

Prediction of gas chromatographic retention indices of toxicologically relevant compounds

Zahra GARKANI-NEJAD¹, Manfred KARLOVITS²
Wilhelm DEMUTH², Thomas STIMPF³
Walter VYUDILIK³, Mehdi JALALI-HERAVI⁴
Kurt VARMUZA^{2*}

- ¹ Valie Asr University of Rafsanjan, Faculty of Science, Rafsanjan, Iran
² Vienna University of Technology, Institute of Chemical Engineering, Laboratory for ChemoMetrics, Vienna, Austria
³ University of Vienna, Institute of Forensic Medicine, Vienna, Austria
⁴ Sharif University of Technology, Department of Chemistry, Tehran, Iran

* Presenting author kvarmuza@email.tuwien.ac.at
www.lcm.tuwien.ac.at

Acknowledgment Austrian Science Fund, project P14792-CHE



Poster Presentation: **Advances in Chromatography and Electrophoresis - Conferentia Chemometrica (ACE & CC 2003)**
27 - 29 October 2003, Budapest, Hungary

Introduction

With 846 organic compounds,

- - most of them relevant in forensic GC-MS analyses, and
- possessing very diverse chemical structures -

multivariate calibration models

have been developed to predict

Kovats GC retention indices from molecular descriptors

Strategy

Straight forward application of widely used and easily available methods from computer chemistry and chemometrics.

Aim

Development of a tool that supports the identification of unknowns in forensic analyses by GC-MS.

- [1] Stimpf T., Demuth W., Varmuza K., Vycudilik W.: *J. Chromatogr. B*, 789, 3-7 (2003). Systematic toxicological analysis: computer-assisted identification of poisons in biological materials.

Data

Database: Mass spectra and GC data of drugs, etc. [2]

↓ Selection of compounds (structure, *RI* available*)

846 organic compounds **: 2D-structures (Molfiles), *RI*

↓ Software **WebLab Viewer** [3] (2D → 3D)

846 compounds with very approximate 3D-structures and explicit H-atoms (Molfiles)

↓ Software **DRAGON** [4] (generation of molecular descriptors)

1497 molecular descriptors for each structure

↓ Basic feature selection (elimination of constant, extreme, and highly correlating features)

529 descriptors for each of the 846 structures
Multivariate data: X (846 * 529) descriptors
 y (846 * 1) *RI*

* *RI*, Kovats retention index

** Hypnotics, insecticides, tranquilizers, analgesics, ..., *n*-alkanes C₁₄ - C₃₀, molecular masses 109-491; *RI* between 1110 and 3870.

[2] K. Pfleger, H. H. Maurer, A. Weber: Mass spectral and GC data of drugs, poisons, pesticides, pollutants and their metabolites, 2nd ed. VCH, Weinheim, Germany, 1992.

[3] Software WebLab Viewer. Accelrys Inc., San Diego, CA, www.accelrys.com, 2002.

[4] R. Todeschini, V. Consonni, A. Mauri, M. Pavan: Software Dragon. University of Milano-Bicocca, and Talete srl., disat.unimib.it/chm/Dragon.htm, 2003.

Methods

Feature Selection

- Selection of features possessing highest correlation coefficients with *RI*
- Stepwise forward feature selection - together with MLR/OLS (software Systat).

Multivariate Calibration

- PLS, PCR, MLR/OLS, ANN (Software Unscrambler, Systat, and ANN [5])
- Training set 700 compounds, prediction set 146 compounds
- Cross validation for PLS and PCR with training set using 10 segments.

Evaluation

- *SEP*, standard error of prediction

$$SEP = \left[\frac{\sum (RI_i^* - RI_i - bias)^2}{n - 1} \right]^{0.5} \quad i = 1 \dots n \quad \text{for prediction set}$$

$$bias = \frac{\sum (RI_i^* - RI_i)}{n} \quad \begin{array}{l} RI_i^* \text{ predicted} \\ RI_i \text{ experimental} \end{array}$$

- Mean of 4 experiments with different random samples for training and prediction set.

[5] M. Jalali-Heravi, Z. Garkani-Nejad: *J. Chromatogr. A*, 927, 211-218 (2001). Prediction of electrophoretic mobilities of sulfonamides in capillary zone electrophoresis using artificial neural network.

Results

| Method | No. of descriptors | Selection method | No. of components | SEP |
|---------|--------------------|------------------|-------------------|-------|
| PLS | 529 | - | 15 | 82 |
| | 100 | max. corr. | 20* | 119 |
| | 15 | stepwise | 8 | 79 |
| PCR | 529 | - | 20* | 133 |
| | 100 | max. corr. | 20* | 159 |
| | 15 | stepwise | 15 | 79 |
| MLR/OLS | 100 | max. corr. | - | 125 |
| | 15 | max. corr. | - | 156 |
| | 15 | stepwise | - | 79 |
| ANN** | 15 | stepwise | - | ca 72 |

* No optimum in cross validation, 20 components used.

** Training of backpropagation neural network with 700 objects, optimization (validation) with 100, prediction with 46 objects.

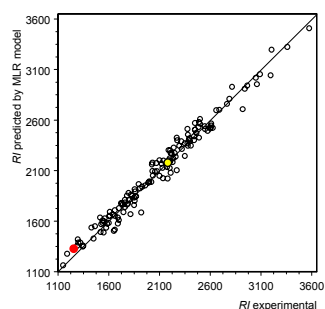
Best models

- PLS with all 529 descriptors (quick model building).
- MLR/OLS (or PLS or PCR) with 15 descriptors obtained by forward stepwise selection by MLR/OLS (slow, easily interpretable model).
- ANN (very slow, not yet extensively verified).

Prediction errors are rather high, probably because of the very diverse structures, crude 3D-structures, and varying GC conditions for retention indices in the used database.



Results

Predicted versus experimental *RI* for prediction set



Model created from 700 compounds by MLR/OLS and forward stepwise feature selection (15 features/descriptors selected). Prediction set with 146 compounds; SEP is 74, corr. coeff. is 0.988.

Examples

| Compound | <i>RI</i> experimental | <i>RI</i> prediction error |
|--|------------------------|----------------------------|
|  CAS Reg.no. 670804, a psychedelic drug | 1260 | 71 |
|  CAS Reg.no. 6463214, a diuretic | 2195 | -9 |

Conclusions

- Improvements necessary and probably possible.
- Promising approach to support the identification of unknown compounds exhibiting very similar mass spectra but different retention indices.