

# Using Multivariate Curve Resolution to Improve Proteomics Biomarker Discovery

Li Chen<sup>1</sup>  
chenli@bii.a-star.edu.sg

<sup>1</sup> Bioinformatics Institute, 30 Biopolis Street, #07-01 Matrix, Singapore 138671

## Abstract

This poster describes a novel proteomic pattern analysis algorithm for biomarker discovery using MALDI-TOF or SELDI-TOF mass spectrometry. The algorithm is based on the combination of Multivariate Curve Resolution with classification methods for the detection of potential biomarkers. It is validated on two datasets from the literature.

**Keywords:** Mass Spectrometry; MCR-marker; Multivariate Curve Resolution; Biomarker Discovery; Proteomic Pattern Analysis

## 1 Introduction

Proteomic Pattern analysis plays an important role for biomarker discovery. Feature selection and learning algorithms have been widely used to define an optimal discriminatory proteomic signature using samples of known diagnosis. This pattern is then used to predict the identity of masked samples. The precise  $m/z$  values or  $m/z$  bins are often used as proteomics biomarkers or discriminatory patterns to classify biological samples in various cancer studies. However, the lack of reproducibility of these discriminatory patterns has been a major criticism.

## 2 Method and Results

A novel proteomic pattern analysis algorithm is proposed for biomarker discovery using MALDI-TOF or SELDI-TOF mass spectrometry. The algorithm (MCR-marker) is based on the combination of **Multivariate Curve Resolution** with classification methods for the detection of potential biomarkers. The MCR-marker algorithm applies singular value decomposition to select differentially expressed  $m/z$  windows. In each selected  $m/z$  window, MCR is applied to decompose experimental mass spectra into pure component spectra and their relative concentration profiles. The significant difference of concentrations between known diagnostic groups can be used as an indicator of a potential biomarker. It is found that such potential biomarkers show better performance than the precise  $m/z$  values or  $m/z$  bins.

Two datasets are used to examine the performance of the MCR-marker algorithm. The first dataset is the mycobacteria MALDI-TOF dataset [1]. The second dataset is an ovarian cancer SELDI-TOF dataset

(4/3/02 low resolution MS) that is downloaded from the Clinical Proteomics Program Databank [2]. For mycobacteria dataset, a previous study [1] reported unique biomarkers were observed only for species, but no strain-specific biomarkers could be identified. MCR-marker is applied to this dataset. Three species-specific and twelve strain-specific biomarkers have been successfully identified for discrimination of mycobacteria at the strain level. For ovarian cancer dataset, five discriminatory patterns are found with 72% sensitivity and 86% specificity in distinguishing ovarian cancer patients from controls. The identified potential biomarkers are not dependent on the selection of MCR methods and they consist of clearly detectable peaks, which may represent identifiable proteins, protein fragments or peptides.

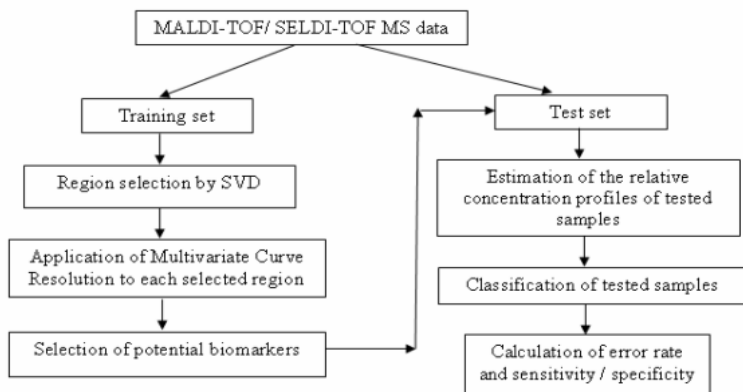


Figure 1 The workflow of the MCR-marker algorithm

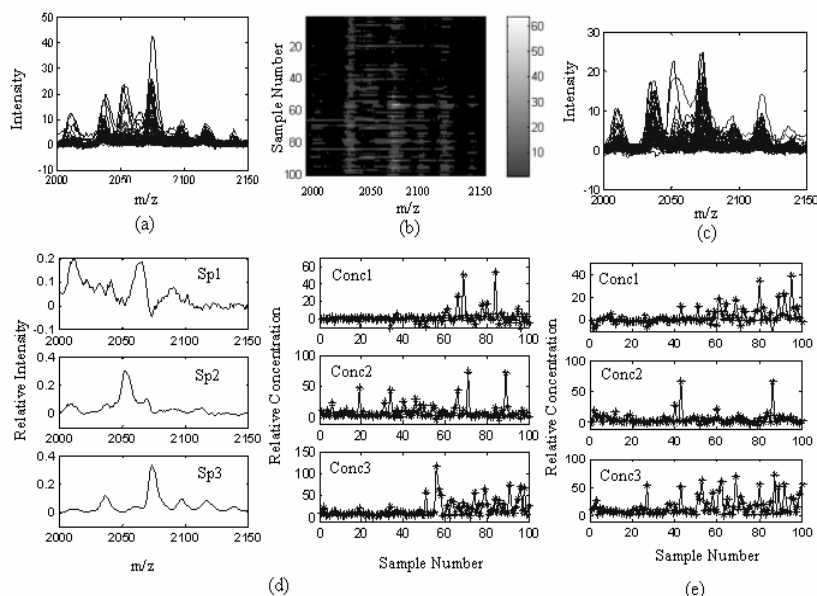


Figure 2 The results of multivariate curve resolution in the selected  $m/z$  window [2000.8 2149.6] from the ovarian cancer dataset. (a) Training spectra, the first fifty are ovarian cancer samples and the last fifty are controls. (b) The intensity image of training spectra. (c) Test spectra. (d) Pure component spectra and their relative concentration profiles of training spectra by SIMPLISMA. (e) The estimated relative concentrations of test spectra. Sp3 is selected as a potential discriminatory pattern because their relative concentrations between controls and cancer samples are significantly different.

## References

- [1] J.M. Hettick, M.L. Kashon, J.E. Slaven, Y. Ma, J.P. Simpson, P.D. Siegel, G. N. Mazurek, D.N. Weissman, *proteomics*, 6: 6416-6425, 2006
- [2] E.F. Petricoin, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, L.A. Iotta, *The Lancet*, 359: 572-577, 2002.