A machine learning approach to generate high-quality metabolite and lipid profiles from mass spectrometry data

Shuji Kawaguchi, Ph.D.
Kyoto-McGill International Collaborative School in Genomic Medicine
Graduate School of Medicine, Kyoto University

Comprehensive measurement of biomaterials provides various information representing the dynamics of biomolecules in the body. Quantitative evaluation of metabolites or lipids may lead to identifying biomarkers associated with disease onset and pathological mechanisms.

Currently, mass spectrometry technology combined with liquid- or gas-chromatography (LC-MS and GC-MS, respectively) is a major means for such measurements because of their throughput and cost-effectiveness. However, mass spectrometry signals suffer from unexpected inconsistencies that result in poor experimental reproducibility and consequently lead to an erroneous data interpretation. The problem arises from intra-batch effects during the chemical reaction process and the fluctuation of the performance, or inter-batch effects due to different experimental conditions and analysis software parameter settings.

To circumvent these shortcomings, we have developed a novel machine learning-based quality control (QC) method of mass spectrometry data which corrects the signal inconsistencies. The developed algorithm detects un-differentiable points among the internal standard samples' signals and normalizes signal intensities with kernel-based regression without using batch information. We defined correction parameters with a sample set used to optimize the correction conditions only. Such an approach minimizes distortions and biases when the correction is separately made for case and control data sets because of the different MS profiles in the two groups. Once the parameters are defined, the comparative analysis can be done between any sample sets as far as they contain internal-standard samples in a common control group. In this talk, I will present signal normalization of ~300 lipids measured by LC-MS for 8,633 plasma samples of the Nagahama Study.