Quantifying Peptide Signal in MALDI-TOF Mass Spectrometry Data*

Timothy W. Randolph‡§, Bree L. Mitchell¶, Dale F. McLerran∥, Paul D. Lampe¶∥, and Ziding Feng

This study addressed the question of which properties in MALDI-TOF spectra are relevant to the task of identifying mass and abundance of a peptide species in human serum. Data of this type are common to biomarker studies, but significant within- and between-spectrum variabilities make quantifying biologically induced features difficult. We investigated this signal content and quantified the existence, or lack, of peptide-induced signal (as manifest in a multiresolution decomposition) by generating spectra from human serum in which the abundance of peptides of specific masses is controlled by a sequence of dilutions. The intensities of the corresponding features were directly proportional to peptide concentration. The primary goal was to exhibit some quantifiable properties of raw spectra from this application of MALDI-TOF mass spectrometry. Although no recommendations are given regarding the best method for processing these data, the results confirm the utility of a simple method, based on wavelets, for defining and quantifying features related to low abundance peptide species in a heterogeneous set of complex spectra. Estimates on lower limits of detectable peptide abundance (in the 20-nmol range) and on the number of features present in a spectrum are made possible by the controlled experimental design, the use of a large external reference data set, and dependence on relatively few modeling assumptions. *Molecular & Cellular Proteomics* 4:1990–1999, 2005.

Despite the apparent success of some early biomarker studies based on mass spectrometry TOF data, the use of this technology for biomarker discovery remains controversial (1–3). Some of the debate revolves around the issue of signal content and which properties of a spectrum are relevant for use in classifying samples in case/control studies. There is, at present, no consensus on a preferred method for distinguishing signal from noise or on which properties of the spectra are truly relevant for use in inferring peptide mass and abundance. This stage of analysis (often referred to as preprocessing) is important as it impacts all subsequent analysis of these data (4).

Peaks in a spectrum are typically used as indications of peptide content in the sample with relative abundance of a particular species of (ionized) peptide corresponding to the height, or volume, of a peak. Identifying peaks (or overlapping peaks) and interpreting intensities is complicated by high frequency “noise” along with global and local trends. In addition to these within-spectrum variabilities, a collection of spectra typically exhibits substantial between-sample heterogeneity in noise level and base-line intensity as well as the existence or nonexistence of biological features. Moreover there is a variable amount of mass shifting inherent in the spectrometer that results in the property that features from different spectra rarely align at exactly the same TOF value.

Studies related to biomarker discovery using mass spectrometry TOF data have been based on a variety of assumptions regarding signal-to-noise ratio, spectrum base line, and the offset of features between spectra (5–8). The analysis in Baggerly et al. (9) moreover reveals a variety of non-biological content in a set of spectra from a MALDI-TOF spectrometer. For a brief survey on current approaches to processing and analyzing TOF data, see Coombes (10).

Because of the potentially rich spectrum obtained from a single serum sample, the methods and techniques used, from the first stages of sample collection to the final stages of data analysis, are likely to have considerable impact on the outcome. The aim in this study was to provide a benchmark for understanding which of this information is biologically relevant. We emphasize that this is not purely a laboratory question: what is found in the data is dependent on the assumptions made when processing the data. One extreme is to impose strict assumptions on the amount of offset between peaks from different spectra, on modeling a global baseline, on window widths in which peaks are sought, and on noise content and the amount a peak must exceed the noise before it is recorded. Tacit in these assumptions is that the notion of a peak is formally defined, but such a definition is not obvious for MALDI-TOF spectra. The other extreme is to accept every TOF measurement as informative and use all (e.g. 50,000 or more) unadjusted intensities in a comparison of samples. In investigating properties in MALDI-TOF spectra that are relevant to the task of identifying mass and abundance of a peptide in a complex serum sample, our goals were 2-fold: 1) to exhibit the utility of a method for defining and quantifying peptide-related features that imposes minimal assumptions...