Novi Sad J. Math.
Vol. 38, No. 3, 2008, 39-46

STATISTICAL METHODS IN PHYSICO-CHEMICAL CHARACTERIZATION OF NEWLY SYNTHESIZED COMPOUNDS

Tatjana Djaković-Sekulić[1], **Zagorka Lozanov-Crvenković**[2], **Nada Perišić-Janjić**[1]

Abstract. Evaluation of the physico-chemical properties that affected the experimental data was performed for 45 newly synthesized compounds, derivatives of 2,4-dioksotetrahydro-1,3-thiazole. In order to quantify molecular structure various molecular descriptors were calculated for each of the 45 compounds. Evaluation of physico-chemical properties was done by the partial least-squares (PLS) regression. Using PLS, some of the descriptors were identified as the most important for the experimental behavior. Proposed PLS model gives a good correlation between the experimental and calculated retention data.

AMS Mathematics Subject Classification (2000): 92E10

Key words and phrases: Molecular descriptors, Quantitative Structure-Retention Relationships (QSRR), Principal component analysis (PCA), Partial Least Squares Regression(PLS)

1. Introduction

It is well known that physical properties of a chemical substance like color, odor, melting point, boiling point, density, electrical conductivity, etc. are closely correlated to the structure of a molecule. Molecular structure predetermines chemical reactivity and biological activity, too. In other words, the architecture or the structure of the molecules is an independent variable, while physical properties, chemical reactivities and biological activities are dependent variables. Therefore, we can specify symbolically such functional dependence:

$$\begin{gathered}\text{Property} = \text{f(structure)}\\ \text{Reactivity} = \text{g(structure)}\\ \text{Activity} = \text{h(structure)}\end{gathered}$$

In practice, physical properties and chemical reactivities are hard to separate; hence the expression physico-chemical properties is in use. Unfortunately, the explicit functional dependence is very hard to give, but it can be only guessed on the basis of accumulated experience on what molecular feature will provide a desired feature. A mathematical model that relates structure to physico-chemical

[1] Department of Chemistry, Faculty of Sciences, University of Novi Sad, Serbia e-mail: tatjana.djakovic-sekulic@ih.uns.ac.rs

[2] Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad, Serbia e-mail: zagorka.lozanov-crvenkovic@dmi.uns.ac.rs

properties of the molecule are known as Quantitative Structure Property Relationship (QSPR), while similarly, Quantitative Structure Activity Relationship (QSAR) relates molecular structure to pharmacological activity. Once established QSPR and QSAR can be used further to predict the activity of related compounds. Hence, a fundamental problem in chemistry is the evaluation of the relationships between the structure of chemical compounds and their physicochemical properties or biological activity. Molecular descriptors are numerical values that characterize properties of molecules, helping predict the properties and activity of molecules.

Molecular descriptors vary in complexity of encoded information and in compute time. They may represent the properties of complete molecules (logP, Molar Refractivity), could be calculated from 2D graphs (Topological Indexes, 2D fingerprints) or may require 3D representations (Pharmacophore descriptors). Today chemical drawing software package enables users not only to draw molecule but to calculate properties, too. Because a number of descriptors are available the question arises how to select the right ones. Selected descriptors and experimental results are the basic elements for training set of data (a large set of data with many variables and cases).

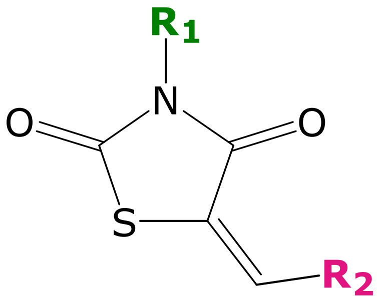
The main problem is how to reduce the number of variables and how to detect structure in the relationships between variables, that is to classify variables. This can be done by various statistical methods of: 1) explorative analysis (PCA - Principal Component Analysis, FA - Factor Analysis or CA - Cluster Analysis); 2) classification methods (Classification Tree (CART), Discriminant Analysis or Neural Networks); 3) regression methods (Linear Regression, MLR - Multivariate Linear Regression, and PLS - Partial Least Squares Regression).

Exploratory data analysis looking for the relationships between samples or relationships between variables. PCA and CA are most often used. Principal component analysis (PCA) involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. Principal components, PC, are formed by combination of the original data in such a way that the first principal component (PC1) covers as much of the variation within the data set as possible. The second principal component (PC2) describes the maximum amount of residual variation after the first PC has been taken into consideration, etc. The results of a PCA are usually discussed in terms of component scores and loadings. Using only a limited number of PCs, the dimensionality of the data space is reduced, thereby simplifying further analysis.

Cluster analysis is used for classification of objects into different groups or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure.

PLS regression is a technique that generalizes and combines features from principal component analysis and multiple regression. It is particularly useful to predict a set of dependent variables from a (very) large set of independent variables (i.e. predictors), and when the matrix of predictors has more variables than observations (multicollinearity). By contrast, standard regression will fail in these cases.

Table 1. The investigated compounds



for 1-12	$R_1 = H$	for 23-33	$R_1 = CO\text{-}2'\text{-}C_4H_3O$
for 13-22	$R_1 = COOC_2H_5$	for 34-45	$R_1 = COO\text{-}CH_2\text{-}C_6H_5$
Cmp.	R_2	Cmp.	R_2
1	C_6H_3-2'-F,6'-Cl	23	C_6H_3-2'-F,6'-Cl
2	1'-$C_{10}H_7$	24	1'-$C_{10}H_7$
3	C_6H_4-4'-OCH_3	25	-
4	C_6H_4-2'-OH	26	C_6H_4-2'-OH
5	2'-C_4H_3S	27	2'-C_4H_3S
6	C_6H_4-4'-$N(CH_3)_2$	28	C_6H_4-4'-$N(CH_3)_2$
7	C_6H_4-4'-Br	29	2'-C_4H_2O-5'-CH_3
8	C_6H_3-3',4'-OCH_3	30	C_6H_3-3',4'-$(OCH_3)_2$
9	C_6H_4-4'-$CH(CH_3)_2$	31	C_6H_4-4'-$CH(CH_3)_2$
10	C_6H_4-4'-OC_2H_5	32	C_6H_3-3',4'-OCH_2O
11	C_6H_5	33	C_6H_5
12	-	34	-
13	C_6H_3-2'-F,6'-Cl	35	C_6H_4-4'-$N(CH_3)_2$
14	1'-$C_{10}H_7$	36	1'-$C_{10}H_7$
15	C_6H_4-4'-OCH_3	37	C_6H_4-2'-OH
16	C_6H_4-2'-OH	38	C_6H_3-3',4'-OCH_2O
17	2'-C_4H_3S	39	C_6H_5
18	C_6H_4-4'-$N(CH_3)_2$	40	3'-C_4H_3S
19	2'-C_8H_5-N-$COOC_2H_5$	41	C_6H_3-3',4'-$(OCH_3)_2$
20	C_6H_3-3',4'-$(OCH_3)_2$	42	C_6H_4-4'-Br
21	C_6H_4-4'-$CH(CH_3)_2$	43	2'-C_4H_2O-5'-CH_3
22	C_6H_5	44	C_6H_4-4'-$OCH_2C_6H_5$
		45	C_6H_4-4'-OC_2H_5

The goal of the present study is to evaluate the physico-chemical properties that affected the experimental data of 45 newly synthesized compounds derivatives of 2,4-dioksotetrahydro-1,3-thiazole (Table 1). In order to quantify molecular structure various molecular descriptors were calculated for each of the 45 compounds. Evaluation of physico-chemical properties was done by the partial least-squares regression.

2. Experimental

Thin-layer chromatography (TLC) has been used as experimental technique. TLC was performed on 20×20 cm glass-backed plates with thin layers of rice starch [3], prepared in our laboratory. The mobile phases were mixture of aqueous ammonia and acetone with volume fraction of acetone in the range 0-24% (v/v); increment 4%. Following development and drying plates, the spots were observed under a UV light at =254 nm. Results of experiments were expressed as R_M value, defined by the Bate-Smith and Westall [1] equation:

$$R_M = log(\frac{1}{R_F} - 1) \tag{1}$$

where R_F is the retardation factor defined as the ratio of the distance traveled by the center of the spot to the distance simultaneously traveled by the mobile phase. Linear extrapolation to zero organic component volume fraction the R_M^0 values were calculated from the equation:

$$R_M = R_M^0 - b\varphi \tag{2}$$

where φ stands for the volume fraction of the organic component in the mobile phase, and b is the slope. The R_M^0 and b values are listed in Table 2. Equations (1) and (2) are the bases for the QSPR studies.

Table 2. The values of R_M^0 and b of linear equation $R_M = R_M^0 - b\varphi$.

Co.	$\mathrm{R}_M{}^0$	b	Co.	$\mathrm{R}_M{}^0$	b	Co.	$\mathrm{R}_M{}^0$	b
1	0.259	3.347	16	-0.198	2.109	31	0.889	4.965
2	0.433	3.875	17	0.439	4.802	32	0.535	3.562
3	0.480	4.265	18	0.508	3.993	33	0.349	3.008
4	-0.321	1.396	19	0.471	4.198	34	-0.533	0.270
5	0.348	3.020	20	0.020	2.247	35	0.472	3.733
6	0.360	3.279	21	0.763	4.449	36	0.553	4.615
7	0.631	3.882	22	0.341	3.697	37	-0.360	0.439
8	0.017	2.884	23	0.335	4.230	38	0.526	4.004
9	0.808	4.997	24	0.625	5.098	39	0.241	2.687
10	0.428	4.047	25	-0.120	1.128	40	0.334	2.906
11	0.241	2.767	26	-0.166	0.849	41	0.050	2.899
12	-0.301	0.359	27	0.425	3.703	42	0.644	3.911
13	0.343	4.812	28	0.528	3.585	43	0.134	1.729
14	0.555	4.906	29	0.182	2.749	44	1.089	5.438
15	0.238	3.449	30	0.177	2.917	45	0.497	4.451

2.1. Molecular descriptors and statistical calculations

Molecular descriptors were calculated using QSAR and SciLogP option of the molecular modeling computer programs ALCHEMY 2000, Interactive Analysis LogP and LogW predictor website, and SC ChemDraw software. Calculations were performed using Statistica software.

3. Results and Discussion

3.1. Descriptor selection - PCA

In order to evaluate the physico-chemical properties of investigated compounds 48 descriptors were calculated and PCA performed for data matrix X(48×45) with 48 descriptors and 45 compounds. To determine the number of principal components (PCs) of data set the modeled variance was used. The first principal component explains up to 63.5894% of the variability, the second accounts for up to 8.4862%, third 7.6498%, and so on. Figure 1 shows scree plot and Figure 2 projection of the variables on the plane (PC1xPC2).

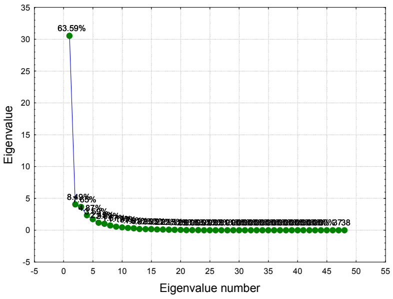


Figure 1: Scree plot.

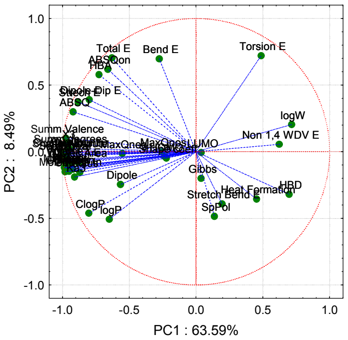


Figure 2: Projection of the variables PC1 vs. PC2.

Some of the descriptors clustered in the 2D and 3D loading plots, which means that they describe similar information. For the subsequent calculations the number of descriptors was reduced by choosing a representative one for each cluster of variables and removing those descriptors related to the same information. After that, the initial number of 48 descriptors was reduced to 8. These 8 selected descriptors were further used in the PLS regression.

3.2. Partial least squares

For PLS analysis the R_M^0 values were employed as the responses (Y-variables) and selected molecular descriptors were used as the predictive X-variables. The non-iterative partial least squares (NIPLAS) algorithm, optimized by cross-validation procedure was employed. Prior to the PLS analysis, the X and Y data were first mean centered and scaled to unit variance (i.e. the mean was subtracted and then divided by the standard deviation of the variable). The PLS modeling yielded two significant PLS component model. Sum of squares of the dependent variables (R2) explained by model were 69.3584%, see Table 3.

Table 3. Partial Least Squares Analysis Summary.

Co.	**R^{2}X**	**R^{2}X(Cum)**	**Eignval.**
1	0.3752	0.3752	2.3090
2	0.1670	0.5422	1.2572
3	0.0925	0.6348	0.4869

Co.	**R^{2}Y**	**R^{2}Y(Cum)**	**Q^2**	**Q^2(Cum)**	**Sigf.**	**Iter.**
1	0.4472	0.4472	0.0395	0.039514	S	1
2	0.2464	0.6936	0.1954	0.227186	S	1
3	0.0399	0.7335	-0.4321	-0.106719	NS	1

The number of components is 3, and 73.3486% of sum of squares of the dependent variables has been explained by all the extracted components.

The obtained PLS model gives a good correlation between the experimental and calculated retention data, as shown in Figure 3.

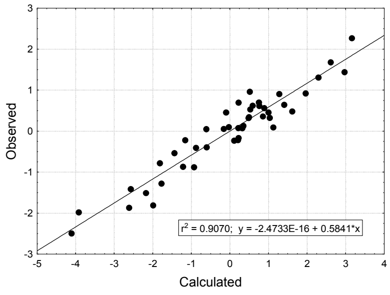


Figure 3: Correlation between experimental TLC data and values predicted by PLS method.

Using the PLS model we were able to identify the most important descriptors and rank them to the significance. Each descriptor is uniquely and independently described by its VIP, which is normalized so that they can be compared. Higher VIPs indicates more significant influence on retention. A variable with a modeling power equal to one is completely relevant for building the PLS model. Variables with lower modeling power are regarded to be less significant. Thus, in the present study, the VIPs were useful to judge the importance of descriptors for the retention of compounds studied in particular mobile phase. Table 4 lists the VIPs of the descriptors of the above models.

Table 4. Variable importance (VIP) of selected descriptors.

Descriptor	**VIP**	**importance**	**Descriptor**	**VIP**	**importance**
Polar	0.840	1	MaxQneg	0.588	5
Dipole	0.750	2	Gibbs	0.398	7
logP	0.694	3	ABSQon	0.287	6
logW	0.623	4	HOMO	0.158	8

In general, physico-chemical properties that highly influence the experimental results are electronic characteristics of the molecule (molecular polarizability and dipole moment), as well as its lipophilic character (logP). This is not surprising since both rice starch support and thiazoles are specific solutes capable for polar interactions [2].

4. Conclusions

An insight into the complex interactions that exist between the solutes, mobile and stationary phase was successfully investigated using partial least square regression. The correlation obtained between chromatographic retention data and structure descriptors for substituted 2,4-dioksotetrahydro-1,3-thiazoles is highly significant and might be used to predict the retention behavior. Also, using PLS model the most influenced descriptors to the retention were identify. The most important for the retention lipophilicity and electronic descriptors identified were those that account for polar interactions of the solutes.

Acknowledgement

This work was performed in the framework of the research projects: "Investigation of the Synthesis, Structure and Activity of Natural and Synthetic Organic Compounds" (Grant number 1420632) and "Numerical linear algebra, stochastic and statistics with applications" (grant number 144025) supported by the Ministry of Science and Technological Development of the Republic of Serbia.

References

[1] Bate-Smith, E.C., Westall, R.G., Chromatographic behavior and chemical structure. I. Some naturally occurring phenolic substances. Biochim. Biophys. Acta 4 (1950), 427-440.

[2] Brzezinska, E., Koska G., Walczynski, K., Application of thin-layer chromatographic data in quantitative structure-activity relationship assay of thiazole and benzothiazole derivatives with H -antihistamine activity. I. J. Chromatogr. A 1007 (2003), 145-155.

[3] Perišić-Janjić, N.U., Djaković, T.Lj., Petrović, S.M., Thin-layer Chromatography of Benzamides on Cellulose and Unconventional Starch and Aminoplast Supports. Chromatographia 40 (1995), 96-98.

Received by the editors September 1, 2008