ProteinChip Data Manager Software	41
Going Beyond ProteinChip Data Manager Software	41
Develop and Test Classification Algorithms.....	41
Export Data for Further Analysis	42
Validate Candidate Biomarkers	42
■ Appendix	45
Glossary	45
Workflow Summary	49

Part I: The Data Analysis Process

The Data Analysis Process.....	1
Overview	1
ProteinChip® SELDI Data Types.....	2



The Data Analysis Process

Mass spectrometry (MS)-based clinical proteomics and biomarker research generate vast and complex data sets. As a result, data analysis can become a critical bottleneck in the research process. This chapter provides an overview of the data analysis workflow used with the ProteinChip SELDI system and introduces the primary types of ProteinChip SELDI data.

Overview

The MS-based techniques used in biomarker research generate vast quantities of data. For example, depending on the type of study and sample source, discovery phase experiments may require 100 or more samples and profile thousands of proteins and peptides, all with varying expression levels. Rigorous data analysis is required to sort through these data to select the most robust candidate biomarkers.

The data analysis workflow used for biomarker discovery with the ProteinChip SELDI system and ProteinChip data manager software involves the following steps:

- Data organization — creation or revision of sample annotations and organization of spectra into folders. Analysis of one condition requires organizing spectra from that condition into the same folder
- Mass calibration — generation or updating of the calibration equation, which converts raw time-of-flight (TOF) data into mass data
- Processing — baseline subtraction, filtering, noise estimation, alignment, and normalization of spectra. These steps optimize the accuracy and reproducibility of peak mass and intensity measurements prior to clustering
- Generation of peak clusters — creation of peak clusters from spectral data, cluster editing, and preparation of peak cluster data for statistical analysis. Individual spectral features or peaks are labeled across all spectra in a folder and then grouped together based on matches between their calculated masses. The result is a list of masses, or peak clusters, that contain the same peaks from each spectrum with associated intensities for each sample
- Selection of candidate biomarkers — analysis of peak clusters by univariate statistical methods to identify candidate biomarkers that distinguish sample groups. The comparison of peak intensities of the same peak across multiple spectra (within a cluster) is a measure of protein expression levels from multiple samples
- Assessment of sample relationships — further evaluation of peak clusters using unsupervised multivariate statistics to reveal desired associations, examine possible sources of experimental variability (preanalytical or analytical bias), and generate more confidence in candidate biomarkers

This process yields lists of candidate biomarkers that, to effectively conclude the biomarker research process, must undergo experimental validation. In addition, biomarker candidates are often identified and more specific clinical assays are developed. For more information about the design and implementation of biomarker research projects, see the sidebar at the end of this chapter and refer to bulletin 5642, Biomarker Discovery Using SELDI Technology: A Guide to Successful Study and Experimental Design.

ProteinChip SELDI Data Types

The data analysis process is represented by the data types it generates.

SELDI Spectra

ProteinChip SELDI analysis begins with laser desorption and ionization of proteins from a ProteinChip array surface and detection by TOF-MS. Inside the ProteinChip SELDI reader, a nitrogen laser illuminates the sample, and the laser energy induces a change of state from the solid, crystalline phase into the gas phase (desorption) with an associated ionization of the protein molecules. Once in the gas phase, the protein ions accelerate away from the metal array upon application of a voltage differential. The voltage differential imparts the same energy to all of the analytes in the sample, resulting in mass-dependent flight times. Once the protein ions strike the detector, the detector records the TOF of the analyte (expressed as a mass-to-charge ratio, or m/z) and the intensity of the response, which is directly related to the amount of that specific analyte on the array surface.

ProteinChip data manager software displays the data as a spectrum, plotting the measured signal intensity on the y-axis and the calculated m/z or mass on the x-axis (Figure 1). (Since the charge on most ionized proteins and peptides is 1, the m/z calculated from the TOF of a particle is equivalent to its molecular mass plus 1 Da.) The SELDI spectrum is the graphical representation of a sample's protein profile.

Each spectrum has several notable features referred to in this guide. They are:

- Matrix attenuation range — region in which most of the signal is generated by the matrix. This signal is suppressed during data collection up to the designated matrix attenuation setting to increase detection sensitivity. This region is excluded from noise calculations and data analysis
- Low-mass range — region <20 kD. Peaks in this range are generally sharp and well-resolved. Data collection and analysis parameters for this range are optimized to maximize peak resolution
- High-mass range — region >20 kD. Peaks in this range are generally of lower intensity and are broader due to chemical heterogeneity. Data collection and analysis parameters for this range are optimized to maximize detection sensitivity

Peaks

Within each spectrum, peaks represent the protein or peptide components of a sample, and their intensities correspond to the amounts of the species in a sample (their expression levels).

Peaks are labeled either manually by the user or automatically by the software. In the manual method, the user applies the **Centroid** function of the software to click on and label peaks in the spectra. In the automatic method, the software identifies peaks by locating changes in intensity versus the noise that are greater than a specified detection threshold.

The software calculates peak intensity using the distance from the top of the peak to the calculated baseline (Figure 2). It calculates the mass by applying a calibration equation generated from a calibration spectrum (see Mass Calibration later in this guide).

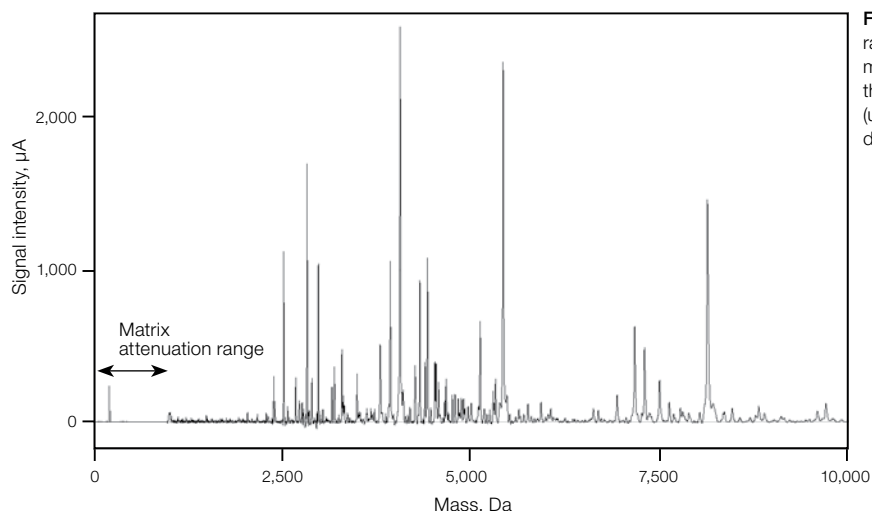


Fig. 1. SELDI spectrum. The 0–10 kD range of this spectrum is plotted with the mass on the x-axis and signal intensity on the y-axis. The matrix attenuation range (up to 1 kD in this example) is set during data acquisition.

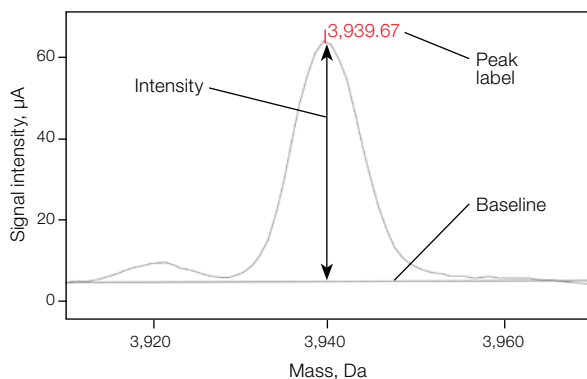


Fig. 2. Anatomy of a peak. Peaks are labeled either manually by the user or automatically by the software. The software calculates peak intensity using the distance from the top of the peak to the calculated baseline.

Peak Clusters

Though spectra and peaks may be two of the most recognizable forms of data generated by the ProteinChip SELDI system, peak clusters are the actual units of analysis used for biomarker discovery.

Peak clusters are groups of peaks of similar mass that are treated as the same protein or peptide across multiple spectra. ProteinChip data manager software defines clusters within the spectra in each data folder (Figure 3). Most of the data processing procedures described in this guide optimize the spectral data to enable detection of the most robust clusters and selection of the most robust biomarker candidates.

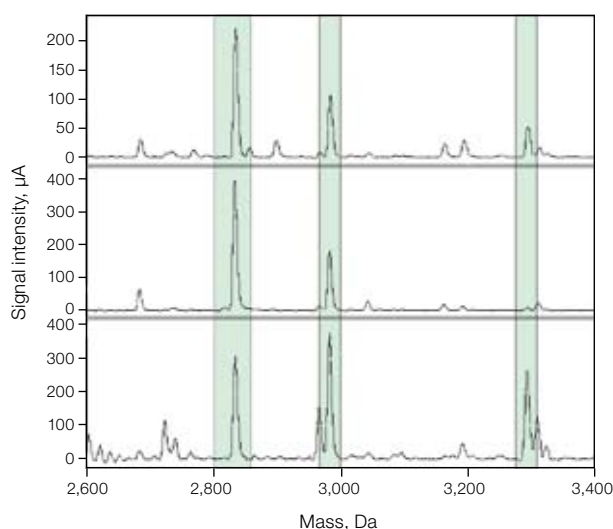


Fig. 3. Example of peak clusters. ProteinChip data manager software identifies peaks within the spectra in each data folder and groups them into clusters. In this example, three clusters are shown by the peaks that are grouped in each shaded box across three spectra.

For biomarker data analysis, the normalized peak intensities from each spectrum within a cluster are used to compare relative expression levels for that substance between different sample groups. The basis of biomarker analysis is searching for a difference in the expression level, or peak intensity, within a peak cluster (univariate analysis) or in a combination of multiple clusters (multivariate analysis).

Final Output

Univariate statistical analyses only consider single features. In biomarker research, this means they consider differences in the intensity of a single peak at a time. The aim is to identify peak clusters with statistically significant differences in intensity between sample groups (Figure 4). Evaluating clusters using univariate statistics (using the P values and areas under receiver-operating curves, or AUC values, calculated by ProteinChip data manager software) indicates their potential as single biomarkers.

The final output from ProteinChip data manager software is a list of candidate biomarkers for further analysis. The next steps in the characterization, confirmation, and application of the biomarker candidates involve experimental validation, identification, and clinical assay implementation as described in more detail in bulletin 5642, Biomarker Discovery Using SELDI Technology: A Guide to Successful Study and Experimental Design.

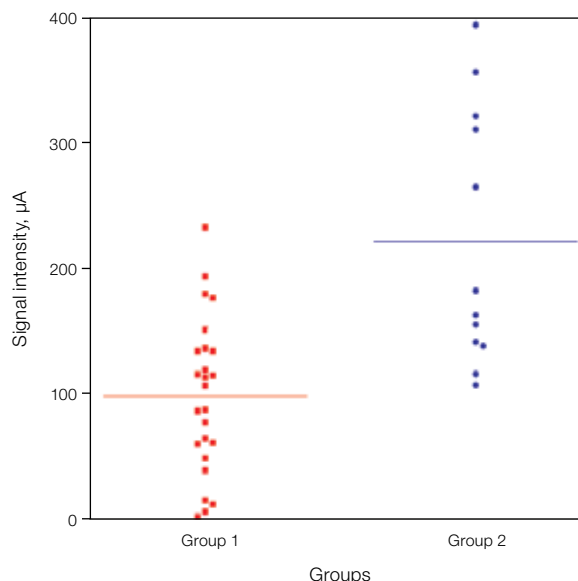


Fig. 4. Sample-group comparison within a cluster. Group scatter plot displaying the intensity values for one peak per spectrum from a selected cluster across two sample groups. The horizontal bars indicate the average intensity value for each group. In this example, the sample group at left contains the protein (peak) at expression levels that are statistically lower than those in the group at right. Such significant expression level changes may indicate the presence of a biomarker.

The Importance of Study and Experimental Design

Any analysis is only as good as the data it uses, and good spectral data require adherence to best practices in study and experimental design. To facilitate data



processing and analysis, and to maximize the potential for obtaining reliable biomarker candidates, follow the guidelines detailed in bulletin 5642, A Guide to Successful Study and Experimental Design. General recommendations are summarized below.

Minimize Preanalytical and Analytical Bias

These biases can have profound effects on the outcome of a study and on the ability to apply biomarker candidates to broader populations in validation studies or clinical assays. Sources of preanalytical bias include any systematic differences in patient populations or sample characteristics, including the procedures used for sample collection, handling, and storage. Sources of analytical bias arise from differences in how the samples are processed and analyzed.

Implement Standard Operating Procedures (SOPs)

Implement SOPs for all phases of an experiment, from sample collection to data analysis. Establishing and optimizing SOPs helps minimize bias and increases the opportunity to discover robust biomarkers.

Collaborate With Specialists

Collaborate with specialists, especially clinicians and biostatisticians, during the study design and data analysis phases of a project. All MS methods, including SELDI, generate vast and complex data sets that must be processed and analyzed properly to generate reproducible results of clinical utility. These profiling techniques generate many peak intensity features per sample, significantly more than the total number of samples in a study. Therefore, employ assistance from a biostatistician during the planning stages of a project and throughout the later phases of data analysis and evaluation of biomarker candidates.

Use the biostatistician's expertise to:

- Calculate the number of samples required for statistical relevance

- Plan data analysis strategies that minimize the risk of false discovery and over fitting of classification models
- Develop solid statistical assumptions
- Apply conservative feature selection and statistical cross-validation within a sample set

Develop Data Analysis Strategies Before Acquiring Data

With help from the biostatistician, plan the data analysis workflow and methods you will use before acquiring data. Take into account the clinical question being asked, the study type, and the methods and success criteria being used for the project. Planning the analysis strategy ahead of time ensures that your experimental design generates the amount and types of data you need to perform the types of analysis you have planned.

Use Appropriate Samples, Controls, and Standards

Select samples and controls using the general principles defined in bulletin 5642. For each phase of a project:

- Include one control sample on each array for quality control (QC sample). The total number of QC sample spots can be reduced with very large sample sets
- Use calibration and reference samples that are good models for the experimental samples. For example, dilute protein standards in a sample that is similar to that being analyzed (serum, plasma, urine, etc.), ensuring calibrant concentrations are within the concentration range expected for the experimental samples

Optimize Acquisition Protocols Before Acquiring Data

The ProteinChip SELDI instrument uses acquisition protocols to acquire data from a spot or portion of a spot on a ProteinChip array. For best results:

- For every experimental condition, test the protocols on a pooled sample before collecting data from study samples
- Optimize the settings for the low- and high-mass ranges separately to ensure sufficient coverage of the entire mass range
- As a starting point, use the collection parameters for general protein profiling recommended in bulletin 5642
- Evaluate spectra during optimization for the maximum number of well-resolved and sharp peaks
- After optimizing laser energy settings, maintain constant data collection settings for all samples associated with that condition