Scientific paper

Modeling of the Mass Spectrometric Response Factors in Non-target Analysis

Gregor Arh,[1] Leo Klasinc,[2] Marjan Veber[1] and Matevž Pompe[1,*]

[1] *Faculty of Chemistry and Chemical Technology, University of Ljubljana, Aškerčeva 5, 1000 Ljubljana, Slovenia*

[2] *Ruđer Bošković Institute, Bijenička 54, HR-10002 Zagreb, Croatia*

* *Corresponding author: E-mail: matevz.pompe@guest.arnes.si*

Received: 28-05-2010

This paper is dedicated to Professor Milan Randić on the occasion of his 80th birthday

Abstract

Experimental MS response factors were measured for 36 different saturated and unsaturated volatile organic compounds (VOC) containing carbon, hydrogen and halogen atoms. Chemical structure was encoded using various molecular descriptors. A quantitative structure-property relationship model was established using the multiple linear regression models. The cross-validation ability of the created model was estimated by leave-one-out cross-validation procedure. Error in the cross-validation of response factors was calculated by cross-validation procedure and was 15%, which is sufficient for the determination of VOCs in the air. The proposed procedure can be used for simultaneous qualitative and quantitative determination of volatile organic compounds in the atmosphere.

Keywords: QSPR, prediction, MLR, MS response factors

1. Introduction

Accurate identification and quantification of organic substances represents one of the key problems during the gas chromatographic analysis. This problem is even more pronounced in cases of non-target analyses, that is, in cases when the number of compounds is unknown. Such type of analyses is often used in environmental, pharmaceutical and food research. The number of components is unknown therefore it is difficult to optimise separation conditions. At the same time qualitative and quantitative determination is hindered due to the lack of standard materials.

Gas chromatography coupled with mass spectrometry is the technique most often used to solve this problem. It offers structural information necessary for accurate identification of individual organic compounds and at the same time it enables quantification of the same substance. The identification of organic compounds is a more or less straightforward procedure and is usually accomplished by library search. On the other hand accurate quantification is almost impossible if standard material is absent, because there is no numerical model that would enable theoretical calculations of response factor for any particular organic compound without experimental measurements. In such cases the theoretical procedures based on molecular similarity or numerical models based on chemical structure can represent a possibility to overcome this problem.

The first such model was proposed in 60s to calculate response factors for FID detector based on effective carbon number.[1] The model was further developed in several consecutive studies.[2–5] However, it was shown that even such developed models can exhibit up to about 25% variation between the calculated and experimental response factor.[6] They have proposed a FID response factor prediction model based on molecular similarity using multiple linear regression models. The main drawback of such models was the inability to model nonlinear correlations. Therefore Jalali-Heravi et al. improved QSPR correlation model by using artificial neural networks (ANNs) as a modeling technique: they used ANNs for the creation of the prediction model for FID,[7] thermal conductivity detector[8] and electron capture detector[9] response factors.

However, none of the mentioned detectors can provide structural information about the measured chemical compound and therefore cannot be used in cases of non-

target analysis. In this study we developed for the quantification of ozone precursors and some chlorinated volatile organic compounds present in the atmosphere a quantitative structure-property relationship (QSPR) model for the prediction of the experimental MS response factors. Since this is just a preliminary study we have applied only simple multiple linear regression model, which is known that can give overestimation of the prediction capabilities. However the main aim of this work was to show that the MS spectrometric response factors can be reliably determined for the longer period of time and that molecular descriptors can encode the main structural features of the modeled molecules which are important for the modeling of the MS response factors.

1. 1. Model Creation and Calidation

The QSPR technique usually involves two stages. In the first step molecular descriptors are found that represent structural features of the molecule. In a second step these descriptors are correlated with a selected property to find a relationship between the structure and its property.

Molecular structures were created by HyperChem™. Afterwards, MOPAC software[10] was used for the geometry optimization and calculation of net atomic charges. Using CODESSA software,[11,12] 367 different topological, geometric, informational, electrostatic, electrotopological and quantum-chemical descriptors necessary for the creation of the models were calculated. The descriptors employed in the study contain information about the connections between atoms, symmetry, shape, branching, distribution of charge, and quantum-chemical properties of the molecules. It is obvious that one cannot use all these descriptors for the creation of a single prediction model, because most of the descriptors do not encode any structural feature that is responsible for the response factor of the mass-spectrometric (MS) detector and such model would be most probably coincidental. Therefore a descriptor reduction procedure had to be applied. The selection of structural descriptors was accomplished by applying the heuristic optimization search in CODESSA software which is described in the literature.[11,12]

The leave-one-out cross-validation procedure performed on the training set was used for the estimation of cross-validation capabilities of MLR models during structural descriptor selection. The root mean square (RMS) error was calculated as well as the correlation coefficient (q^2) for the linear dependence of cross-validated vs. experimental values. The model with the best cross-validation parameters was chosen for further studies.

1. 2. Determination of MS Response Factors

The experimental data were obtained by two gas chromatographs, Varian Star 3400 CX and Varian Star 3600 CX. Both of them were equipped for measurements of volatile organic compounds in air. The first one was used in conjunction with a flame ionization detector and the other was coupled with the Saturn 2000 MS detector. Both systems were equipped with a 10-way VICI Valve (Valco Instruments Co. Inc.) and the cryotrap.[13] Samples were injected into a cryotrap, which was cooled with liquid nitrogen (–196 °C), by using a helium 6.0 carrier gas. All connection tubes were made out of stainless steel and were heated to approximately 100 °C in order to prevent compounds from liquefying already in the analytical instrument. For the separation of compounds a Restek RTX-5MS column (l = 60 m, 2r = 250 μm, d = 5 μm) was used. Temperature program was as follows: initial temperature 3 °C (hold time 10 min), temperature gradient 2 °C/min to 140 °C and 20 °C/min to 250 °C (hold time 10 min).

Two multicomponent standard mixtures, Restek VOC AB-18475 and Matheson Toxi-Mat TO-14 VOC,

Table 1. Experimental and calculated MS response factors (x10.000)

Name	Experimental RF	Cross-validated RF
Pentene	2.05	1.96
Isoprene	1.96	2.68
cis-2-Pentene	2.28	2.06
2,3-Dimethylbutane	2.58	3.47
2-Methylpentene	2.63	3.30
3-Methylpentene	2.64	3.27
1-Hexene	2.59	1.96
n-Hexane	3.54	3.46
Methylcyclopentane	3.36	3.60
2,4-Dimethylpentane	2.82	3.33
Cyclohexane	3.29	4.39
Benzene	3.99	3.89
2-Methylhexane	2.53	3.35
2,3-Dimethylpentane	4.13	3.43
3-Methylhexane	3.18	3.32
2,2,4-Trimethylpentane	3.67	3.40
n-Heptane	3.32	3.57
Methylcyclohexane	4.18	3.74
2,3,4-Trimethylpentane	4.14	3.52
2-Methylheptane	3.73	3.46
3-Methylheptane	4.30	3.42
n-Octane	4.90	3.67
Trichlorofluoromethane	7.71	8.37
1,1-Dichloroethene	4.38	4.04
Methylene chloride	7.72	7.13
1,1-Dichloroethane	3.91	3.41
cis-1,2-Dichloroethene	3.63	4.34
Chloroform	10.4	9.91
1,1,1-Trichloroethane	4.02	4.13
1,2-Dichloroethane	4.23	4.97
Trichloroethylene	5.54	5.06
Toluene	3.51	2.58
1,1,2-Trichloroethane	4.19	5.31
1,2-Dibromoethane	7.85	7.96
Tetrachloroethylene	10.6	10.1
Chlorobenzene	3.42	3.35

were used for measuring the response factors on GC-MS system. The certified accuracy of these standards was 10%. Therefore additional calibration of the standards was performed using a mixture of C1-C6 n-paraffins (Fluka 80311). The certified accuracy of the latter standard was 2%. The experimental response factors for 36 organic compounds are presented in Table 1.

2. Results and Discussion

It is known that the MS response factors depend on the total ionization cross section and the ionizing path length. These parameters are correlated by molecular geometry and can be modeled by atom additive procedures.[14] It can be expected that besides pure geometric parameters, the electronic configuration of the molecule plays important role in its ionization processes as well. Therefore we have considered modeling of the electron impact MS response factors using molecular similarity QSPR procedures. The response factors were experimentally obtained for 36 organic compounds. The relative reproducibility of the experimental response factors was around 10%. The MS response factors remained constant for at least 14 days if no auto-tuning procedure was performed within this period.

The chemical structures were encoded by various molecular descriptors present in the CODESSA software. The MLR models with up to 5 parameters were selected using the step-wise selection procedure. The selected models based on the best cross-validation results are shown in Table 2. We can see that topological, electrostatic and geometrical descriptors have been selected to the final model. They represent information about the size, shape and electronic properties of the individual molecule. Although the best cross-validation results were obtained for the 5-parameter MLR model, we prefer here the 4-parameter MLR model to keep the ratio between the parameter and the data point numbers around 10 and avoid overfitting of the prediction model. The cross-validated results of the final 4-parameter MLR model are presented in Fig. 1 and Table 1.

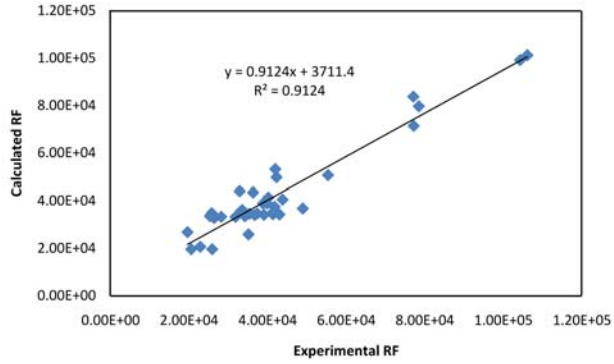

Fig 1. Experimental vs. cross-validated MS response factors

The regression parameters reported were obtained from the model constructed from the whole data set of 36 compounds. An extensive model validation was done in order to prevent insignificant variables to participate in the created prediction model. Inter-correlation coefficients were calculated among the descriptors that were included in the model. The highest value was lower than 0.8, so we can say that the descriptors which were included in the model were not highly inter-correlated. Nevertheless, in order to prevent chance correlation, a t-test[15] was done for every descriptor included in the model. The critical value for the Student's t-distribution for $\alpha = 0.01$ and 35 degrees of freedom is around 2.8. We can see from Table 3 that all descriptors are above this value and are therefore retained in the model.

A close inspection of the selected descriptors gives some insight into the structural features that determine the modeled property. In our QSPR study a chemical structure was represented by a four-dimensional vector, the com-

Table 2. Selection of best *n*-parameter MLR model

	Descriptors	RMS (cv)	R^2	q^2	F	s
1	Relative molecular weight	15900	0.684	0.625	75.8	11800
2	Relative molecular weight Average information content (order 0)	12300	0.813	0.770	73.7	9220
3	Relative molecular weight RNCG (relative negative charge) (QMNEG/QTMINUS) (Semi-MO PC) Maximal net atomic charge	10200	0.875	0.831	76.8	7650
4	Relative molecular weight RNCG relative negative charge (QMNEG/QTMINUS) (Semi-MO PC) Minimal atomic orbital electronic population Principal moment of inertia B / # of atoms	7520	0.912	0.885	83.3	6490
5	Relative molecular weight RNCG relative negative charge (QMNEG/QTMINUS) (Semi-MO PC) Minimal atomic orbital electronic population Principal moment of inertia B / # of atoms Wiener index	6380	0.940	0.913	96.4	5480

Table 3. Model equation for the 4 parameter model

	X	DX	t-test
Intercept	3.2E + 05	5.8E + 04	–5.4
Relative molecular weight	4.1E + 03	2.6E + 02	15
RNCG (relative negative charge) (QMNEG/QTMINUS) (Semi-MO PC)	7E + 04	1.2E + 04	–5.6
Minimal atomic orbital electronic population	3.7E + 05	6.3E + 04	5.9
Principal moment of inertia B / # of atoms	9E + 05	2.4E + 05	3.7

ponents of which were two electrostatic, one geometric and one topological descriptor. It is almost impossible to completely separate structural features that are encoded by individual descriptors however we can say that relative molecular weight and "principal moment of inertia B/number of atoms" encodes the size and the shape of the molecule. On the other hand, the remaining descriptors encode electronic configuration of the organic species. It must be mentioned that the structural interpretation of the model represents only qualitative information about the factors that are determining the distribution of electron impact MS response factors.

The main reason why the influence of individual factors on any modeled property cannot be quantitatively evaluated is the mutual correlation of the descriptors involved because it is usually very difficult to obtain orthogonal descriptors that would at the same time create a very good prediction model.

The cross-validated results allow an estimation of prediction capabilities of the final model: the q^2 and RMS_{cv} parameters were 0.885 and 7521, respectively. The visual inspection of the linear model does not reveal any outliers. The differences in the individual experimental MS response factors were almost one order of magnitude. Therefore we have considered the relative RMS error as a parameter for the evaluation of the prediction error. The relative RMS error for the obtained model was around 15%. This is a bit higher than the experimental uncertainty which was evaluated at 10%, however it is still within the estimated inter-laboratory GC/MS response factor precision, which was found to be around 19%.[16] These results proved that the described calibration procedure can be used for the determination of volatile organic compounds in the atmosphere when standard substances are not available for each individual organic compound.

3. Conclusions

Experimental MS response factor data were determined for 36 different saturated and unsaturated organic compounds and multiple linear regression models were used for the creation of the prediction model. Due to the relative the small number of data points, the prediction abilities of the created model were estimated by leave-one-out cross-validation. The error in the cross-validation of the response factors was calculated by a leave-one out cross-validation procedure and was around 14% based on the relative RMS error which is slightly less than the experimental error. Such errors are acceptable for the determination of VOCs in the air.

The proposed procedure can be used for simultaneous qualitative and quantitative determination of volatile organic compounds in the atmosphere. Further study will be needed to prove that the described procedure can be used also for the identification and quantification of oxygenated organic species that are formed in the atmosphere by photochemical oxidation. It should be mentioned that this is just preliminary study where we wanted to test if molecular descriptors can be applied for the modeling of MS response factors. The described problem is rank deficient, so in order to use such modeling procedure for the routine determination of VOC in non-target analysis and we will have to develop more robust modeling procedure, such as, principal component regression, partial least-square regression or ridge regression.

4. Acknowledgements

The authors acknowledge the financial support by the Ministry of Education, Science and Sport of the Republic of Slovenia (Grant P1-0153) and the Ministry of Science, Education and Sports of the Republic of Croatia (Grant 098-0982915-2947). The authors would like to acknowledge the valuable suggestions from the reviewer about robust modeling, which will be implemented in our subsequent study.

5. References

1. Sternberg, J. C., Gallaway, W. S., Jones, D. T. L., Gas chromatography, Brenner, N., Callen, J. E., Weiss, M. D., Eds., Academic Press: New York, 1962, chapter 18.
2. Scanlon, J. T., Willis, D. E., *J. Chrom. Sci.* **1985**, *23*, 333.
3. Tong, H. Y., Karasek, F. W., *Anal. Chem.* **1984**, *56*, 2124.
4. Jorgensen, A. D., Picel, K. C., Stamoudis, V.C., *Anal. Chem.* **1990**, *62*, 683.
5. Yieru, H., Qingyu, O., Weile, Y., *Anal. Chem.* **1990**, *62*, 2063.
6. Katritzky A. R., Lobanov, V. S., Karelson, M., *Chem. Soc. Rev.* **1995**, *24*, 279.

7. Jalali-Heravi, M., Fatemi, M. H., *J. Chromatogr. A* **1998**, *825*, 161.
8. Jalali-Heravi, M., Fatemi, M. H., *J. Chromatogr. A* **2000**, *897*, 227.
9. Jalali-Heravi, M., Noroozian, E., Mousavi, M., *J. Chromatogr. A* **2004**, *1023*, 247.
10. Stewart J. J. P. Special issue – MOPAC – A semiempirical molecular-orbital program, *J. Comput. Aided Mol. Des.* **1990**, *4*, 1.
11. A. R. Katritzky and E. V. Gordeeva, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 835.
12. M. Karelson, V. S. Lobanov, A. R. Katrizky, *Chem.Rev.* **1996**, *96*, 1027.
13. Veber, M., Pompe, M., Baša-Češnik, H., *Journal of the Institute of Science and Technology of Balikesir University* **2001**, *3*, 125–147.
14. W. L. Fitch, A. D. Sauter, *Anal. Chem.* **1983**, *55*, 832.
15. Neter, J.; Kutner, M. H.; Nachtshein, C. J.; Wasserman, W. Applied Linear Statistical Models, 4th ed., McGraw-Hill, Boston, MA, **1996**; p. 268.
16. A. D. Sauter, P. E. Mills, W. L. Fitch, R. Dyer, *J. High Res. Chrom.* **1982**, *5*, 27.

Povzetek

Eksperimentalno smo določili faktorje odziva za masno selektivni detektor za 36 nasičenih in nenasičenih hlapnih organskih spojin, ki vsebujejo ogljikove vodikove in včasih tudi halogenidne atome. Kemijsko strukturo smo kodirali z različnimi molekularnimi deskriptorji. Z uporabo multiple linearne regresije smo izdelali kvantitativni model, ki povezuje kemijsko strukturo z modelirano lastnostjo. Za oceno napovedne sposobnosti modela smo uporabili navzkrižni validacijski test. Tako ocenjena napaka modela za določevanje faktorjev odziva za MS detektor je bila okoli 15 %, kar zadostuje pri določevanju hlapnih organskih snovi v zraku. Predlagana procedura se lahko uporablja za kvalitativno in kvantitativno določevanje hlapnih organskih snovi v atmosferi.