J Incl Phenom Macrocycl Chem (2012) 74:57–66
DOI 10.1007/s10847-010-9881-6

ORIGINAL ARTICLE

Application of GA-MLR method in QSPR modeling of stability constants of diverse 15-crown-5 complexes with sodium cation

Shahin Ahmadi

Received: 16 August 2010 / Accepted: 28 September 2010 / Published online: 19 October 2010

Abstract A genetic algorithm based multiple linear regressions (GA-MLR) method was applied for quantitative structure property relationship (QSPR) modeling of stability constants for 65 complexes of 1,4,7,10,13-pentaoxacyclopentadecane ethers (15C5) with sodium cation (Na^+). The best subset of molecular descriptors was selected with genetic algorithm subset selection procedure, to a variety of theoretical molecular descriptors, calculated by the Dragon software. The MLR model was developed with particular attention to external validation and applicability domain (AD). The validation was performed on the internal and external validation sets. The QSPR model presented in this study showed most accurate predictions with the leave one out cross validated variance ($Q^2_{loo-cv} = 0.88$) and the external-validated variance ($Q^2_{ext} = 0.82$). The AD of the models was analysed by the leverage approach.

Keywords 15-Crown-5 ethers · Stability constant · QSPR · Genetic algorithm · MLR · Applicability domain

Introduction

Crown ethers are an important class of macrocyclic compounds in supramolecular chemistry, and have gained much attention for their ability to form stable complexes with metal ions within their central cavity by right of ion–dipole interaction. Furthermore, crown ethers exhibit fine complex selectivity for metal and ammonium ions, and it is decided by complementary size matching of ionic radii and the ligand cavity and the crown structure, leading to excellent recognition properties for metal ions [1–6].

The selective complexing properties of crown ethers with metal ions have enabled them to be extensively used in many areas, such as metal ion adsorption and separation, preconcentration and determination [7, 8], selective transport [9], preparing ion selective electrode [10–12] phase-transfer catalysis in organic synthesis [13–15], neutral carrier in constructing of ion sensors [16] and so on.

QSPR approach is based on the assumption that the behavior of a compound, expressed by any simple and complex physicochemical properties, is correlated with numerical descriptors of the compound. In the development of a mathematical relationship between the structure descriptors and the property, the commonly used linear methods, such as multiple linear regression (MLR) [17], principal component regression (PCR) [18] and partial least squares (PLS) [19, 20], and non-linear models, such as radial basis function neural network (RBFNN) [21, 22] can be used.

It should be noted that there are not many publication on structure–property modeling of crown-ether complexes with metals. The previous quantitative structure–properties relationship (QSPR) studies on the complexes of macrocyclic polyethers with alkali cations including neural network modeling of Gakh et al. [23] were reported on the limited set of simple crown ether complexes and this simple computational scheme had an average accuracy of 0.3 logK units. Multiple linear regression studies by Shi et al. [24] on 314 cation-macrocycle-solvent systems resulted in the standard errors in logK range from 1.42 in the largest system to 0.36 in the smallest. The substructural molecular fragments by Varnek et al. [25] was applied to assess stability constants of the complexes of crown ethers

S. Ahmadi (✉)
Department of Chemistry, Kermanshah Branch,
Islamic Azad University, Kermanshah, Iran
e-mail: ahmadi_shahin@iauksh.ac.ir

with alkali cations in methanol; several model including different fragment sets coupled with linear or nonlinear fitting equations were applied for the data sets. They used the additional descriptor named cyclicity and obtained the results for standard error of predicted logK range from 0.16 to 0.22 for sodium ion cation. Recently QSPR model for the stability constants of 58 complexes of 15C5 with metal ion K^+ was established with the CODESSA program by Ghasemi et al. [26]. The proposed model had an average accuracy of 0.10 logK units.

The primary purpose of this work is establish a accurate QSPR modeling between the molecular descriptors of 65 15C5 ethers and the stability constants measured in different labs. To perform this analysis, as a powerful tool, genetic algorithm based multivariate linear regression (GA-MLR) is applied as variable selection method [27, 28]. The most important aspect of the proposed QSPR model is verifying the chemical applicability domain by the leverage approach. Finally, finding the important structural descriptors influencing the binding constant will give us some invaluable information for future research.

Materials and methods

Data set

The chemical structures and experimental value for the stability constants of 65 sodium ion complexes of 15C5 derivatives selected from the literature [29] are presented in Tables 1 and 2, respectively. Since the temperature and solvent also affect the stability constants, we used only data obtained at 298 K and just in methanol. The data set was randomly split into the calibration, prediction and validation sets (35 calibration samples, 15 prediction samples and 15 validation samples). The calibration and prediction sets were used to build and optimize the QSPR model and the external validation set was used to evaluate the prediction power of the obtained model. Experimental logK values vary from 2.72 to 3.89, 2.79 to 3.91, and 2.74 to 3.9 for calibration, prediction, and validation sets respectively.

Molecular optimization and descriptor calculation

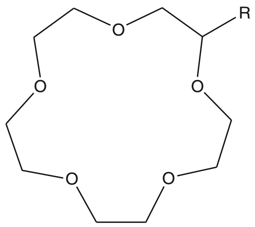
All calculations were run on a Pentium IV personal computer with windows XP as operating system. The structures were drawn in HyperChem 7.5 [30] and the geometrical structure of ligand molecules was optimized using semi-empirical quantum method Austin Method 1 (AM1) [31] using the Polack–Rabiere algorithm until the root mean square gradient was 0.01 within the MOPAC [32] program package. The geometry and other information from the output of quantum chemical calculations were inserted into the Dragon [33] program, and descriptors for ligands were calculated. All these descriptors are derived solely from molecular structure and do not require experimental data to be calculated. More than 466 molecular descriptors is derived to properly characterize the chemical structure of the 65 15C5 derivatives, involving variables of the type Constitutional, Topological, GETAWAY (GEometry, Topology and Atoms-Weighted AssemblY), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), Aromaticity Indices.

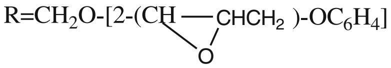
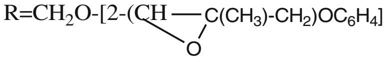
Variable selection with genetic algorithm

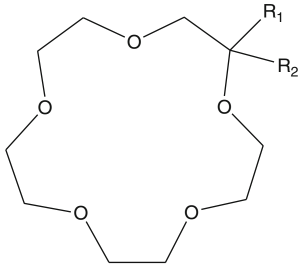
The GA is a powerful search technique based on the evolution of biological systems [34]. It is used to find approximate solutions of combinatorial optimization problems based on Darwinian biological evolution principle. In the GA, possible solutions of a given problem are represented by bit strings, and it is optimized toward better solutions by applying genetic operators such as selection, crossover and mutation. In each generation, individuals are decoded and fitness values are calculated using an objective function. Then individuals having high objective values are randomly selected from the current population and are modified to generate a new population. In QSPR studies, it is important to obtain a model with a few numbers of structure-based molecular descriptors because this will lead to a simple and interpretable model. One of the most commonly used methods for variable selection is the GA-MLR. GA is an evolutionary method widely used for complex optimization problems in several fields such as QSPR [20, 35–38].

In this work, we used the GA-MLR algorithm for variable selection. In order to calculate GA-MLR, a program was written based on MATLAB software. A total of 466 descriptors were initially calculated by Dragon software for the entire data set of 65 compounds. The total number of descriptors was reduced to 403 descriptors, by eliminating the collinear descriptors (correlation coefficient is less than 0.1). The GA applied to the variable selection in this work uses a binary representation as the coding technique for the given problem; the presence or absence of a descriptor in a chromosome is coded by 1 or 0 [39–41]. The GA performs its optimization by variation and selection via the evaluation of the fitness function η. The fitness function that we used was the one that was proposed by Depczynski et al. [40]. As described in the data set section, the samples were randomly selected to the calibration, prediction and validation sets (35 calibration samples, 15 prediction samples and 15 validation samples). The root-mean-square errors of calibration (RMSEC) and prediction (RMSEP) were calculated and the fitness function was calculated as Eq. 1.

Table 1 Chemical structures of 65 15C5 ethers

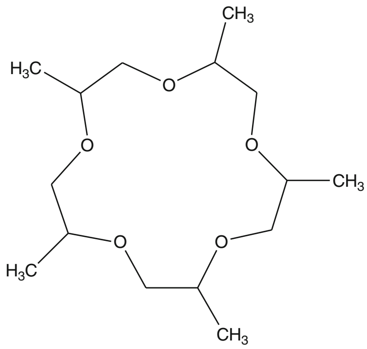


No.	Structure	No.	Structure
1	R=H	19	R=CH$_2$O–[2-CH$_2$=C(CH$_3$)–CH$_2$OC$_6$H$_4$]
2	R=C$_6$H$_{13}$	20	 R=CH$_2$O-[2-(CH—CHCH$_2$)-OC$_6$H$_4$] O
3	R=C$_8$H$_{17}$	21	 R=CH$_2$O-[2-(CH—C(CH$_3$)-CH$_2$)OC$_6$H$_4$] O
4	R=C$_{10}$H$_{21}$	22	R=CH$_2$OCH$_2$CH(OH)CH$_3$
5	R=CH$_2$OH	23	R=CH$_2$OCH$_2$CH(OH)CH$_2$OH
6	R=CH$_2$OCH$_3$	24	R=CH$_2$OCH$_2$CH(OC$_6$H$_{13}$)–CH$_2$OC$_6$H$_{13}$
7	R=CH$_2$OCH$_2$CH=CH$_2$	25	R=CH$_2$OCH$_2$CH$_2$OCH$_3$
8	R=CH$_2$O(CH$_3$)$_3$	26	R=CH$_2$OCH$_2$CH$_2$OC$_4$H$_9$
9	R=CH$_2$OC$_8$H$_{17}$	27	R=CH$_2$OCH$_2$CH$_2$OC$_8$H$_{17}$
10	R=C$_6$H$_5$	28	R=CH$_2$O(CH$_2$CH$_2$O)$_2$CH$_3$
11	R=CH$_2$OCH$_2$C$_6$H$_5$	29	R=CH$_2$O(CH$_2$CH$_2$O)$_2$C$_8$H$_{17}$
12	R=CH$_2$O–[2-CH$_3$OC$_6$H$_4$]	30	R=CH$_2$O(CH$_2$CH$_2$O)$_3$H
13	R=CH$_2$O–[3-CH$_3$OC$_6$H$_4$]	31	R=CH$_2$O(CH$_2$CH$_2$O)$_3$CH$_3$
14	R=CH$_2$O–[4-CH$_3$OC$_6$H$_4$]	32	R=CH$_2$NHC(CH$_3$)$_3$
15	R=CH$_2$OCH$_2$–[2-CH$_3$OC$_6$H$_4$]	33	R=CH$_2$NHC$_6$H$_{13}$
16	R=CH$_2$O–[2-NO$_2$C$_6$H$_4$]	34	R=CH$_2$NHC$_6$H$_5$
17	R=CH$_2$O–[4-NO$_2$C$_6$H$_4$]	35	R=CH$_2$NHCH$_2$CH$_2$NH$_2$
18	R=CH$_2$O–[2-CH$_2$=CH–CH$_2$OC$_6$H$_4$]		



36	R$_1$=CH$_3$; R$_2$=CH$_2$Br	55	R$_1$=C$_6$H$_{13}$; R$_2$=CH$_2$(OCH$_2$CH$_2$)$_2$OCH$_3$
37	R$_1$=CH$_3$; R$_2$=CH$_2$OCH$_2$CH$_2$OH	56	R$_1$=C$_6$H$_{13}$; R$_2$=CH$_2$(OCH$_2$CH$_2$)$_3$OCH$_3$
38	R$_1$=CH$_3$; R$_2$=CH$_2$(OCH$_2$CH$_2$)$_2$OH	57	R$_1$=C$_6$H$_{13}$; R$_2$=CH$_2$OC$_6$H$_{13}$
39	R$_1$=CH$_3$; R$_2$=CH$_2$(OCH$_2$CH$_2$)$_3$OH	58	R$_1$=C$_6$H$_{13}$; R$_2$=CH$_2$OC$_8$H$_{17}$
40	R$_1$=CH$_3$; R$_2$=CH$_2$OCH$_2$CH$_2$OCH$_3$	59	R$_1$=C$_6$H$_{13}$; R$_2$=CH$_2$OCH$_2$CH$_2$OC$_8$H$_{17}$
41	R$_1$=CH$_3$; R$_2$=CH$_2$O(CH$_2$CH$_2$O)$_2$CH$_3$	60	R$_1$=C$_6$H$_{13}$; R$_2$=CH$_2$(OCH$_2$CH$_2$)$_2$OC$_8$H$_{17}$
42	R$_1$=CH$_3$; R$_2$=CH$_2$O(CH$_2$CH$_2$O)$_3$CH$_3$	61	R$_1$=C$_8$H$_{17}$; R$_2$=CH$_2$Br
43	R$_1$=CH$_3$; R$_2$=CH$_2$O(CH$_2$)$_3$OCH$_3$	62	R$_1$=C$_8$H$_{17}$; R$_2$=CH$_2$OCH$_2$CH$_2$OCH$_3$
44	R$_1$=CH$_3$; R$_2$=CH$_2$OC$_6$H$_{13}$	63	R$_1$=C$_8$H$_{17}$; R$_2$=CH$_2$(OCH$_2$CH$_2$)$_2$OCH$_3$
45	R$_1$=CH$_3$; R$_2$=CH$_2$OC$_8$H$_{17}$	64	R$_1$=C$_8$H$_{17}$; R$_2$=CH$_2$(OCH$_2$CH$_2$)$_3$OCH$_3$

Table 1 continued

46	R_1=CH_3; R_2=$CH_2OCH_2CH_2OC_8H_{17}$	65	
47	R_1=CH_3; R_2=$CH_2(OCH_2CH_2)_2OC_8H_{17}$		
48	R_1=CH_3; R_2=$CH_2OC_{12}H_{25}$		
49	R_1=CH_3; R_2=$CH_2OCH_2CH_2OC_{12}H_{25}$		
50	R_1=CH_3; R_2=$CH_2(OCH_2CH_2)_2C_{12}H_{25}$		
51	R_1=CH_3; R_2=CH_2OCH_2–[2-C_5H_4N]		
52	R_1=C_6H_{13}; R_2=CH_2O[2-$CH_3OC_6H_4$]		
53	R_1=C_6H_{13}; R_2=CH_2Br		
54	R_1=C_6H_{13}; R_2=$CH_2OCH_2CH_2OCH_3$		

$$\eta = \left\{\left[(m_c - n - 1)RMSEC^2 + m_pRMSEP^2\right] / (m_c + m_p - n - 1)\right\}^{1/2} \quad (1)$$

where m_c and m_p are the number of compounds in the calibration and prediction set, respectively, and n represents the number of selected variables.

In this paper, the size of the population is 100, the probability of crossover is 0.7 in two points, the probability of mutation is 0.01 and the number of evolution generations is 100. The validity of the resulted regression model was evaluated by predicting the property of the molecules in the validation set.

Applicability domain

A crucial problem of a QSPR model is the applicability domain (AD). Not even a robust, significant, and validated QSPR model can be expected to reliably predict the modeled property for the entire universe of chemicals. In fact, only the predictions for chemicals falling within this domain can be considered reliable and not model extrapolations.

A way of defining the AD of a QSPR model is according to the leverage of a compound. The leverage h [42] of a compound measures its influence on the model. The leverage of a compound in the original variable space is defined as:

$$h_i = x_i^T\left(X^TX^{-1}\right)x_i \quad (i = 1, \ldots, n) \quad (2)$$

where x_i is the descriptor vector of the considered compound and X is the model matrix derived from the calibration set descriptor values. The warning leverage h* is defined as follows:

$$h* = 3 \times \sum_i h_i/n = 3 \times p'/n \quad (i = 1, \ldots, n) \quad (3)$$

where n is the number of calibration compounds and p′ is the number of model parameters.

To visualize the AD of a QSPR model, the plot of standardized residuals versus leverage values (h) can be used for an immediate and simple graphical detection of both the response outliers (i.e., compounds with cross validated standardized residuals greater than 2.5 standard deviation units, $>2.5\sigma$) and structurally influential chemicals in a model ($h > h*$).

Through the leverage approach, it is possible to verify whether a new chemical will lie within the structural model domain or outside the domain. A compound with high leverage in a QSPR model would reinforce the model if the compound is in the calibration set, but such a compound in the prediction and validation sets could have unreliable predicted data, the result of substantial extrapolation of the model [43].

Validation of the model

Leave one out cross validation (LOO-CV) is one of the QSPR model internal validation. The predictability of the QSPR model is determined using the LOO-CV method. The cross-validated explained variance (Q_{cv}^2) is calculated by the following equation:

$$Q_{cv}^2 = 1 - \frac{\sum_{i=1}^{cal}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{cal}(y_i - \bar{y})^2} \quad (4)$$

where y_i, $\hat{y}_i$ and $\bar{y}$ are, respectively, the measured, predicted, and averaged (over the entire data set) values of the dependent variable, respectively; the summations cover all the compounds in the validation set.

The LOO-CV approach is not sufficient to assess robustness and predictivity. The QSPR model developed using only calibration set chemicals is then applied to the external validation set chemicals to verify, more reliably, the predictive ability of the model.

Table 2 Molecular descriptors, experimental and predicted logK of 15C5 ethers complexes

No.	GATS6e	H1u	HATS8m	R7u	nCrR2	Experimental	Predicted	Residual
1	1.151	1.895	0.022	0.997	0	3.27	3.02	0.25
2	1.207	2.041	0.034	1.176	0	3.2	3.13	0.07
3	1.184	2.14	0.028	1.072	0	3.2	3.13	0.07
4	1.166	2.216	0.024	1.064	0	3.18	3.2	−0.02
5	1.112	1.955	0.042	1.093	0	3.1	3.03	0.07
6	1.171	1.968	0.04	1.162	0	3.03	3.07	−0.04
7	1.202	2.067	0.047	1.084	0	3.12	2.97	0.15
8	1.414	2.126	0.046	1.144	0	2.95	2.9	0.05
9	1.287	2.208	0.032	1.053	0	3.18	3.04	0.14
10	1.161	2.249	0.062	1.249	0	3.3	3.13	0.17
11	1.257	2.313	0.049	1.017	0	2.97	2.96	0.01
12	1.185	2.34	0.063	1.142	0	3.25	3.05	0.2
13	1.27	2.319	0.052	0.967	0	2.89	2.88	0.01
14	1.393	2.094	0.044	0.997	0	3	2.77	0.23
15	1.287	2.34	0.038	0.933	0	3.04	2.94	0.1
16	1.157	2.241	0.072	0.984	0	2.83	2.82	0.01
17	0.996	2.109	0.052	0.949	0	2.72	2.98	−0.26
18	1.237	2.414	0.029	0.892	0	3.07	3.03	0.04
19	1.191	2.46	0.028	0.974	0	3.04	3.17	−0.13
20	1.217	2.488	0.044	1.043	0	3.03	3.12	−0.09
21	1.245	2.56	0.03	0.962	0	3.02	3.15	−0.13
22	1.095	2.229	0.051	1.057	0	3.14	3.07	0.07
23	1.075	2.153	0.049	1.096	0	3	3.1	−0.1
24	1.384	2.621	0.026	0.935	0	2.97	3.08	−0.11
25	1.158	2.102	0.042	1.072	0	3.05	3.04	0.01
26	1.237	2.125	0.027	1.014	0	3.09	3.04	0.05
27	1.257	2.189	0.022	1.004	0	3.22	3.07	0.15
28	1.161	2.12	0.03	1.015	0	3.13	3.07	0.06
29	1.242	2.289	0.019	0.952	0	3.23	3.1	0.13
30	1.139	2.254	0.029	0.971	0	3.04	3.11	−0.07
31	1.164	2.341	0.024	0.949	0	3.09	3.14	−0.05
32	1.44	2.111	0.047	1.178	0	2.79	2.9	−0.11
33	1.233	1.963	0.028	1.056	0	2.82	3	−0.18
34	1.233	2.131	0.058	1.072	0	2.91	2.89	0.02
35	1.185	2.056	0.046	1.161	0	2.92	3.06	−0.14
36	1.125	2.134	0.153	1.115	1	2.86	2.8	0.06
37	1.044	2.442	0.064	1.359	1	3.88	3.82	0.06
38	1.057	2.51	0.051	1.122	1	3.88	3.7	0.18
39	1.069	2.535	0.04	1.134	1	3.73	3.79	−0.06
40	1.077	2.516	0.058	1.24	1	3.87	3.76	0.11
41	1.086	2.475	0.039	1.103	1	3.89	3.73	0.16
42	1.094	2.538	0.033	1.093	1	3.87	3.78	0.09
43	1.18	2.502	0.053	1.203	1	3.48	3.67	−0.19
44	1.261	2.361	0.041	1.138	1	3.57	3.57	0
45	1.246	2.522	0.036	1.195	1	3.54	3.74	−0.2
46	1.21	2.526	0.029	1.102	1	3.75	3.73	0.02
47	1.196	2.431	0.03	1.114	1	3.88	3.7	0.18
48	1.224	2.417	0.03	1.136	1	3.42	3.69	−0.27

Table 2 continued

No.	GATS6e	H1u	HATS8m	R7u	nCrR2	Experimental	Predicted	Residual
49	1.185	2.5	0.029	1.188	1	3.75	3.81	−0.06
50	1.169	2.495	0.029	1.188	1	3.89	3.82	0.07
51	1.121	2.387	0.059	1.199	1	3.58	3.62	−0.04
52	1.032	2.361	0.058	1.24	1	3.79	3.72	0.07
53	1.182	2.154	0.208	1.282	1	2.74	2.55	0.19
54	1.161	2.519	0.062	1.315	1	3.9	3.74	0.16
55	1.152	2.577	0.047	1.213	1	3.91	3.77	0.14
56	1.147	2.533	0.042	1.203	1	3.71	3.78	−0.07
57	1.373	2.568	0.042	1.181	1	3.56	3.61	−0.05
58	1.369	2.576	0.035	1.161	1	3.39	3.64	−0.25
59	1.311	2.616	0.031	1.11	1	3.62	3.68	−0.06
60	1.28	2.581	0.031	1.14	1	3.75	3.72	0.03
61	1.172	2.14	0.15	1.218	1	2.79	2.88	−0.09
62	1.157	2.512	0.052	1.242	1	3.82	3.74	0.08
63	1.146	2.541	0.046	1.258	1	3.86	3.81	0.05
64	1.138	2.56	0.044	1.277	1	3.75	3.86	−0.11
65	1.217	2.109	0.054	1.744	0	3.34	3.55	−0.21

The formula for the calculation of Q^2_{ext} is:

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{valid} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{valid} (y_i - \bar{y}_{cal})^2} \quad (5)$$

where y_i and $\hat{y}_i$ are respectively the measured and predicted (over the validation set) values of the dependent variable, and $\bar{y}_{cal}$ is the averaged value of the property for the calibration set; the summations cover all the compounds in the validation set.

The Q^2 value is good tests for evenly distributed data, but they are not always reliable for unevenly distributed datasets; instead RMSEs (Root Mean Squared Errors) provide a more reliable indication of the fitness of the model, independently of the applied splitting. Other useful parameters to be considered are the RMSEs calculated on different sets: on calibration (RMSEC), prediction (RMSEP) and validation (RMSEV). RMSE is calculated as in Eq. 6:

$$RMSE = \left(\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n} \right)^{1/2} \quad (6)$$

where y_i and $\hat{y}_i$ are respectively the measured and predicted values of the property; n is the number of compounds in each set of data.

It is important to note that RMSE values must not only be low but also as similar as possible for the calibration, prediction and validation sets: this suggests that the proposed model has both predictive ability (low values) as well as sufficient generalizability (similar values) [44].

After model formation, statistical tests were performed using the best subset of descriptors to find any outliers that existed in the data set.

Furthermore, a variance inflation factor (VIF) analysis was performed to see if multicollinearities existed between the descriptors in the model. The VIF value is calculated from $1/1 - r^2$, where r^2 is the multi correlation coefficient of one descriptor's effect regressed on the remaining molecular descriptors. Models are not accepted if they contained descriptors with VIFs over a value of 10. This ensured that the squared multicollinearity coefficient for each descriptor in the model did not exceed 0.90 [45, 46]. Finally, the model was validated using the external validation set.

Results and discussion

To select important variables for describing the stability constants, GA-MLR was applied. For obtaining best QSPR model, first, the best one molecular descriptors model was obtained (the model with high fitness function value). Then the best two molecular descriptors model was obtained. This procedure was repeated to obtain the best three, four, five and so on molecular descriptors model. The best multivariate linear model has five parameters because increase in the number of molecular descriptors has no significant effect on the accuracy of the best model. The best significant relationship for the logK 15C5 ethers has been deduced to be

$$\begin{aligned}\text{Log K} = {} & 2.26(\pm 0.63) - 0.75(\pm 0.20)\text{GATS6e} \\ & + 0.44(\pm 0.19)\text{H1u} - 6.71(\pm 0.96)\text{HATS8m} \\ & + 0.94(\pm 0.31)\text{R7u} + 0.43(\pm 0.091)\text{nCrR2}\end{aligned} \tag{7}$$

$$\begin{aligned}n_{\text{Calibration}} = {} & 35;\ n_{\text{Prediction}} = 15;\ n_{\text{Validation}} = 15 Q^2_{\text{Calibration}} \\ = {} & 0.93;\ Q^2_{\text{LOO}} = 0.88;\ Q^2_{\text{Prediction}} \\ = {} & 0.84;\ Q^2_{\text{Validation}} = 0.82;\ \text{RMSEC} \\ = {} & 0.104;\ \text{RMSECV} = 0.133;\ \text{RMSEP} \\ = {} & 0.140;\ \text{RMSEV} = 0.153\end{aligned}$$

Table 3 indicates the linear correlation-coefficient matrix for logK and five descriptors in the MLR model. The MLR model results are given in Table 4; b and S_b are the non-standardized coefficient of descriptors and standard error of coefficient, respectively, and b_s is the standardized regression coefficient. The molecular descriptors, descriptor type, definition of descriptors, and coefficient of descriptors are presented in Table 4. In Fig. 1, the plot of predicted logK by the MLR model employed against the experimental logK is represented. MLR outlier and leverage are indicated in Fig. 2.

GATS6e is one of the Geary Autocorrelation descriptors. However, the general index of spatial autocorrelation that is applied to a molecular graph, can be defined as:

$$\begin{aligned}\text{GATSdw} &= ((n-1)/2)(A)/(B) \\ A &= (1/\Delta)\left(\sum_{i=1}^{n}\sum_{g=1}^{n}\delta_{ij}(w_i - w_j)^2\right), \\ B &= \sum_{i=1}^{n}(w_i - \bar{w})^2\end{aligned} \tag{8}$$

where w_i is any atomic property, $\bar{w}$ is its average value on the molecule, A is the atom number, d is the considered topological distance (i.e. the lag in the autocorrelation terms), δ_{ij} is a Kronecker delta ($\delta_{ij} = 1$ if $d_{ij}, = d$, zero otherwise). Δ is the sum of the Kronecker deltas, i.e. the number of vertex pairs at distance equal to d [47].

H1u, HATS8m, and R7u descriptors are in GETAWAY types descriptors. GETAWAY types of descriptors have been designed with the main purpose of matching the 3D-molecular geometry and are derived from the elements h_{ij} of the Molecular Influence matrix (H), obtained through the values of atomic Cartesian coordinates. The diagonal elements of H (h_{ii}) are called leverages, and are considered to represent the influence of each atom on the shape of the molecule. For instance, the mantle atoms always have higher h_{ii} values than atoms near the molecule center, while each off-diagonal element h_{ij} represents the degree of accessibility of the jth atom to interactions with the ith atom. The influence/distance matrix (R) involves a combination of the elements of the H matrix with those of the Geometric Matrix (G) [48, 49].

Table 3 Linear correlation-coefficient matrix for the five descriptors and logK in the MLR model

	GATS6e	H1u	HATS8m	R7u	nCrR2	Log K
GATS6e	1.00					
H1u	0.01	1.00				
HATS8m	−0.19	−0.17	1.00			
R7u	−0.17	0.12	0.33	1.00		
nCrR2	−0.23	0.66	0.27	0.48	1.00	
Log K	−0.31	0.69	−0.25	0.45	0.77	1.00

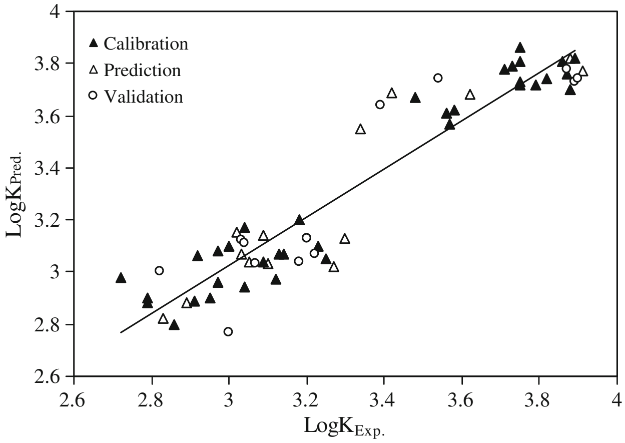


Fig. 1 The plot of the predicted versus experimental logK values for calibration, prediction, and validation sets of complexes of 15C5 with Na^+

nCrR2 is one of the chemical functional group descriptors and indicates the number of ring quaternary C.

The negative sign of GATS6e and HATS8m in MLR model shows that increasing these two parameters cause the logK decreases. Meanwhile, H1u, R7u, and nCrR2 have a positive effect on logK.

The standardized regression coefficients reveal the significance of an individual descriptor presented in the regression model. Obviously, in Table 4, the effect of the number of ring quaternary C on logK of the 15C5 complexes is more significant than that of the other descriptors. The order of significance of the other descriptors is HATS8m > R7u > H1u > GATS6e.

The inter-correlation of the descriptors used in the MLR model (Table 3) was low (below 0.66) which is in conformity to the study that for a statistically significant model, it is necessary that the descriptors involved in the equation should not be inter-correlated with each other [50]. To further check the inter-correlation of descriptors VIF analysis was performed. The VIF for each descriptor is summarized in Table 4. As one can see, the VIF values are

Table 4 The MLR model results

Variable	Descriptor type	Definition	b	S_b	b_s	VIF
Intercept	–		2.2561	0.6331		
GATS6e	2D autocorrelation	Geary autocorrelation-lag 6/weighted by atomic sanderson electronegativities	−0.7506	0.2049	−0.1982	1.17
H1u	GETAWAY descriptors	H autocorrelation of lag 1/unweighted	0.4390	0.1941	0.2064	3.31
HATS8m	GETAWAY descriptors	Leverage-weighted autocorrelation of lag 8/weighted by atomic masses	−6.7153	0.9615	−0.4800	1.88
R7u	GETAWAY descriptors	R autocorrelation of lag 7/unweighted	0.9432	0.3101	0.2360	2.39
nCrR2	Functional groups	Number of ring quaternary C(sp3)	0.4237	0.0908	0.5462	5.45

all less than 6.0, indicating the stability of the equations constructed (according to statistics principle, a value of 1.0 is indicative of no correlation, while a value of under 10.0 is statistically satisfactory) [45, 46].

On analyzing the model AD in the Williams plot of MLR model (Fig. 2) only compound number 48 in the prediction set was identified as an outlier but it belongs to the model AD, while compound numbers of 36 and 61 of calibration set, compound number 65 of prediction set, and compound number 53 of validation set are chemicals with high leverage.

Almost all the high-leverage compounds contain a bromine atom. According to these observations, the model appears to be applicable and predictive to diverse 15C5 ethers; however, particular attention should be paid to halogenated compounds, since they might fall outside the structural AD of the model. Outlier compound belongs to the model AD, this erroneous prediction could probably be attributed to wrong experimental data rather than to molecular structure.

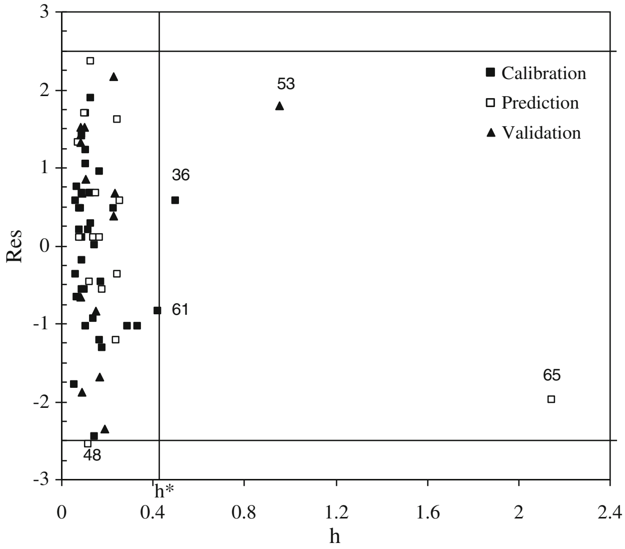


Fig. 2 MLR outlier and leverage plot. Compounds falling outside the applicability domain of the model (h* > 0.43) as well as outliers for response (standardized residuals > 2.5σ) are labeled with the respective ID number

Conclusions

A new reliable and accurate MLR–QSPR model was proposed for prediction of the stability constants (logK) of the complexes of 15C5 ethers with Na^+. The model was strongly verified for its predictive power using different internal and external validation techniques. The selection of the best variables from among the available descriptors was performed by MLR-GA, and resulted in the combination of GATS6e, H1u, HATS8m, R7u, and nCrR2 Dragon descriptors. The predictive ability of this combination of variables was with high Q^2_{ext} (0.82) and low RMSEV (0.15), which highlights the importance of these variables in modeling the studied property.

The selected variables for QSPR model are confirmed to be the most related to the studied property, as already observed in the literature. Even in spite of the model shows a large AD, particular attention should be paid when this QSPR is applied to halogenated 15C5 ethers.

Acknowledgments This work is supported by Islamic Azad University, Kermanshah Branch, Kermanshah, Iran.

References

1. Saleh, M.I., Kusrini, E., Fun, H.K., Yamin, B.M.: Structural and selectivity of 18-crown-6 ligand in lanthanide–picrate complexes. J. Organomet. Chem. **693**, 2561–2571 (2008)
2. Moriuchi-Kawakami, T., Aoki, R., Morita, K., Tsujioka, H., Fujimori, K., Shibutani, Y., Shono, T.: Conformational analysis of 12-crown-3 and sodium ion selectivity of electrodes based on bis(12-crown-3) derivatives with malonate spacers. Anal. Chim. Acta **480**, 291–298 (2003)
3. Takeda, Y., Yasui, A., Katsuta, S.: Extraction of sodium and potassium perchlorates with dibenzo-18-crown-6 into various organic solvents. Quantitative elucidation of anion effects on the extraction-ability and -selectivity. J. Incl. Phenom. Macrocycl. Chem. **50**, 157–164 (2004)

4. Kim, H.S., Chi, K.W.: Monte Carlo simulation study for QSPR of solvent effect on the selectivity of 18-crown-6 between Gd^{3+} and Yb^{3+} ions. J. Mol. Struct. Theochem. **722**, 1–7 (2005)
5. Kim, H.S.: QSPR analysis of solvent effect on selectivity of 18-crown-6 between Nd^{3+} and Eu^{3+} ions: a Monte Carlo simulation study. Bull. Korean Chem. Soc. **27**, 2011–2018 (2006)
6. Kim, H.S.: QSPR analysis of solvent effect on selectivity of 18-crown-6 between Nd^{3+} and Eu^{3+} ions: a Monte Carlo simulation study. Abstr. Pap. Am. Chem. Soc. **230**, U1327–U1328 (2005)
7. Yazdi, A.S., Mofazzeli, F., Es'haghi, Z.: Determination of 3-nitroaniline in water samples by directly suspended droplet three-phase liquid-phase microextraction using 18-crown-6 ether and high-performance liquid chromatography. J. Chromatogr. A **1216**, 5086–5091 (2009)
8. Parat, C., Betelu, S., Authier, L., Potin-Gautier, M.: Determination of labile trace metals with screen-printed electrode modified by a crown-ether based membrane. Anal. Chim. Acta **573**, 14–19 (2006)
9. Raut, D.R., Mohapatra, P.K., Ansari, S.A., Sarkar, A., Manchanda, V.K.: Selective transport of radio-cesium by supported liquid membranes containing calix[4]crown-6 ligands as the mobile carrier. Desalination **232**, 262–271 (2008)
10. Heng, L.Y., Hall, E.A.H.: Assessing a photocured self-plasticised acrylic membrane recipe for Na^+ and K^+ ion selective electrodes. Anal. Chim. Acta **443**, 25–40 (2001)
11. Han, W.S., Lee, Y.H., Jung, K.J., Ly, S.Y., Hong, T.K., Kim, M.H.: Potassium ion-selective polyaniline solid-contact electrodes based on 4′,4″(5″)-di-tert-butyldibenzo-18-crown-6-ether ionophore. J. Anal. Chem. **63**, 987–993 (2008)
12. Zeng, X.S., Han, X.X., Chen, L.X., Li, Q.S., Xu, F.B., He, X.W., Zhang, Z.Z.: The first synthesis of a calix[4](diseleno)crown ether as a sensor for ion-selective electrodes. Tetrahedron Lett. **43**, 131–134 (2002)
13. Pozzi, G., Quici, S., Fish, R.H.: Perfluorocarbon soluble crown ethers as phase transfer catalysts. Adv. Synth. Catal. **350**, 2425–2436 (2008)
14. Xia, L.X., Jia, Y., Tong, S.R., Wang, J., Han, G.X.: Interfacial behavior of phase transfer catalysis of the reaction between potassium thiocyanate and *p*-nitrobenzyl bromide with crown ethers as catalysts. Kinet. Catal. **51**, 69–74 (2010)
15. Jaszay, Z., Pham, T.S., Nemeth, G., Bako, P., Petnehazy, I., Toke, L.: Asymmetric synthesis of substituted alpha-amino phosphonates with chiral crown ethers as catalysts. Synlett. **9**, 1429–1432 (2009)
16. Seki, A., Motoya, K., Watanabe, S., Kubo, I.: Novel sensors for potassium, calcium and magnesium ions based on a silicon transducer as a light-addressable potentiometric sensor. Anal. Chim. Acta **382**, 131–136 (1999)
17. Katritzky, A.R., Chen, K., Maran, U., Carlson, D.A.: QSPR correlation and predictions of GC retention indexes for methyl-branched hydrocarbons produced by insects. Anal. Chem. **72**, 101–109 (2000)
18. Ghasemi, J.B., Ahmadi, S., Brown, S.D.: A quantitative structure-retention relationship study for prediction of chromatographic relative retention time of chlorinated monoterpenes. Environ. Chem. Lett. (2009) (in press)
19. Fang, L., Huang, J., Yu, G., Li, X.: Quantitative structure-property relationship studies for direct photolysis rate constants and quantum yields of polybrominated diphenyl ethers in hexane and methanol. Ecotoxicol. Environ. Saf. **72**, 1587–1593 (2009)
20. Ghasemi, J., Ahmadi, S.: Combination of genetic algorithm and partial least squares for cloud point prediction of nonionic surfactants from molecular structures. Ann. Chim. Rome **97**, 69–83 (2007)
21. Tetko, I.V., Solov'ev, V.P., Antonov, A.V., Yao, X., Doucet, J.P., Fan, B., Hoonakker, F., Fourches, D., Jost, P., Lachiche, N., Varnek, A.: Benchmarking of linear and nonlinear approaches for quantitative structure–property relationship studies of metal complexation with ionophores. J. Chem. Inf. Model. **46**, 808–819 (2006)
22. Yao, X.J., Fan, B.T., Doucet, J.P., Panaye, A., Liu, M.C., Zhang, R.S., Zhang, X.Y., Hu, Z.D.: Quantitative structure property relationship models for the prediction of liquid heat capacity. QSAR Comb. Sci. **22**, 29–48 (2003)
23. Gakh, A.A., Sumpter, B.G., Noid, D.W., Sachleben, R.A., Moyer, B.A.: Prediction of complexation properties of crown ethers using computational neural networks. J. Incl. Phenom. Mol. Recognit. Chem. **27**, 201–213 (1997)
24. Shi, Z.G., Mccullough, E.A.: A computer simulation statistical procedure for predicting complexation equilibrium constants. J. Incl. Phenom. Mol. Recognit. Chem. **18**, 9–26 (1994)
25. Varnek, A., Wipff, G., Solov'ev, V.P., Solotnov, A.F.: Assessment of the macrocyclic effect for the complexation of crown-ethers with alkali cations using the substructural molecular fragments method. J. Chem. Inf. Comput. Sci. **42**, 812–829 (2002)
26. Ghasemi, J., Saaidpour, S.: QSPR modeling of stability constants of diverse 15-crown-5 ethers complexes using best multiple linear regression. J. Incl. Phenom. Macrocycl. Chem. **60**, 339–351 (2008)
27. Leardi, R., Boggia, R., Terrile, M.: Genetic algorithms as a strategy for feature-selection. J. Chemom. **6**, 267–281 (1992)
28. Todeschini, R., Consonni, V., Mauri, A., Pavan, M.: Detecting "bad" regression models: multicriteria fitness functions in regression analysis. Anal. Chim. Acta **515**, 199–208 (2004)
29. Izatt, R.M., Pawlak, K., Bradshaw, J.S., Bruening, R.L.: Thermodynamic and kinetic data for macrocycle interaction with cations and anions. Chem. Rev. **91**, 1721–2085 (1991)
30. Hyperchem, v.7.5. Hypercube Inc. http://www.hyper.com (2002)
31. Dewar, M.J.S., Zoebisch, E.G., Healy, E.F., Stewart, J.J.P.: AM1—a new general purpose quantum mechanical molecular model. J. Am. Chem. Soc. **107**, 3902–3909 (1985)
32. Stewart, J.J.P.: Mopac 6.0, Quantum chemical program exchange. (1990)
33. Talete, S.: Dragon for windows (software for molecular descriptor calculations), version 5.4. http://www.talete.mi.it (2006)
34. Goldberg, D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, Reading, MA (1989)
35. Goodarzi, M., Freitas, M.P., Wu, C.H., Duchowicz, P.R.: pKa modeling and prediction of a series of pH indicators through genetic algorithm-least square support vector regression. Chemom. Intell. Lab. **101**, 102–109 (2010)
36. Cho, S.J., Hermsmeier, M.A.: Genetic algorithm guided selection: variable selection and subset selection. J. Chem. Inf. Comput. Sci. **42**, 927–936 (2002)
37. Gharagheizi, F., Alamdari, R.F.: Prediction of flash point temperature of pure components using a Quantitative Structure–Property Relationship model. QSAR Comb. Sci. **27**, 679–683 (2008)
38. Rogers, D., Hopfinger, A.J.: Application of genetic function approximation to quantitative structure–activity relationships and quantitative structure–property relationships. J. Chem. Inf. Comput. Sci. **34**, 854–866 (1994)
39. Hemmateenejad, B., Miri, R., Akhond, M., Shamsipur, M.: QSAR study of the calcium channel antagonist activity of some recently synthesized dihydropyridine derivatives. An application of genetic algorithm for variable selection in MLR and PLS methods. Chemom. Intell. Lab. **64**, 91–99 (2002)
40. Depczynski, U., Frost, V.J., Molt, K.: Genetic algorithms applied to the selection of factors in principal component regression. Anal. Chim. Acta **420**, 217–227 (2000)

41. Jouanrimbaud, D., Massart, D.L., Leardi, R., Denoord, O.E.: Genetic algorithms as a tool for wavelength selection in multivariate calibration. Anal. Chem. **67**, 4295–4301 (1995)
42. Atkinson, A.C.: Plots, Transformations and Regression. Clarendon Press, Oxford (1985)
43. Gramatica, P.: Principles of QSAR models validation: internal and external. QSAR Comb. Sci. **26**, 694–701 (2007)
44. Guha, R., Serra, J.R., Jurs, P.C.: Generation of QSAR sets with a self-organizing map. J. Mol. Graph. Model. **23**, 1–14 (2004)
45. Jaiswal, M., Khadikar, P.V., Scozzafava, A., Supuran, C.T.: Carbonic anhydrase inhibitors: the first QSAR study on inhibition of tumor-associated isoenzyme IX with aromatic and heterocyclic sulfonamides. Bioorg. Med. Chem. Lett. **14**, 3283–3290 (2004)
46. Shapiro, S., Guggenheim, B.: Inhibition of oral bacteria by phenolic compounds—Part 1. QSAR analysis using molecular connectivity. Quant. Struct. Act. Relatsh. **17**, 327–337 (1998)
47. Geary, R.C.: The contiguity ratio and statistical mapping. Incorp. Statist. **5**, 115–145 (1954)
48. Consonni, V., Todeschini, R., Pavan, M.: Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. J. Chem. Inf. Comput. Sci. **42**, 682–692 (2002)
49. Consonni, V., Todeschini, R., Pavan, M., Gramatica, P.: Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies. J. Chem. Inf. Comput. Sci. **42**, 693–705 (2002)
50. Deswal, S., Roy, N.: Quantitative structure activity relationship studies of aryl heterocycle-based thrombin inhibitors. Eur. J. Med. Chem. **41**, 1339–1346 (2006)