STUDIA UBB CHEMIA, LX, 1, 2015 (p. 265-280)
(RECOMMENDED CITATION)

MODELLING AND PREDICTION OF LIPOPHILICITY FOR NATURAL COMPOUNDS WITH STRONG BIOLOGICAL ACTIVITY

COSTEL SÂRBU[a], RODICA DOMNICA NAŞCU-BRICIU[a,*]

ABSTRACT. The goal of this study was to develop high statistical significant models for lipophilicity estimation for a group of 60 compounds with increased toxicity, belonging to alkaloids and mycotoxins. The multiple linear regression modelling was made by means of genetic algorithms as a function of 972 molecular descriptors, computed by ChemOffice and Dragon Plus software and completed by internet available module for Log P computation, ALOGPS 2.1. The compounds classification has been realized using principal component analysis and hierarchical cluster analysis. Data evaluation has been realized by various correlation matrices and relevant graphs. The modelling was made on the basis of 26 compounds with known log P_{exp} values and the results were validated by means of additional models developed for a series of 20 compounds. The other 6 compounds, which were excluded from the modelling process, were used afterwards as test set for prediction and comparison. The most descriptive models were those retaining four descriptors, and in all models were selected at least one computationally expressed log P value (miLogP, KOWWIN and ALOGP most often). All the obtained results are highly suggestive and offer a very pertinent idea regarding the lipophilicity range of natural compounds of increased toxicity. The models were validated considering various statistical parameters and different correlation matrices defined by high statistical significance.

Keywords: *modelling, alkaloids, mycotoxins, toxicity*

INTRODUCTION

The importance of natural products in diseases prevention is known from ancient times and used up to early 1900s, when the “Synthetic Era” began and an increased tendency to replace the natural product drugs with synthetic

[a] *Babeş-Bolyai University, Faculty of Chemistry and Chemical Engineering, 11 Arany Janos Str., RO-400028, Cluj-Napoca, Romania*
** Corresponding author: rodicab2003@yahoo.com*

ones has been observed [1]. However, up today they are still intensively used in prevention and treatment of cancer and infectious diseases [2, 3]. On the other side, despite of beneficent effects, some of the most controversial and toxic compounds are also of natural origin. Amongst them there may be mentioned the alkaloids and the mycotoxins. The alkaloids are substances synthesized by plants [4] or even animals [5], which are defined by slightly basic properties, usually associated to the content of nitrogen atoms. These compounds are sophisticatedly combining the beneficent and toxic activities. For examples, the cinchona alkaloids are generally used for malaria treatment, but in overdose it may cause the disease called cinchonism, or the morphine is well known as a strong analgesic drug, but it may cause addiction and also in overdose is leading to asphyxia and death. In the majority of cases the alkaloids are central nervous system stimulants and they are inducing a large variety of biological effects [6]. On the other side the mycotoxins, are natural compounds biosynthesized by moulds, which unlike the alkaloids, do not have any biological beneficent effect. They have strong toxic effects over human and animal health, causing cancer, infertility, liver failure, cirrhosis, etc [7-10].

Most of the biological effects are interconnected to compounds lipophilic character. A typical example is represented by heroin and morphine. These compounds are differing just by the nature of the functional groups (methoxyl or hydroxyl). The heroin, which has a higher lipophilicity induced by the methoxyl groups, may easily crossover the biological membranes and exercises a very strong biological activity after hydrolysis. On the other side, the morphine is crossing the biological membranes more difficult because of the hydroxyl groups and implicitly its activity is lower [11, 12]. The lipophilicity is a property often used in strategies proposed to enhance the passive internalization of drugs into cells [13]. In fact, lipophilicity is an important endpoint used extensively in medicinal chemistry and environmental toxicology in predicting biological and hazardous effects of chemicals [14]. The lipophilicity is experimentally determined as partition coefficient (log P) between two immiscible phases (usually octanol-water), but it may be also computationally expressed [15]. Furthermore, it is a major experimental and theoretical tool in numerous disciplines, including medicinal chemistry, toxicology, pharmacology, and environmental monitoring [16-18]. The lipophilicity of a solute controls its distribution among body fluids, liquid-rich phases, and tissue proteins. For this reason, a quantitative assessment of lipophilicity is of great importance in quantitative structure–activity (or –property) relationship (QSAR/QSPR) studies [19-21]. This methodology is highly advantageous because by means of a generated mathematical model, many effects may be predicted [22-24]. The ideal goal of investigations concerned with the QSPR/QSAR is to predict the behaviour of chemical species from a minimal set of input data [14]. There are many possibilities of modelling,

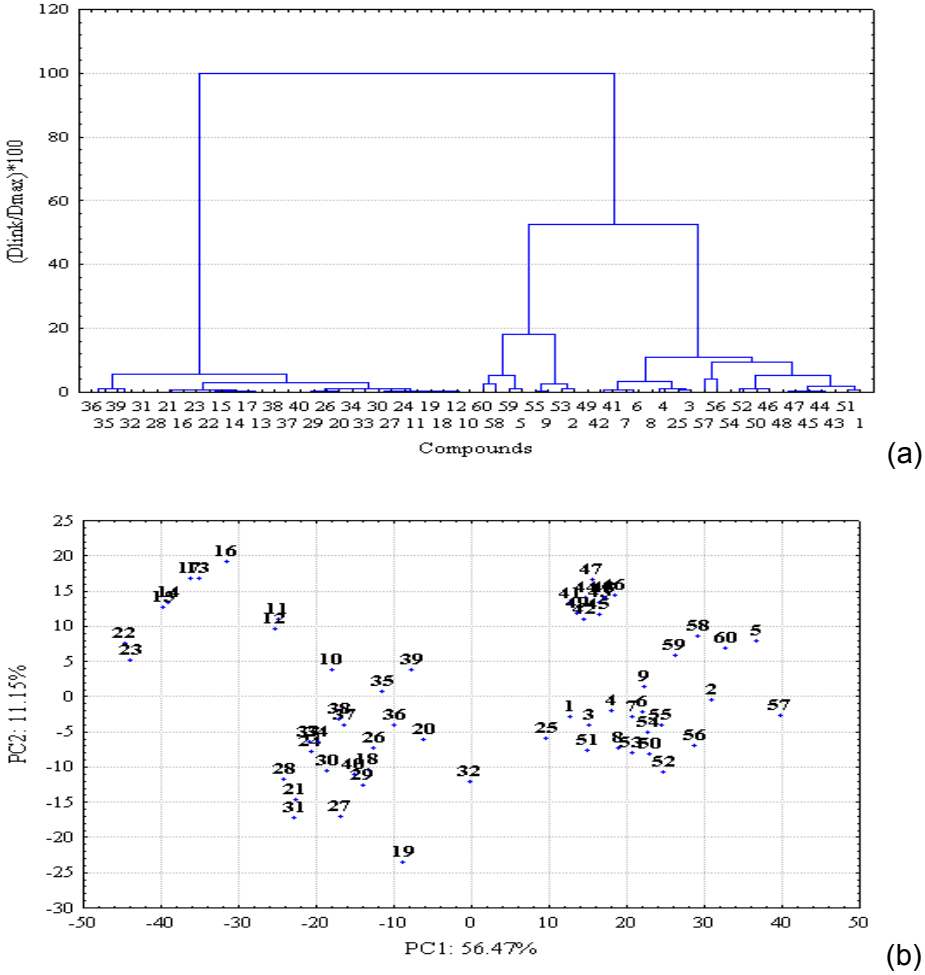
but one of the most preferred techniques is multiple linear regression (MLR), which is also completed nowadays by partial least square (PLS), principal component regression (PCR), genetic algorithms (GA) and even fuzzy clustering [25-27].

In the view of above considerations, our goal was to construct simple models for predicting the lipophilicity of various alkaloids and mykotoxins of increased toxicity, based on the summation of a contribution value for various physicochemical and structural features. In this work we have attempted to determine the important factors influencing the lipophilicity of toxic compounds and give them quantitative values. The chemometric techniques involved in this study are MLR, PCA, GA, PCR and cluster analysis (CA). The actual study is built on 60 compounds belonging to alkaloids and mycotoxins. Compounds selection for the analysis was based on the known toxic effects and the large diversity concerning their physicochemical characteristics and biological activity. The experimentally determined partition coefficient (Log P_{exp}) was employed as dependent variable. All the validation procedures strongly support the reliability and quality of results obtained.

RESULTS AND DISCUSSION

The selected compounds, belonging to the mycotoxins and alkaloids, are always presenting complex structures, with nitrogen and/or oxygen heteroatoms involved in the aromatic system. In order to reveal the similarity and differences between the selected compounds some exploratory investigations have been performed, by CA (Figure 1(a)) and PCA (Figure 1(b)). The multivariate exploratory methods were applied on 980 descriptors computed with Chem 3DUltra, Dragon Plus and ALOGPS 2.1. According to Figure 1(a) the investigated compounds are forming two major groups, one formed by narcotics, mycotoxins, and quinine derivatives, while the second is formed by nicotine and caffeine derivatives. Moreover, if investigating Figure 1 (b) there may be observed that the first principal component (PC1) is linearly distributing the investigated compounds and along of this direction the compounds are not separated in groups. Once again the aflatoxins group is considered highly similar with narcotics, quinine derivatives and also ochratoxins. Moreover, the caffeine and nicotine derivatives exhibit high similarities. The literature data are indicating that compounds **18**, **19** and **20** (cytisine and its derivatives) are nicotinic acetylcholine receptor agonist, which means that they interact with the same receptors in the brain, and the cytisine intoxication is similar to nicotine poisoning [11]. The fact that all these compounds are highly related in the PCA chart is an indicative that the entire phenomenon is dictated by their chemical structure and induced lipophilicity.

The selected group of compounds is defined by a large variety of chemical structures, which makes it to be potently used for lipophilicity prediction. The importance of such a study occurs from the high impact of the selected compounds over human and animal health. In this view the experimental lipophilicity (Log P_{exp}) of the selected compounds has been modelled in two different ways: firstly all Log P_{exp} presented in Table 1 were used to develop the models (group A), while in the second situation, groups of 6 values were eliminated and then predicted by means of the obtained models. Afterwards,



(a)

(b)

Figure 1. Cluster analysis (a) and principal component analysis (b) of the investigated compounds considering all descriptors.

the obtained values were compared to the observed ones. In the group B were eliminated the values that were covered very well the lipophilicity range, but there was never selected the extremes. In the groups C and D the selected compounds which were left out were different by those selected in group B and finally in the case of group E, the eliminated compounds were those with the maximum (3 compounds) and minimum (3 compounds) lipophilicity level. The obtained models along some statistical parameters (determination coefficient: R^2, F value, standard deviation: s, cross validated determination coefficient: Q^2, standard deviation error in calculation: SDEC, standard deviation error of prediction: SDEP, the prediction sum of squares: PRESS or root sum

Table 1. The lipophilicity indices of the investigated compounds

No.	$LogP_{exp}$	ALOGPs	AC logP	miLogP	KOWWIN	XLOGP2	XLOGP3	ALOGpS	AC logS	Hy	MLOGP	ALOGP
1	2.30	1.97	2.07	2.39	2.17	2.31	2.30	-1.78	-2.55	-0.76	2.41	2.12
2	1.58	2.30	2.25	1.61	1.80	1.69	1.49	-3.14	-3.55	-0.77	2.75	1.80
3	0.76	0.99	1.53	1.10	0.72	0.76	0.76	-1.45	-2.55	0.28	1.93	1.39
4	1.14	1.20	1.72	1.41	1.28	1.08	1.14	-2.72	-2.86	-0.35	2.17	1.64
5		2.00	2.49	2.82	1.97	2.49	2.71	-3.36	-3.40	-0.73	1.51	3.01
6		2.78	2.42	1.74	2.24	1.48	2.06	-2.68	-2.84	-0.82	2.32	1.80
7	1.01	1.06	1.21	0.90	0.88	0.12	0.95	-1.73	-2.47	-0.33	1.38	0.78
8	0.92	1.81	1.70	1.54	1.83	0.71	1.68	-1.89	-2.61	-0.81	2.17	1.78
9	2.95	4.19	3.50	3.52	3.69	3.00	3.94	-4.42	-4.23	-0.79	2.44	3.50
10	-0.07	-1.22	-0.77	0.87	-1.96	-0.60	-1.32	-0.47	-0.21	-0.56	0.29	-0.35
11	-0.78	-0.46	-0.51	-0.95	-0.05	-0.64	-0.78	-1.27	-1.51	0.02	0.40	-0.31
12	-0.02	-0.24	-0.09	0.00	-0.39	-0.48	-0.02	-0.80	-1.48	0.02	-0.00	-0.58
13	-0.73	-0.65	-0.25	-1.09	-1.15	-1.08	-0.73	-1.49	-1.87	1.84	-0.33	-0.72
14	-1.11	-0.85	-0.44	-0.73	-1.19	-0.24	-0.49	-1.43	-1.57	0.94	0.18	-0.78
15	-0.55	-0.55	0.10	-0.27	-0.55	0.07	0.08	-1.02	-1.90	0.94	0.58	-0.10
16	-2.17	-2.00	-3.69	-1.44	-3.61	-2.11	1.63	-1.48	-1.59	2.78	-1.23	-1.76
17		-0.29	0.31	-0.17	-0.28	0.40	0.06	-1.20	-1.87	1.84	0.48	0.28
18		0.56	0.79	0.27	0.60	0.24	0.18	-1.25	-1.30	-0.24	1.40	-0.34
19		1.66	1.64	2.00	2.12	1.22	1.63	-1.05	-1.09	-0.84	2.19	1.42
20		1.39	2.00	1.01	1.03	0.85	0.79	-3.38	-2.31	-0.29	1.11	1.04
21	1.17	0.87	1.24	1.09	1.00	1.13	1.17	-0.24	-0.79	-0.81	1.27	1.24
22	0.36	0.29	0.43	0.27	0.69	0.39	0.36	-0.17	-0.83	0.00	0.12	0.28
23	-0.34	-0.45	-0.11	-0.48	-0.45	-0.34	-0.37	-0.39	-0.91	0.84	-0.29	-0.32
24		0.55	1.00	0.47	0.38	0.68	0.68	-0.30	-1.24	-0.75	0.75	0.46
25		3.10	3.11	2.96	2.95	2.95	3.08	-3.89	-3.19	-0.37	2.04	3.02
26		-0.96	1.23	-1.45	-1.55	1.70	-0.34	-2.71	-2.97	-0.70	-2.61	0.36
27		1.52	1.67	1.47	1.49	1.55	1.54	-0.52	-1.09	-0.82	1.56	1.59
28		2.02	1.49	1.51	3.74	1.15	0.96	-2.48	-2.22	-0.81	1.45	1.39

No.	$LogP_{exp}$	ALOGPs	AC logP	miLogP	KOWWIN	XLOGP2	XLOGP3	ALOGpS	AC logS	Hy	MLOGP	ALOGP
29		-0.01	-0.06	0.00	-0.06	0.23	0.17	-0.87	-0.52	0.48	0.73	0.05
30		0.89	0.99	0.86	0.72	1.04	0.86	-0.70	-1.39	0.53	1.52	1.04
31	0.87	0.90	1.27	0.45	1.28	1.25	0.97	-1.21	-1.42	-0.24	1.27	1.16
32		1.61	1.70	0.30	1.72	1.38	1.42	-3.39	-1.86	-0.33	1.54	1.30
33	0.07	0.39	0.63	0.41	0.34	0.39	-0.32	-0.18	-0.77	-0.75	0.75	0.30
34		0.41	0.73	0.48	0.34	0.48	0.28	-0.13	-0.80	-0.75	0.75	0.73
35		0.05	-0.53	-0.24	-0.48	-0.56	-0.58	-1.07	-0.41	-0.17	0.17	-0.41
36		0.05	0.18	0.06	-0.80	-0.29	-0.42	-1.54	-0.86	-0.73	0.43	-0.42
37	-1.54	-0.66	0.36	-0.84	-1.80	1.66	-1.31	-1.24	-3.16	-0.70	-1.17	-0.55
38	-1.45	-0.32	-0.20	-0.50	-1.20	0.06	-0.87	-0.61	-0.38	-0.17	-0.06	-0.51
39		-0.64	-0.69	-0.95	-1.32	-0.33	-0.78	-1.11	-0.39	0.53	-0.33	-0.90
40		0.35	0.48	0.56	-0.05	0.30	0.43	-0.64	-0.44	-0.24	0.73	0.34
41		1.61	1.81	1.48	-0.38	0.78	1.62	-2.69	-3.27	-0.73	1.88	1.24
42		1.52	2.16	1.57	-0.17	0.44	1.33	-2.68	-3.04	-0.73	1.96	1.63
43		0.81	1.81	0.91	-1.71	-0.10	0.81	-2.03	-2.86	-0.26	1.60	1.03
44		1.76	0.97	1.52	-1.12	1.00	1.77	-2.53	-2.98	-0.70	1.93	1.25
45		1.81	1.32	1.61	-0.91	0.66	1.47	-2.56	-2.75	-0.70	2.01	1.64
46		0.81	0.97	0.95	-2.44	0.13	0.95	-1.75	-2.57	-0.24	1.67	1.04
47		1.17	1.06	0.90	-1.88	0.09	0.54	-2.18	-2.76	-0.26	1.12	0.63
48		1.06	1.41	1.00	-1.67	-0.51	0.25	-1.97	-2.54	-0.26	1.20	0.75
49		1.69	1.64	1.14	-0.16	1.08	1.22	-2.47	-3.21	-0.28	1.69	1.28
50	3.44	2.82	2.84	3.06	3.29	2.60	2.88	-2.99	-3.10	-0.38	2.19	2.73
51	2.68	3.20	2.94	3.03	3.21	2.69	2.68	-2.89	-3.08	-0.40	2.50	2.75
52		3.36	3.13	3.30	3.43	2.97	3.69	-3.02	-3.16	-0.38	2.28	3.13
53		2.77	3.25	3.22	3.49	2.92	2.99	-4.76	-4.23	-0.38	2.38	2.92
54		2.85	2.92	3.40	4.24	2.89	3.37	-3.48	-3.83	-0.82	2.38	3.13
55		1.36	2.47	4.53	3.60	4.04	3.96	-4.95	-4.50	-0.79	2.71	3.95
56	1.93	1.68	1.48	1.84	1.85	0.57	1.93	-2.29	-3.55	-0.83	2.90	1.15
57	0.98	1.85	1.27	1.46	1.49	0.13	0.98	-3.09	-3.59	-0.79	2.31	1.11
58	4.74	3.18	2.56	1.74	4.41	3.67	4.74	-4.25	-4.22	0.95	3.21	3.35
59		3.72	1.95	1.13	3.77	3.05	4.11	-4.32	-3.48	0.95	2.72	2.68
60		3.49	3.01	3.81	4.70	3.99	5.07	-4.71	-4.34	0.26	3.43	3.60

square: RSS) are enlisted in Table 2. Analysing the obtained models there may be remarked that in each model at least one of the computed log P values was selected (mostly KOWWIN, ALOGPs, miLogP, ALOGP and XLOGP3). These values are computed by fragmental (KOWWIN and miLogP), atomistic (ALOGP-using Ghose and Crippen algorithm), topological (ALOGPs) and also empirical (XLOGP3) algorithm. The selection of these lipophilicity descriptors in the best models is not randomly since during the last years, they have proved to be the most descriptive computed lipophilicity indices [28, 29]. Moreover, according to the selected descriptors, the log P_{exp} values are a consequence of

Table 2. The GA linear multiple regression models for lipophilicity prediction

No	Models	Q^2	R^2	s	F	SDEP	SDEC	PRESS	RSS
Models build on 26 compounds (Group A)									
1	Log P_{exp} = 0.1395 + 0.8255 KOWWIN	0.88	0.91	0.52	230	0.56	0.50	8.21	6.41
2	Log P_{exp} = - 0.2349 + 0.6514 KOWWIN + 0.0125 Vs	0.91	0.93	0.45	155	0.48	0.42	6.03	4.67
3	Log P_{exp} = 0.1091 + 1.0709ALogPs – 0.9774Mor20m + 4.7996Mor32m	0.96	0.97	0.32	213	0.33	0.29	2.77	2.26
4	Log P_{exp} = 1.2519 + 0.3954miLogP + 0.5413KOWWIN + 3.4885Mor32m – 5.7300HATS1u	0.96	0.98	0.26	255	0.32	0.23	2.64	1.37
5	Log P_{exp} = - 0.6331 + 1.1622ALogPs + 0.2518X1v – 1.6898MATS3e – 0.2352DP09 + 4.5783Mor32m	0.98	0.99	0.22	276	0.23	0.19	1.41	0.97
Models build on 20 compounds (Group B; without compounds **4**, **13**, **23**, **33**, **51**, **57**)									
1	Log P_{exp} = 0.1816 + 0.83271KOWWIN	0.86	0.90	0.58	161	0.64	0.55	8.19	6.06
2	Log P_{exp} = - 0.2941 + 0.6324KOWWIN + 0.0298As	0.91	0.93	0.50	112	0.53	0.46	5.71	4.23
3	Log P_{exp} = - 0.9193 – 1.9805MATS3e + 0.0912RDF055m + 1.1346ALOGP	0.94	0.96	0.37	140	0.41	0.33	3.40	2.20
4	Log P_{exp} = -0.6413 + 0.8452ALogPs – 1.6617MATS3m + 3.5942Mor32m + 0.4381MLOGP	0.96	0.98	0.31	157	0.34	0.37	2.25	1.41
5	Log P_{exp} = 3.8334 – 0.9217ALogPs + 0.8379miLogP + 0.8175KOWWIN – 3.3724Ovality + 0.0005PMIY	0.96	0.98	0.26	178	0.35	0.22	2.42	0.93
Models build on 20 compounds (Group C; without compounds **1**, **7**, **21**, **31**, **37**, **56**)									
1	Log P_{exp} = - 0.2090 + 1.0901ALOGP	0.87	0.91	0.55	174	0.61	0.52	7.40	5.49
2	Log P_{exp} = - 0.4625 + 0.5522KOWWIN + 0.0116D/D	0.91	0.94	0.46	129	0.50	0.53	5.02	3.62
3	Log P_{exp} = - 3.3935 + 2.4782GATS3p + 0.0163Vs + 1.0494ALOGP	0.96	0.98	0.28	240	0.32	0.25	2.11	1.27
4	Log P_{exp} = - 4.7179 + 3.5208GATS3e + 0.7891Mor10v + 0.0106Vs + 1.2385ALOGP	0.98	0.99	0.18	426	0.22	0.16	0.97	0.51
5	Log P_{exp} = 0.0973 + 0.6455XLOGP2 – 3.2650MATS3e + 0.0660RDF015u – 8.0145G1m + 0.5829ALOGP	0.99	0.99	0.15	536	0.16	0.12	0.51	0.30
Models build on 20 compounds (Group D; without compounds **2**, **9**, **22**, **38**, **50**, **58**)									
1	Log P_{exp} = - 0.1992 + 0.9829miLogP	0.89	0.91	0.39	188	0.41	0.37	3.31	2.68
2	Log P_{exp} = - 0.0780 + 0.6053miLogP + 0.3286KOWWIN	0.95	0.96	0.27	207	0.28	0.24	1.61	1.21
3	Log P_{exp} = - 0.9215 + 0.3604miLogP + 2.2192DISPp + 0.4643MLOGP	0.95	0.97	0.24	171	0.26	0.22	1.40	0.93
4	Log P_{exp} = - 0.01782 + 0.7195miLogP + 0.3036 KOWWIN + 0.7766 MATS4m – 0.2757Mor08u	0.97	0.98	0.18	234	0.21	0.16	0.92	0.49
5	Log P_{exp} = - 0.8269 + 0.7924miLogP + 0.3270KOWWIN + 0.6819MATS4e – 0.3205Mor08e + 12.6679R4u+	0.99	0.99	0.13	352	0.14	0.11	0.39	0.24

No	Models	Q^2	R^2	s	F	SDEP	SDEC	PRESS	RSS
Models build on 20 compounds (Group E; without compounds **9**, **16, 37**, **38**, **50**, **58**)									
1	Log P_{exp} = - 0.05856 + 0.8986miLogP	0.89	0.91	0.33	177	0.34	0.31	2.30	1.94
2	Log P_{exp} = - 0.0541 + 0.5530miLogP + 0.4046XLOGP3	0.94	0.95	0.24	171	0.26	0.22	1.31	1.00
3	Log P_{exp} = 0.5240 + 0.4764miLogP + 0.3904XLOGP3 – 4.0970HATS3m	0.95	0.97	0.21	195	0.24	0.18	1.12	0.58
4	Log P_{exp} = 0.9268 + 0.4889miLogP + 0.4650XLOGP3 – 2.0570Mor27p – 14.8156R3m+	0.98	0.99	0.15	246	0.16	0.13	0.53	0.32
5	Log P_{exp} = 1.7449 + 0.3402miLogP + 0.5925XLOGP3 – 0.6149piPC03 – 0.1954Mor05u – 0.5576Mor30e	0.98	0.99	0.14	209	0.16	0.12	0.50	0.28

the molecular conformation/configuration and also of the atoms nature, since the most selected descriptors were from the 2D autocorrection and 3D MoRSE descriptors. These are completed by WHIM descriptors which are also very descriptive in partition coefficient prediction, since they are 3D dimensional descriptors based on the calculation of principal component axes computed from a weighted covariance matrix obtained by the molecule geometrical coordinates. They contain information concerning, size, symmetry, shape and distribution of the molecular atoms [30]. Other descriptors were belonging to the following categories: steric, Randic molecular profiles, RDF descriptors, topological descriptors, connectivity indices, geometrical descriptors and GETAWAY descriptors. The most selected descriptor (in all groups) was miLogP (12 times), followed by KOWWIN (10 times), ALOGP (5 times) and XLOGP3 (4 times). All the results and even the selected class of descriptors are in fair agreement with observations made by our group in previous studies [20, 21] and also by Benfenati [31].

Furthermore, as can be seen from the Table 2 the statistical parameters of quality are increasing while the number of selected variables is bigger. However, this doesn't means that the predictive capacity is higher, since the coefficients level may be a consequence of over-fitting. Even if the Q^2 has an increased value, the prediction capacity may be low, especially for compounds not included in the training set. In order to observe, which models are more valuable, a correlation matrix has been made between computed and experimental Log P values and the predicted ones (Table 3).

There may be observed that Log P_{exp} is strongly correlating to ALOGP and to KOWWIN, which is in fact expected since they were often selected in the best models obtained. In addition, it is strongly correlated to the predicted values. Investigating the correlation of the predicted values with experimental ones, there may be observed that the highest correlations were obtained for

Table 3. The correlation matrix between the experimental and computed log P values and the predicted ones (bold values indicate correlation coefficients higher than 0.90, while italic bolded values indicate correlation coefficients between 0.80 and 0.89)

Group	Predicted Log P	ALOGPs	AC logP	miLogP	KOWWIN	XLOGP2	XLOGP3	MLOGP	ALOGP	ALOGpS	AC logS	Log P_{exp}
A	LogP 1	***0.83***	0.78	0.78	**1.00**	***0.84***	***0.80***	0.74	***0.84***	-0.62	-0.53	**0.95**
	LogP 2	***0.88***	***0.81***	***0.82***	**0.98**	***0.88***	***0.87***	0.79	***0.89***	-0.72	-0.64	**0.97**
	LogP 3	**0.94**	***0.85***	***0.86***	***0.87***	***0.83***	***0.89***	***0.86***	**0.91**	-0.72	-0.65	**0.98**
	LogP 4	***0.84***	***0.80***	***0.88***	**0.95**	***0.86***	***0.88***	***0.80***	**0.91**	-0.69	-0.62	**0.99**
	LogP 5	**0.94**	***0.85***	***0.85***	***0.83***	0.79	***0.87***	***0.87***	***0.89***	-0.69	-0.66	**0.99**
B	LogP 1	***0.83***	0.78	0.78	**1.00**	***0.84***	***0.80***	0.74	***0.84***	-0.62	-0.53	**0.95**
	LogP 2	**0.91**	***0.84***	***0.85***	**0.95**	***0.88***	***0.89***	***0.82***	**0.92**	-0.76	-0.69	**0.96**
	LogP 3	***0.88***	***0.87***	**0.90**	***0.80***	***0.88***	**0.91**	***0.82***	**0.97**	-0.79	-0.79	**0.98**
	LogP 4	**0.93**	***0.82***	***0.85***	***0.81***	0.77	***0.88***	**0.91**	***0.89***	-0.69	-0.65	**0.99**
	LogP 5	0.75	0.73	***0.85***	**0.92**	***0.85***	***0.84***	0.74	***0.86***	-0.67	-0.60	**0.99**
C	LogP 1	**0.91**	**0.92**	**0.93**	***0.84***	**0.91**	**0.91**	***0.82***	**1.00**	-0.77	-0.75	**0.95**
	LogP 2	***0.86***	0.79	***0.81***	**0.95**	***0.89***	***0.88***	0.76	***0.89***	-0.74	-0.64	**0.96**
	LogP 3	***0.87***	***0.84***	***0.88***	0.77	***0.87***	**0.92**	0.79	**0.96**	-0.79	-0.78	**0.97**
	LogP 4	***0.84***	***0.83***	***0.86***	0.67	***0.81***	***0.87***	0.78	**0.93**	-0.76	-0.78	**0.98**
	LogP 5	***0.84***	***0.85***	***0.86***	0.78	**0.93**	***0.89***	0.74	**0.95**	-0.77	-0.77	**0.95**
D	LogP 1	***0.86***	***0.85***	**1.00**	0.78	0.79	***0.84***	***0.84***	**0.93**	-0.68	-0.67	**0.90**
	LogP 2	**0.90**	***0.87***	**0.96**	**0.93**	***0.86***	***0.87***	***0.85***	**0.94**	-0.69	-0.65	**0.95**
	LogP 3	***0.89***	***0.85***	**0.94**	***0.80***	0.77	***0.84***	**0.92**	**0.90**	-0.64	-0.61	**0.92**
	LogP 4	***0.89***	***0.85***	**0.96**	***0.89***	***0.83***	***0.87***	***0.87***	**0.93**	-0.67	-0.64	**0.95**
	LogP 5	***0.89***	***0.85***	**0.96**	***0.88***	***0.80***	***0.85***	***0.86***	**0.92**	-0.66	-0.63	**0.94**
E	LogP 1	***0.86***	***0.85***	**1.00**	0.78	0.79	**0.84**	***0.84***	**0.93**	-0.68	-0.67	**0.90**
	LogP 2	**0.90**	***0.84***	**0.97**	***0.82***	***0.84***	***0.95***	***0.85***	**0.96**	-0.76	-0.73	**0.93**
	LogP 3	**0.91**	***0.85***	**0.97**	***0.83***	***0.84***	***0.94***	***0.85***	**0.96**	-0.74	-0.71	**0.93**
	LogP 4	**0.90**	***0.82***	**0.94**	***0.81***	***0.82***	***0.94***	***0.85***	**0.94**	-0.73	-0.70	**0.93**
	LogP 5	***0.86***	0.77	**0.90**	***0.87***	***0.86***	***0.96***	0.79	**0.92**	-0.73	-0.67	**0.93**
Log P_{exp}		**0.92**	***0.87***	**0.90**	**0.95**	***0.84***	***0.86***	***0.89***	**0.95**	-0.72	-0.65	1.00

groups A and B, while the lowest is obtained in case of group E. This observation is indicating that for a more accurate prediction a very important step is the selection of the training set. The correlation with log P values obtained for group A was expected because there was no elimination made and each particular value has been contributed to the final model. On the other side, in case of group B, the elimination was made in such a manner that the entire range was described, without eliminating the extreme values, which allowed them to have a significant contribution to the final developed models. However, when the eliminated values were the extreme ones, the correlation was lower because the model was built in a reduced range and then used

for outside values prediction. On the basis of this observation there may be concluded that the selection of training set in case of prediction models generation is crucial and it may affect the reliability of the entire following results. Moreover, it is indicating that the elimination of the values, which would be used further for prediction evaluation (test set), must cover as much as possible the training, but the extreme values must be kept. The viability of the models which leaded to the best correlation with experimental values (models with 4 descriptors) is illustrated by the representation of the observed vs. predicted values (Figure 2). The statements above are also very well supported by histograms and normal distribution and box and whisker plot presented in Figure 3 and 4.

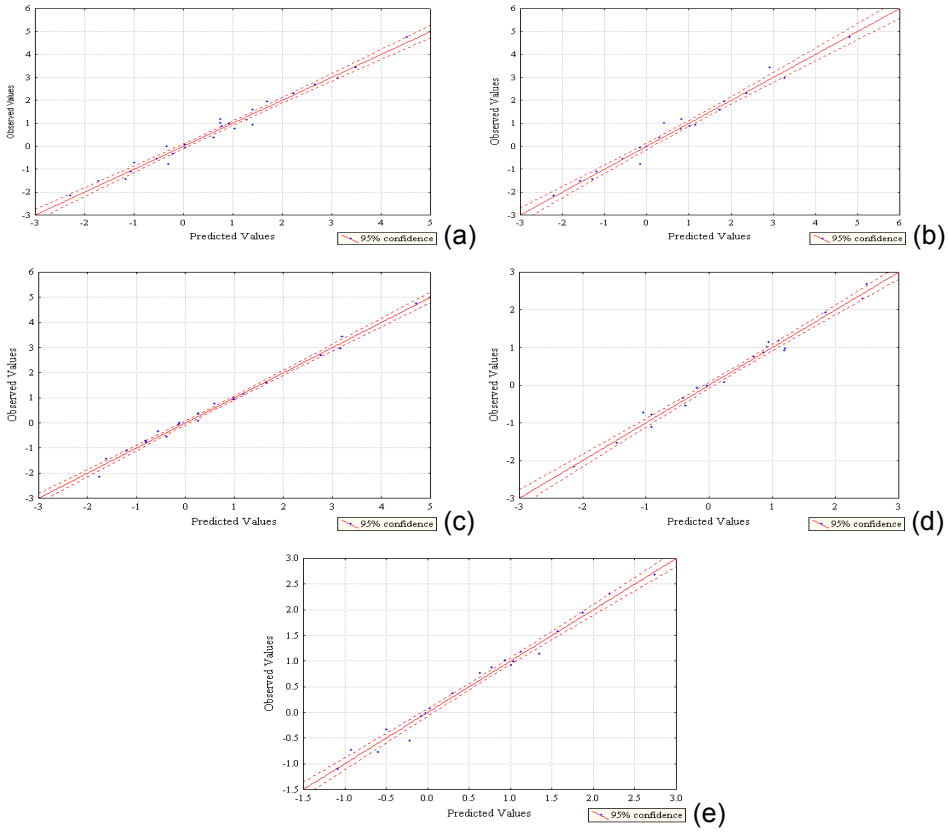


Figure 2. Graphs of predicted vs observed values corresponding to the best models: group A (a), group B (b), group C (c), group D (d), and group E (e).

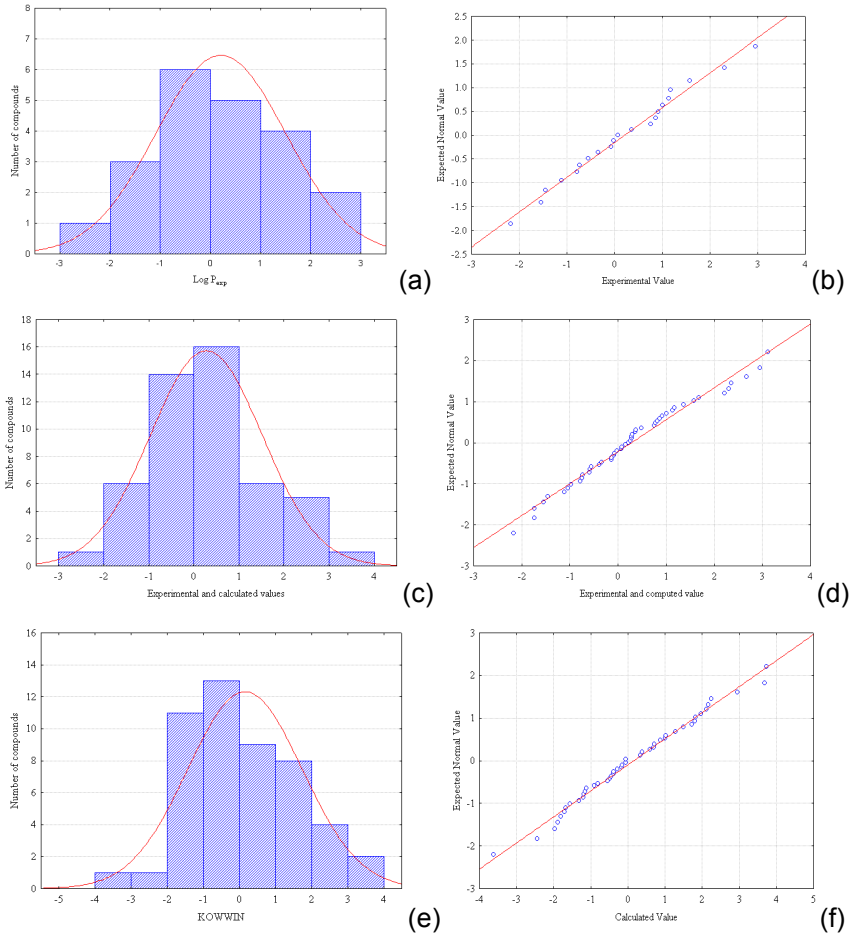


Figure 3. Histograms and normal probability plot of data corresponding to experimental (a,b), experimental including estimated values by model E (5)-(c,d), and KOWWIN values (e,f).

In order to observe the ability of the obtained models to predict the lipophilicity the list of Log P_{exp} has been completed with the predicted values (were missing) and correlated with the computed ones (Table 4). The maximum value of the correlation coefficient (r) in this case is 0.98, and it is observed that once again the best correlations are obtained with ALOGP and KOWWIN. In all cases the log S values have led to lower statistical correlations, which mean that the solubility descriptors are not describing very well the lipophilicity.

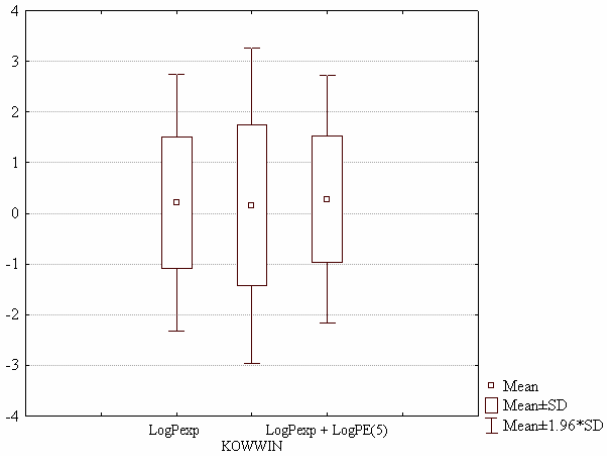


Figure 4. Box and whisker plot of data corresponding to experimental, experimental including estimated values by model E (5), and KOWWIN values.

In many situations the PCR combined with GA for property modelling is leading to more appropriate results. However, in our case the obtained models had lower statistical quality and prediction capacity. The highest Q^2 value that was reached was 0.91. If comparing to the models presented in Table 2, there is no doubt that the classical modelling has leaded to better results. This may be explained by the fact that the PCA has the ability to concentrate in the first few components the entire information contained in the initial matrix. This is in many cases advantageous, but as can be seen from the models presented in Table 2, only few descriptors are correctly defining the lipophilicity, while the others have no positive contribution. The inefficient prediction models obtained by PCR-GA may be a consequence of the large of useless information retained in the scores of the first principal components involved later in the modelling. The principal components are a consequence of all the descriptors and most of them do not describe correctly the lipophilicity, which is finally leading to the lower quality models.

As can be observed from Table 1, the lipophilicity range is generally between -3 to 3, which means that for a high biological activity is very important to have a certain level of lipophilicity, in order to cross the biological membranes. This range is indicating that for an increased activity the lipophilicity level doesn't have to be too high, when appear the situation of total absorption in the adipose tissue (i.e. skin discoloration in case of carotene overdose), or the opposite situation when the compounds are unable to cross in simple form the biological membranes (i.e. water biological membrane crossing by means of an aquapurine protein). These observations are indicating that the partition phenomena take place at cellular level when the lipophilicity is in an optimal range.

Table 4. The Correlation matrix between the experimental and predicted lipophilicity values vs. the computed lipophilicity indices (bold values indicate correlation coefficients higher than 0.90, while italic bolded values indicate correlation coefficients between 0.80 and 0.89)

Group	Log P*	ALOGPs	AC logP	miLogP	KOWWIN	XLOGP2	XLOGP3	MLOGP	ALOGP	ALOGpS	AC logS
A	1	***0.81***	0.75	0.79	**0.98**	***0.84***	***0.82***	0.73	***0.84***	-0.62	-0.52
	2	***0.86***	***0.80***	***0.83***	**0.96**	***0.87***	***0.87***	0.78	***0.89***	-0.69	-0.60
	3	**0.93**	***0.84***	***0.86***	***0.86***	***0.82***	***0.89***	***0.86***	**0.90**	-0.71	-0.65
	4	***0.84***	***0.80***	***0.88***	**0.94**	***0.86***	***0.88***	***0.80***	**0.90**	-0.69	-0.62
	5	**0.93**	***0.84***	***0.85***	***0.82***	0.79	***0.88***	***0.87***	***0.89***	-0.69	-0.65
B	1	***0.81***	0.76	0.79	**0.98**	***0.84***	***0.82***	0.73	***0.84***	-0.63	-0.52
	2	***0.88***	***0.82***	***0.86***	**0.94**	***0.87***	**0.90**	***0.81***	**0.92**	-0.73	-0.65
	3	***0.87***	***0.87***	**0.90**	***0.80***	***0.87***	**0.91**	***0.81***	**0.97**	-0.77	-0.76
	4	**0.93**	***0.82***	***0.85***	***0.80***	0.77	***0.88***	**0.90**	***0.89***	-0.68	-0.65
	5	0.74	0.72	***0.85***	**0.92**	***0.85***	***0.84***	0.72	***0.86***	-0.66	-0.59
C	1	***0.89***	***0.89***	**0.91**	***0.84***	***0.88***	**0.91**	***0.82***	**0.97**	-0.75	-0.73
	2	***0.84***	0.78	***0.82***	**0.94**	***0.87***	***0.87***	0.77	***0.88***	-0.71	-0.61
	3	***0.87***	***0.85***	***0.88***	0.78	***0.85***	**0.91**	***0.81***	**0.95**	-0.77	-0.76
	4	***0.85***	***0.84***	***0.87***	0.68	0.79	***0.87***	0.79	**0.93**	-0.75	-0.78
	5	***0.85***	***0.85***	***0.88***	***0.80***	***0.89***	**0.90**	0.78	**0.95**	-0.78	-0.74
D	1	***0.87***	***0.85***	**0.94**	***0.82***	***0.81***	***0.87***	***0.86***	**0.93**	-0.69	-0.68
	2	***0.88***	***0.84***	**0.91**	**0.92**	***0.86***	***0.88***	***0.84***	**0.93**	-0.69	-0.64
	3	***0.89***	***0.84***	**0.90**	***0.82***	0.79	***0.86***	***0.89***	**0.90**	-0.66	-0.63
	4	***0.89***	***0.84***	**0.93**	***0.89***	***0.84***	***0.89***	***0.87***	**0.93**	-0.69	-0.65
	5	***0.89***	***0.84***	**0.93**	***0.89***	***0.82***	***0.88***	***0.87***	**0.93**	-0.68	-0.66
E	1	***0.87***	***0.86***	**0.94**	***0.82***	***0.81***	***0.87***	***0.86***	**0.93**	-0.69	-0.68
	2	**0.90**	***0.87***	**0.93**	***0.84***	***0.86***	**0.91**	***0.86***	**0.96**	-0.73	-0.70
	3	**0.90**	***0.88***	**0.93**	***0.85***	***0.85***	**0.90**	***0.86***	**0.95**	-0.71	-0.68
	4	**0.91**	***0.87***	**0.91**	***0.84***	***0.84***	**0.90**	***0.87***	**0.95**	-0.70	-0.67
	5	***0.89***	***0.85***	**0.90**	**0.90**	***0.89***	**0.91**	***0.83***	**0.94**	-0.71	-0.65

* Log P_{exp} known + predicted

CONCLUSIONS

The quantitative structure-lipophilicity relationships of 60 structurally diverse bioactive compounds have been investigated in order to develop a sound predictive model for the lipophilicity estimation of natural compounds with strong biological activity. The lipophilicity expressed as Log P was found to be significantly influenced by a series of descriptors coded as 2D autocorrection,

3D MoRSE or WHIM. The modelling process validated through different methodologies of training set selection, has been supported by highly relevant statistical parameters and graphs. The obtained results were highly descriptive and in a very good agreement with the computed lipophilicity indices. The best models were those with four retained descriptors, but the models with three, two or one descriptor produce also relevant results in a very good agreement with computed lipophilicity indices. In addition, the results obtained in this study for a large number of important natural compounds add a real and useful contribution to the data collection concerning their characteristics and might be a starting point for future investigation concerning the optimal lipophilicity range.

EXPERIMENTAL SECTION

The selected compounds for this study belong to alkaloids and mycotoxins group, as follows: cocaine (**1**), heroine (**2**) morphine (**3**), codeine (**4**), noscapine (**5**), thebaine (**6**), oxycodone (**7**), hydromorphone (**8**), papaverine (**9**), caffeine (**10**), theobromine (**11**), theophylline (**12**), xanthine (**13**), hypoxanthine (**14**), allopurinol (**15**), uric acid (**16**), oxypurinol (**17**), cytisine (**18**), sparteine surrogate (**19**), varenicline (**20**), nicotine (**21**), nicotinic acid (**22**), nicotinamide (**23**), n-formyl nornicotine (**24**), brevicolline (**25**), pyridine,3-(1-methyl-1-oxido-2-pyrrolidinyl)-,1-oxide (**26**), n-ethyl nornicotine (**27**), anabaseine (**28**), rac-trans 3'-aminomethyl nicotine (**29**), rac-2-amino nicotine (**30**), anabasine (**31**), anatalline (**32**), cotinine (**33**), ortho-cotinine (**34**), rac-trans-cotinine carboxylic acid (**35**), rac-trans-cotinine carboxylic acid methyl ester (**36**), cotinine n-oxyde (**37**), 3-hydroxy cotinine (**38**), rac 3'-hydroxy cotinine 3-carboxylic acid (**39**), rac3'hydroxy-methyl nicotine (**40**), aflatoxin B1 (**41**), aflatoxin B2 (**42**), aflatoxin B2a (**43**), alatoxin G1 (**44**), aflatoxin G2 (**45**), aflatoxin G2a (**46**), aflatoxin M1 (**47**), aflatoxin M2 (**48**), aflatoxicol (**49**), quinine (**50**), cinchonine (**51**), hydroquinine (**52**), quinotoxine (**53**), quininone (**54**), bromoquinotoxine (**55**), strychnine (**56**), brucine (**57**), ochratoxin A (**58**), ochratoxin B (**59**) and ochratoxin C (**60**). The Log Pexp for 26 of the above mentioned compounds has been obtained from different databases (www.chemspider.com; http://esc.syrres.com/esc/est_kowdemo.htm; www.vcclab.org; and http://www.biobyte.com). Both experimental and computed lipophilicity indices are enlisted in Table 1. In order to obtain the desired information, the first two web pages had required the CAS registry number of the compounds, while the last two had required the SMILE formula, which was obtained on www.molinspiration.com.

The molecular descriptors for the selected compounds were computed with ALOGPS 2.1 Internet module, Chem Office 8.0 and Dragon Plus 5.4 software. The ALOGPS 2.1 was able to compute six log P values (ALOGPs,

AC logP, miLogP, KOWWIN, XLOGP2 and XLOGP3) and two Log S values (ALOGpS and AC logS) on the basis of compounds SMILE formula. Previous of descriptors computation, the structure of each molecule has been drawn in Chem Draw 8.0 application, and the obtained structures were further energetically optimised by means of molecular mechanics force field procedure included in Hyperchem version 8.0 (www.hyper.com) and the resulting geometries were further refined by means of the semi-empirical method Parametric Method-3 using the Fletcher–Reeves algorithm and a gradient norm limit of 0.009 kcal Å-1. The optimized geometries were loaded by the above presented software in order to calculate the molecular descriptors. The Chem Office 8.0 software through the application Chem 3DUltra allows the computation of 30 descriptors classified as electronic, steric and thermodynamic ones, while Dragon Plus 5.4 allows the computation of 942 descriptors belonging to the following groups: constitutional descriptors, walk and path counts, information indices, edge adjacency indices, Randic molecular profiles, RDF descriptors, WHIM descriptors, functional groups counts, charge descriptors, topological descriptors, connectivity indices, 2D autocorrelation, Burgen eigenvalues, eigenvalue based indices, geometrical descriptors, 3D-MoRSE descriptors, GETAWAY descriptors, atom centred fragments and molecular properties. The correlations, graphs, PCA and CA were realized by Statistica 8.0 software, while GA was made by MobyDigs 1.0 software. The developed models have been designed to retain 1 to 5 descriptors. The modelling has been realized on the basis of the matrix formed by the computed descriptors and also on the basis of the scores obtained by applying PCA on the matrix formed by the computed descriptors.

ACKNOWLEDGMENTS

This work was possible with the financial support offered by Romanian Ministry of Education, Research, Youth and Sport through research grant PN-II-ID-PCE-2011-3-0366 (Project Manager: C. Sârbu).

REFERENCES

1. J. D. McChesney, S. K. Venkataraman, J. T. Henri, *Phytochemistry,* **2007**, *68*, 2015.
2. D. J. Newman, G. M. Cragg, K. M. Snader, *Journal of Natural Products,* **2003**, *66* (7), 1022.
3. G. M. Cragg, D. G. I. Kingston, D. J. Newman, “Anticancer Agents from Natural Products”; CRC Press, Boca Raton, **2005**.
4. K. E. Panter, L. F. James, *Journal of Animal Science,* **1990**, *68*, 892.

5. R. A. Saporito, M. A. Donnelly, T. F. Spande, H. M. Garraffo, *Chemoecology,* **2012**, *22*, 159.
6. R. Andraws, P. Chawla, D. L. Brown, *Progress in Cardiovascular Diseases,* **2005**, *47*, 217.
7. A. Yiannikourisa, J.-P. Jouany, *Animal Research,* **2002**, *51*, 81.
8. N. W. Turner, S. Subrahmanyam, S. A. Piletsky, *Analytica Chimica Acta*, **2009**, *632*, 168.
9. C. A. Robbins, L. J. Swenson, M. L. Nealley, R. E. Gots, B. J. Kelman, *Applied Occupational and Environmental Hygiene,* **2000**, *15* (10), 773.
10. M. Z. Zheng, J. L. Richard, J. Binder, *Mycopathologia*, **2006**, *161*, 261.
11. S. Berger, D. Sicker, „Classics in Spectroscopy“, Wiley-VCH, Weinheim, **2009**.
12. K. C. Nicolaou, T. Morgan, “Molecules that changed the world*”,* Wiley-VCH, Weinheim, **2008**.
13. R. Pignatello, S. Guccione, S. Forte, C. Di Giacomo, V. Sorrenti, L. Vicari, G. U. Barretta, F. Balzanoc, G. Puglisi, *Bioorganic & Medicinal Chemistry,* **2004**, *12*, 2951.
14. A. A. Toropov, A. P. Toropova, *Journal of Molecular Structure: Theochem*, **2001**, *538*, 197.
15. I. V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. A. Palyulin, E. V. Radchenko, N. S. Zefirov, A. S. Makarenko, V. Y. Tanchuk, V. V. Prokopenko, *Journal of Computer-Aided Molecular Design,* **2005**, *19*, 453.
16. B. Testa, P. A. Carrupt, P. Gaillard, F. Billois, P. Weber, *Pharmaceutical Research,* **1996**, *13*, 335.
17. M. J. Waring, *Expert Opinion on Drug Discovery,* **2010**, *5*, 235.
18. Q. Du, G. A. Artec, *Journal of Computer-Aided Molecular Design,* **1996**, *10*, 133.
19. M.Kompany-Zareh, *Medicinal Chemistry Research,* **2009**, *18*, 143.
20. C. Sârbu, C. Onişor, M. Posa, S. Kevresan, K. Kuhajda, *Talanta*, **2008**, *75*, 651.
21. C. Onişor, M. Poša, S. Kevrešan, K. Kuhajda, C. Sârbu, *Journal of Separation Science* **2010**, *33*, 3110.
22. S. Trapp, R. W. Horobin, *European Biophysics Journal,* **2005**, *34*, 959.
23. M. Jaiswal, P. V. Khadikar, C. T. Supuran, *Bioorganic & Medicinal Chemistry Letters,* **2004**, *14*, 5661.
24. G. Caron, G; Ermondi, A. Damiano, L. Novaroli, O. Tsinman, J. A. Ruell, A. Avdeef, *Bioorganic & Medicinal Chemistry,* **2004**, *12*, 6107.
25. T. Takagi, M.Sugeno, *IEEE Transactions on Systems, Man, and Cybernetics,* **1998**, *15*, 116.
26. C. Sârbu, D. Casoni, A. Kot-Wasik, A. Wasik, J. Namieśnik, *Journal of Separation Science,* **2010**, *33*, 2219.
27. D. Casoni, J. Petre, V. David, C. Sârbu, *Journal of Separation Science,* **2011**, *34* (3), 247.
28. R. D. Briciu, A. Kot-Wasik, A. Wasik, J. Namieśnik, C. Sârbu, *Journal of Chromatography A*, **2010**, *1217,* 3702.
29. C. Sârbu, R. D. Naşcu-Briciu, D. Casoni, A. Kot-Wasik, A. Wasik, J. Namieśnik, *Journal of Chromatography A*, **2012**, *1266*, 53.
30. R. Todeschini, M. Lasagni, E. Marengo, *Journal of Chemometrics,* **1994**, *8*, 263.
31. E. Benfenati, “Quantitative Structure Activity Relationships for Pesticide Regulatory Purposes”, Elsevier, Amsterdam, **2007**, chapter 8.