Metabolome data mining of mass spectrometry measurements with random forests

Chihiro Higuchi *and Shigeo Takenaka †

Metabolome analysis is expected to become a leading technology for rapid discovery of novel biomarkers, which are key components for successful drug development. Nuclear magnetic resonance (NMR) and mass spectrometry (MS) are frequently employed as effective tools for metabolome measurements, and when it comes to analysis, the principal component analysis and the partial least square methods have been the methods of choice for mining of metabolome data.

In the present study, we have investigated the application of the random forests machine learning method (Breiman 2001) for analysis of metabolme data. The data comprised FT-ICR-MS measurements of urine from rats, which have been administered the antiarrhythmic agent amiodarone. Amiodarone treated rats will exhibit lipidosis and phenylacetylglycine (PAG) can be measured in the urine.

Unsupervised classification applied to these data with the random forests approach clearly separated the groups, that is before and after amiodarone treatment, and the separation was superior to that of the principal component analysis method. The supervised classification with the random forests approach furthermore suggested several class discriminating MS peaks, which were selected by the importance value generated by the random forests machine learning method. These MS peaks were assigned biomarker candidates and ranked by the loading values from the principal component analysis.

This analysis was carried out with the randomForest, amap and Heatplus packages of R 2.4.1 on Linux (kernel 2.6.21) operating system.

*Genomic Science Laboratories, Dainippon Sumitomo Pharma Co., Ltd., Suita-city, Osaka 564-0053, Japan

†Graduate School of Life and Environmental Sciences, Osaka Prefecture University, Sakai-city, Osaka 599-8531, Japan