On nonlinear machine learning methods for Quantitative structure-retention relationships modeling in proteomics

Petar Zuvela, 유 준†, Katarzyna Macur[1], Tomasz Bączek[1]
부경대학교; [1]Medical University of Gdańsk
(jayliu@pknu.ac.kr†)

RP-LC-MS/MS is a powerful method widely used in proteomics. Here, proteins are broken into peptides and their spectra are matched with theoretical ones. Retention time is dependent on molecular structure. Its prediction is gaining increasing attention for simultaneous qualitative and quantitative profiling, with Quantitative Structure-Retention Relationships (QSRR) used for their prediction. Since more than 4000 molecular descriptors can be calculated, variable selection is crucial. It was shown that a Genetic Algorithm (GA) coupled with Partial Least Squares (PLS) was superior for developing QSRR models. However, it gave inadequate predictions for large peptides for which relationship between retention time and molecular structure is non-linear. In this work, machine learning methods: Support Vector Regression (SVR), Artificial Neural Networks (ANN), and kernel Partial Least Squares (kPLS) were compared in respect to their predictive ability. GA was used for variable selection. Final models: GA-SVR, GA-ANN, and GA-kPLS were constructed out of subsets of ten variables and compared to previously obtained results. They were also thoroughly validated and their applicability domain was defined.