Identifying Differentially Expressed Proteins by a Binary Threshold Model

Sebastian Gibb and Korbinian Strimmer

Institute for Medical Informatics, Statistics and Epidemiology (IMISE),
University of Leipzig,
Härtelstr. 16-18, D-04107 Leipzig, Germany.

Abstract. Proteomics-based analyzes using mass spectrometry are becoming routine in clinical diagnostics, for example to monitor cancer biomarkers using blood samples. For preprocessing mass spectrometry data a wide variety of algorithms is available. Accordingly, we have recently compiled a standard analysis pipeline that includes methods for baseline correction, normalization and alignment of spectra in the freely available R package MALDIquant [1, 2]. However, identifying differential expression and ranking of proteomic features remains challenging due to the simultaneously discrete (absence-presence) and quantitative nature of protein expression.

Here, we present a simple yet effective approach using a binary threshold model for ranking and identifying differentially expressed proteins. This approach works by peak-wise data-adaptive thresholding of protein expression and subsequent ranking of the dichotomized features using a suitable entropy measure. Our framework may be viewed as a generalization of the ‘peak probability contrast’ approach of [3] and works, in contrast to [3], both in the two-group and the multi-group setting. Using data from a recent pancreas cancer study conducted at the University of Leipzig [4] we are able to identify biological relevant and statistically predictive marker peaks unrecognized in the original analysis.

Keywords: Differential expression, proteomics, mass spectrometry, peak ranking, cancer biomarker.

References