Development of a statistical variable selection method to identify useful biomarkers in metabolomic profiling

Seri KITADA[1)] and Shuji KAWAGUCHI[2)]
Faculty of Medicine, Kyoto University[1)] and Center of Genomic Medicine, Kyoto University Graduate School of Medicine[2)]

The multi-omics approach plays an essential role in elucidating unrevealed links between genotypes and phenotypes. In particular, the high-throughput quantitative metabolomics has widely been applied to probe into underlying molecular mechanisms of diseases and other biological systems. However, it is challenging to thoroughly identify significant factors from high-dimensional data consisting of a large number of metabolome profiles. Therefore, a robust statistical approach is required to obtain an interpretable and beneficial outcome such as a diagnostic classification model or a biomarker discovery.

When constructing a model to identify biomarkers in a case-control study, a mathematical model that clearly explains disease pathology with a small subset of variables is preferred to high discrimination performance with a larger number of parameters. Analytical methods such as principal component analysis, partial least squares regression, support vector machine, and random forest are commonly used in the field for their data visualization capacity or high classification performance. These approaches, however, do not fulfill the above-mentioned requirement of variable selection. On the other hand, the least absolute shrinkage and selection operator (LASSO) is a variable selection technique that performs sparse modeling of high-dimensional data. Nevertheless, even LASSO is not fully functional in selecting significant parameters or biomarker candidates when applied to metabolomic profiling data that often show strong inter-correlation between associated metabolites.

With this background, we developed a LASSO-based variable selection method to identify biomarker candidates from metabolomic profiling data. The technique successfully selected a small subset of biomarker candidate metabolites among a LASSO estimator by sequentially narrowing down the influential variables for classification. To evaluate its performance, we tried identifying biomarker candidates of eosinophilic esophagitis. We used plasma samples of 327 patients and 5,751 individuals in the Nagahama prospective genome cohort study as cases and controls, respectively. We quantified 268 lipids by liquid chromatography-mass spectrometry (LC/MS) and used them for classification analysis and variable selection with participants' age and sex as covariates. As a result, we succeeded in selecting two or more lipids as biomarker candidates from approximately 50 lipids in the resulting LASSO model without a considerable decrease (falling below 80%) of classification accuracy. The performance was verified with other rare diseases, further demonstrating the usefulness of the method.