Deconvolution of Complex Tandem Mass Spectra of Intact Proteins: A Combinatorial Approach

Xiaowen Liu¹, Yuval Inbar¹, Pieter C. Dorrestein², Puneet Souda³, Julian P. Whitelegge³, Vineet Bafna¹ and Pavel A. Pevzner¹

¹Department of Computer Science and Engineering, UCSD
²Department of Pharmacology, Chemistry and Biochemistry, UCSD
³Pasarow Mass Spectrometry Laboratory, NPI-Semel Institute, UCLA

Abstract  Top-down proteomics is a mass spectrometry-based approach for analyzing protein sequences and their post translational modifications (PTMs). In top-down spectra, fragment ions of the same chemical formula and charge state are represented by an isotopomer envelope. Spectral deconvolution focuses on grouping spectral peaks into isotopomer envelopes. We present MS-Deconv, a combinatorial algorithm for spectral deconvolution. The algorithm generates a large set of candidate envelopes and efficiently selects a high scoring subset of them using a dynamic programming approach. The algorithm employs an intensity-split scoring model that allows peaks to be assigned to multiple envelopes. We demonstrate MS-Deconv improves on the widely used Thrash software tool. In addition, MS-Deconv is an order of magnitude faster than Thrash.

We study both the envelope selection problem without sharing peaks and the envelope selection problem with sharing peaks. For the second problem, a peak can be assigned to several envelopes, which is different from most common approaches. We assume the contribution of each peak to the score of an envelope is independent, and define a scoring function between a peak and a set of assigned envelopes. We construct a directed graph based on the candidate envelopes. We prove that the two versions of envelope selection problem can be reduced to the problem of finding a heaviest path in the constructed graph, which can be solved in polynomial time using a dynamic programming approach.

We compared MS-Deconv with Thrash on a collection of 6 CAD spectra of intact proteins with known protein sequence. Since Thrash outputs a list of mono-isotopic masses that are not explicitly assigned to envelopes, we compared them based on the output mono-isotopic mass lists only. Given the protein sequence from which the spectrum is acquired, we generate a mono-isotopic mass list of theoretical ions. Each list consists of the mono-isotopic masses of the most abundant ions plus mass offsets. For each mass in the output mass list we assume it is from a true positive envelope if nearest theoretical mass value is within a given PPM error tolerance. We ran Thrash with the default r-value 0.9, as well as with several other values. The numbers of output envelopes of MS-Deconv were designated as an input parameter. Prior to the evaluation we performed the same recalibration procedure for both Thrash and MS-Deconv results. The comparison shows that MS-Deconv improves upon Thrash in the number of true positive envelopes by 10% on average. The software is available as a web-based tool at NIH/NCRR Center for Computational Mass Spectrometry at UCSD.