MassQuest: a comprehensive proteomic pipeline for biomarker discovery in different MS platforms

Karin Noy[a], Fadi Towfic[b], Gayle M. Wittenberg[a] and Daniel Fasulo[a]

[a]Integrated Data System Department, Siemens Corporate Research, 755 College Road East, Princeton, NJ 08540, USA. [b]Computer Science Department, Iowa State University, Ames, Iowa 50011-1040

Using mass spectrometry (MS) for proteomic analysis is still challenging due to the large dynamic range and the high complexity of the molecules of interest. In addition, there are problems related to variations between the biological samples, resolution and accuracy of the MS instruments, and limitations in finding low-abundance molecules. As a result, independent MS studies to date have failed to reproduce the same biomarkers. All of these problems can be greatly reduced by improving existing analytical tools.

We present MassQuest, a Java package that provides a comprehensive pipeline for the analysis of large-scale proteomic data obtained from different MS platforms. It is important to note that by different platforms we also mean low- and high-resolution 1D MS as well as 1D and 2D MS. The main theme in this pipeline is to rely upon robust nonparametric methods when simple assumptions about the data cannot be made, yet also incorporate prior knowledge about the data and mass spectrometer when applicable.

MassQuest includes multiple components with different algorithms. First, a low-level signal processing component can be applied to 1D data for all MS platforms. Second, peptides are extracted in 1D by a model-based, platform-independent feature extraction algorithm, together with their monoisotopic mass, intensity, and charge information. Finally, matching of the corresponding peptides across multiple MS data is performed by clique finding and optimization methods. For liquid chromatography mass spectrometry (LC-MS) data, the feature extraction step is augmented with a robust estimation method specific to LC-MS feature extraction. Local alignment based on a nonparametric kernel-type regression estimator is used to align the multiple LC-MS maps prior to matching.

We tested the algorithms implemented in MassQuest on different data and MS platforms and compared the results with the current state-of-the-art software. Using our approach, we found more true features in 1D MS platforms such as SELDI, MALDI and even tandem MS. In 2D MS, we found more features, more matches, and better correlation between replicates in LC-MS experiments.