STUDIA UBB CHEMIA, LXIV, 2,Tom I, 2019 (p. 243-254)
(RECOMMENDED CITATION)
DOI:10.24193/subbchem.2019.2.20

Dedicated to Professor Florin Dan Irimie on the Occasion of His 65th Anniversary

MODELING AND PREDICTION OF AMINO ACIDS LIPOPHYLICITY USING MULTIPLE LINEAR REGRESSION COUPLED WITH GENETIC ALGORITHM

ALEXANDRINA GUIDEA[a], COSTEL SÂRBU[a*]

ABSTRACT. Quantitative structure-retention relationships (QSRR) approach was used to model chromatographic lipophilicity of sixteen proteinogenic amino acids using molecular descriptors computed with DRAGON and ALCHEMY software packages. Modeling was performed applying multiple linear regression (MLR) coupled with genetic algorithms (GA) methodology (MLR-GA). The most important descriptors, highly significant in the predictive models of amino acids lipophilicity (R_{M0}), were related to atomic polarizabilities (MATS3p; Ap; H1p), atomic van der Waals volume (MATS3v), Sanderson electronegativity (RDF070e) and Randic molecular profiles (DP11; DP12) calculated with Dragon software. The internal statistical evaluation procedure highlighted some appropriate models for the chromatographic lipophilicity prediction. Moreover, the statistical parameters of regression in order to evaluate the relationship between experimental and predicted values, in case of the test set (four amino acids), revealed three statistically valid models (model A, E and F) that can be successfully used in lipophilicity prediction of amino acids.

Keywords: *chromatographic lipophilicity, amino acids, multiple linear regression, genetic algorithm, molecular descriptors, modeling, prediction*

INTRODUCTION

Quantitative structure-activity/property relationships (QSAR/QSPR) describe how the molecular structure, in terms of descriptors – lipophilic, electronic and steric – affects the biological activity or any other property of

[a] *Babeş-Bolyai University, Faculty of Chemistry and Chemical Engineering, 11 Arany Janos str. No 11, RO-400028, Cluj-Napoca, Romania*
** Corresponding author: csarbu@chem.ubbcluj.ro*

a compound [1-3]. Similarly, a quantitative structure-retention relationship (QSRR) relates these descriptors to chromatographic retention [4, 5]. Finally, the quantitative retention-property relationships (QRPR) imply that conclusions concerning various properties including also a biological activity can be based on chromatographic experiments. Regarding the QRPR approach, it can be considered that the same basic molecular interactions determine the behavior of chemical compounds in both biological and chromatographic environments. As a direct consequence, the chromatographic approach has been quite successful applied, for example, in duplicating lipophilicity data derived by traditional shake-flask technique or other procedures [6-8].

Lipophilicity (hydrophobicity) is a fundamental molecular property defined as the logarithm of the octanol-water partition coefficient ($\log P_{OW}$), which practically reflects the non-ionized compound partition between two phases usually octanol and water [9,10].

Different chromatographic methods were applied and continue to be used with success in order to estimate the physico-chemical characteristics of chemical compounds, of which lipophilicity seems to be the most important [11]. In many cases, this molecular parameter strongly correlates with the biological activity of chemicals, as well as with other important physico-chemical properties. Studies on the relationships between chromatographic retention and the structure are helpful not only for the molecular design/template synthesis of chemical compounds with controlled properties, but also to better understand the biochemical and biophysical processes. In addition, the chromatographic methods have significant advantages in comparison with other physico-chemical methods because (a) they are fast and relatively simple, (b) only small amounts of any compound are needed, (c) the compound should not be very pure because it is purified during the chromatographic process, (d) the process is dynamic and can be very easy modeled [12-15].

The modeling of the lipophilicity for twenty of the proteinogenic amino acids investigated in this study as well as the prediction of this parameter with different molecular descriptors calculated using performant software as Alchemy and Dragon will allow a better understanding of the relationships between the structure and their physico-chemical and biochemical properties. Highly statistical significant multiple linear regression-genetic algorithms (MLR-GA) models were successfully applied to the prediction of chromatographic lipophilicity (R_{M0}) of some amino acids.

RESULTS AND DISCUSSION

A comprehensive investigation was carried out for QSRR of twenty proteinogenic amino acids using MLR-GA methodology. Because the major goal still is to improve the predictive power of the models and the creation of a

''more general'' QSRR model, which can be applied over a wide range of amino acids, a data set of lipophilicity values (R_{M0} indices) obtained in previous determinations for some amino acids (Alanine-**Ala**, Arginine-**Arg**, Asparagine-**Asn**, Aspartic Acid-**Asp**, Cysteine-**Cys**, Glutamic acid-**Glu**, Glycine-**Gly**, Histidine-**His**, Leucine-**Leu**, Lysine-**Lys**, Methionine-**Met**, Phenylalanine-**Phe**, Proline-**Pro**, Serine-**Ser**, Tyrosine-**Tyr**, Valine-**Val**) [16], was used in our investigations. Because at that time, the R_{M0} of four amino acids (Glutamine-**Gln**, Isoleucine-**Ile**, Threonine-**Thr**, Tryptophan-**Trp**) were not been determined they were used here as a test set and also for the external validation of the obtained models. The descriptors that generated the most statistically significant MLR models were selected using GA methodology. The best predictive models for lipophilicity estimation were chosen considering the following regression parameters (goodness of fit): the determination coefficient (R^2), Fisher function (F), residual sum of squares (RSS), standard error of estimate (s), and leave-one-out cross-validation coefficient (Q^2), predictive error sum of squares ($PRESS$) and standard deviation error of prediction ($SDEP$) obtained in the cross-validation process. The retained descriptors from both used software packages are summarized in Table 1. The models obtained using descriptors retained from Dragon and Alchemy with the highest predictive ability and related statistical parameters are shown below in Table 2**.**

The statistical parameters corresponding to the all regression models retained with three, four, and respectively five independent variables (descriptors) illustrate a high to moderate statistically significant prediction power. Furthermore, it is easy to observe that the most powerful models contain four or five descriptors. The „goodness of model" is given by its robustness, prediction ability, and the applicability domain. The determination coefficient of fitting power (R^2) was higher than 89% in the case of Alchemy models and higher than 99% in the case of Dragon models, respectively. By a careful examination, one may be observed that the most informative Alchemy descriptors were molecular polarizability (Polar) and specific polarizability (Sp.Pol), the sum of absolute values of the charges on the nitrogen and oxygen atoms in the molecule ($ABSQ_{ON}$), and Wiener index (WienI) (Table 2). All these retained descriptors appeared to be important in describing the chromatographic lipophilicity. The molecular polarizability increases the lipophilicity, but the charges decrease it [17].

The most significant descriptors calculated with Dragon are related to atomic polarizability (MATS3p, Ap, H1p), atomic van der Waals volume (MATS3v), atomic Sanderson electronegativity (RDF070e) and Randic molecular profiles (DP11/12).

Table 1. Descriptors of amino acids calculated with Alchemy[2000] and Dragon Plus 5.4 software packages and selected by GA methodology

Abv.	Exp. Data*	Descriptor																
		ALCHEMY							DRAGON									
	R_{M0}	Volume	ABSQon	MaxQ-	Polar	Sp.Pol	$^{0}\chi_{v}$	Wienl	MATS3p	MATS3v	DP11	DP12	RDF070e	Mor12u	Ap	HATS2u	SP13	H1p
Ala	-1.14	164.62	1.92	-0.35	17.49	0.11	6.73	247.00	-0.27	-0.34	0.02	0.01	0.00	-0.55	2.66	0.61	0.002	0.40
Arg	-0.60	165.16	1.92	-0.36	17.87	0.11	6.71	247.00	-0.06	-0.06	6.73	6.42	5.70	-1.43	11.46	0.32	6.06	0.51
Asn	-1.19	114.90	1.66	-0.42	11.88	0.10	4.70	96.00	-0.31	-0.32	1.50	0.95	0.25	-1.01	4.76	0.43	0.66	0.43
Asp	-1.26	112.39	1.57	-0.33	11.04	0.10	4.57	96.00	-0.36	-0.38	1.42	0.89	0.00	-0.82	4.22	0.42	0.60	0.42
Cys	-0.90	101.85	0.94	-0.33	11.48	0.11	4.56	46.00	0.01	-0.04	0.43	0.19	0.00	-0.52	3.69	0.55	0.10	0.56
Gln	-	132.91	1.66	-0.42	13.71	0.10	5.41	136.00	-0.21	-0.23	3.30	2.73	1.42	-1.03	5.74	0.39	2.28	0.47
Glu	-1.18	129.38	1.57	-0.33	12.87	0.10	5.28	136.00	-0.16	-0.18	3.38	2.81	1.92	-1.20	6.39	0.39	2.37	0.46
Gly	-1.07	68.43	0.95	-0.33	6.64	0.10	2.64	18.00	0.34	0.23	0.01	0.004	0.00	-0.34	1.45	0.77	0.001	0.24
His	-0.59	137.90	1.53	-0.33	15.43	0.11	5.82	165.00	-0.01	-0.01	2.74	2.11	3.22	-1.06	7.31	0.42	1.65	0.68
Ile	-	135.50	0.94	-0.33	13.86	0.10	5.80	92.00	0.14	0.11	1.50	0.98	2.46	-0.77	6.61	0.41	0.72	0.55
Leu	-0.47	135.82	0.94	-0.33	13.86	0.10	5.79	96.00	0.19	0.15	0.90	0.47	1.61	-1.09	6.46	0.45	0.24	0.65
Lys	-0.93	148.91	1.27	-0.33	15.21	0.10	5.92	143.00	0.08	0.06	5.10	4.69	3.21	-1.15	8.26	0.37	4.29	0.48
Met	-0.55	137.48	0.94	-0.33	15.02	0.11	6.15	102.00	0.24	0.19	3.79	3.27	3.52	-0.52	7.25	0.51	2.83	0.56
Phe	-0.02	157.90	0.94	-0.33	18.14	0.11	6.60	212.00	0.33	0.30	3.96	3.40	4.33	-1.04	10.37	0.36	2.94	0.86
Pro	-0.90	108.70	0.93	-0.33	11.24	0.10	4.55	62.00	0.21	0.15	0.28	0.11	0.00	-0.97	4.06	0.55	0.05	0.53
Ser	-1.21	93.57	1.33	-0.39	9.12	0.10	3.66	46.00	-0.40	-0.46	0.23	0.09	0.00	-0.68	2.98	0.54	0.04	0.39
Thr	-	110.39	1.33	-0.39	10.83	0.10	4.54	65.00	-0.15	-0.19	0.27	0.11	0.00	-0.56	4.14	0.45	0.06	0.48
Trp	-	186.60	1.22	-0.33	22.32	0.12	8.10	396.00	0.19	0.18	5.87	5.48	6.88	-1.01	16.43	0.33	5.26	0.89
Tyr	-0.44	165.64	1.34	-0.39	18.78	0.11	6.97	268.00	0.11	0.08	5.39	4.99	4.14	-1.20	10.69	0.33	4.63	0.73
Val	-0.68	118.81	0.94	-0.33	12.02	0.10	5.09	65.00	0.01	-0.01	0.30	0.12	0.00	-0.74	5.15	0.46	0.07	0.51

*data values for lipophilicity parameters (R_{M0}) obtained on RP-18W chromatographic plates, according to the reference [16]

Table 2. The linear multiple regression models for lipophilicity prediction obtained by applying genetic algorithms on *Alchemy* and D*ragon* descriptors

ID	Size	Model	R^2 %	F	s	RSS	SDEC	Q^2 %	PRESS	SDEP
Alchemy										
A	5	*R_{MD}=14.01–0.14***Volume**+1.68***MaxQ**-+1.54***Polar** – 155.52***Sp.Pol** –0.008***WienI** ** R_{MD} =–11.08***Volume**+0.15***MaxQ**-+15.01***Polar**–2.61***Sp.Pol** –1.691***WienI**	89.19	16.51	0.143	0.206	0.113	76.37	0.450	0.168
B	4	* R_{MD} =13.38–0.14***Volume**+1.56***Polar**–156.11***Sp.Pol**–0.008***WieIn** ** R_{MD} = – 11.06***Volume** + 15.14***Polar** – 2.62***Sp.Pol** – 1.86***WieIn**	87.33	18.95	0.148	0.242	0.123	73.03	0.514	0.179
C	3	* R_{MD} = -0.97 – 0.68***$ABSQ_{ON}$** + 0.18***Polar** – 0.28*$^{\chi^{V}_{0}}$ ** R_{MD} = - 0.70***$ABSQ_{ON}$** + 1.79***Polar** – 0.94*$^{\chi^{V}_{0}}$	83.45	20.16	0.162	0.315	0.140	71.50	0.543	0.184
D	2	* R_{MD} =-1.17 – 0.66***$ABSQ_{ON}$** + 0.09***Polar** ** R_{MD} = - 0.68***$ABSQ_{ON}$** + 0.86***Polar**	80.95	27.62	0.167	0.363	0.151	68.96	0.591	0.192
Dragon										
E	5	*R_{MD}=-1.36+0.36***MATS3p**–0.22***DP11**+0.09***RDF070e**+0.37***Mor12u** + 0.20***Ap** ** R_{MD}=0.24***MATS3p**–1.37***DP11**+0.50***RDF070e**+0.32***Mor12u** +1.69***Ap**	99.83	1172.90	0.018	0.003	0.014	99.54	0.009	0.023
F	4	* R_{MD} = -3.05 – 0.22***DP12**+0.31***Mor12u**+0.32***Ap**+2.07***HATS2u** ** R_{MD} = -1.33***DP12** + 0.27***Mor12u** + 2.73***Ap** + 0.69***HATS2u**	**99.06**	290.00	0.040	0.018	0.034	**98.17**	0.035	0.047
G	3	* R_{MD} =-3.32 – 0.24***SP13** + 0.32***Ap** + 2.85***HATS2u** ** R_{MD} = -1.30***SP13** + 2.71***Ap** + 0.95***HATS2u**	**97.51**	156.50	0.063	0.047	0.055	**95.56**	0.085	0.073
H	2	* R_{MD} = -1.64 + 0.61***MATS3v** + 1.61***H1p** ** R_{MD} = 0.40***MATS3v** + 0.68***H1p**	88.45	49.80	0.130	0.220	0.117	84.97	0.286	0.134

**– non standardized coefficients*

*** –standardized coefficients*

Usually, the prediction ability of a model can be better characterized by an internal validation assessing the correlation coefficient between the experimental and predicted values. To evaluate the result of a large number of samples over the whole measurement range, usually, the regression analysis is preferred. The statistical parameters to evaluate the linear relationship between the experimental and predicted lipophilicity parameters for the training set of studied amino acids revealed that the models obtained with Dragon's descriptors have a good predictive capacity for R_{M0}, $Q^2 > 99\%$ and, in case of Alchemy's descriptors, $Q^2 > 76\%$.

Although satisfactory model robustness is a necessary condition to have a high prediction power, the real prediction ability of the model is assessed with the help of the external test set never used to build the models. So, the validation strategies should check the reliability of the developed models for their possible real application on a new set of data, and confidence of prediction can be judged.

In order to observe the ability of the obtained models to predict the lipophilicity the list of R_{M0} values both calculated with the model (Table 3) and predicted in the validation process (Table 4) has been compared with the experimental ones. Good correlation values were found for the training (sixteen amino acids) and training and test set (twenty amino acids) in the case of models using Dragon descriptors (Figure 1 c, d) and Alchemy descriptors (Figure 1 a, b), respectively. Based on the prediction criteria, the best model for lipophilicity could predict 99.83% of the variance in the case of Dragon descriptors and 76.37% in case of Alchemy descriptors. The test set used to validate models revealed that model C and D, although with better statistics, show a limited applicability. It may be also a good argument for robustness of model E that performs better predictability for both sets: training, and test, respectively (see Table 2). The model F has a better prediction for test sets. This is supported by the statistical parameters of regression ($R^2 = 99.06\%$, $Q^2 = 98.17\%$, $F = 290$, $s = 0.04$).

The predictive power of the models obtained with Alchemy descriptors is lower comparing with Dragon models because as we mentioned above, the set of amino acids do not form a homologous series. This is explaining by the fact that the retained descriptors do not contain sufficient information to describe the repartition behavior of all amino acids in thin layer chromatography (TLC) analysis. Lower prediction capacity is observed for amino acids with aliphatic side chain (Ala, Leu, Met, and Val) and for basic side chain (Arg and Lys) in comparison with experimental data. In Table 5 (the red marked) one can be observed that model E have a better prediction almost for all amino acids: basic side chain (Arg, His, Lys), hydrophobic side chain (aromatic) (Phe, Trp, Tyr), polar neutral side chain (Asn, Cys), unique amino acids (Gly, Pro). The most important selected descriptors indicate that the following descriptors are highly significant in the predictive lipophilicity models developed in this study: (a)

molecular descriptors obtained by radial basis functions centred on different interatomic distances (RDF descriptors – RDF070e); (b) molecular descriptors calculated by summing atoms weights viewed by a different angular scattering function (3D-MoRSE descriptors – Mor12u); (c) molecular descriptors obtained as statistical indices of the atoms projected onto the 3 principal components obtained from weighted covariance matrices of the atomic coordinates (WHIM descriptors – Ap); (d) molecular descriptors calculated from the molecular graph by summing the products of atom weights of the terminal atoms of all the paths (2D correlation – MATS3p, MATS3v); (e) molecular descriptors derived from the distance distribution moments of the geometry matrix (RMP descriptors – DP11, DP12).

Table 3. The R_{M0} values calculated with all MLR-GA models based on Alchemy and Dragon descriptors

Abv.	*$R_{M0\ exp}$	Model ID							
		Alchemy				Dragon			
		A	B	C	D	E	F	G	H
Ala	-1.138	-0.958	-0.981	-0.913	-0.891	-1.137	-1.093	-1.115	-1.201
Arg	-0.598	-0.741	-0.732	-0.838	-0.859	-0.589	-0.566	-0.553	-0.861
Asn	-1.187	-1.216	-1.085	-1.212	-1.218	-1.203	-1.147	-1.116	-1.144
Asp	-1.255	-1.191	-1.231	-1.272	-1.235	-1.263	-1.267	-1.304	-1.195
Cys	-0.896	-0.909	-0.919	-0.754	-0.777	-0.904	-0.909	-0.971	-0.770
Gln	-	-	-	-	-	-	-	-	-
Glu	-1.185	-1.237	-1.279	-1.130	-1.074	-1.156	-1.176	-1.109	-1.003
Gly	-1.071	-1.070	-1.122	-1.118	-1.208	-1.079	-1.085	-1.048	-1.005
His	-0.586	-0.624	-0.661	-0.778	-0.819	-0.607	-0.624	-0.567	-0.556
Ile	-	-	-	-	-	-	-	-	-
Leu	-0.472	-0.695	-0.697	-0.658	-0.567	-0.457	-0.466	-0.407	-0.507
Lys	-0.930	-0.817	-0.825	-0.668	-0.666	-0.945	-1.012	-1.009	-0.841
Met	-0.554	-0.309	-0.303	-0.541	-0.463	-0.532	-0.547	-0.607	-0.630
Phe	-0.017	-0.056	-0.089	-0.093	-0.187	-0.033	-0.037	-0.055	-0.072
Pro	-0.897	-0.911	-0.930	-0.786	-0.788	-0.895	-0.923	-0.855	-0.691
Ser	-1.214	-1.134	-1.058	-1.210	-1.246	-1.211	-1.188	-1.214	-1.301
Thr	-	-	-	-	-	-	-	-	-
Trp	-	-	-	-	-	-	-	-	-
Tyr	-0.439	-0.461	-0.409	-0.347	-0.393	-0.443	-0.389	-0.418	-0.414
Val	-0.681	-0.793	-0.800	-0.801	-0.729	-0.666	-0.691	-0.772	-0.829

*data values for lipophilicity parameters (R_{M0}) obtained on RP-18W chromatographic plates, according to the reference [16]

Table 4. The predicted (in cross-validation process) R_{M0} values with all MLR-GA models based on Alchemy and Dragon descriptors

Abv.	*$R_{M0\ exp}$	Model ID							
		Alchemy				Dragon			
		A	B	C	D	E	F	G	H
Ala	-1.138	-0.858	-0.899	-0.822	-0.796	-1.136	-1.083	-1.110	-1.215
Arg	-0.598	-0.784	-0.773	-0.941	-0.965	-0.583	-0.540	-0.522	-0.879
Asn	-1.187	-1.262	-1.070	-1.217	-1.225	-1.207	-1.137	-1.100	-1.136
Asp	-1.255	-1.172	-1.225	-1.276	-1.231	-1.267	-1.275	-1.326	-1.179
Cys	-0.896	-0.947	-0.988	-0.731	-0.759	-0.906	-0.912	-0.980	-0.761
Gln	-	-1.170	-1.040	-1.069	-1.056	-1.271	-1.310	-1.287	-1.030
Glu	-1.185	-1.259	-1.311	-1.117	-1.058	-1.145	-1.173	-1.090	-0.986
Gly	-1.071	-1.066	-1.214	-1.166	-1.278	-1.094	-1.112	-1.008	-1.178
His	-0.586	-0.635	-0.678	-0.807	-0.844	-0.619	-0.629	-0.565	-0.550
Ile	-	-0.666	-0.665	-0.658	-0.567	-0.382	-0.510	-0.578	-0.690
Leu	-0.472	-0.785	-0.788	-0.742	-0.582	-0.446	-0.462	-0.378	-0.512
Lys	-0.930	-0.776	-0.787	-0.646	-0.643	-0.954	-1.045	-1.042	-0.831
Met	-0.554	-0.170	-0.162	-0.536	-0.446	-0.511	-0.540	-0.619	-0.642
Phe	-0.017	-0.090	-0.144	-0.172	-0.268	-0.044	-0.051	-0.080	-0.110
Pro	-0.897	-0.914	-0.935	-0.768	-0.771	-0.894	-0.936	-0.849	-0.664
Ser	-1.214	-1.078	-1.006	-1.209	-1.254	-1.209	-1.183	-1.214	-1.334
Thr	-	-1.092	-1.008	-1.134	-1.092	-0.854	-0.978	-1.120	-0.985
Trp	-	0.262	0.152	0.071	-0.004	0.955	1.425	1.287	-0.095
Tyr	-0.439	-0.479	-0.390	-0.314	-0.380	-0.445	-0.373	-0.412	-0.408
Val	-0.681	-0.821	-0.829	-0.837	-0.736	-0.655	-0.694	-0.791	-0.839

*data values for lipophilicity parameters (R_{M0}) obtained on RP-18W chromatographic plates, according to the reference [16]

The predictive power of the models obtained with Alchemy descriptors is lower comparing with Dragon models because as we mentioned above, the set of amino acids do not form a homologous series. This is explaining by the fact that the retained descriptors do not contain sufficient information to describe the repartition behavior of all amino acids in thin layer chromatography (TLC) analysis. Lower prediction capacity is observed for amino acids with aliphatic side chain (Ala, Leu, Met, and Val) and for basic side chain (Arg and Lys) in comparison with experimental data. In Table 5 (the red marked) one can be observed that model E have a better prediction almost for all amino acids: basic side chain (Arg, His, Lys), hydrophobic side chain (aromatic) (Phe, Trp, Tyr), polar neutral side chain (Asn, Cys), unique amino acids (Gly, Pro). The most important selected descriptors indicate that the

following descriptors are highly significant in the predictive lipophilicity models developed in this study: (a) molecular descriptors obtained by radial basis functions centred on different interatomic distances (RDF descriptors – RDF070e); (b) molecular descriptors calculated by summing atoms weights viewed by a different angular scattering function (3D-MoRSE descriptors – Mor12u); (c) molecular descriptors obtained as statistical indices of the atoms projected onto the 3 principal components obtained from weighted covariance matrices of the atomic coordinates (WHIM descriptors – Ap); (d) molecular descriptors calculated from the molecular graph by summing the products of atom weights of the terminal atoms of all the paths (2D correlation – MATS3p, MATS3v); (e) molecular descriptors derived from the distance distribution moments of the geometry matrix (RMP descriptors – DP11, DP12).

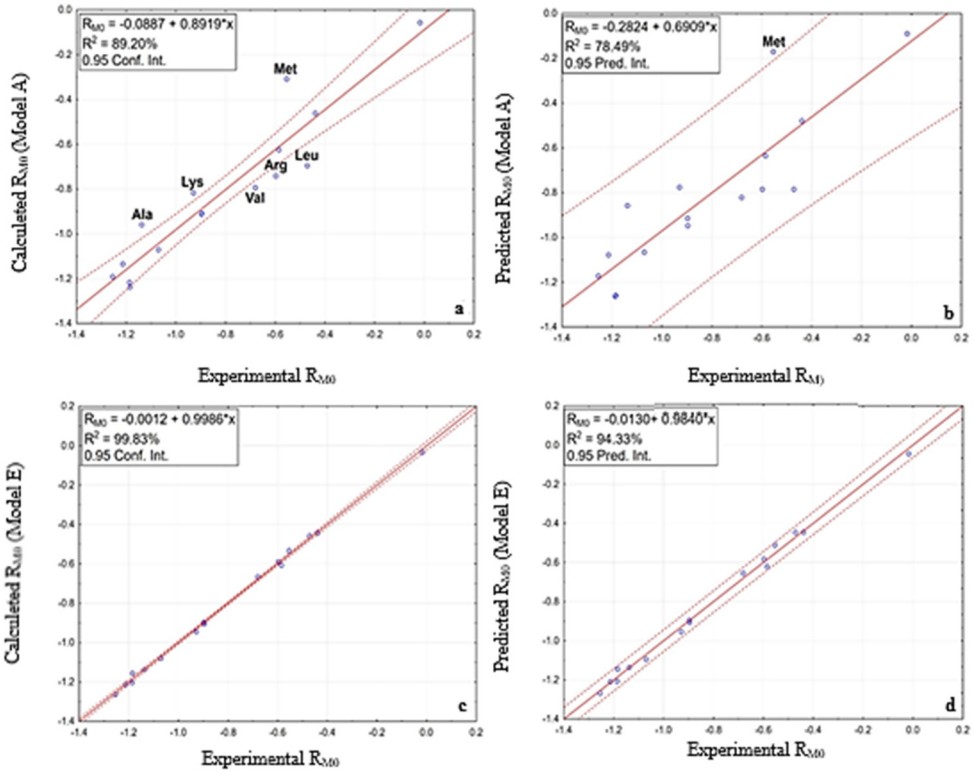

Figure 1. Calculated and predicted versus experimental R_{M0} values of amino acids for the training set (a, c); training and test set (b, d) for the best models developed using Dragon and Alchemy descriptors, respectively.

The most important descriptors in these models, accounting for 2D and 3D aspects of the molecular structure, can be classified as RDF (Radial Distribution Function), Randic molecular profiles, WHIM and GETAWAY signals. The selected RDF descriptors are related to the atomic van der Waals volumes (v) and atomic Sanderson electronegativities (e). The GETAWAY descriptors are related to the atomic Sanderson electronegativities (e). Also, the use of WHIM descriptors and GETAWAY descriptors show that atomic polarizabilities (p) and atomic van der Waals volumes (v) are the most important properties responsible for repartition coefficient of amino acids in TLC.

CONCLUSION

The chromatographic retention data for a set of proteinogenic amino acids have been modeled by a wide set of computational molecular descriptors using multiple linear regression and genetic algorithms methodologies. The best models, internally validated by leave-one-out procedure, revealed that only a small number of descriptors seem to be necessary in order to obtain statistically significant prediction models. The models derived from Dragon descriptors are more efficient comparing to the Alchemy descriptors. The descriptors selected as the best combinations correlated to the different lipophilicity response are not easily interpretable concerning the complex underlying lipophilicity mechanism. However, the most important descriptors, highly significant in the predictive lipophilicity models of amino acids, were related to the atomic polarizabilities, atomic Sanderson electronegativities and atomic van der Waals volumes of the molecules.

EXPERIMENTAL SECTION

Computation of the molecular descriptors

There are three common methods for structure representation: whole molecule 1D descriptors, 2D descriptors, and 3D descriptors. 1D descriptors attempt to express chemical information in a simple 1D molecular code and are designed for compact storage of information. 2D descriptors are calculated from a chemical structure which is represented as a connection table or a molecular graph. In the graphical representation of molecular structures, atoms in the molecular structure are represented as vertices while bonds are represented as edges. 3D molecular descriptors provide molecular information about the 3D arrangement of structural features and general molecular surfaces and volumes. There are many thousands of descriptors defined in a comprehensive handbook [18].

Dragon Plus version 5.4 (*www.talete.mi.it/dragon.htm*) [19] is widely used to calculate molecular descriptors for QSAR/QSPR/QSRR modeling. Generally, the Dragon calculated descriptors encoding the molecular structure of an analyte are categorized in 22 different types: constitutional (1D), molecular properties (1D), atom-centred fragments (1D), functional group counts (1D), charge (1D), information indices (2D), walk and path counts (2D), topological (2D), topological charge indices (2D), connectivity indices (2D), eigenvalue-based indices (2D), Burden eigenvalues (2D), 2D edge adjacency indices (2D), autocorrelation (2D), 2D binary fingerprints (2D), 2D frequency fingerprints (2D), geometrical descriptors (3D), Radial Distribution Function (RDF) descriptors (3D), Randic molecular profiles (3D), GETAWAY (Geometry, Topology and Atoms-Weighted Assembly) descriptors (3D), 3D-MoRSE (3D Molecular Representation of Structure based on Electron diffraction) descriptors (3D), and WHIM (Weighted Holistic Invariant Molecular) descriptors (3D). For this study were used 1056 descriptors.

The second set of descriptors related to charge dependent, 3D-structure-dependent parameters, topological and descriptors related to atom properties, formal and delocalized charge and molecular surface based on molecular mechanics for optimizing models, were computed using Alchemy^{2000} [20] (*http://www.tripos.com*). The descriptors used (19) are: the partition coefficient (ScilogP), the first-order ($^{1}\chi$) and third-order ($^{3}\chi$) connectivity indexes, the zero-order ($^{0}\chi^{V}$) and first-order ($^{V}\chi^{1}$) valence order connectivity indexes, the third-order shape index for molecule ($^{3}K_{a}$), the Wiener (WienI) index based on the graph of the molecule, the volume (Volume), the dipole moment (Dipole), the molecular polarizability (Polar), the specific molar polarizability (Sp.Pol), the largest positive/negative charges over the atoms in molecule, in electrons ($MaxQ^{+}$/$MaxQ^{-}$), the sum of absolute values of the charges on each atom of the molecule, in electrons (ABSQ), the sum of absolute values of the charges on the nitrogen and oxygen atoms in molecule, in electrons ($ABSQ_{ON}$), the surface area, the ovality (Ovality) of the molecule.

Chemometric methods

Multiple linear regression-genetic algorithm analysis [21, 22] was performed using the MobyDigs v.1.0 package [23]. Genetic algorithm procedure [24-26] was used to select the most significant variables. Models predictive performance [27, 28] was described by means of statistical parameters related to model goodness of fit (the determination coefficient R^2, Fisher function F, residual sum of squares RSS, standard error of estimate *s*, and predictive capability (cross-validation coefficient Q^2, predictive error sum of squares *PRESS*, and standard deviation error of prediction SDEP).

REFERENCES

1. C. Hansch, A. Leo (eds), Exploring QSAR: fundamentals and applications in chemistry and biology, American Chemical Society, Washington, D.C., **1995**.
2. M. Karelson, Molecular Descriptors in QSAR/QSPR, John Wiley & Sons, New York, **2000**.
3. M. Karthikeyan, V. Renu, Practical Bioinformatics, Springer, India, **2014**.
4. R. Kaliszan, Quantitative structure–chromatographic retention relationships, Wiley–Interscience, New York, **1987**.
5. R. Kaliszan, Structure and Retention in Chromatography. A Chemometric Approach, Harwood Academic Publishers, Amsterdam, **1997**.
6. Test No. 107, Partition coefficient (n-octanol/water), Shake-flask method, OECD, Paris, **1995**.
7. Test No. 123, Partition coefficient (n-octanol/water), Slow-stirring method, OECD, Paris, **2005**.
8. Test No. 117, Partition coefficient (n-octanol/water), HPLC method, OECD, Paris, **2004**.
9. A. Leo, C. Hansch, D. Elkins, *Chem. Rev.*, **1971**, *71*, 525–616.
10. J. Sangster, Octanol-water partition coefficients: fundamentals and physical chemistry, John Wiley & Sons, Inc., New York, **1997**.
11. H. Van de Waterbeemd, M. Kansy, B. Wagner, H. Fischer, Lipophilicity measurement by high-performance liquid chromatography (RP-HPLC). In: V. Pilska, B. Testa, H. Van de Waterbeemd (eds), Lipophilicity in drug action and toxicology, VCH, Weinheim, **1996**.
12. C. Sârbu, B. Malawska, *J. Liq. Chromatogr. Relat. Technol.*, **2000**, 23, 2143-2154.
13. C. Sârbu, R. D. Naşcu-Briciu, *Studia UBB CHEMIA*, **2015**, LX, *1*, 265-280.
14. R. D. Briciu, C. Sârbu, *Studia UBB CHEMIA*, **2010**, LV, *3*, 105-118.
15. D. Casoni, C. Sârbu, *J. Sep. Sci.*, **2012**, *35*, 915-921.
16. D. Casoni, C. Sârbu, *Studia UBB CHEMIA*, **2011***,* LVI, *1*, 45-61.
17. L. Xing, R.C. Glen, *J. Chem. Inf. Comput. Sci.*, **2002**, *47*, 796-805.
18. R. Todeschini, V. Consonni, Handbook of Molecules Descriptors, Wiley, Weinheim, **2000**.
19. Talete SRL, DRAGON for windows (software for molecular descriptor calculations). Version 5.4-2006. http://www.talete.mi.it
20. SciQSAR Application, Version 3.0, 1998 SciVision, Method: *Alchemy 2000* (software for molecular descriptor calculations). http://www.tripos.com
21. R. Kaliszan, *J. Chromatogr. A,* **1993**, *656*, 417-435.
22. R. Put, Y. Vander Heyden, *Anal. Chim. Acta*, **2007,** *602*, 164-172.
23. R. Todeschini, Moby Digs Academic version software for variable subset selection by genetic algorithms, Rel. 1.0 for Windows, Talete, Milan, **2004**.
24. R. Leardi, R. Boggia, M. Terrile, *J. Chemom.*, **1992**, *6*, 267–281.
25. S. Riahi, M.R. Ganjali, E. Pourbasheer, P. Norouzi, *Chromatographia*, **2008**, *67*, 917-922.
26. J. Devillers, Genetic algorithms in molecular modeling, Academic Press, Inc., San Diego, **1996**.
27. H. Kubinyi, *Quant. Struct.-Act. Relat.,* **1994**, *13*, 285–294.
28. P. Gramatica, A. Sangion, *J. Chem. Inf. Model.,* **2016**, *56*, 1127−1131.