COMPARING FALSE DISCOVERY RATE ESTIMATION METHODS FOR USE IN THE ANALYSIS OF MASS SPECTROMETRIC PROTEOMIC PROFILING EXPERIMENTS

David A. Cairns[1]

[1]*Section of Oncology and Clinical Research & Section of Epidemiology and Biostatistics, Leeds Institute of Molecular Medicine, Leeds, UK*

Proteomics is one of the complementary group of “-omics” disciplines which have the potential to provide a wealth of information regarding a patient’s diagnosis, prognosis and response to treatment through the discovery of new biomarkers (measurable biological traits which are associated with e.g. the onset or progression of disease). Proteomic profiling experiments are a mechanism for the discovery of such biomarkers in the proteome, the protein complement of a system. These experiments are observational in nature in that they compare samples of body fluids or tissue from, for example, cancer cases and controls, using technology that will separate out the proteins or peptides from the biological sample and give some measure of their mass and abundance. One such technology used for this procedure is mass spectrometry (MS). The two most common MS platforms for profiling are matrix assisted laser desorption/ionisation time-of-flight (MALDI-TOF-MS) and surface enhanced laser desorption time-of-flight (SELDI-TOF-MS). These platforms produce megavariate data in spectral form representing the measured mass and abundance of the molecule. Standardly some form of data-reduction or feature extraction is performed which reduces the number of variables to the order of hundreds rather than tens of thousands. This step has a biological basis as it is thought that the peaks reflect the biologically relevant entities and also a statistical basis as it is a heuristic method of removing some of the correlation in spectra.

In proteomic profiling experiments which are costly both in terms of expenditure and the use of valuable biological samples, it is important to extract as much information as possible. This means that traditional methods of assessing significance in a multiple testing situation such as the Bonferroni correction are not appropriate as they are overly conservative and assume independence. However, there has as yet not been wide use of other methods of assessing significance in proteomic profiling. A recent paper by Karp *et al*.[1] has highlighted the importance of careful analysis in minimising the number of false discoveries while simultaneously maximising the potential of these experiments by advocating the use of the *q*-value[2]. In this paper we evaluate a number of the available methods for controlling and/or estimating false discovery rate (FDR) in proteomic profiling experiments undertaken by MS. These include both parametric and non-parametric approaches from the literature based on the Simes inequality, permutation, empirical Bayes and the optimal discovery procedure[3]. Discussion of the way in which such disparate methods with numerous tuning parameters can be compared will be described. This comparison is undertaken using a novel simulation model for proteomic mass spectra which mimics the biological dependence that can be observed in peaks. These methods are evaluated in terms of empirical estimates of power, FDR and the estimated proportion of truly null proteins in conjunction with estimates of computational expense. Although no one method is shown to be universally preferable some methods are shown to perform more poorly than might be anticipated.

1. Mol Cell Proteomics 6:1354-64, 2007
2. JRSS B, 64:479-498, 2002
3. Biostatistics 8:414-32, 2007