

A Statistical Method for Significant Analysis of Comparative Proteomics Based on LC-MS/MS Experiments

Pei Wang, Brian Piening, Martin McIntosh, Amanda G. Paulovich

One of the essential challenges of comparative proteomics using LC-MS/MS instrumentation is the need to assess quantitative differences between two samples using only qualitative information, such as the presence or absence of a peptide in one or another type of samples. In this paper, we propose a statistical method SASPECT (significant analysis of peptide counts) for quantitatively identifying proteins differentially expressed between two groups based on the tandem mass spectral experiment results.

SASPECT employs the commonly used “spectral-count” assumption: the probability of a protein’s being observed in one LC-MS/MS experiment is proportional to its abundance in the complex sample. However, in contrast to spectral counting, SASPECT uses the Boolean values of whether the spectral count is greater than zero instead of the raw values of spectral counts, for the latter are more subjected to the changes of various experimental factors. In addition, by properly controlling the false discovery rates (FDR), SASPECT provides quantitative guidance in peptides and proteins selection.

Compared with other similar approaches in the microarray literature (searching for differentially expressed genes), the LC-MS/MS problem has several additional sources of variability. To address these challenges, an Expectation-Maximization (EM) model is developed in SASPECT to infer differential level of proteins based on peptides’ observations, and at the same time to account for various sources of variability. The model makes explicit use of peptide probability assignments (e.g. peptide prophet scores) to account for database search errors. It employs a rescaling factor to remove the artificial effect due to the limitation of total CID number in each LC-MS/MS experiment. The permuted null distribution of test statistics is used to estimate the FDR. The performance of the algorithm is illustrated through searching for putative biomarkers in biological samples. The approach is provided freely in an open-source program.