Peptide Identification in Whole-Sample Mass Spectrometry Proteomics

Richard Pelikan[1,3], Milos Hauskrecht, PhD[1,2,3]
[1]Intelligent Systems Program; [2]Department of Computer Science; [3]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA

Abstract

Peptide identification for whole-sample mass spectrometry (MS) proteomics is in its infancy. While sophisticated tandem MS/MS instrumentation exists for accurate peptide identification after sample separation, there are few options for those who produce data from intact protein samples. We present a novel algorithm which uses available information from the literature and online protein databases to provide reliable labeling of features in whole-sample MS proteomic data.

Introduction

Mass spectrometry (MS) is a popular tool for discovery of surrogate biomarkers for various diseases. MS profiles display relative abundances of molecules detected in the sample. The identification of these molecules is a necessary step for the conclusion of the discovery phase; the measured ion species are accompanied only by a mass-to-charge (*m/z*) ratio.

Current methods for biomarker identification use tools to search databases for telltale fragmentation patterns after a set of molecules is isolated and further analyzed. However, this process is cost and time-intensive, and may never consider the quantitative aspect of the data when proteins are left intact.

Methods

We propose a new protein identification method that attempts to find the most probable assignment of protein labels to prominent peaks in the whole-sample MS data. The method relies on the information about the expected location of a peak for a protein or its modification in the spectra and the information about its expected abundance (intensity) in the specimen. Labels must fit the criteria of a good match to both the location and intensity aspect simultaneously. The dynamic programming technique is devised and applied to find the most probable assignment of labels to peaks, in a fashion similar to sequence alignment.

Results

We preliminarily tested our method on data simulated from a virtual MALDI-TOF mass spectrometer[1]. Using only the location aspect to assign labels, the method achieves very low precision. By incorporating knowledge about relative abundance and using the intensity aspect, we are able to improve the precision of the labeling procedure as a tradeoff for sensitivity

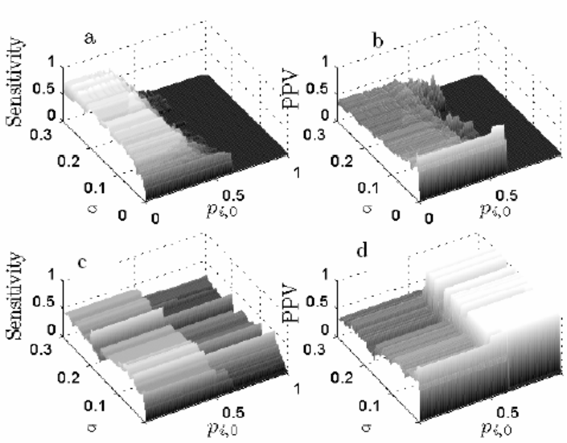

Figure 1. Sensitivity and Precision (PPV) of the method when using only the location aspect (a, b) and when combined with the intensity aspect (c,d).

Conclusion

With the addition of relative abundance information, our peak labeling procedure became more reliable. Incorporating additional information can help to improve the protein labeling procedure.

Acknowledgement

This research was supported by an NLM Training grant.

References

1. Morris, J. S., Coombes, K. R., Koomen, J., Baggerly, K. A., and Kobayashi, R. (2005). Feature extraction and quantification for mass spectrometry in biomedical applications using

the mean spectrum. Bioinformatics, 21(9), 1764–1775