

STATISTICAL APPROACHES FOR DIFFERENTIAL EXPRESSION IN LC/GC-MS DATA

Mark D. Robinson,^{1,2} and Terence P Speed¹

¹*Bioinformatics Division, Walter & Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia 3050 and*

²*Department of Medical Biology, University of Melbourne, Parkville, Victoria, Australia 3050*

Proteomics and metabolomics studies using liquid or gas chromatography (LC/GC) coupled to a mass spectrometer (MS) are now becoming mainstream approaches. Often the goal of such an analysis is *difference detection*: to determine the peptides or metabolites that are significantly different between experimental conditions or classes of patients, sometimes known as biomarker discovery. Our work focuses on unlabeled and gel-free experiments, where there is replication.

There are now a myriad of methods and software available for processing LC/GC-MS data, which can be separated into two categories: peak-based or signal-based approaches. Peak-based methods extract “features” from the raw MS data and all operations (e.g. alignment, normalization) are performed on this discrete subset of the data, with the hope that important aspects are not missed through the feature selection step. In contrast, signal-based methods operate directly on the mass spectra, or some simple transformation of it (e.g. binned matrices of m/z by retention time). Treating the data in this manner loses no information, and in fact, it may improve our ability to find true differences.

Since many of the statistical methods developed for gene expression analysis can be applied to peak-based feature tables, we focus on the determination of differences with signal-based processing. We investigate various statistical approaches, such as peak finding on a matrix of t -statistics, or model-based methods to find differential peaks.

The methods are benchmarked using two datasets: a high-resolution designed spike-in experiment and a low-resolution public dataset that has been previously analysed.