A method for estimating the number of peaks in liquid chromatography-mass spectrometry data sets

T. Szczepińska[1] S. Piersma[2] M. Codrea[1] J. Heringa[1] E. Marchiori[1]

[1] Centre for Integrative Bioinformatics VU (IBIVU), VU University Amsterdam
[2] Cancer Center Amsterdam, VU University Amsterdam

Background: Liquid chromatography-mass spectrometry (LC-MS) has become a key technology for comparing biosamples. It allows to broadly survey the peptide or protein constituents of the samples and hence it provides tremendous opportunities for biomarker-related clinical applications. In a LC-MS system peptides are subjected to liquid chromatography separation and then each fraction is analysed by a mass spectrometer. The resulting spectra consist of one intensity measurement for each pair of molecular mass-to-charge ratio (m/z) and retention time (RT) values. Computational comparative analysis of the LC-MS data is a challenging task, due to the high dimensionality and complexity of the data [1]. Crucial steps of the analysis are peak detection and alignment, in order to group together peaks generated by the same peptide but detected in different samples.

Method: We developed a method for estimating the number of peaks, to be used in algorithms for simultaneous peak detection and alignment based on clustering. The method consist of the following two main steps:

1. Estimate the number of peaks on each sample run separately by means of an unsupervised machine learning technique.

2. Merge the results across the runs using suitable criteria to extract the final number of peaks, which correspond to the signals of the molecules.

Results: The method has been embedded in a clustering based simultaneous peak detection and alignment (PDA) algorithm [2] and has been tested on variability mixtures of known proteins whose concentrations were designed to change between the two mixtures (data set prepared for the test of a platform for marker discovery [3]). The results substantiate the utility of the proposed method for peak detection and alignment with LC-MS data.

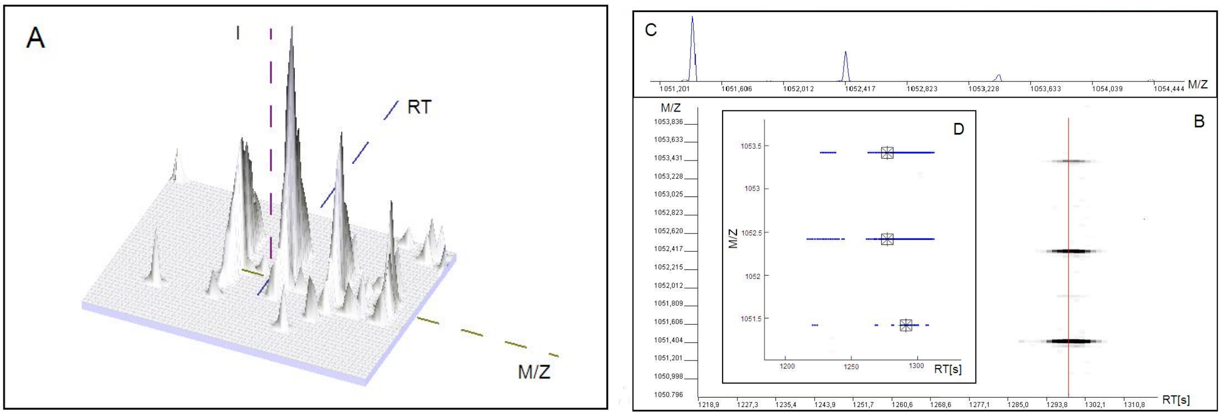

Figure 1: (A) 3D view of a raw data from a single run with m/z on x-axis, retention time on y-axis and intensity on z-axis; (B) Zoomed-in 2D map of a raw data from a single run; (C) Mass spectrum for selected retention time; (D) Peaks from the same region identified from 50 runs, squares indicate local maxima with the highest intensity within peaks

References

[1] Listgarten J and Emili A. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol Cell Proteomics*, 2005.

[2] M.C. Codrea, C.R. Jimenez, S. Piersma, J. Heringa, and E. Marchiori. Robust peak detection and alignment of nanolc-ft mass spectrometry data. In *Proceedings of The Fifth European Conference on Evolutionary Computation, Machine Learning and Datamining in Bioinformatics, EvoBIO'07, Valencia*, 2007.

[3] Jaffe JD, Mani DR, Leptos KC, Church GM, Gillette MA, and Carr SA. Pepper, a platform for experimental proteomic pattern recognition. *Mol Cell Proteomics*, 2006.