High through put analysis of MALDI time-of-flight MS/MS and LCMS/MS data.

Richard J. Jacob[1], Peter R. Baker[1], Michael A. Baldwin[1], Armin Graber[2], A. L. Burlingame[1]

[1]University of California, San Francisco, CA, [2]Applied Biosystems, Framingham, MA

Recent advances in mass spectrometry instrumentation allow smaller quantities of samples to be analyzed more accurately and rapidly than ever before. The development of new instruments, contemporaneous with the completion of the Human genome project and the emergence of proteomics as a viable tool for the analysis of the cell state, have lead to huge increases in raw data production. Here we present an approach to the management of large data sets that combines instrumentation control via feedback from the search results of spectra, automated spectral processing, batch searching and organization then documentation of results into a laboratory information management system (LIMS).

An evaluation of differential protein expression in cell lysates separated by 2-D gels may require analysis of many thousands of spots per gel, particularly to identify proteins below the level sensitive to silver of fluorescent staining. To achieve such analyses within a reasonable time, robotic sample preparation and automated MS operation, data analysis and interpretation are essential. Comprehensive and well organized sample management, archiving of data and results, such as can be achieved with a LIMS, are also critically important.

In this approach, slices cut by hand or by robotic spot pickers from 1-D or 2-D gels must be handled entirely by robots, i.e. digested with trypsin, extracted and spotted with α–cyano-4-hydroxycinnamic acid onto MALDI target plates. The digests will then be analyzed by MALDI time-of-flight MS/MS using the following strategy. Each target plate will be subjected to three passes. The first pass is to identify proteins by mass mapping, the second to confirm these assignments by CID-MS/MS of selected ions, and the third to sequence peptides that are not attributable to proteins identified thus far. With the laser operating at 200 Hz, mass spectra of individual spots will be acquired at a rate of approximately one per second during the first pass of the MALDI sample plate, collected in batches of 50 and submitted to MS-Fit in real-time. For a 400-spot sample plate, the total acquisition should take approximately 7 minutes. A higher frequency laser (i.e. kHz) would require a proportionately shorter time. By the time the first pass is completed, the results from the initial MS-Fit searches will be available to control the CID experiments in the second pass, starting again with spot #1. For each protein identified by MS-FIT, one peak will be selected for CID-MS/MS to confirm the assignment using MS-Tag. During the second pass, all peaks from the first pass that can be attributed to assigned proteins will be used to create an exclusion list for the third pass. During the third pass the remaining unmatched peaks that meet the required criteria of “goodness” will be selected for CID-

MS/MS. CID data will be batched and submitted to MS-Tag while the target plate is moved to the next spot. In general, peaks selected for CID MALDI time-of-flight MS/MS second pass analysis should be of above average intensity and no other peptides should be within +/- 0.25% mass of the peak. For the third pass, peaks should match the criteria set for second pass selection. Furthermore, the mass should verify the identity of a protein, particularly masses potentially associated with a second or third protein in a sample mixture or not yet be attributed to any protein by MS-Fit searching. Alternatively, the peptide sequence should determine the origin of any peptide that, on the basis of mass alone, could be assigned to more than one protein.

As an alternative to MALDI time-of-flight MS/MS for samples judged to be of the greatest interest, a nanobore-HPLC-MS/MS IDA analysis of a digest from a single spot will yield more comprehensive data. The raw data will be processed, including determination of the charge state of each isotopic peak cluster, and the picked peaks will be submitted for batch-wise database searching with protein Prospector.

Sample information about the origin, procedures and conditions used during sample preparation will be stored in a relational database. SQL*LIMS from PE Informatics is based on the Standard Query Language (SQL) database from Oracle. It has a graphical user interface that allows easy access to all aspects of the protein analysis procedure. The database supports one to many and many to many connections, linking single spots or pixels on a 2D gel to its preparation and workup information and its MS data and search results. This is designed for integration into a proteomics laboratory, sending and receiving information with the laboratory's instrumentation.

Both the MALDI TOF/TOF and QSTAR can rapidly provide large quantities of high quality sequence data, in fact too much for manual human analysis. Therefore automated analysis is essential. A LIMS becomes vital for keeping the data organized especially when working with multiple unrelated samples on one MALDI time-of-flight MS/MS sample plate. Furthermore, a LIMS should also enable an analyst to quickly drill down to the data to check the quality and the interpretation of the results.

This work was supported by NIH NCRR 01614.