QSPR model development to simplify compound identification in complex matrix analysis

Elyette Martin, Pavel Pospisil

Philip Morris International R&D, Philip Morris Products S.A., Quai Jeanrenaud 5, 2000 Neuchâtel, Switzerland (part of Philip Morris International group of companies)

Introduction

In order to assess and evaluate the toxicity of new products in a wide range of industrial settings (e.g., food and beverage, cosmeceutical industries), it is important to understand their chemical composition. Non-targeted screening of small molecules in complex matrices can be performed using various analytical techniques such as gas chromatography coupled to mass spectrometry (GC-MS). However, compound identification using a conventional mass spectral library search alone, e.g. NIST MS Search, generally does not provide sufficient confidence regarding the proposed structures.

The application of cheminformatics provides analytical chemists with tools to increase the accuracy for identifying compound structures and to accelerate and standardize the identification process. QSPR (Quantitative Structure-Property Relationship) models can be used to predict retention times (RT) or retention indices (RI) for all constituents potentially present in the complex matrix. These predicted retention times/indices may then be used to enhance the level of confidence in the correct assignment of compounds to determined mass spectra.

This poster presents QSPR models, which have been developed using different software algorithms, including ACD/ChromGenius, RapidMiner, Dragon, and Pipeline Pilot, and describes the improvement afforded by such tools for elucidating the chemical composition of complex aerosol matrices at Philip Morris International (PMI).

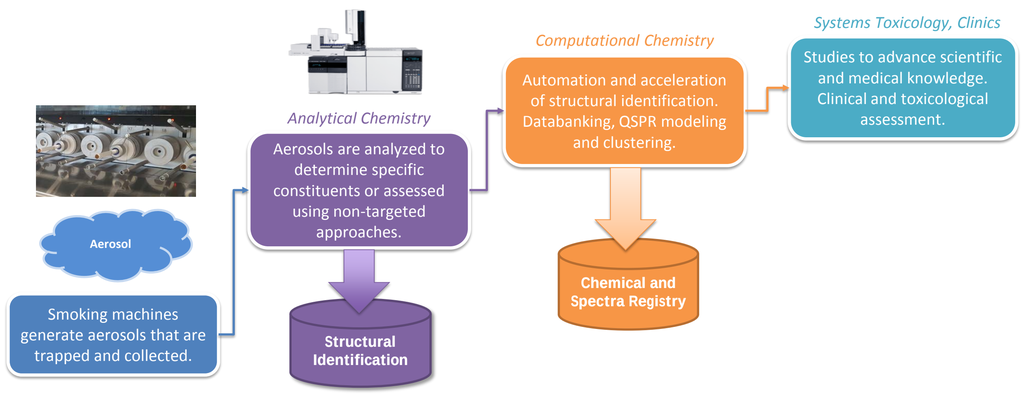


Figure 1. Aerosol analysis workflow at PMI

Methods

The first steps for QSPR modeling consisted of cleaning the chemical data and splitting the molecules into training and test sets. For this purpose a Biovia Pipeline Pilot (PP) protocol was developed:

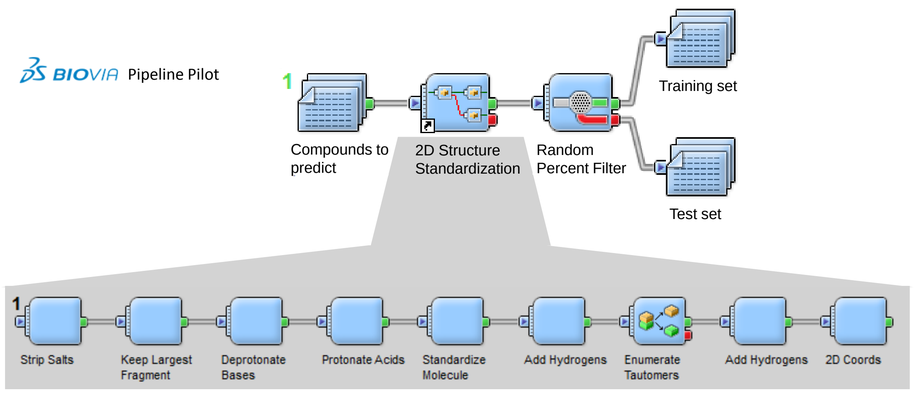


Figure 2. Pipeline Pilot protocol for standardization and splitting into training and test sets

Two different approaches were developed for two different gas chromatographic techniques:

- For two-dimensional gas chromatography with time-of-flight mass spectrometry (GCxGC-TOFMS), retention times for each dimension are projected on an x/y diagram. A Computer-Assisted Structure Identification (CASI) approach was developed at PMI to enhance the process for structural identification[1]. The confidence in correct identification is increased using the following process:

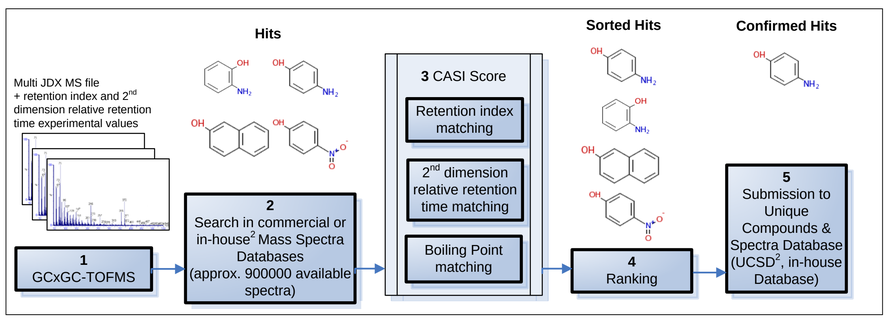


Figure 3. Enhanced identification of aerosol constituents using the CASI platform with GCxGC-TOFMS analysis

- For gas chromatography (single dimension) with mass spectrometry (GC-MS) or with high resolution mass spectrometry (GC-HR-MS), two different workflows were used to build QSPR models.

  Workflow A: Based on structural descriptors calculated using Dragon and algorithms using Pipeline Pilot and RapidMiner software

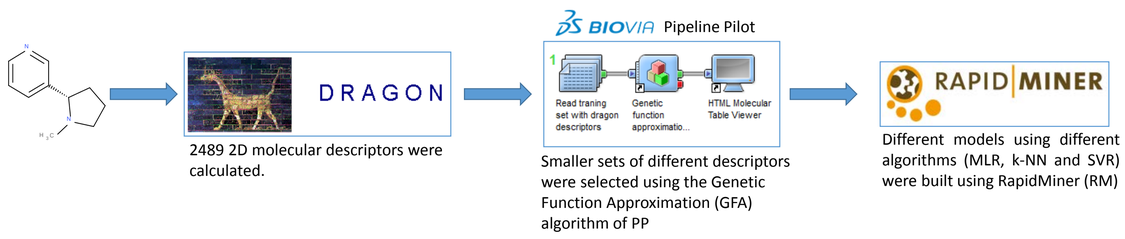
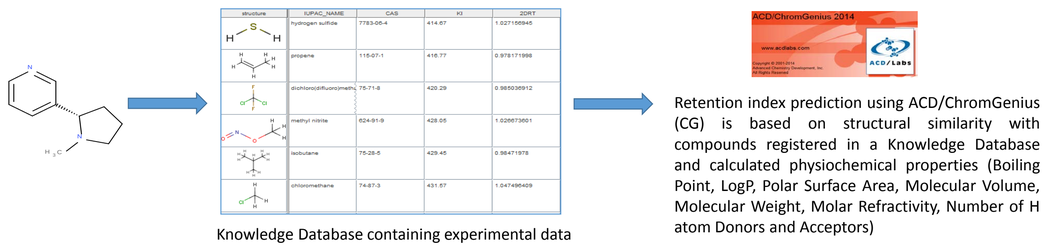


Figure 4. QSPR modeling workflow using descriptors calculated from the chemical structure

Workflow B: Based on structural similarity and physicochemical properties using ACD/ChromGenius software



Knowledge Database containing experimental data

Retention index prediction using ACD/ChromGenius (CG) is based on structural similarity with compounds registered in a Knowledge Database and calculated physiochemical properties (Boiling Point, LogP, Polar Surface Area, Molecular Volume, Molecular Weight, Molar Refractivity, Number of H atom Donors and Acceptors)

Figure 5. QSPR modeling workflow using structural similarity and calculated physicochemical properties.

Results

Several QSPR models have been developed using:

- different training and test sets, randomly generated (Figure 2)
- different algorithms (RapidMiner with MLR, k-NN and SVM in Workflow A (Figure 4) and ACD/ChromGenius in Workflow B (Figure 5))
- data from 3 different chromatographic instruments (GC-MS, GC-HR-MS and GCxGC-TOFMS) linked to different columns (volatile, non-polar and polar compounds)

In the following table, only models for a single method per instrument are shown. In each case, only the best combination of training and test sets leading to the best predictive models are presented.

Table 1: Selection of algorithm for producing models with highest correlations.

		Sets	Workflow A (RapidMiner)			Workflow B (ChromGenius)	
GC-HR-MS (volatile, semi-volatile)	LRI	TR: 400 TS: 151	**MLR 20 descriptors q^2=0.960**	k-NN 25 descriptors q^2=0.875	SVR all descriptors q^2=0.959	15 most similar structures r^2_{test}=0.976	GC-(HR)-MS approach
GC-MS (volatile)	LRI	TR: 183 TS: 82	**MLR 25 descriptors q^2=0.982**	k-NN 20 descriptors q^2=0.878	SVR 15 descriptors q^2=0.978	25 most similar structures r^2_{test}=0.963	
GCxGC-TOFMS (polar)	RI	TR: 98 TS: 28	**MLR 25 descriptors q^2=0.928**	k-NN 10 descriptors q^2=0.709	SVR 10 descriptors q^2=0.866	–	GCxGC-TOFMS approach (CASI platform)
	2DRT		**MLR 25 descriptors q^2=0.909**	k-NN 20 descriptors q^2=0.569	SVR 25 descriptors q^2=0.907	–	

Legend: GC-MS – gas chromatography mass spectrometry, HR – high resolution, GCxGC-TOFMS – two dimensional GC time-of-flight MS, LRI – linear retention index, RI – retention index, 2DRT - 2nd dimension retention time, TR – number of compounds in training set, TS – number of compounds in test set, MLR – multilinear regression, k-NN – k nearest neighbor, SVR – support vector regression, q^2 – cross-validation squared correlation (LMO – leave many out), r^2_{test} – squared correlation test set, bold font – best models.

Examples of correlations demonstrating excellent RI/LRI prediction capability:

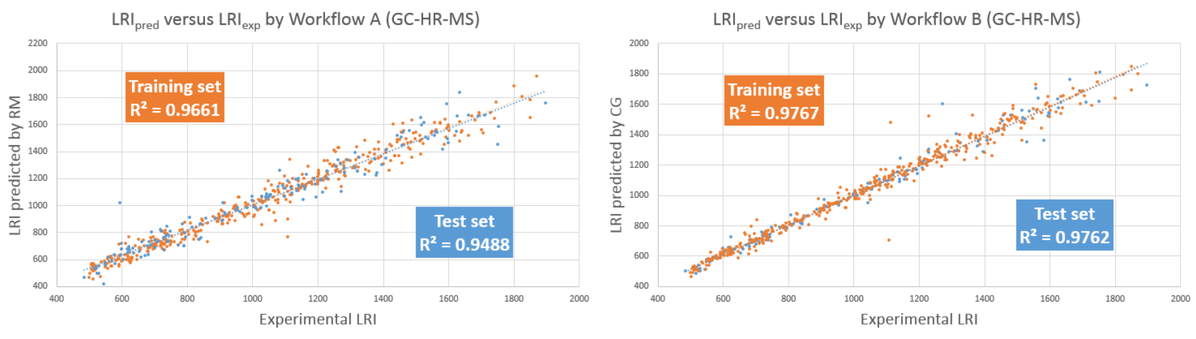


Figure 6. Correlation between experimental linear retention indices (LRI_{exp}) determined using GC-HR-MS (volatile/semi-volatile method) and LRI values predicted (LRI_{pred}) by Workflow A (left) and by Workflow B (right) .

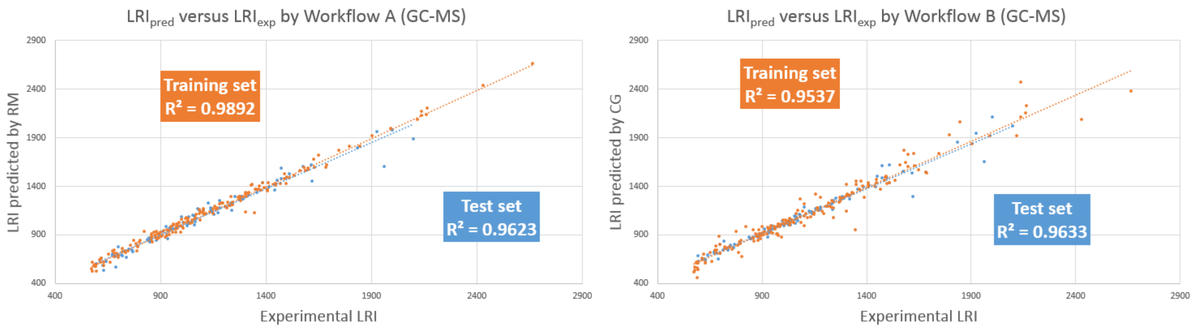


Figure 7. Correlation between experimental linear retention indices (LRI_{exp}) using GC-MS (volatile method) and LRI values predicted by Workflow A (left) and by Workflow B (right).

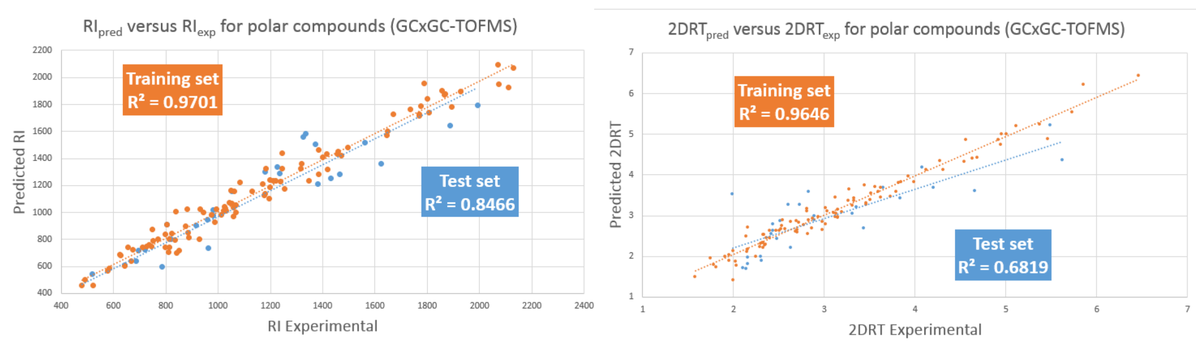


Figure 8. Prediction models developed in CASI for retention index (left) and 2nd dimension retention time (right) using GCxGC-TOFMS (polar method)

The confidence in correct identification of unknown compounds is enhanced when mass spectral comparisons are combined with predicted RI/LRI values using the models that have been developed.

Conclusions

- Several approaches using QSPR prediction of retention indices improve the compound identification process
- The methodology is suitable for several GC techniques (GC-MS, GC-HR-MS, GCxGC-TOFMS), for a wide range of compounds
- The confidence in correct identification is higher for GC-(HR)-MS when both models predict LRI values in close agreement (i.e. Workflow A and B)
- The establishment of a confidence score using the output from both LRI prediction models is planned
- Ideally, the selection for the combination of best algorithm leading to the best models should be automated

References

1. Knorr, A., Monge, A., Stueber, M., Stratmann, A., Arndt, D., Martin, E., Pospisil, P., Computer-Assisted Structure Identification (CASI) - An Automated Platform for High-Throughput Identification of Small Molecules by Two-Dimensional Gas Chromatography Coupled to Mass Spectrometry, Anal Chem. 2013, 85(23):11216-24.
2. Martin E., Monge A, et al., Building an R&D chemical registration system, Journal of Cheminformatics 2012 4:11, DOI: 10.1186/1758-2946-4-11

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL
www.pmiscience.com

International Conference and Exhibition on Cheminformatics and Computational Chemical Biology

Brisbane, Australia

July 11-12, 2016

Competing Financial Interest

The research described in this poster was funded by Philip Morris International