Maximizing proteomic information from MS data: Enhancements to Protein Prospector, a suite of programs for mining genomic databases.

Richard J Jacob, Peter R Baker, Lan Huang, Michael A Baldwin and A.L. Burlingame.

Mass Spectrometry Facility, University California, San Francisco, CA 94143-0446

Protein Prospector is a suite of tools to interrogate genomic databases that is freely available via the UCSF Mass Spectrometry Facility website. Since its inception in 1995 it has continually improved in both breadth and functionality. Here we document enhancements and additions to the next public release and we describe current research on the development of several new tools. It is a selection of CGI programs written in C++ with an HTML interface. It can be served from either Windows NT or Sun Solaris webservers. All of the computational work was carried out within this web server framework. MALDI peptide mass maps were obtained from a Voyager DE STR reflectron TOF mass spectrometer (PE Biosystems, Framingham). MS/MS data came from a PE Sciex QSTAR (PE Sciex, Concord, Ontario) equipped with a nano ElectroSpray source (Protana, Odense). Nanoflow LC was performed with an LC-Packings Ultimate chromatography system flowing at 250 nanoliters/minute with an LC-Packings 75 micron diameter C18 PepMap column. The gradient used was 5% B to 50 % B in 30 minutes where Solvent A was water with 0.05 % formic acid and Solvent B was acetonitrile with 0.04 % formic acid. For LCMS the Ultimate HPLC was coupled to the PE- Sciex QSTAR mass spectrometer equipped with a Sciex MicroIonSpray source.

An LCMS run can be acquired within less than one hour whereas the analysis of the data is a time consuming task taking many hours. A new application, MS-Centroid, was written for Protein Prospector to aid in the automation of LCMS data analysis, and of MS data in general. An Applescript was written to process LCMS data that was acquired with the QSTAR interfaced to an Apple Macintosh. The Applescript divided the chromatogram into 10 sections, saving raw MS data into text files on a Protein Prospector server running on a Windows NT PC. A script written in the perl language submitted each of the text files to MS-Centroid for centroiding and reduction to monoisotopic format. It then picked the peaks, saving them to individual files, generated a complete list of all the peaks picked, and removed duplicates and known trypsin and keratin contaminants. The resulting list was saved to a text file and submitted to MS-Fit for peptide fingerprint database searching. In this way data analysis was reduced to an automated task taking approximately 30 minutes per chromatogram.

Mass mapping is limited to searches for homologous proteins with conserved tryptic peptides. When this fails, peptide sequence information derived by collision induced dissociation (CID) and tandem mass spectrometry can be used to identify proteins in a genomic data base using the sequence tag approach; eg by using MS-Tag. Peptide sequences can be obtained by *de Novo* sequencing based on known fragmentation rules but such sequences are frequently incomplete or may contain ambiguities. MS-Homology combines the results from *de Novo* sequencing experiments on several peptides derived from the same protein in a single search. It allows a substantial degree of sequence variation between the sequences presented and remote homologs in the database. Sequences can be entered in a syntax similar to that developed to describe regular expressions. MS-Homology attempts to align the sequences entered against each protein sequence in the database in turn. The final score is calculated by adding together the scores for the individual peptide alignments. This facilitates the identification of a number of proteins, which may be unachievable with other methods.

MS-Bridge was developed to help assign disulfide links between proteins. It calculates both singly and multiply disulfide linked peptides, allowing for both inter- and intra-peptide linkages. After specifying the target protein and digest conditions, MS-Bridge calculates all the possible disulfide liked peptides that match the query peptide masses within a user definable mass window. MS-Bridge was used to help assign a disulfide bridge in Rabbit moysin heavy chain. Rabbit moysin heavy chain was submitted to CNBr digest and resulting peptides were analyzed by MALDI. The peptide fingerprint was searched

against the NCBInr database with MS-Fit and the majority of the peaks were assigned to the moysin heavy chain while a few were assigned to a contaminating protein, moysin light chain. The remaining unassigned masses were submitted to a search with MS-Bridge. Each of the masses that had one or more potential disulfide linked peptides was submitted to Nanoflow ES low energy CID MS/MS (figure 1). Of those peptides, one with a mass of 7835.03 Da could have only one possible disulfide linked peptide structure. The fragment ions were interpreted and found to match the theoretical prediction of the structure. MS-Bridge has now been extended to allow for chemical crosslinking agents in addition to disulfide links.

This work was supported by NIH NCRR 01614

Figure 1 MS-Bridge displays a predicted disulfide linked peptide for the mass 7853.03Da