EJAC
Eurasian Journal of Analytical Chemistry

Eurasian J Anal Chem 8(2): 50-63, 2013

Linear and nonlinear quantitative structure linear retention indices relationship models for essential oils

Hadi Noorizadeh*

Chemometrics Lab, Department of Chemistry, Faculty of Science, Islamic Azad University, Ilam Branch, Iran

Received: *30/07/2011****; Accepted:*** *07/10/2011*

Abstract

Genetic algorithm and multiple linear regression (GA-MLR), partial least square (GA-PLS) and kernel PLS (GA-KPLS) techniques were used to investigate the correlation between linear retention indices (LRI) and descriptors for 101 diverse compounds in essential oils of six Stachys species which obtained by gas chromatography/electron impact mass spectrum (GC-EIMS). The correlation coefficient LGO-CV (Q^2) between experimental and predicted LRI for training and test sets by GA-MLR, GA-PLS and GA-KPLS was 0.936, 0.942 and 0.967 (for 80 compounds), 0.860, 0.871 and 0.919 (for 21 compounds), respectively. This indicates that GA-KPLS can be used as an alternative modeling tool for quantitative structure–retention relationship (QSRR) studies.

Keywords:

Essential oils; QSRR; Genetic algorithm-kernel partial least squares

1. Introduction

An essential oil is a volatile mixture of organic compounds derived from odorous plant material by physical means [1]. The composition of essential oil has been extensively investigated because of its commercial interest in the fragrance industry (soaps, colognes, perfumes, skin lotion and other cosmetics), in aromatherapy (relaxant), in pharmaceutical preparations for its therapeutic effects as a sedative, spasmolytic, antiviral and antibacterial agent [2]. Recently it has also been employed in food manufacturing as natural flavouring for beverages, ice cream, candy, baked goods and chewing gum. The constituents of an essential oil may be classified into two principal groups: (a) hydrocarbons (terpenes, sesquiterpenes and diterpenes); (b) oxygenated compounds derived from these hydrocarbons including alcohols, aldehydes, esters, kethons, phenols, oxides, etc [1]. Salvia genus has about 900 species, and is widespread throughout the world. This genus is represented in Turkish flora, by 89 species and 97 taxa, 45 of which are endemic [3]. From Turkish Salvia species many antibacterial [4], cytotoxic [5], antioxidant [6] and antituberculous [7] compounds, as well as cardioactive terpenoids, have been isolated. The constituents of the essential oils of three Salvia species from Turkish flora includes: monoterpene hydrocarbons, oxygenated monoterpenes, sesquiterpene hydrocarbons, diterpenes, not iso-prenoid compounds and oxygenated sesquiterpenes. These entire compounds have been identified by gas chromatography/electron impact mass spectrum (GC/EIMS).

GC/EIMS is the main method for identification of these volatile plant oils. To increase the reliability of the MS identification, comprehensive two-dimensional GC–MS can be used.

* **Corresponding Author**
E-mail: noorizadehadi@yahoo.com
ISSN: 1306-3057

This technique is based on two consecutive GC separations, typically according to boiling point and polarity [8]. The compounds are identified by comparison of retention indices with those reported in the literature and by comparison of their mass spectra with libraries or with the published mass spectra data [9]. Chromatographic retention for capillary column gas chromatography is the calculated quantity, which represents the interaction between stationary liquid phase and gas-phase solute molecule. This interaction can be related to the functional group, electronic and geometrical properties of the molecule [10, 11].

Mathematical modeling of these interactions helps chemists to find a model that can be used to obtain a deep understanding about the mechanism of interaction and to predict the retention indices of new or even unsynthesized compounds [12]. Building retention prediction models may initiate such theoretical approach, and several possibilities for retention prediction in GC. Among all methods, quantitative structure-retention relationships (QSRR) are most popular. In QSRR, the retention of given chromatographic system was modeled as a function of solute (molecular) descriptors. A number of reports, dealing with QSRR retention indices calculation of several compounds, have been published in the literature [13-15].

The QSRR models apply to multiple linear regression (MLR) and partial least squares (PLS) methods often combined with genetic algorithms (GA) for feature selection [16-17]. Because of the complexity of relationships between the property of molecules and structures, nonlinear models are also used to model the structure–property relationships. In the recent years, nonlinear kernel-based algorithms as kernel partial least squares (KPLS) have been proposed [18, 19]. The basic idea of KPLS is first to map each point in an original data space into a feature space via nonlinear mapping and then to develop a linear PLS model in the mapped space. According to Cover's theorem, nonlinear data structure in the original space is most likely to be linear after high-dimensional nonlinear mapping [20]. Therefore, KPLS can efficiently compute latent variables in the feature space by means of integral operators and nonlinear kernel functions. Compared to other nonlinear methods, the main advantage of the kernel based algorithm is that it does not involve nonlinear optimization. It essentially requires only linear algebra, making it as simple as the conventional linear PLS. In addition, because of its ability to use different kernel functions, KPLS can handle a wide range of nonlinearities.

2. Experimental methods

2.1. Data set

Linear retention indices of the essential oils of three different Salvia species [Salvia aucheri var. aucheri (endemic), Salvia aramiensis and Salvia pilifera (endemic)] was studied by GC–EIMS, which contains 101 compounds[21] (see Table.1). This set were measured at the same condition with the DB-5 capillary column (30 m$\times$0.25 mm; coating thickness 0.25 μm) and a Varian Saturn 2000 ion trap mass detector. The linear retention indices of these compounds were decreased in the range of 1672 and 802 for both Bulnesol and Hexanal, respectively.

In order to evaluate the generated models, we used leave-group-out cross validation (LGO-CV). This methodology systematically removed one group data at a time from the data set. A QSRR model was then constructed on the basis of this reduced data set and subsequently used to predict the removed data set. This procedure was repeated until a complete set of predicted was obtained.

Table 1. The data set and the corresponding observed and predicted LRI values by GA-KPLS for the training and test set.

No	Name	LRI_{Exp}	LRI_{Cal}	RE	AbsE
	Training set				
1	Hexanal	802	841	4.86	39
2	Tricyclene	927	953	2.80	26
3	α -Pinene	939	957	1.92	18
4	α -Fenchene	953	922	3.25	31
5	Sabinene	975	929	4.72	46
6	β -Pinene	978	986	0.82	8
7	1-Octen-3-ol	979	1028	5.01	49
8	3-Octanol	990	1039	4.95	49
9	2-Octanol	995	1022	2.71	27
10	p-Mentha-1,7(8)-diene	1004	987	1.69	17
11	α -Terpinene	1017	1044	2.65	27
12	Limonene	1029	1038	0.87	9
13	β -Phellandrene	1030	1041	1.07	11
14	1,8-Cineole	1031	1023	0.78	8
15	(Z)- β -Ocimene	1037	1023	1.35	14
16	γ -Terpinene	1060	1048	1.13	12
17	cis-Sabinene hydrate	1070	1097	2.52	27
18	trans-Linalool oxide	1073	1086	1.21	13
19	cis-Linalool oxide	1087	1092	0.46	5
20	Terpinolene	1089	1049	3.67	40
21	trans-Sabinene hydrate	1098	1084	1.28	14
22	3-Octyl acetate	1113	1067	4.13	46
23	trans-Thujone	1114	1099	1.35	15
24	cis-p-Menth-2-en-1-ol	1122	1147	2.23	25
25	trans-p-Mentha-2,8-dien-1-ol	1123	1128	0.45	5
26	trans-p-Menth-2-en-1-ol	1141	1180	3.42	39
27	Camphor	1146	1132	1.22	14
28	Pinocarvone	1165	1181	1.37	16
29	δ -Terpineol	1166	1257	7.80	91
30	Borneol	1169	1206	3.17	37
31	Terpinen-4-ol	1177	1138	3.31	39
32	p-Cymen-8-ol	1183	1201	1.52	18
33	α -Terpineol	1189	1187	0.17	2
34	Myrtenal	1194	1169	2.09	25
35	cis-Piperitol	1196	1221	2.09	25
36	γ -Terpineol	1199	1296	8.09	97
37	cis-p-Mentha-1(7),8-diene-2-ol	1231	1252	1.71	21
38	Piperitenone	1253	1242	0.88	11
39	trans-Myrtanol	1261	1272	0.87	11
40	Bornyl acetate	1289	1334	3.49	45
41	trans-Sabinyl acetate	1291	1324	2.56	33
42	Carvacrol	1299	1319	1.54	20
43	Piperitenone	1343	1428	6.33	85
44	α -Cubebene	1351	1363	0.89	12
45	Eugenol	1359	1397	2.80	38
46	α -Ylangene	1375	1392	1.24	17
47	α -Copaene	1377	1371	0.44	6
48	β -Bourbonene	1388	1369	1.37	19

Table 1 (continiued)

No	Name	LRI Exp	LRI Cal	RE	AbsE
49	Z-Jasmone	1393	1358	2.51	35
50	(E)-Caryophyllene	1419	1440	1.48	21
51	Aromadendrene	1441	1454	0.90	13
52	(Z)-b-Farnesene	1443	1427	1.11	16
53	α -Humulene	1455	1476	1.44	21
54	Geranylacetone	1456	1583	8.72	127
55	γ -Muurolene	1480	1471	0.61	9
56	ar-Curcumene	1481	1503	1.49	22
57	(E)- β -Ionone	1489	1609	8.06	120
58	β -Selinene	1490	1537	3.15	47
59	epi-Cubenol	1494	1387	7.16	107
60	Valencene	1496	1480	1.07	16
61	α -Selinene	1498	1409	5.94	89
62	α -Muurolene	1502	1493	0.60	9
63	γ -Muurolene	1506	1519	0.86	13
64	γ -Cadinene	1514	1541	1.78	27
65	δ -Cadinene	1523	1567	2.89	44
66	cis-Calamenene	1540	1534	0.39	6
67	Selina-3,7(11)-diene	1547	1567	1.29	20
68	Elemol	1550	1564	0.90	14
69	Spathulenol	1578	1634	3.55	56
70	Caryophyllene oxide	1583	1575	0.51	8
71	Gleenol	1587	1565	1.39	22
72	Salvial-4(14)-en-1-one	1595	1634	2.45	39
73	Guaiol	1601	1631	1.87	30
74	Benzophenone	1628	1524	6.39	104
75	γ -Eudesmol	1632	1652	1.23	20
76	α -Muurolol	1646	1657	0.67	11
77	Cubenol	1647	1680	2.00	33
78	β -Eudesmol	1651	1641	0.61	10
79	α -Eudesmol	1656	1716	3.62	60
80	Bulnesol	1672	1703	1.85	31
	Test set				
81	α -Thujene	930	951	2.26	21
82	Camphene	954	901	5.56	53
83	Myrcene	991	982	0.91	9
84	p-Cymene	1025	1051	2.54	26
85	(E)- β -Ocimene	1036	1047	1.06	11
86	(Z)-2-Hexenal	1085	1184	9.12	99
87	Linalool	1097	1106	0.82	9
88	trans-Pinocarveol	1139	1214	6.58	75
89	Naphthalene	1181	1125	4.74	56
90	Myrtenol	1195	1207	1.00	12
91	trans-Carveol	1217	1191	2.14	26
92	Thymol	1290	1178	8.68	112
93	α -Terpinyl acetate	1349	1501	11.27	152
94	Geranyl acetate	1381	1330	3.69	51
95	β -Ylangene	1421	1496	5.28	75
96	Germacrene D	1485	1524	2.63	39

Table 1 (continiued)

No	Name	LRI Exp	LRI Cal	RE	AbsE
97	Bicyclogermacrene	1500	1512	0.80	12
98	α -Calacorene	1546	1574	1.81	28
99	β -Calacorene	1566	1531	2.24	35
100	Humulene epoxide II	1608	1772	10.20	164
101	α -Cadinol	1654	1682	1.69	28

2.2. Descriptor calculation

All structures were drawn with the HyperChem software (version 6). Optimization of molecular structures was carried out by semi-empirical AM1 method using the Fletcher-Reeves algorithm until the root mean square gradient of 0.01 was obtained. Since the calculated values of the electronic features of molecules will be influenced by the related conformation. In the current research an attempt was made to use the most stable conformations. Some electronic descriptors such as polarizability, dipole moment and orbital energies of LUMO and HOMO were calculated by using the HyperChem software. Also optimized structures were used to calculate 1497 descriptors by DRAGON software Version 3 [22].

2.3. Genetic Algorithm

A detailed description of the genetic algorithm (GA) can be found in the literature [23, 24]. Genetic algorithm is simulated methods based on ideas from Darwin's theory of natural selection and evolution (the struggle for life). In GA a chromosome (or an individual) can be defined as an enciphered entity of a candidate solution, which is expressed as a set of variables. GA consist of the following basic steps: (1) A chromosome is represented by a binary bit string and an initial population of chromosomes is created in a random way; (2) A value for the fitness function of each chromosome is evaluated; (3) Based on the values of the fitness functions, the chromosomes of the next generation are produced by selection, crossover and mutation operations. The fitness function was proposed by Depczynski *et al* [25]. The parameter algorithm reported in Table 2. The root-mean-square errors of calibration (RMSEC) and prediction (RMSEP) were calculated and the fitness function was calculated by Eq. (1).

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

Where m_c and m_p are the number of compounds in the calibration and prediction set and n represent the number of selected variables, respectively. The parameter algorithm reported in Table 2.

Table 2. Parameters of the genetic algorithm

Population size: 30 chromosomes
On average, five variables per chromosome in the original population
Regression method: MLR, PLS, KPLS
Cross validation: leave-group-out
Number subset: 4
Maximum number of variables selected in the same chromosome: (MLR, 10), (PLS, 30)
Elitism: True
Crossover: multi Point
Probability of crossover: 50%
Mutation: multi Point
Probability of mutation: 1%
Maximum number of components: (PLS, 10)
Number of runs: 100

2.4. Linear models

2.4.1. Multiple linear regression

A major step in constructing the QSRR model is finding a set of molecular descriptors that represent variation in the structural property of the molecules. The modeling and prediction of the physicochemical properties of organic compounds is an important objective in many scientific fields [26, 27]. MLR is one of the most modeling methods in QSRR. MLR method provides an equation that links the structural features to the LRI of the compounds:

$$LRI = a_0 + a_1\mathbf{d}_1 + \cdots + a_n\mathbf{d}_n \quad (2)$$

Where a_0 and a_i are intercept and regression coefficients of the descriptors, respectively. $\mathbf{d}_i$ has the common definition, variable or descriptor in this case, the elements of this vector are equivalent numerical values of descriptors of the molecules. The greater absolute value of a coefficient, caused to the greater weight of variable in the model. The positive sign of corresponding regression coefficient between LRI and descriptors indicates that LRI increases with increasing the magnitude of descriptors. The negative sign of the corresponding regression coefficient between LRI and descriptors indicates that LRI increase with decreasing the magnitude of descriptors.

2.4.2. Partial least squares

PLS is a linear multivariate method for relating the process variables X with responses Y. PLS can analyze data with strongly collinear, noisy, and numerous variables in both X and Y [28]. PLS reduces the dimension of the predictor variables by extracting factors or latent variables that are correlated with Y while capturing a large amount of the variations in X. This means that PLS maximizes the covariance between matrices X and Y. In PLS, the scaled matrices X and Y are decomposed into score vectors (t and u), loading vectors (p and q), and residual error matrices (E and F):

$$X = \sum_{i=1}^{a} t_i p_i^T + E$$

$$Y = \sum_{i=1}^{a} u_i q_i^T + F \quad (3)$$

Where a is the number of latent variables. In an inner relation, the score vector t is linearly regressed against the score vector u.

$$U_i = b_i t_i + h_i \quad (4)$$

Where b is regression coefficient that is determined by minimizing the residual h. It is crucial to determine the optimal number of latent variables and cross validation is a practical

and reliable way to test the predictive significance of each PLS component. There are several algorithms to calculate the PLS model parameters. In this work, the NIPALS algorithm was used with the exchange of scores [29].

2.5. Nonlinear model

2.5.1. Kernel partial least squares

The KPLS method is based on the mapping of the original input data into a high dimensional feature space $\Im$ where a linear PLS model is created. By nonlinear mapping $\Phi : x \in \Re^n \rightarrow \Phi(x) \in \Im$, a KPLS algorithm can be derived from a sequence of NIPALS steps and has the following formulation [30]:

1. Initialize score vector w as equal to any column of Y.

2. Calculate scores $u = \Phi\Phi^T w$ and normalize u to ||u|| = 1, where Φ is a matrix of regressors.

3. Regress columns of Y on u: $c = Y^T u$, where c is a weight vector.

4. Calculate a new score vector w for Y: w = Yc and then normalize w to ||w||=1.

5. Repeat steps 2–4 until convergence of w.

6. Deflate $\Phi\Phi^T$ and Y matrices:

$$\Phi\Phi^T = (\Phi - uu^T\Phi)(\Phi - uu^T\Phi)^T \quad (5)$$

$$Y = Y - uu^T Y \quad (6)$$

7. Go to step 1 to calculate the next latent variable.

Without explicitly mapping into the high-dimensional feature space, a kernel function can be used to compute the dot products as follows:

$$k(x_i, x_j) = \Phi(x_i)^T \Phi(x_j) \quad (7)$$

$\Phi\Phi^T$ represents the (n×n) kernel Gram matrix K of the cross dot products between all mapped input data points $\Phi(x_i), i = 1,...,n$. The deflation of the $\Phi\Phi^T = K$ matrix after extraction of the u components is given by:

$$K = (I - uu^T)K(I - uu^T) \quad (8)$$

Where I is an m-dimensional identity matrix. Taking into account the normalized scores u of the prediction of KPLS model on training data $\hat{Y}$ is defined as:

$$\hat{Y} = KW(U^T KW)^{-1} U^T Y = UU^T Y \quad (9)$$

For predictions on new observation data $\hat{Y}_t$, the regression can be written as:

$$\hat{Y}_t = K_t W(U^T KW)^{-1} U^T Y \quad (10)$$

Where K_t is the test matrix whose elements are $K_{ij} = K(x_i, x_j)$ where x_i and x_j present the test and training data points, respectively.

2.6. Software and programs

A Pentium IV personal computer (CPU at 3.06 GHz) with windows XP operational system was used. Geometry Optimization was performed by HyperChem (Version 7.0 Hypercube, Inc.), Dragon software was used to calculate of descriptors. MLR analysis was performed by the SPSS Software (version 13, SPSS, Inc.) by using enter method for model building. MINITAB software (version 14, MINITAB) was used for the simple PLS analysis. Cross validation, GA-MLR, GA-PLS, GA-KPLS and other calculation were performed in the MATLAB (Version 7, Mathworks, Inc.) environment.

3. Results and discussion

3.1. Linear models

3.1.1. GA-MLR analysis

To reduce the original pool of descriptors to an appropriate size, the objective descriptor reduction was performed using various criteria. Reducing the pool of descriptors eliminates those descriptors which contribute either no information or whose information content is redundant with other descriptors present in the pool. From the variable pairs with $R > 0.90$, only one of them was used in the modeling, while the variables over 90% and equal to zero or identical were eliminated. With the use of these criteria, 1137 out of 1497 original descriptors were eliminated and remaining descriptors were employed to generate the models with the GA-MLR program. In order to minimize the information overlap in descriptors and to reduce the number of descriptors required in regression equation, the concept of non-redundant descriptors was used in our study. The best equation is selected on the basis of the highest multiple correlation coefficient leave-group-out cross validation (LGO-CV) (Q^2), the least RMSECV, absolute error (AbsE) and relative error (RE) of prediction and simplicity of the model. These parameters are probably the most popular measure of how well a regression model fits the data. Among the models proposed by GA-MLR, one model had the highest statistical quality and was repeated more than the others. This model had five molecular descriptors including constitutional descriptors (rotatable bond fraction) (RBF), topological descriptor (information content index (neighborhood symmetry of 1-order)) (IC1), RDF descriptors (Radial Distribution Function - 4.5 / weighted by atomic Sanderson electronegativities) (RDF045e) and electronic descriptor (dipole moment (μ) and lowest unoccupied molecular orbital (LUMO)). The best QSRR model obtained is given below together with the statistical parameters of the regression in Eq. (11).

$$LRI = 141.068 (\pm 52.871) - 238.311 (\pm 109.969)\, RBF + 150.028 (\pm 54.587)\, IC1 - 2.901 (\pm 0.487)\, RDF045e + 41.373 (\pm 4.699)\, \mu - 27.025 (\pm 12.430)\, LUMO \quad (11)$$

The greater absolute value of a coefficient, caused to the greater weight of variable in the model. The RBF coefficient is bigger in the equation, thus it is very important descriptor compared to the other descriptors in the model. The RBF, RDF045e and LUMO displays a negative sign which indicates that when these descriptors increase the LRI decreases. The IC1 and μ displays a positive sign which indicates that the LRI is directly related to these descriptors. The statistical parameters of this model, constructed by the selected descriptors, are depicted in Table .3.

3.1.2. GA-PLS analysis

The colinearity problem of the MLR method has been overcome through the development of the partial least-squares projections to latent structures (PLS) method. For this reason, after eliminating descriptors that had identical or zero values for greater than 90% of the compounds, 1010 descriptor were remained. These descriptors were employed to generate the models with the GA-PLS and GA-KPLS program. The best PLS model contained 6

selected descriptors in 2 latent variables space. These descriptors were obtained constitutional descriptors (number of multiple bonds) (nBM), topological descriptors (3D-Balaban index) (J3D), RDF descriptors (Radial Distribution Function - 12.5 / weighted by atomic Sanderson electronegativities) (RDF125e), atom-centred fragments (CH_2R_2 (C-002)) and electronic descriptors (polarizibiloity, dipole moment and high occupied molecular orbital (HOMO)). For this in general, the number of components (latent variables) is less than the number of independent variables in PLS analysis. The obtained statistic parameters of the GA-PLS model were shown in Table 3. The data confirm that higher correlation coefficient and lower prediction error have been obtained by PLS in relative to MLR and these reveal that PLS method produces more accurate results than that of MLR. The PLS model uses higher number of descriptors that allow the model to extract better structural information from descriptors to result in a lower prediction error.

3.2. Nonlinear model

3.2.1. GA-KPLS analysis

With the aim of improving the predictive performance of nonlinear QSRR model, GA-KPLS modeling was performed. The leave-group-out cross validation has been performed. The n selected descriptors in each chromosome were evaluated by fitness function of PLS and KPLS based on the Eq. (1). In this paper a radial basis kernel function, $k(x,y)= \exp(\|x-y\|^2/c)$, was selected as the kernel function with $c = rm\sigma^2$ where r is a constant that can be determined by considering the process to be predicted (here r set to be 1), m is the dimension of the input space and σ^2 is the variance of the data [31]. It means that the value of c depends on the system under the study. The 6 descriptors in 3 latent variables space chosen by GA-KPLS feature selection methods were contained constitutional descriptors (mean atomic van der Waals volume (scaled on Carbon atom)) (Mv) and (number of double bonds)(nDB), GETAWAY descriptors (R autocorrelation of lag 3 / weighted by atomic masses)(R3m), atom-centred fragments (H attached to $C1(sp^3)/C0(sp^2)$)(H-047) and electronic descriptors (LUMO and polarizibility). The predicted values of LRI are plotted against the experimental values for training and test set in Fig. 1. Obviously, there is a close agreement between the experimental and predicted LRI and the data represent a very low scattering around a straight line with respective slope and intercept close to one and zero.

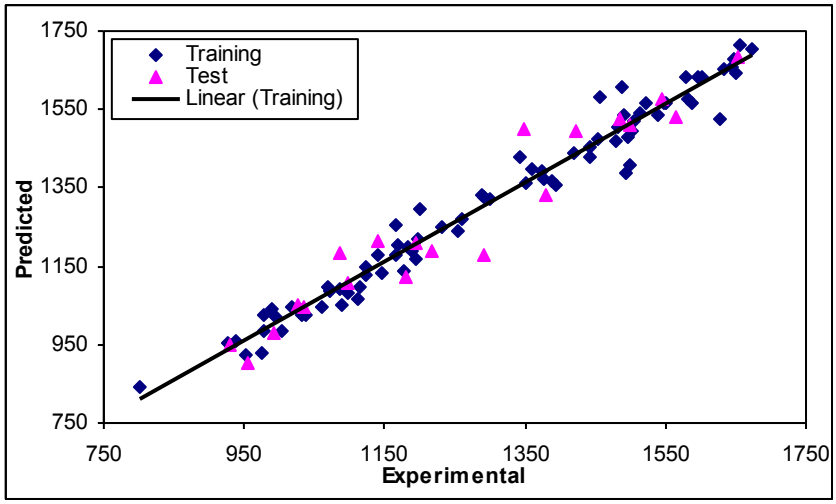

Fig. 1. Predicted vs. experimental LRI by GA-KPLS

The values of experimental, calculated and percent relative error are shown in Table 1. The statistical parameters obtained by this model for the training and test sets are summarized

in Table 3. For the constructed model, five general statistical parameters were selected to evaluate the prediction ability of the model for the LRI. Table 3 shows the statistical parameters for the compounds obtained by applying models to training and test sets. Each of the statistical parameters mentioned above were used for assessing the statistical significance of the QSRR model. The data presented in Table .3 indicate that the GA-PLS and GA-MLR linear model have good statistical quality with low prediction error, while the corresponding errors obtained by the GA-KPLS model are lower.

Table 3. The statistical parameters of different constructed QSRR models.

Model	Training set						Test set					
	R^2	Q^2	RE	RMSE	AbsE	N	R^2	Q^2	RE	RMSE	AbsE	N
GA-PLS	0.935	0.936	3.92	52.26	38.54	80	0.858	0.860	8.02	86.74	68.19	21
GA-KPLS	0.942	0.942	3.47	49.07	37.84	80	0.870	0.871	7.63	82.51	63.70	21
GA-KPLS	0.967	0.968	2.43	41.61	31.27	80	0.919	0.919	4.04	68.59	52.04	21

The Q^2, which is a measure of the model fit to the cross validation set, can be calculated as:

$$R_{cv}^2 \equiv Q^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - y_i^{\wedge})^2}{\sum_{i=1}^{n} (y_i - y^{-})^2} \quad (12)$$

Where y_i , $y_i^{\wedge}$ and y^{-} were respectively the experimental, predicted, and mean LRI values of the samples**.** The accuracy of cross validation results is extensively accepted in the literature considering the Q^2 value. In this sense, a high value of the statistical characteristic ($Q^2 > 0.5$) is considered as proof of the high predictive ability of the model [32]. However, several authors suggest that a high value of Q^2 appears to be a necessary but not sufficient condition for a model to have a high predictive power and consider that the predictive ability of a model can only be estimated using a sufficiently large collection of compounds that was not used for building the model [33].

We believe that applying only LGO-CV is not sufficient to evaluate the predictive ability of a model. Thus we employed a two-step validation protocol which contains internal (LGO-CV) and external (test set) validation methods. The data set was randomly divided into training (calibration and prediction sets) and test sets after sorting based on the LRI values. The training set consisted of 80 molecules and the test set, consisted of 21 molecules. The training set was used for model development, while the test set in which its molecules have no role in model building was used for evaluating the predictive ability of the models for external set.

The statistical parameters obtained by LGO-CV for GA-KPLS and the linear QSRR models are compared in Table .3. Inspection of the results of the table reveals a higher R^2 and Q^2 values and lower RMSE and RE for GA-KPLS model for the training and test sets compared with their counterparts for other models. This clearly shows the strength of GA-KPLS as a nonlinear feature selection method. Result indicates that the LRI of essential oils possesses some nonlinear characteristics.

3.2.2. Description of models descriptors

In the chromatographic retention of compounds in the nonpolar or low polarity stationary phases two important types of interactions contribute to the chromatographic retention of the compounds: the induction and dispersion forces. The dispersion forces are related to steric factors, molecular size and branching, while the induced forces are related to the dipolar moment, which should stimulate dipole-induced dipole interactions. For these reasons, constitutional descriptors, functional group and electronic descriptors are very important.

Electronic descriptors were defined in terms of atomic charges and used to describe electronic aspects both of the whole molecule and of particular regions, such atoms, bonds, and molecular fragments. This descriptor calculated by computational chemistry and therefore can be consider among quantum chemical descriptor.

As expected, the model included HOMO and LUMO energies to quantify electronic effects of drugs. HOMO energy is a useful descriptor that presents information on the distribution of π electron and explains $\pi - \pi$ charge transfer interactions of unsaturated compounds. HOMO energy plays a very important role in nucleophilic behaviour and it represents molecular reactivity as a nucleophile. Good nucleophiles are those in which electrons reside in high lying orbitals. Electron affinity was also shown to greatly influence the chemical behaviour of compounds, as demonstrated by its inclusion in the QSPR/QSRR. The eigenvalues of LUMO and HOMO and their energy gap reflect the chemical activity of the molecule. LUMO as an electron acceptor represents the ability to obtain an electron, while HOMO as an electron donor represents the ability to donate an electron. The lowest unoccupied molecular orbital (LUMO) energy can be interpreted as a measure of charge transfer interactions and/or of hydrogen bonding effects [34, 35].

Constitutional descriptors are most simple and commonly used descriptors, reflecting the molecular composition of a compound without any information about its molecular geometry. The most common Constitutional descriptors are number of atoms, number of bound, absolute and relative numbers of specific atom type, absolute and relative numbers of single, double, triple, and aromatic bound, number of ring, number of ring divided by the number of atoms or bonds, number of benzene ring, number of benzene ring divided by the number of atom, molecular weight and average molecular weight.

The number of H atoms attached to C1 (sp3)/C0 (sp2) (H-047) is an atom-centered descriptor calculated by knowing the molecular composition and atom connectivities. This descriptor encodes information about the hybridization and oxidation state of the carbon atoms.

Topological descriptors are based on a graph representation of the molecule. They are numerical quantifiers of molecular topology obtained by the application of algebraic operators to matrices representing molecular graphs and whose values are independent of vertex numbering or labeling. They can be sensitive to one or more structural features of the molecule such as size, shape, symmetry, branching and cyclicity and can also encode chemical information concerning atom type and bond multiplicity. Balaban index is a variant of connectivity index, represents extended connectivity and is a good descriptor for the shape of the molecules and modifying biological process. Nevertheless, some of chemists have used this index successfully in developing QSPR/QSRR models.

The radial distribution function (RDF) descriptors are based on the distances distribution in the geometrical representation of a molecule and constitute a radial distribution function code. The RDF descriptors can be restricted to specific atom types or distance ranges to represent specific information in a certain three-dimensional structure space, e.g. to describe steric hindrance or structure/activity properties of a molecule.

The GETAWAY (geometry, topology, and atom-weights assembly) descriptors try to match 3Dmolecular geometry provided by the molecular influence matrix and atom relatedness by molecular topology, with chemical information by using different atomic weights (atomic mass, polarizability, van der Waals volume, and electronegativity. GETAWAY descriptors are quickly computed from the atomic positions of the molecule atoms (hydrogens included) [36].

4. Conclusion

In this study, an accurate QSRR model for estimating the linear retention indices (LRI) of essential oils of three different Salvia species which obtained by GC–EIMS was developed by employing the two linear models (GA-MLR and GA-PLS) and one nonlinear model (GA-KPLS). The most important molecular descriptors selected represent the constitutional descriptors, functional group and electronic descriptors that are known to be important in the retention mechanism of essential oils. Three models have good predictive capacity and excellent statistical parameters. A comparison between these models revealed the superiority of the GA-KPLS to other models. It is easy to notice that there was a good prospect for the GA-KPLS application in the QSRR modeling. This indicates that LRI of essential oils possesses some nonlinear characteristics. It can also be used successfully to estimate the LRI for new compounds or for other compounds whose experimental values are unknown.

References

1. Heath HB (1978) In flavor technology. Westport, Connecticut: AVI Publishing company, Inc.
2. Kim NS, Lee DD (2002) Comparison of different extraction method for the analysis of fragrance from Lavandula species by gas chromatography-mass spectrometry, J. Chromatogr. A. 982: 31.
3. Guner A, Ozhatay N, Ekim T, Baser KHC (2000) Flora of Turkey and the East Aegean Islands, Edinburgh: Edinburgh University Press.
4. Güner A, Özhatay N, Ekim T, K.H.C. Baser, Flora of Turkey and the East Aegean Islands, Edinburgh: Edinburgh University Press, 2000.
5. Topçu G, Tan N, Kökdil G (1997) A. Ulubelen, Phytochemistry, Terpenoids from Salvia glutinosa, Phytochemistry 45: 1293.
6. [6] Topçu G, Ertas A, Kolak U, Öztürk M (2007) Antioxidant activity tests on novel triterpenoids from Salvia macrochlamys, Arkivoc 7: 195.
7. Ulubelen A, Topçu G, Bozok-Johansson B (1997) Cardioactive and antibacterial terpenoids from some Salvia species, J. Nat. Prod 60: 1275.
8. Peters R, Tonoli D, van Duin M, Mommers J, Mengerink Y, Wilbers ATM, van Benthem R, Koster CHD, Schoenmakers PJ, derWal SJV (2008) Low-molecular-weight model study of peroxide cross-linking of ethylene-propylene (-diene) rubber using gas chromatography and mass spectrometry I. Combination reactions of alkanes, J. Chromatogr. A 1201: 141.
9. Jennings W, Shibamoto T (1980) Quantitative Analysis of Flavor and Fragrance Volatile by Glass Capillary Column Gas Chromatography, Academic Press, New York.
10. Ong VS, Hites RS (1991) Relationship between gas chromatographic retention indexes and computer calculated physical properties of four compound classes, Anal. Chem. 63: 2829.

11. Peng CT, Ding SF, Hua RL, Yang WC (1988) Prediction of retention indexes: I. structure-retention index relationship on apolar column, J. Chromatogr 436: 137.

12. Kaliszan R (1997) Structure and Retention in Chromatography, Harwood, Amsterdam.

13. Qin LT, Liu SHSH, Liu HL, Tong J (2009) Comparative multiple quantitative structure–retention relationships modeling of gas chromatographic retention time of essential oils using multiple linear regression, principal component regression, and partial least squares techniques, J. Chromatogr A 1216: 5302.

14. Riahi S, Pourbasheer E, Ganjali MR, Norouzi P (2009) Investigation of different linear and nonlinear chemometric methods for modeling of retention index of essential oil components: Concerns to support vector machine, J. Hazard. Mater 166: 853-859.

15. Bombarda I, Dupuy N, Le Van Da JP, Gaydou EP (2008) Comparative chemometric analyses of geographic origins and compositions of lavandin var. Grosso essential oils by mid infrared spectroscopy and gas chromatography, Analytica Chimica Acta 613: 31.

16. Olivero J, Gracia T, Payares P, Vivas R, Diaz D, Daza E, Geerlings P (1997) Molecular structure and gas chromatographic retention behavior of the components of Ylang-Ylang oil, J. Pharm. Sci 86: 625.

17. Kaliszan R, (1993) Quantitative structure-retention relationships applied to reversed-phase high-performance liquid chromatography, J. Chromatogr. A 656: 417.

18. Scholkopf B, Smola AJ, Muller KR (1998) Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput 10: 1299.

19. Rosipal R, Trejo LJ (2001) Kernel partial least squares regression in reproducing kernel Hilbert space, J. Mach. Learning Res 2: 97.

20. Haykin S (1999) Neural Networks, Prentice-Hall, New Jersey.

21. Kelen M, Tepe B (2008) Chemical composition, antioxidant and antimicrobial properties of the essential oils of three Salvia species from Turkish flora, Bioresour. Technol 99: 4096.

22. Todeschini R, Consonni V, Mauri A, Pavan P (2003) DRAGON-Software for the calculation of molecular descriptors; Version 3.0 for Windows.

23. Cai W, Xia B, Shao X, Guo Q, Maigret B, Pan Z (2001) Molecular interactions of - cyclodextrin inclusion complexes using a genetic algorithm, J. Mol. Struct. (Theochem.) 535:115.

24. Goldberg DE (2000) Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley–Longman, Reading, MA, USA.

25. Depczynski U, Frost VJ, Molt K (2000) Genetic algorithms applied to the selection of factors in principal component regression, Anal. Chim. Acta 420: 217.

26. Citra M (1999) Estimating the pK_a of phenols, carboxylic acids and alcohols from semi-empirical quantum chemical methods, Chemosphere 38: 191.

27. Booth TD, Azzaoui K, Wainer IW (1997) Prediction of Chiral Chromatographic Separations Using Combined Multivariate Regression and Neural Networks, Anal. Chem 69: 3879.

28. Wold S, Sjostrom M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst 58: 109.

29. Geladi P, Kowalski BR (1986) Partial least-squares regression: a tutorial, Anal. Chim. Acta. 185: 1.

30. Rosipal R, Trejo LJ (2001) Kernel partial least squares regression in reproducing kernel Hilbert space, J. Mach. Learning Res. 2: 97.

31. Kim K, Lee JM, Lee IB (2005) A novel multivariate regression approach based on kernel partial least squares with orthogonal signal correction, Chemom. Intell. Lab. Syst. 79: 22.

32. Wold S (1991) validation of QSARs, Quant. Struct-Act. Relat. 10: 191.

33. Golbraikh A, Tropsha A (2002) Beware of q^2, J. Mol. Graph. Model. 20: 269.

34. Booth TD, Azzaoui K, Wainer IW (1997) Prediction of Chiral Chromatographic Separations Using Combined Multivariate Regression and Neural Networks, Anal. Chem. 69: 3879.

35. Azzaoui K, Morin-Allory L (1996) Comparison and quantification of chromatographic retention mechanisms on three stationary phases using structure-retention relationships, Chromatographia. 42: 389.

36. Todeschini R, Consonni V (2000) Handbook of molecular descriptors, Wiley-VCH, Weinheim.