

Biomarker discovery from mass spectral profiles – a chemometric approach

Tarja Rajalahti
Haukeland University Hospital/
University of Bergen, Norway



"Every attempt to employ mathematical methods in the study of chemical questions must be considered profoundly irrational and contrary to the spirit of chemistry..."

If mathematical analysis should ever hold a prominent place in chemistry – an aberration which is happily almost impossible – it would cause a rapid and widespread degeneration of that science."

Auguste Comte, *Philosophie Positive* (1830)

Contents

This work presents a new method for biomarker detection in complex spectral profiles.

The method is validated by spiking cerebrospinal fluid (CSF) with peptide standard at different concentration levels.

The method is applied for whole mass spectral profiles from patients with multiple sclerosis (MS) and orthopedic patients (controls).

What do we want to do?

Find new biomarkers to improve early diagnosis of neurological diseases, e.g. multiple sclerosis.

Tools

Spectral profiling using MALDI-TOF (*matrix-assisted laser desorption ionization time of flight*) mass spectrometry

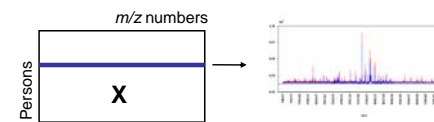
Multivariate analysis using latent variable methods, e.g. partial least squares (PLS) regression and target projection (TP)

Selectivity ratio (SR) plot

How to do it?

Collect samples (body fluid) from one group of healthy persons (controls) and one group of persons with a certain disease (patients).

Subject the samples from each person to experimental work-up and analyze with a mass spectrometer, i.e. each person is represented by a multivariate profile of intensities in a chosen mass-to-charge (m/z) range and with a certain m/z resolution.



How to do it?

Pretreat the spectral profiles to correct for non-compositional variation

- baseline effects
- small shifts in m/z number (alignment)
- concentration differences (normalization)
- noise structure (heteroscedasticity, i.e. noise increasing with signal size)

Search for the m/z regions that discriminate between controls and patients by decomposing the pretreated matrix \mathbf{X} using latent variable methods to obtain scores \mathbf{T} and loadings \mathbf{P} .

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E}$$

Latent variable methods

Partial least squares discriminant analysis (PLS-DA)

Create a response \mathbf{y} : a vector of zeros and ones for controls and patients, respectively.

Calculate a PLS regression model (spectral data matrix as \mathbf{X}) with optimal predictive performance to obtain the regression coefficients \mathbf{b}_{PLS} .

Target projection (TP)

Project the spectral data matrix onto the regression vector.

This gives a single latent variable (the target-projected component) that explains the covariances between the spectral variables and the response.

The target component represents the axis of optimal discrimination between controls and patients.

Selectivity ratio

The variance explained ($v_{\text{expl},i}$) by the target component can be calculated for each spectral variable i and compared with the residual variance ($v_{\text{res},i}$) for the same variable. The ratio between explained and residual variance, called the selectivity ratio (SR_i), represents a measure of the ability of a spectral variable to discriminate two groups of samples (e.g. controls from patients).

$$SR_i = v_{\text{expl},i} / v_{\text{res},i}$$

High SR means that the variable in question has an excellent ability to separate controls from patients.

When calculated for each spectral variable, the ratio can be displayed similarly to a spectrum and can be used directly for biomarker detection without extensive expertise in multivariate analysis.



Sampling

Cerebrospinal fluid (CSF) samples were taken by a standard lumbar puncture procedure from patients at the Haukeland University Hospital (Bergen, Norway).

About 10 ml of CSF was immediately placed on ice before freezing at $-80\text{ }^{\circ}\text{C}$.



Simulating the pathogenesis by spiking

The CSF pool was divided into four 500 μL aliquots:

SP0 – no spiking (reference sample)

SP1 – 400 pM peptide standard

SP2 – 800 pM peptide standard

SP3 – 1600 pM peptide standard

Peptide standard m/z value	Name
1047.20	angiotensin II
1297.51	angiotensin I
1348.66	substance P
1620.88	bombesin
2094.46	ACTH clip 1-17
2466.73	ACTH clip 18-39
3149.61	somatostatin 28



MALDI-TOF data

Samples were fractionated using standard fractionation protocol and spotted three times each on the MALDI target and analyzed using an AutoFlex (Bruker Daltonics) MALDI-TOF mass spectrometer in the low molecular weight proteome range ($740\text{-}9000\text{ Da}$; described by 44 403 m/z values).

This will give a dataset consisting of 36 spectral profiles (3 fractionations spotted 3 times for SP0, SP1, SP2, and, SP3).

Ref: Berven, F.S. *et al.* Proteomics Clin. Appl. 2007,1, 699-711.



Data pretreatment

Baseline correction by FlexAnalysis software from Bruker Daltonics.

Removal of regions with negative intensities by shifting the profiles independently by the absolute value of the largest negative intensity in each profile.

Smoothing by moving average using a 10 point window.

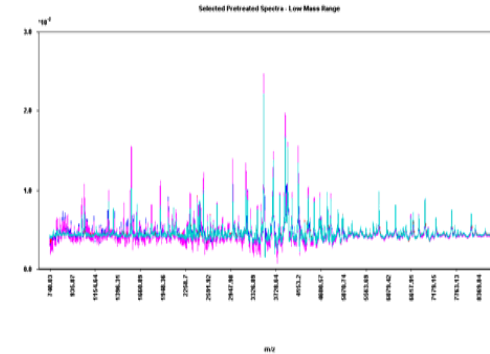
Alignment using a window size of 20.

Transform by square root to compensate for structured noise and create a homoscedastic noise structure.

Normalization to unit length.

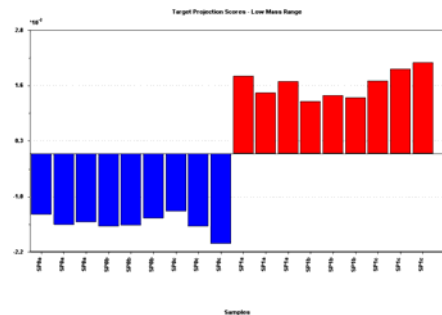
Ref: Arneberg, R. *et al.* Anal. Chem. 2007, 79, 7014-7026.

Pretreated MALDI-TOF profiles



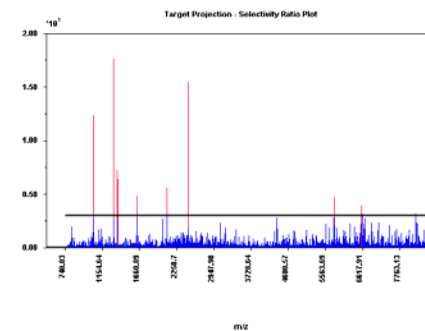
Reference sample and three different concentration levels for added peptide standard (740-9000 Da, 44403 m/z numbers).

Target projection scores (0 pM vs 400 pM)

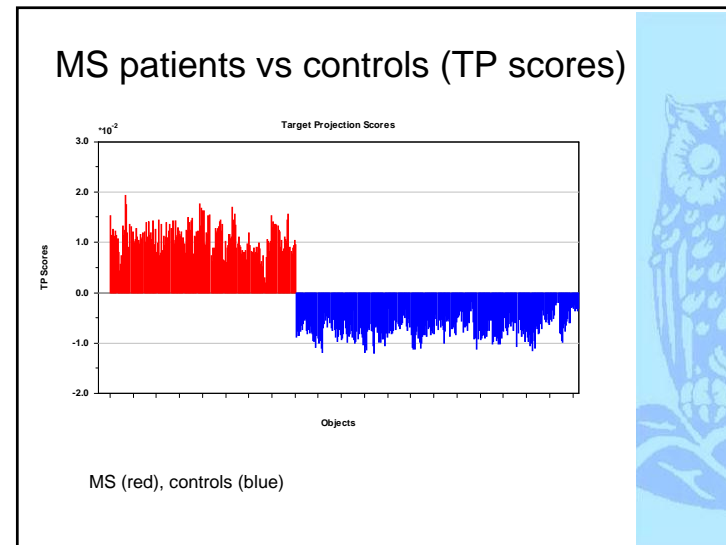
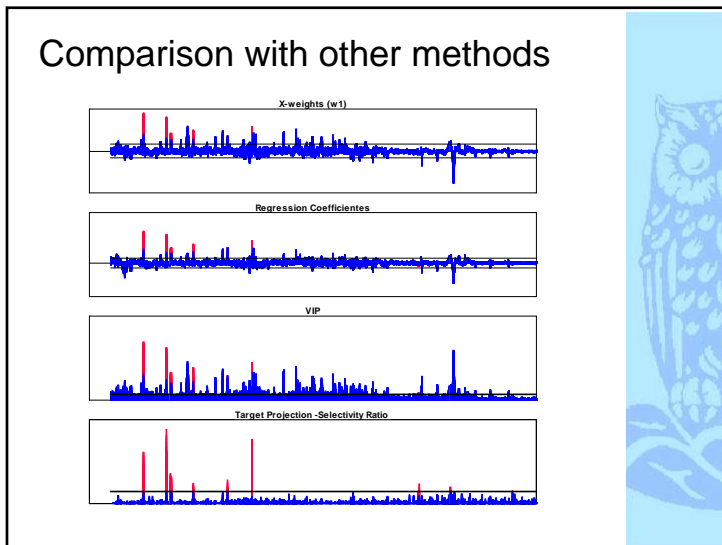
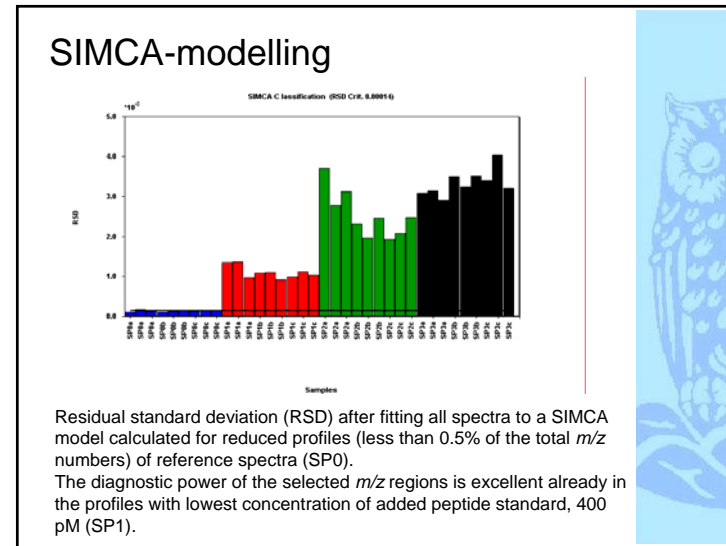
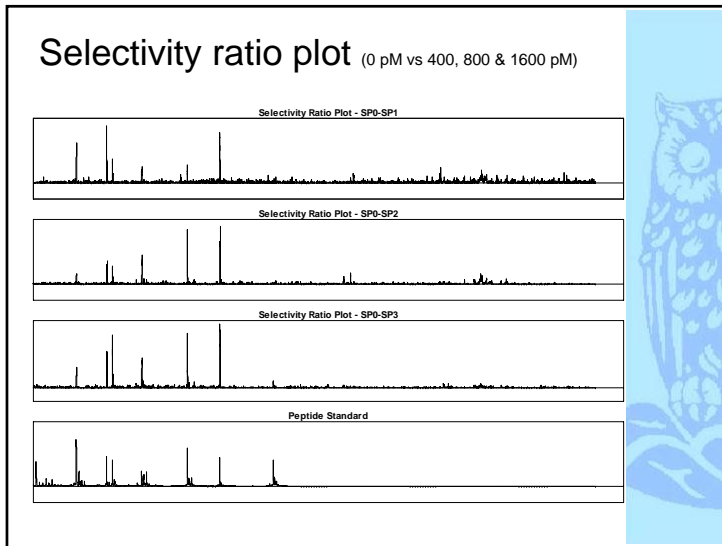


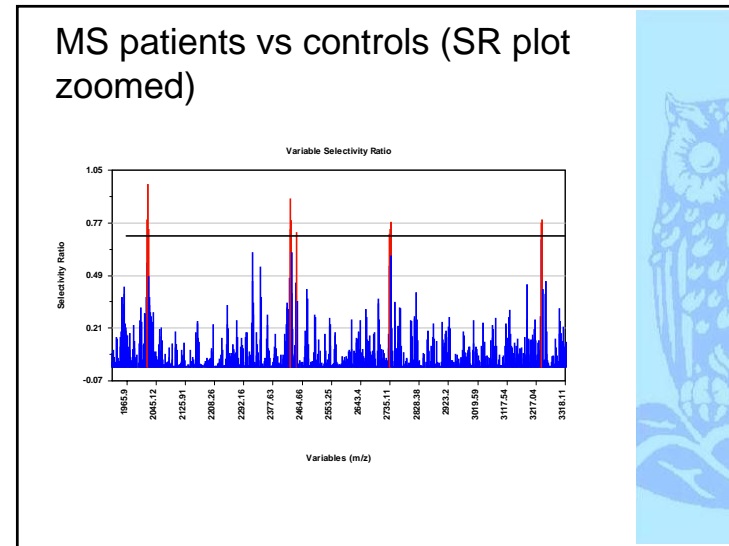
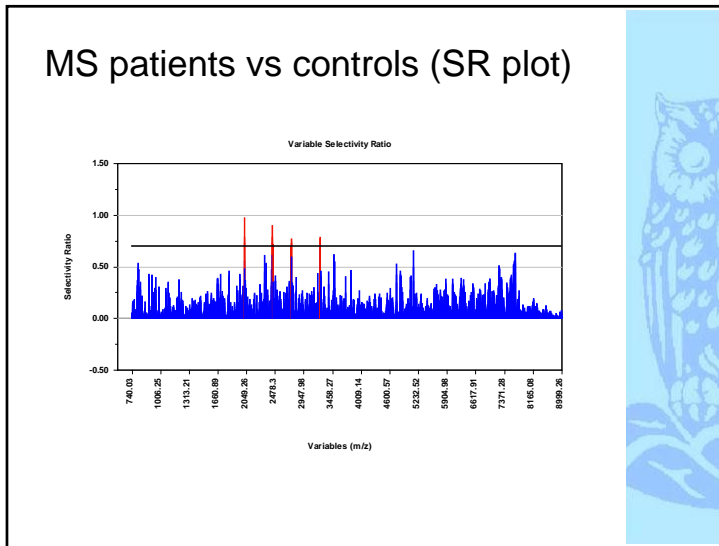
Scores on target-projected component (a three-component PLS-DA model rotated to the best single component model) for reference sample (blue) and sample spiked with 400 pM of peptide standard (red).

Selectivity ratio plot (0 pM vs 400 pM)



A cut-off ratio of three (corresponding to 75% explained variance in a spectral variable) is marked by the black horizontal line. All m/z numbers with a selectivity ratio exceeding this limit (marked in red) are possible biomarker candidates





Summary

The pathogenesis of a disease was simulated by adding known peptide standard to a CSF reference sample and then revealed the differences in composition in the spectral profiles by using a new approach to biomarker detection.

Target projection can provide a univariate scale with diagnostic value from multivariate spectral profiling of complex samples.

The great advantage of target projection, is that the discriminating part of the spectral profile is represented on a single loading vector. This simplifies interpretation.

Biomarker candidates are easily and reliably identified using this approach. It is also shown to reduce the risk of finding false candidates.

Co-workers

Prof. Kjell-Morten Myhr and prof. Christian A. Vedeler, Dept. of Neurology, Haukeland University Hospital, Bergen

Prof. Rune J. Ulvik, Laboratory of Clinical Biochemistry, Haukeland University Hospital, Bergen

Prof. Olav M. Kvalheim, Dept. of Chemistry, University of Bergen

Reidar Arneberg, Pattern Recognition Systems AS

PhD Frode Berven and master students, Proteomics Unit at University of Bergen (PROBE)

Thank you for your attention!



Tarja.Rajalahti@kj.uib.no

11th Scandinavian Symposium on Chemometrics
(SSC11)

June 8-11, 2009
Hotel Alexandra, Loen/Stryn, Nordfjord, Norway



www.kj.uib.no/ssc11