Bull. Aust. Math. Soc. **94** (2016), 345–346
doi:10.1017/S0004972716000502

STATISTICAL ANALYSIS OF PROTEOMIC MASS SPECTROMETRY DATA FOR THE IDENTIFICATION OF BIOMARKERS AND DISEASE DIAGNOSIS

TYMAN E. STANFORD

(Received 6 May 2016; first published online 16 August 2016)

2010 *Mathematics subject classification*: primary 62P10; secondary 62H30, 62H35, 62J12.

Keywords and phrases: mass spectra, proteomics, biomarkers, image analysis, pre-processing, missing values, signal smoothing, baseline correction, normalisation, peak detection, peak alignment, linear models, supervised learning.

Proteomic spectra obtained from matrix-assisted laser desorption ionisation (MALDI) time-of-flight mass spectrometry (TOF-MS) are generated from the proteins and peptides present in serum obtained from blood. By ionising the proteins and resolving them in the mass spectrometer, data on the expression of proteins can be obtained, realised from the amplitude of the signal for different mass to charge ratios. Of primary interest is the biological signal, in particular the expression of proteins related to disease [2]. In common with many 'omic' technologies, the raw spectra suffer from systematic errors due to technological artefacts and batch effects, in addition to sample and biological variability [1]. To negate these effects, novel applications of genetic microarray pre-processing and analysis methods to proteomic TOF-MS data are presented. However, there are important differences between microarray and TOF-MS data which require consideration and nontrivial modifications to be successfully applied. One important difference between MALDI TOF-MS data and other high-throughput data, seldom addressed, is the high proportion of missing values.

The pre-processing of raw proteomic TOF-MS data needs to be undertaken prior to analysis and remains a mathematical and statistical challenge. Performed in distinct steps, pre-processing consists of signal smoothing, baseline correction, spectra normalisation, peak detection and peak alignment. An argument as to why the order of these steps is highly important is presented. Standard and novel data pre-processing methods are investigated and compared to optimise the process. Each step is given due consideration since the cumulative effects of substandard pre-processing can render subsequent statistical analysis highly unreliable.

Thesis submitted to The University of Adelaide in August 2015; degree approved on 3 November 2015; supervisors Patty Solomon and Christopher Bagley.

Ultimately, the aim of proteomic MS is to analyse the protein profiles. Two different but related approaches to the analysis are undertaken. The first approach is to identify biological markers (biomarkers) that exhibit differential expression between disease groups [3]. Identifying potential biomarkers for further research requires appropriate exploratory, visual and statistical modelling, which is addressed in detail here. The second approach is to perform statistical discrimination between groups, a classical supervised learning problem. The ability of mathematical models to predict disease groups using differential biological signals provides insight into the plausibility of diagnostic tests. Methodologically, supervised learning is a multifaceted problem given that feature selection, model parameter optimisation and the handling of the training and test data all contribute to the inference that can be made from the results. Empirical appraisal of the methods applied to the proteomic data is provided with the outcome of discrimination error as a quantitative benchmark.

A number of proteomic TOF-MS datasets with differing characteristics are used throughout this thesis to assess the validity of the methods presented. The detailed analysis of a murine model MALDI TOF-MS dataset has facilitated the discovery of potential biomarkers for gastric cancer. Correct classification of spectra to their respective disease groups (gastric cancer or control mice) as high as 97.4% was achieved using supervised learning. The thorough treatment of all the differently behaved datasets contained in this thesis, starting from the raw data pre-processing steps through to the challenging process of identifying potential biomarkers, provides a comprehensive and best-practice pipeline to analyse real-world proteomic MS data.

Acknowledgements

My sincere gratitude goes to my supervisors Professor Patty Solomon and Dr Chris Bagley, whose guidance and wisdom made my PhD dissertation possible. I would also like to acknowledge the influence of my two generous and kind grandmothers, Pauline and Rita, who have always been exceedingly supportive. My PhD candidature was financially supported by an Australian Postgraduate Award scholarship.

References

[1] J. Albrethsen, 'Reproducibility in protein profiling by MALDI-TOF mass spectrometry', *Clin. Chem.* **53**(5) (2007), 852–858.

[2] G. L. Hortin, 'The MALDI-TOF mass spectrometric view of the plasma proteome and peptidome', *Clin. Chem.* **52**(7) (2006), 1223–1237.

[3] V. Kulasingam and E. P. Diamandis, 'Strategies for discovering novel cancer biomarkers through utilization of emerging technologies', *Nat. Clin. Pract. Oncol.* **5**(10) (2008), 588–599.

TYMAN E. STANFORD, School of Mathematical Sciences,
The University of Adelaide, SA 5005, Australia
e-mail: tyman.stanford@adelaide.edu.au