Enhanced Protein Identification with the Agilent Spectrum Mill MS Proteomics Workbench Version A.03.01

Technical Overview

Jose Meza and John Chakel
Agilent Technologies

Introduction

Mass spectrometry (MS) is widely used for protein identification. As MS techniques have become easier and more automated, the quantity of data generated has increased exponentially, necessitating high-volume data processing tools such as the Agilent Spectrum Mill MS proteomics workbench. The Spectrum Mill workbench is a comprehensive set of high-throughput software tools for identifying proteins from MS and MS/MS spectra. This software can extract the high-quality peptide spectra from raw data files, search the spectra against protein databases, and summarize the results in ways that are meaningful to biologists and protein chemists. The summary capabilities allow ready comparison of multiple samples, each with multiple fractions, and include the ability to calculate relative protein concentrations either with or without use of isotope labels. Also included are homology mode searches, *de novo* sequencing, and tools and utilities to speed proteomic research. The Spectrum Mill workbench is an open platform that can process both Agilent and non-Agilent data.

The latest A.03.01 version of the Spectrum Mill workbench extends the range of instruments for which the workbench can be used, and adds a number of new capabilities. This version of the Spectrum Mill workbench is compatible with the Microsoft® Windows® Server 2003 and Windows 2000 Server operating systems. This technical overview discusses the new features of Spectrum Mill workbench version A.03.01.

Agilent Technologies

New Features

The A.03.01 version of the Spectrum Mill workbench includes the following new functions:

- Processing of Agilent LC/MSD TOF peptide mass fingerprinting (PMF) data, including data generated with the following interfaces:
  - Atmospheric pressure matrix-assisted laser desorption/ionization (AP-MALDI)
  - Infusion-electrospray (infusion-ESI)
  - Liquid chromatography-electrospray (LC-ESI)
- Processing for new isotopically labeled modifications, including light/heavy ratio calculations for:
  - Lysine imidazole D_0/D_4
  - C-terminal methyl ester D_0/D_3
  - N-terminal propionyl D_0/D_5
- Processing of MS-only (PMF) data for Agilent LC/MSD Trap and Applied Biosystems/MDS SCIEX oMALDI QSTAR
- Mixture scoring in PMF Search, to score and order possible mixtures along with individual proteins
- Dynamic mass tolerance to compensate for mass error in externally calibrated data
- Searching of the International Protein Index (IPI) protein database
- Integration with the Agilent Synapsia Informatics Workbench, to allow a Synapsia user to import a Spectrum Mill summary report
- Sequence map feature that allows rapid assessment of b- and y-series ion coverage
- Tool Belt utility for more convenient termination of Spectrum Mill processes
- Check boxes to remove prior results in Data Extractor, MS/MS Search, and PMF Search
- Customized handling in *de novo* sequencing of amino acid pairs that have the same or very close monoisotopic masses (isoleucine/leucine, lysine/glutamine)
- Additional flexibility for customization of amino acid modifications
- Use of Peak Picker for MS data
- The ability to run on both Windows Server 2003 (IIS 6.0) and Windows 2000 Server
- Java Runtime Environment (JRE) 1.4.x support

A number of these new capabilities are discussed in detail in the following material.

PMF Data Processing

Peptide mass fingerprinting is a technique for the identification of proteins using the MS-only spectra from peptides produced by proteolytic digestion. Because the digestion enzymes cleave at specific amino acids, a mass spectrum of the cleavage products results in a unique “fingerprint” for the protein. The protein can often be identified via search of the spectrum versus a protein database. The A.03.01 version adds significant capabilities for processing MS-only (PMF) data from the Agilent LC/MSD Trap (ion trap) and TOF (time-of-flight) mass spectrometers, as well as the Applied Biosystems/MDS SCIEX oMALDI QSTAR. Figure 1 shows a diagram of the overall workflow for PMF analysis with the Spectrum Mill workbench, using the new and enhanced capabilities highlighted.

Processing of Agilent TOF PMF Data

With excellent sensitivity, resolution, and mass measurement accuracy, the Agilent LC/MSD TOF time-of-flight mass spectrometer is well-suited for acquisition of peptide molecular weights via peptide mass fingerprinting. The A.03.01 version of the Spectrum Mill MS proteomics workbench offers easy processing of all types of Agilent TOF data, including AP-MALDI-TOF, infusion-ESI-TOF, and LC-ESI-TOF data. This version of the software includes an Agilent TOF data extractor that automatically extracts mass lists for PMF search from the raw data files. This saves time compared to manual spectral averaging and peak list export. Once the data extraction parameters are set (Figure 2), multiple files can be processed with a single mouse click.

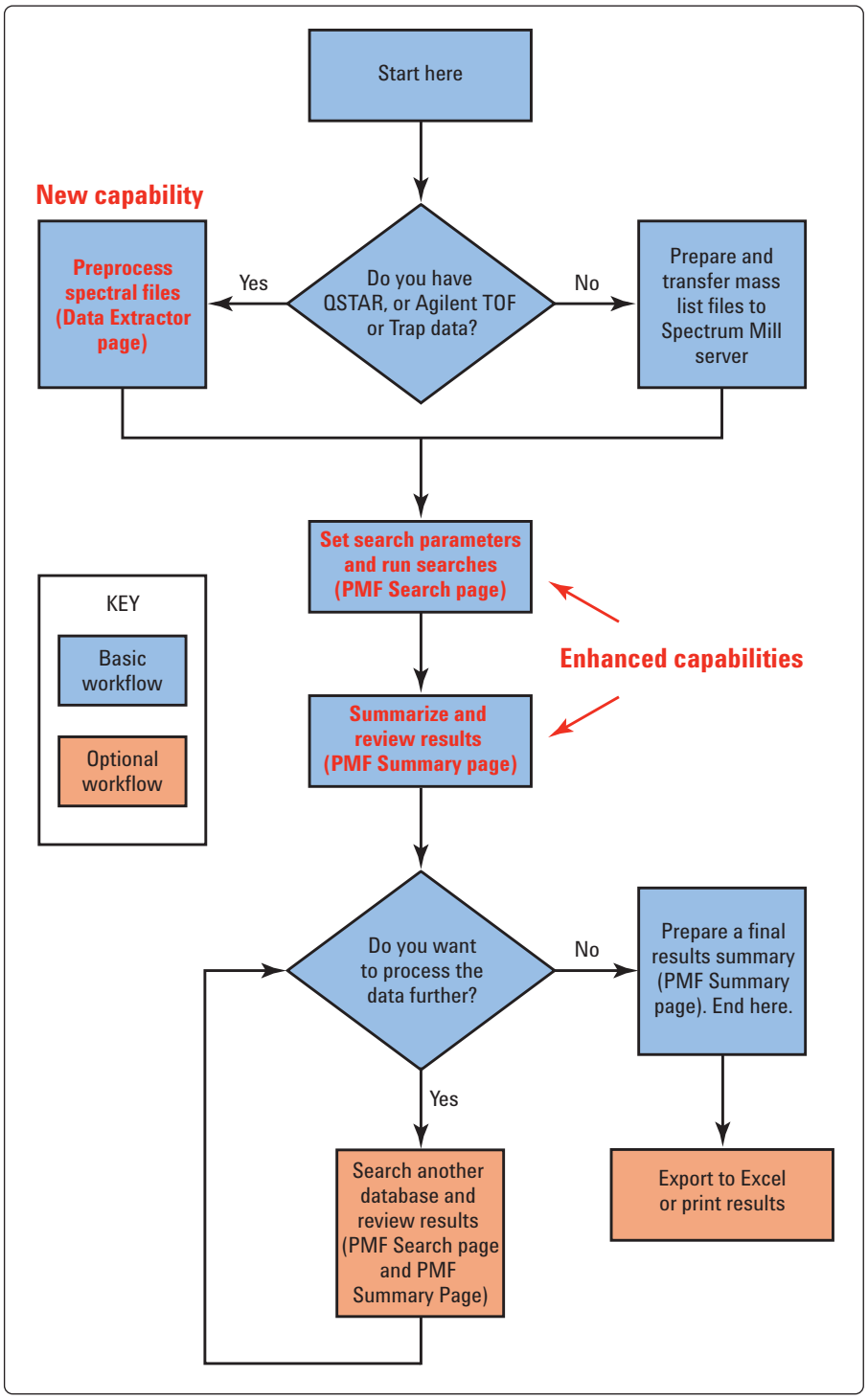

Figure 1. Flowchart for processing MS-only data with the Spectrum Mill workbench

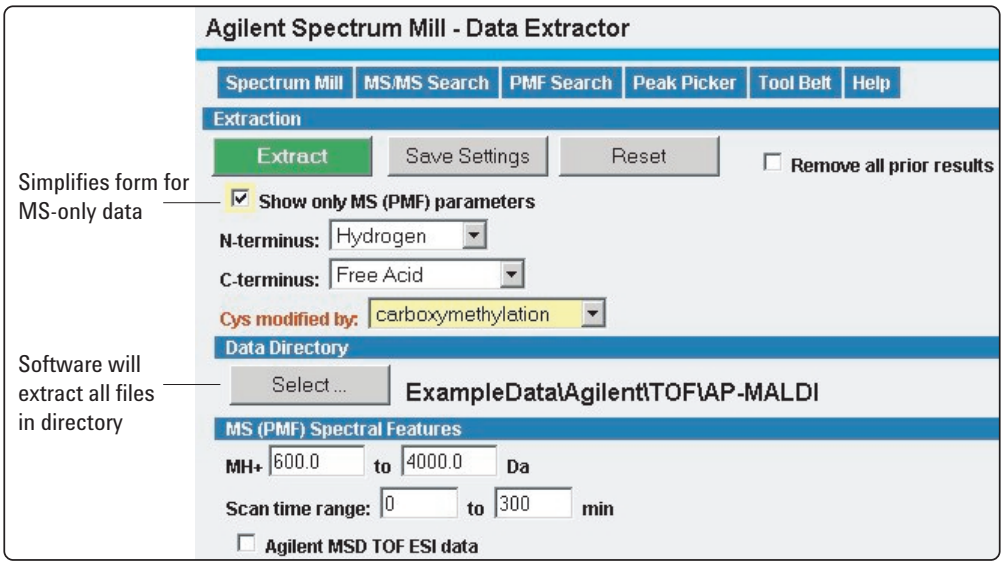

Figure 2. Data Extractor form set up for AP-MALDI-TOF data

Because the data extractor must handle both chromatographic data and non-chromatographic data, it is customized for each type of TOF data. For example, for LC-ESI-TOF data, the spectra are extracted only for the chromatographic peaks, and a user-specified baseline region is used for background subtraction. For AP-MALDI-TOF, where there are no peaks, the spectra are extracted from the entire file, and no background subtraction is performed. The background subtraction is accomplished later during the PMF search, when the user specifies a list of contaminant masses for subtraction.

To reduce false positives, the A.03.01 version of the Spectrum Mill workbench allows PMF searching of TOF data with low-ppm peptide mass match tolerances. The software includes convenient summary and export of PMF search results.

Data processing for new isotopically labeled modifications

Studies with isotopically labeled derivatives contribute to understanding of relative protein expression levels between different cell states. The first release of the Spectrum Mill workbench supported light/heavy ratio calculations for isotope-coded affinity tag (ICAT) quantitation, with support for both the original ICAT (D_0/D_8) and the newer cleavable ICAT ($^{12}C/^{13}C$) reagents. The new release adds support for:

- N-terminal propionyl-D_0, propionyl-D_5, and propionyl mix
- C-terminal methyl ester-D_0, methyl ester-D_3, and methyl ester mix
- C-terminal lysine imidazole-D_0, lysine imidazole-D_4, and lysine imidazole mix (Agilent's lysine mass tagging reagent)

The Agilent lysine mass tagging reagent increases sensitivity by five to twenty times for AP-MALDI analysis of lysine-containing peptides. The reagent also increases the intensity of y-series ion MS/MS fragments, which enhances confidence in search results and makes it easier to perform *de novo* sequencing. Since the reagent is available in unlabeled (D_0) and labeled (D_4) form, it can be used for quantitative studies. The A.03.01 release of the Spectrum Mill workbench automatically calculates the light/heavy ratios.

Mixture Scoring in PMF Search

Samples such as gels spots frequently contain mixtures of proteins that must be identified simultaneously by PMF searching. The new mixture scoring feature in the Spectrum Mill workbench version A.03.01 allows both mixtures and individual components to be scored on a common probability scale, making it easy to evaluate the results (see Figure 3). When a potential mixture is detected, the mass spectrum is color-coded to differentiate peaks from each component.

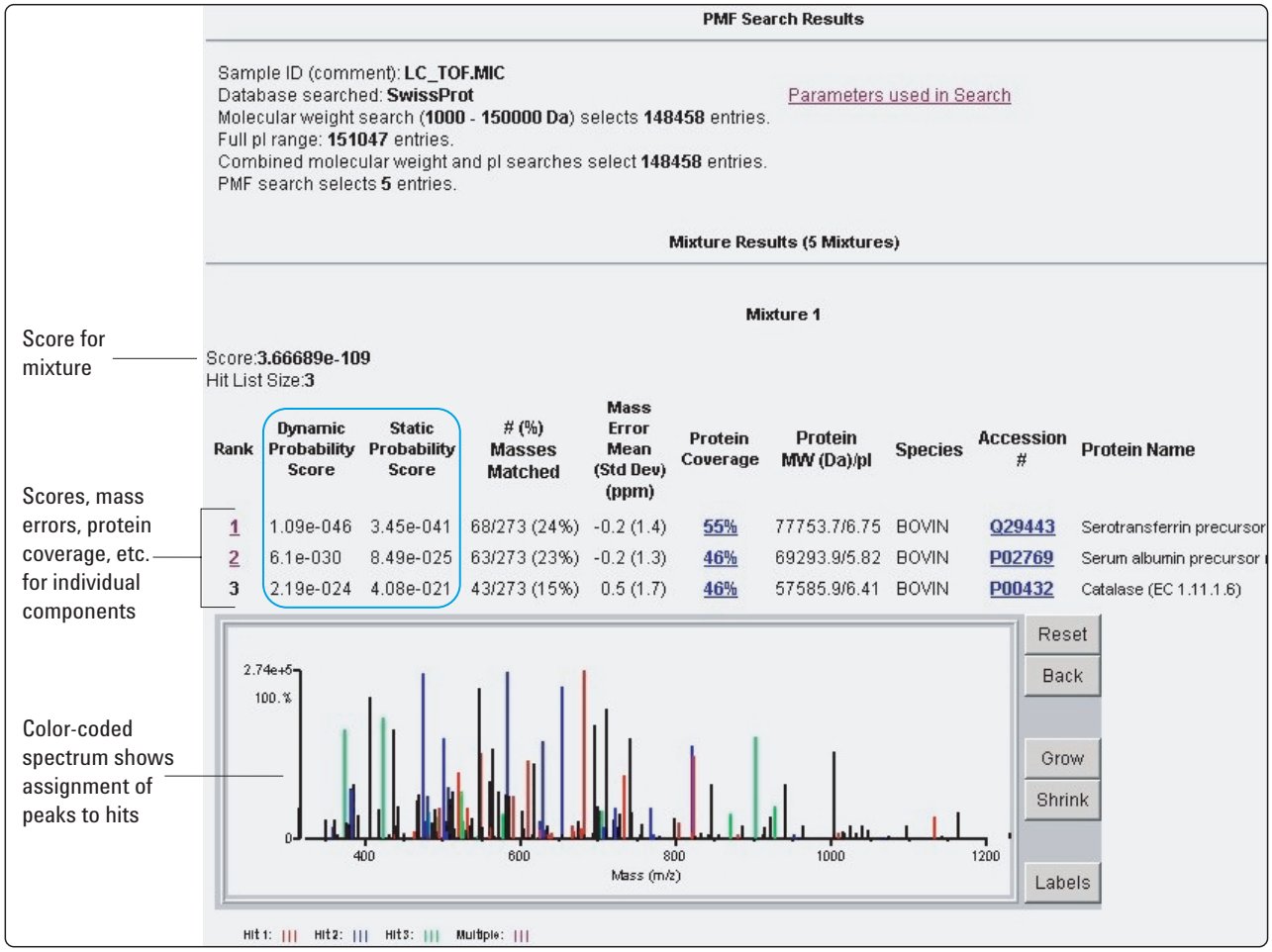

Figure 3. Mixture scoring enables meaningful comparison of mixture and individual protein scores

When mixture scoring is enabled, the scoring method changes significantly. For example, with mixture scoring disabled, the mass spectral peaks that are not assigned to the single tentative hit (protein identification) are assumed to be noise and count against the score. With mixture scoring enabled, the mass spectral peaks assigned to *all* components of the mixture contribute to a better score; none are considered to be noise. Also, there are bonus points for the peaks being mutually exclusive (for example, no overlap of peaks among components).

The mixture scoring feature is especially useful for protein combinations where one component dominates the spectrum. In that case, the top hits are often the dominant protein and various precursors of the dominant protein. With mixture scoring, the likelihood of identifying the less-abundant proteins is increased.

Dynamic Mass Tolerance

Dynamic mass tolerance makes it easier to process externally calibrated MALDI or AP-MALDI data. Without internal mass calibration, it is not easy to choose the appropriate peptide mass tolerance (the acceptable difference in mass between the experimental mass peak and the theoretical mass peak) when conducting a PMF database search. If the chosen setting is too restrictive, it is possible that no matches will be found. On the other hand, if the setting is not restrictive enough, false positive matches may emerge.

Dynamic mass tolerance overcomes this difficulty because it allows users to set a larger mass window for PMF searching without increasing the likelihood of false positives. This is because the probability scores are calculated based on the actual deviation of the mass data from theoretical. An example is shown in the highlighted area of Figure 3. In this example, a 10-ppm mass tolerance window was used for PMF searching. Two sets of scores were returned, the static probability score and the dynamic probability score. The static probability score was based on the peptide mass tolerance set in the PMF Search form. The dynamic probability score was based on the actual peptide mass deviations determined from the experimental data. Since the actual data was more accurate than the mass tolerance set in PMF Search, the dynamic probability score was better (smaller number) than the static probability score.

Data Recalibration

The Spectrum Mill PMF data recalibration capability saves time in the event that samples are inadvertently analyzed using an MS with a poorly calibrated mass axis. In that situation, the samples would normally be rerun. This software feature allows recalibration of a data set based on results of an initial database search. During the initial search, the peptide mass tolerance is set larger than the calibration error. Then recalibration is enabled in PMF Summary so that a slope and intercept are calculated based on deviation of the experimental masses from theoretical masses for the top database hit. To avoid recalibration based on an incorrect database match, the user examines the data to determine a consistent slope and intercept among data files. The data are searched again with this recalibration applied. The result is better scores and greater confidence without the need to reanalyze the samples.

Searching of the International Protein Index (IPI) protein database

The International Protein Index (IPI) protein database is a non-redundant database that cross-references human, mouse, and rat proteins from other databases. The database provides stable identifiers so that sequences can be tracked from one release of the database to the next. The Spectrum Mill workbench version A.03.01 includes the capability to search this database.

Conclusions

The new features in the A.03.01 version of the Agilent Spectrum Mill MS proteomics workbench extend the applicability of this software. The new data extractors, mixture scoring, dynamic probability scoring, and data recalibration make PMF data processing faster, easier, and more reliable. Processing flexibility is extended with the capability to search the IPI database, to perform light/heavy calculations for Agilent's lysine tagging reagent and to export summary reports to Agilent's Synapsia informatics workbench. These and other features build upon the original Spectrum Mill workbench to increase productivity and ease-of-use.

Authors

Jose Meza and ***John Chakel*** are scientists at Agilent Technologies in Santa Clara, California U.S.A.

www.agilent.com/chem

Printed in the U.S.A. June 23, 2004
5989-1351EN

Agilent Technologies