Oral

Bringing machine learning into tile-based processing for complex GC×GC-TOFMS datasets

Pierre-Hugues Stefanuto* (1) Meriem Gaida (1) Caitlin Cain (2) Robert Synovec (2) Jef Focant (1)

1: OBiAChem Group Liège University Belgium; 2: Department of Chemistry University of Washington USA

In the quest of making multidimensional gas chromatography (MDGC) a robust method for untargeted screening of small molecules one of the key remaining challenges is the data processing meaning transforming raw data into pertinent information [1]. To reach this objective important analytical aspects such as data integrity workflow transferability and comparability need to be tackled. For MDGC the tile-based approach developed by the Synovec Group has gained popularity to conduct chromatogram comparison [2]. The success of this method relies on its large applicability its commercial availability and its lower computing power requirements.

The tile-based approach is mostly used in combination with Fisher Ratio Analysis (FRA). This analysis of variance (ANOVA) is performed in order to identify the tiles which display a significant difference between classes. FRA is a highly efficient univariate statistical test. The statistical background is easy to grasp and the link with the chemical information is easy to establish. Nevertheless FRA suffers for some limitations. The variance within the different classes should be equivalent the univariate dimension could underestimate the importance of correlated variables and it is not suitable for unbalance data set. Over the years several complementary methods have been tested to overpass the first limitation [3]. In this study we aim to tackle the two others. We have developed and evaluated a new processing approach combining tile-based image comparison and machine learning-based feature selection. Our approach is using the four-grid tiling scheme with random forest (RF) algorithm. RF was selected for its capacity to work with unbalanced data and its low dependency on the data pre-processing methods [4].

Nevertheless several challenges needed to be overcome for the implementation of such tool. The used of RF required the development of new feature selection metrics and the set-up of cross-validation protocols for the training and the testing of the workflow. We have combined out-of-bag error m/z purity and fold change calculation to evaluate our workflow output.

The study was conducted on a data set based on whole stool research grade test materials (RGTMs) prepared by NIST for interlaboratory studies. The RGTMs contain two diets vegan and omnivore and two sample formats liquid vs lyophilized. In this study we have focused on two groups: omnivore liquid and omnivore lyophilized. This data set was processed using the classical FRA approach and our RF workflow.

Our findings suggest that the RF approach is a promising method for identifying class-distinguishing analytes in settings characterized by both high between-class variance and high within-class variance making it an advantageous method in the study of complex biological matrices.

These promising results open the door to future investigations where other machine learning algorithms could be used for the feature selection. Moreover these RGTM data set will also be