Budapest University of Technology and Economics

Orsolya Farkas

QSAR studies on retention indices and biological activities

Ph.D. Theses

2007.

Consulent:

Károly Héberger

1. Introduction

In the last decades the number of chemical data has been considerably increased. Data evaluation with traditional univariate methods is difficult or sometimes even impossible. Hence, multivariate chemometric methods are acquiring new significance.

Quantitative structure activity relationship (QSAR) belongs to the chemometric methods, which describes a mathematical relationship between structural attributes and a property of a set of chemical compounds. Application of such mathematical relationships to predict certain targeted properties for an extensive set of chemicals without expensive and labor-intensive experimental measurements is possible. Furthermore, QSAR is a useful tool for a better understanding of activity mechanisms and structural properties of compounds.

Classification is also a data analysis method aiming to predict class memberships. In this predictive method the response (dependent variable) is a category (grouping) variable. The purpose of the analysis is to predict which category a new sample belongs to. During the present PhD study chemical and biological problems have been solved using QSAR and classification methods.

Variable selection is perhaps the most interesting and also the most delicate procedure of QSAR because there is no precise prescription for the selection of relevant variables from a large descriptor (independent variables) set, thus each chemical (or biological) problem requires a different approach. Several chemometric modeling methods have been compared on the basis of their variable selection efficiency. Different set of compounds (alcohols, fatty acid-methyl esters and heterocycles containing oxygen, nitrogen or sulfur atoms) have been used to compare the effectiveness of variable selection methods. The dependent variable in the models was Kováts retention index (RI). Using quantitative structure-retention relationship (QSRR)

models identification of isomeric compounds with similar mass spectra could be achieved. Screening of retention index databases with QSRR models is also possible.

Importance of QSRR has suggested by earlier reviews [e.g. Kaliszan R., Quantitative Structure-Chromatographic Retention Relationships, Wiley, New York, **1987***;* Kaliszan R., Structure and Retention in Chromatography - A Chemometric Approach, Harwood, Amsterdam, **1997**]. Actuality of this topic has been shown in recent publications [e.g. Kaliszan R. QSRR: Quantitative Structure-(Chromatographic) Retention Relationships *Chem.Rev.* **2007** 107 (7) 3212-3246; Héberger K. Quantitative structure-(chromatographic) retention relationships *J. Chromatogr. A* **2007** 1158 (1-2) 273-305.]

Classification of antidepressant candidates have been performed on the basis of their ability of blocking a cardiac K^+ channel encoded by human *ether-a-go-go* related gene (hERG). Blocking of this channel can cause cardiac arrhythmia and sudden death, therefore testing of this property is crucial. Using the classification models design of antidepressants without cardiac side-effects could be achieved.

Flavonoids are well-known because of their numerous positive health-effects which are generally associated with their antioxidant properties. More than 4000 flavonoids are known. Many aspects are known about the structural characteristic related to their antioxidant properties. Nevertheless, quantitative models to predict antioxidant activity are scarce.

2. Methods

Kováts retention indices (RI) stems from the literature and the measurements of our partner Prof. I.G. Zenkevich. Antioxidant activity data has been taken from the literature. Remain hERG activity values of antidepressants from have released from Laboratorios Dr. Esteve.

Geometry optimization has been performed using HyperChem program package [HyperChem 7.0 program package HyperCube Inc., Canada] with the AM1 semi-empirical method. Descriptors have been calculated using the Dragon program package [Todeschini R., Consonni V., Pavan, M. Dragon Software Version 2.1, **2002**.].

Chemometric methods in the calculations were the following: pair-wised correlation method (PCM), forward selection (FS), ridge regression (RR), Lasso method, best subset selection (BSS), a multiple linear regression (MLR), genetic algorithm (GA), partial least squares projection of latent structures (PLS), linear discriminant analysis (LDA), classification and regression tress (CART). These methods have been used for variable selection as well as parameter estimation of the models. Statistical calculations have been performed using the Statistica program package [Statistica 5.5 and 6 Software Package, StatSoft Inc., Tulsa, OK, USA]. PCM has been calculated using a Microsoft Excel macro written by Róbert Rajkó. Genetic algorithm calculations have been performed using the MobyDigs program package [Todeschini R., Ballabio D., Consonni V., Mauri A., Pavan M. MobyDigs Professional version 1.0 **2004** MilanoChemometrics and QSAR Research Group]. Lasso calculations for methyl-esters has made by Forrest Stout, PLS regression coefficient variance calculation in case of methyl-esters has been made by Károly Héberger.

3. Results and application possibilities

1. Stable and valid models have been built for Kováts-index prediction of alcohols. These models contain shape- and size-related WHIM descriptors.

   The best performance for variable selection was exhibited by the best subset selection and the forward selection methods for the alcohols investigated.

   Pair-correlation method turned out to provide good variables for description, but it is not suitable for prediction purposes.

   Ridge regression and partial least squares projection of latent structures were not able to describe the retention mechanism and to produce good predictive models for Kováts retention indices in case of the above mentioned compounds.

2. Useful predictive models have been built for Kováts-index prediction of fatty acid methyl-esters.

   Pair-correlation method gave excellent prediction of Kováts-indices of methyl-esters. Partial least squares method is superior to other techniques in its optimal performance. Applying the principle of parsimony the Lasso and especially the PLS methods were not able to produce good predictive models for Kováts retention indices in case of methyl-esters.

   The results presented in this work show that MLR models with variables selected based on PLS and Lasso may not be best for MLR.

   Prediction of retention indices for 37 fatty acid methyl esters for which measured Kováts-indices are not available have been also fulfilled for identification purposes. Topological indices and descriptors encoding flexibility are useful in the RI prediction of methyl-esters.

3. Reliable models have been built for Kováts-index prediction of heterocyclic compounds.

   Boiling point value is necessary to describe Kováts retention indices of saturated heterocycles, but is not suitable (enough) to predict RI on its own.

   Three-dimensional descriptors such as GETAWAY and WHIM seem to be useful to predict retention indices. Two-dimensional molecular profiles also play an important role in encoding the structures of heterocycles.

   Partial least squares method failed as a variable selection method. Combination of best subset selection for variable selection and PLS for model building is a useful way to predict RI in case of heterocyclic compounds.

3. Useful models have been built for the classification of antidepressants based on their hERG activity.

   The models provide a proper classification (nearly 95 %) of the structurally similar compounds. The classification efficiency for the diverse set of antidepressants was more than 80 %.

   Best models were generated with the combination of genetic algorithm, forward selection and linear discriminant analysis.

   Reasonable classification can be achieved using one GETAWAY descriptor and with the number of carbonyl groups present in the molecules in case of the diverse compound set.

   Classification of similar compounds could be managed using one geometrical descriptor (sum of oxygen-oxygen bond distances in the molecule).

5. Efficient model has been built for antioxidant activity description and prediction of a diverse set of flavonoids. The partial least squares regression model can also be used for classification of different flavonoid groups.

4. Publications

4.1. Publication related to the theses:

1. **Farkas O**., Héberger K., Zenkevich I. G. Quantitative structure – retention relationships XIV. Prediction of gas chromatographic retention indices for saturated O-, N- and S-heterocyclic compounds *Chemometrics and Intelligent Laboratory Systems* **2004**, 72, 173-184. IF: 2,45

2. **Farkas O**., Jakus J., Héberger K. Quantitative structure-antioxidant activity relationships of flavonoid compounds *Molecules* **2004**, 9, 1079-1088. IF: 0,84

3. **Farkas O**., Héberger K. Comparison of ridge regression, partial least squares, pair-wise correlation, forward- and best subset selection methods for prediction of retention indices of aliphatic alcohols *J. Chem. Inf. Model*. **2005**, 45, 339-346. IF: 3,42

4. **Farkas O**., Zenkevich I.G., Stout F., Kalivas J.F., Héberger K. . Prediction of Retention Indices for Fatty Acid Methyl Esters *J. Chromatogr. A* (submitted)

5. **Farkas O**., Héberger K. Chemometric analysis of antidepressants based on their hERG K^+ channel activity
(manuscript in preparation)

4.2. Other publications:

1. **Farkas O.**, Gere-Paszti E., Forgacs E. Study of the interaction of structurally similar bioactive compounds by thin-layer chromatography *Journal of Chromatographic Science* **2003**, 41 (4) 169-172. IF: 0,88

2. Gere-Paszti E., **Farkas O**., Prodan M., Forgacs E. Molecular mapping of interaction between cholesterol and model drugs traced by reversed phase bioaffinity chromatography *Chromatographia* **2003**, 57 (9/10) 599-604. IF: 1,17

3. Prodan M., Gere-Paszti E., **Farkas O**., Forgacs E. Validation and simultaneous determination of caffeine and paracetamol in pharmaceutical preparations *Chemia Analityczna - Chemical Analysis* 2003, 48 (6) 901-907. IF: 0,56

4. Forgacs E., Cserhati T., **Farkas O**., Eckhardt A., Miksik I., Deyl Z· Interaction between cholesterol and non-ionic surfactants studied by thin-layer chromatography *Journal of liquid chromatography & related technologies* **2004**, 27 (13) 1981-1992. IF: 0,88

5. Jakus J., **Farkas O**. Photosensitizers and antioxidants: a way to new drugs? *Photochem. Photobiol. Sci*. **2005**, 4, 694-968. IF: 2,42

4.3. Presentations related to the theses:

1. **Farkas O**., Héberger K. Comparison of variable selection methods: Prediction of retention indices of saturated alcohols *Hungarian Chemometric Workshop Kemometria* Tata, Hungary, 29.Sept.-1.Oct. 2002. (poster)

2. **Farkas O**., Héberger K., Zenkevich I. G. Quantitative structure – retention relationships XIV. Prediction of gas chromatographic retention indices for cyclic compounds containing nitrogen and sulphur atoms *Advances in Chromatography and Electrophoresis – Conferentia Chemometrica 2003* Budapest, Oct. 27-29. 2003. (poster)

3. **Farkas O**., Jakus J., Héberger K. Quantitative structure–antioxidant activity relationship of flavonoids Computer Applications and Chemometrics in Analytical Chemistry Balatonfüred, Hungary, Aug.31-Sept. 3. 2004. (poster)

4. **Farkas O**., Héberger K. Classification of antipsychotic drug candidates on the basis of their hERG activity *Workshop on Chemometrics and Molecular Modelling* Szeged, Hungary, Apr. 6-8., 2005. (oral presentation in Hungarian)

5. **Farkas O**., Stadler K., Héberger K. Classification of antidepressants based on their herg activity *Conferentia Chemometrica 2005 – Chemometrics VII* Hajdúszoboszló, Hungary, Aug. 28-30., 2005. (poster)

6. **Farkas O**., Zenkevich I.G., Stout F., Kalivas J.H., Héberger K. Prediction of gas chromatographic retention indices for fatty acid methyl esters Comparison of variable

selection methods *Conferentia Chemometrica 2005 – Chemometrics VII* Hajdúszoboszló, Hungary, Aug. 28-30., 2005. (poster)

7. **Farkas O**., Zenkevich I.G., Stout F., Kalivas J.H., Héberger K. Hogyan hasonlítsunk össze változó-kiválasztási módszereket? „Legjobb" vagy egyszerű modellek *KeMoMo* Szeged, Hungary, Apr. 27-28., 2006. (oral presentation in Hungarian)

8. **Farkas O**., Zenkevich I.G., Stout F., Kalivas J.H., Héberger K. Prediction of Kováts indices for fatty acid methyl esters Optimal or parsimonious models? *Conferentia Chemometrica* , Budapest, Hungary, Aug. 31-Sept 3., 2007. (poster)