Integrated and Automated Data Processing System for Protein Identification, Characterization and Quantification

Lan Huang[1],Peter R. Baker[1],Robert J. Chalkley[1],Kirk Hanson[1], Nadia P. Allen[2],Michael Rexach[2],A.L. Burlingame[1].

[1]UCSF, San Francisco, CA, United States, [2]Stanford University, Stanford, CA, United States.

With the recent completion of several genomes and advances in new technology, large scale protein identification and quantitation has become technically feasible. These developments are essential to further our understanding of protein contents, protein abundance changes and their biological functions under different biological conditions in a given cultured cell line or organism. Thousands of mass spectra can easily be obtained in a short time (hours) using the automated data acquisition platform associated with various mass spectrometers. The development of integrated and automated data processing systems for protein identification, characterization and quantification is obviously the next challenge for dealing with such large mass spectrometric data sets for database searching and generating confident results with minimal human intervention. Protein Prospector (http://prospector.ucsf.edu) developed in our laboratory is a proteomics software package specifically designed for mining protein sequence databases in conjunction with mass spectrometric data. In order to deal with the large data sets generated by LCMSMS analysis, a general strategy is described in Figure 1. After automated data extraction, thousands of MSMS (with or without MS spectra) generated from large data sets of single or multiple LC MSMS runs are submitted to the **LCBatch Tag** program (combined MS-Tag and MS-Fit program) for simultaneous and automatic database searching and unambiguous protein identification at high speed (~mins.). In the **LCBatch Tag** program, the extracted data files with the peak lists are further filtered through various data processing options before searching. During the database searching, multiple search options can be set for flexible database searches. In addition, comprehensive summary results derived from the **LCBatch Tag** search displays the identified protein hits and their corresponding LC MSMS run information. The detailed peptide results are displayed by the **MS-Match** program, which provides a list of the identified peptides associated with each of the unique protein hits and each peptide sequence is linked directly to the **MS-Product** program which displays a graph showing the MSMS spectra along with the matching theoretical fragment ions for that particular sequence.

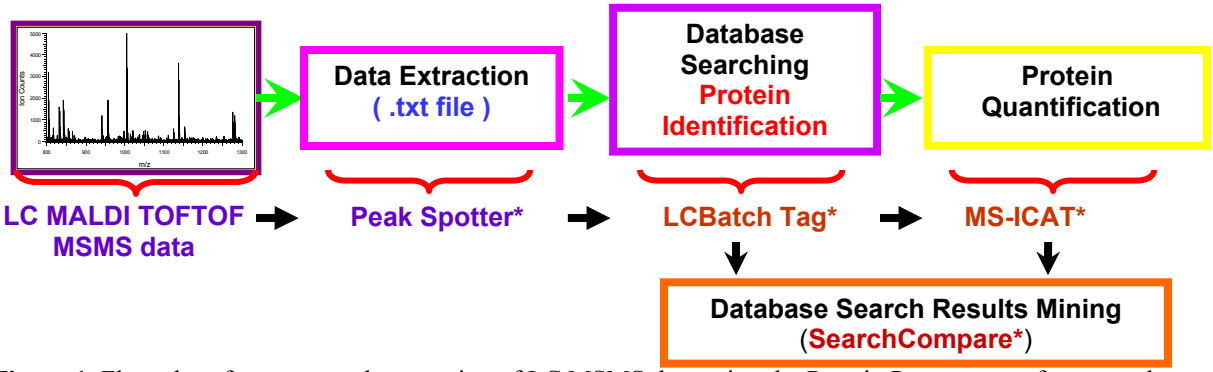

Figure 1. Flow chart for automated processing of LC MSMS data using the Protein Prospector software package.

Recently quantitative proteomics has been carried out using isotope-coded affinity tag (ICAT) technology and mass spectrometric analysis. Quantitative changes of protein abundance under different conditions are evaluated based on the mass spectrometric profile changes of the ICAT labeled peptides (ICAT ratio). **MS-ICAT Batch** was developed to automatically calculate the ICAT ratios for all the ICAT labeled peptide pairs found in a protein which were identified by **LC-Batch Tag** database searches. Both the peak area and intensity ratios of ICAT labeled peptides from the first three isotopes of an MS peak profile can be calculated automatically. The program corrects for the intrinsic ICAT ratio between the light and the heavy peptide caused by the presence of the C13 in the heavy peptide. A detailed report allows graphical evaluation to explain anomalous ratios caused by contaminant isotope profiles or calibration errors.

The newly developed **SearchCompare** program has been specifically designed for comparing and contrasting different search results obtained from different search engines (Protein Prospector and Mascot), as well as different instruments (QSTAR or MALDI TOFTOF) (Figure 2) and for comprehensive results summary, comparison and validation to facilitate protein identification on a large scale. Multiple search results can be compared at both the protein and peptide level. Comprehensive protein or peptide summary reports can be displayed and sorted by various parameters. Links from the matching peptide sequences to MS-Product facilitates protein identification and characterization from large data sets.

In recent years, we have made great efforts to study *S. cerevisiae* nucleoporin interacting proteins on a large scale using various mass spectrometric approaches to understand the mechanism of nucleocytoplasmic trafficking. The *S. cerevisiae* nuclear pore complex (NPC) is a supramolecular assembly of twenty-nine nucleoporins that cooperatively facilitate nucleocytopalsmic transport. Thirteen nucleoporins that contain FG peptide repeats (FG Nups) are proposed to function as stepping stones in karyopherin-mediated transport pathways, but the actual mechanism of protein transport across the NPC is unknown. Here, protein interactions that occur at individual FG Nups were sampled with immobilized nucleoporins and yeast extracts at various check points during the cell cycle. Previously we have identified hundreds of nucleoporin interacting proteins by the 1-D Gel/MS approach using high throughput technology[1]. In order to obtain complimentary results, the 2-dimensional liquid chromatography(SCX-Nanoflow RPLC)- MSMS approach has also been employed to compare the nucleoporin interacting protein contents at various check points during the cell cycle and ICAT technology has been utilized to monitor their relative abundance changes. MALDI-TOFTOF and QSTAR have been coupled with a nanoflow LC system for LCMSMS data acquisition. The raw LCMSMS data were automatically extracted, transferred and submitted for database searching using the approach shown in Figure 1. The comprehensive summary of the searched results provides both detailed protein and peptide reports for protein identification and characterization and in addition the protein relative abundance changes (ICAT ratios) for quantification. The results have allowed us to further our understanding of the transport mechanism and its dynamics.

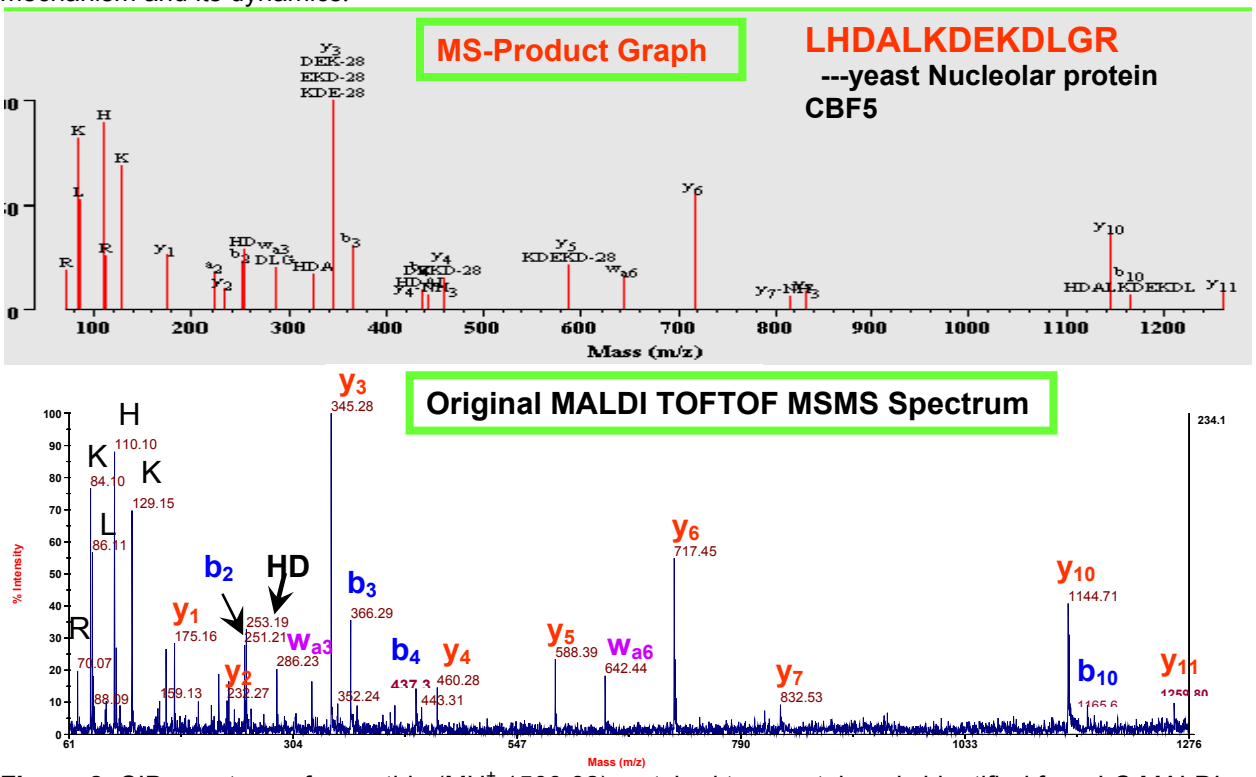

Figure 2. CID spectrum of a peptide (MH^+ 1509.82) matched to a protein only identified from LC MALDI TOFTOF MSMS based on the **SearchCompare** results of LC QSTAR MSMS vs. LC MALDI TOFTOF MSMS.

This work was supported by NIH grants NCCR 01614 and RR14606.

1. Huang L, *et al. Mol Cell Proteomics*. **2002**, 1(6):434-50.