BUPID: Probability-Based Protein Identification by Searching Sequence Databases Using Peptide Mass Fingerprint Data

Tong Weiwei (1,2), Mark E. McComb (1), David Perlman (1), Hua Huang (1), Peter B. O'Connor (1), Catherine E. Costello(1), Zhiping Weng (2),

(1) Cardiovascular Proteomics Center, Boston University School of Medicine. (2), Boston University, Boston, MA.

Introduction: Several software programs have now been developed for MS data interpretation by querying a sequence database with peptide masses obtained from the MS experiment. We present results from a new search algorithm; Boston University Protein Identifier (BUPID), a robust and accurate statistical model for protein identification using MS data. The algorithm offers a number of new features in comparison to presently used algorithms: 1. Using log-likelihood ratio as scoring function, the algorithm can better distinguish correctly assigned peptides from incorrect assignments. 2. Matching peaks with a background-dependent threshold offers more flexibility and accuracy than the traditional mass error window. 3. The statistical model is shown to provide equivalent/ better results in comparison to other models.

Methods: We use a log-likelihood ratio to calculate the probability that a protein is present in the sample. The model distinguishes two hypotheses: H0; that a set of peaks in the spectrum is generated by the random background; and HA; that the same set of peaks is generated by peptides corresponding to a specific protein. A peak is included in the set if the probability that it is produced by the protein is more significant than that it is otherwise produced by the random background. Final results are ranked by the E-value of their probability score using the sequence information of the protein.

Results: We compared the performance of the BUPID server and several other publicly accessible web-based database search engines. Peptide map data sets (consisting of MALDI MS and ESI MS data obtained on proteolytic digests of proteins) were obtained from existing ongoing projects at BUSM. The following parameters are used for the database search - Taxonomy: Human; Enzyme: Trypsin; Maximum Missed Cleavage: 1; Peptide Mass Error Tolerance: 0.2 Dalton. In an example data set (30 samples) obtained from the MS analysis of human blood samples, BUPID and MASCOT (Matrix Science) were used to characterize hemoglobin proteins by searching the peptide mass data against a local copy of the SwissProt database. The BUPID database search results had on average 27% more true positives in top 10 predictions in comparison to results from the MASCOT search. Within the top 100 predictions, BUPID showed 27% more true positives as compared to MASCOT. In addition, BUPID was able to identify all five commonly occurring human hemoglobin chains (α,β,γ,δ,ε) in 6 cases within the top 20 proteins identified while MASCOT succeeded in only one case. When using peaks with higher than 5% relative intensity, the spreads were 28% and 24% within top 10 and 100 predictions, respectively. A comparison was also made with 5 commonly used search engines. MALDI MS data was obtained on a tryptic digest of E2F1 protein. All five search engines tested returned similar results. Aldente (ExPASy proteomics server, Swiss Institute of Bioinformatics) found E2F1 as the top hit with p-Value equaled to 4.6E-09. BUPID found E2F2 as top hit with score equaled to 99.417922 (three times the score of second best hit). In Mascot, the score was 123 with an E-value equal to 2.6E-08. MS-Fit (ProteinProspector, UCSF) found E2F1 as the second hit, with MOWSE score1.298E+05. ProFound (PROWL, Rockefeller University) ranked E2F1 number 1, with a probability of 1.0E+00.

BUPID also provides various data visualizations tools that are found useful by many users, including combined view of a protein mixture, mass spectrum of shared or similar peptides in different proteins, etc.

Acknowledgements: This project was funded by NHLBI contract N01 HV-28178.

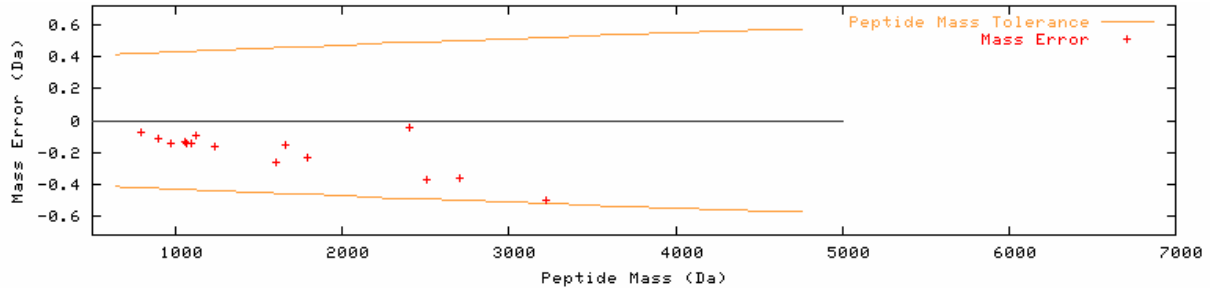
Supplementary Information: Log-likelihood ratio is calculated as

$$score = \ln \frac{likelihood\ H_A}{likelihood\ H_0} = \ln \frac{li(spectrum\ data \mid predicted\ \ protein)}{li(spectrum\ data \mid random\ background)}$$

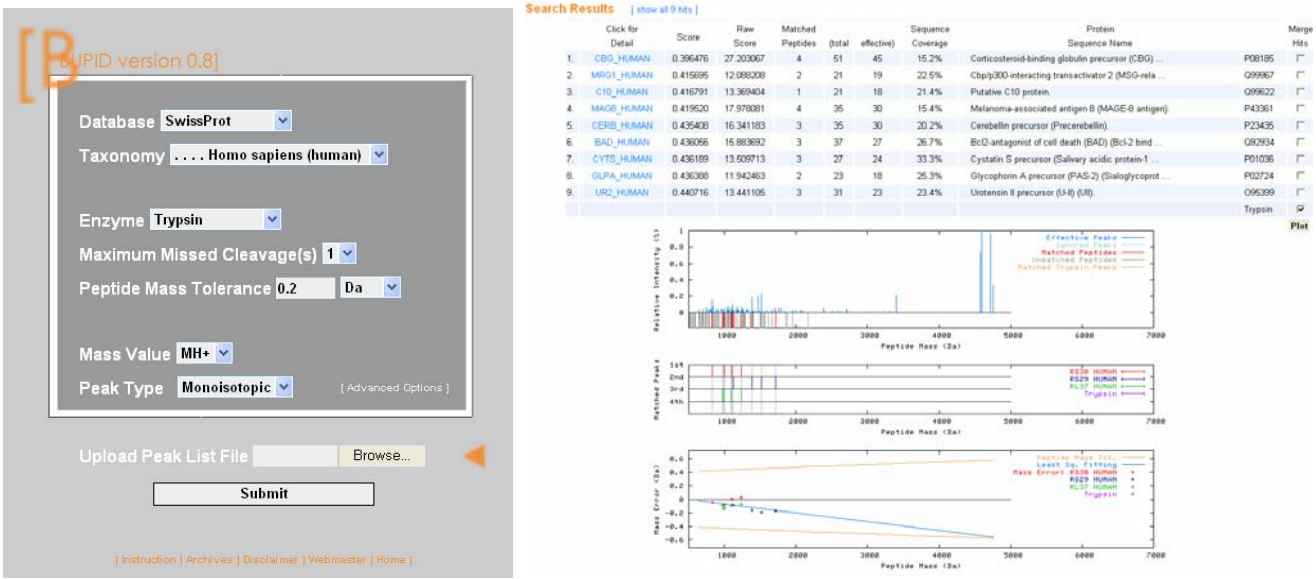
where

$$li(spectrum\ data \mid prediction) = \prod_{all\ \ peaks\ in\ the\ spectrum} [li(peak \mid peptides)] = \prod_{all\ \ peaks\ in\ the\ spectrum} \left[\prod_{all\ \ peptide\ in\ the\ \ protein} [li(peak \mid peptide)] \right]$$

A peak is considered a match if the likelihood of the peak generated by a specific peptide is greater than the background noise (log-likelihood ratio > 0) otherwise it is a miss-match. Automatically, the log-likelihood ratio is maximized if and only if all matched peaks and no miss-matched peaks are included in the calculation.



The effective “peptide mass tolerance” in BUPID varies according to both the mass of the peptide and number of peaks in the spectrum, as shown in the figure. Due to the non-uniform background, the log-likelihood ratio of a match is affected by both the difference and the absolute mass of the peak-peptide pair. The threshold for “matches” rises as peptide mass becomes larger. Additionally, the more peaks a spectrum has, the easier it is to obtain a match and a high score. Thus the background probability is adjusted according to the number of peaks in the spectrum. This serves as an internal quality-control scheme.



BUPID offers a standard interface and an expert interface (with advanced options). The server is expected to be available to the public online in the summer of 2005.