Definition of Valid Proteomic Biomarkers: Bayesian Solutions to a Currently Unmet Challenge

Keith Harris
Glasgow University

Abstract

Clinical proteomics is suffering from high hopes generated by reports on apparent biomarkers, most of which could not be later substantiated via validation. This has brought into focus the need for improved methods of finding a panel of clearly defined biomarkers. To examine this problem, urinary proteome data was collected from healthy adult males and females, and analysed to find biomarkers that differentiated between genders. We believe that models that incorporate sparsity in terms of variables are desirable for biomarker selection, as proteomics data typically contains a huge number of variables (peptides) and few samples making the selection process potentially unstable. This suggested the application of the two-level hierarchical Bayesian probit regression model that Bae and Mallick (2004) proposed for variable selection, which used three different priors for the variance of the regression coefficients (inverse Gamma, exponential and Jeffreys) to incorporate different levels of sparsity in their model. We have also developed an alternative method for biomarker selection that combines model based clustering and sparse binary classification. By averaging the features within the clusters obtained from model based clustering, we dene "superfeatures" and use them to build a sparse probit regression model, thereby selecting clusters of similarly behaving peptides, aiding interpretation.