Identifying Biomarkers from SELDI-TOF Protein Profiles: An Integrative Approach

Yaping Cheng[1,2], Stephen O. Nyangoma[2], Philip J. Johnson[2]

[1] Centre for Epidemiology and Biostatics, University of Leeds, UK

[2] Cancer Studies, Medical School, University of Birmingham, UK

The biomarkers identified from surface-enhanced laser desorption/ionisation time-of-flight (SELDI-TOF) mass spectrometry data have potential in early cancer detection and in selecting therapeutic regimens for cancer patients. However, it is challenging to identifying these biomarkers due to the complexity and diversity of protein patterns. Many statistical approaches, including univariate and multivariate feature selection methods, have been employed to extract informative protein features. However, different approaches often result in identification of different sets of proteins. None of the previous studies address which methods should be used for identifying protein biomarkers from SELDI protein profiles.

We propose a general algorithm for biomarker discovery from SELDI protein profiles by integrating feature selection methods and classification approaches. Three feature selection methods, the Wilcoxon test, receiver operating characteristic and the random forest, and three machine learning techniques (support vector machine, random forest, and k-nearest neighbour) are applied to publicly available prostate and ovarian cancers datasets. The results show that there is no overlap between proteomic features detected by different methods, and that 72% of the features are common between methods, for the prostate cancer dataset, and only 49% are common for the ovarian cancer dataset. By applying the proposed algorithm, 18 proteomic biomarkers have been found for the prostate cancer dataset, and 39 for the ovarian cancer dataset.

Feature selection methods for identifying protein biomarkers from SELDI datasets are not robust, which indicates that different feature selection methods should be used complementarily for a given proteomic dataset. The proposed algorithm could be useful in the identification of the optimal set of protein biomarkers from SELDI data sets.