World Applied Sciences Journal 14 (4): 603-606, 2011
ISSN 1818-4952

QSPR Studies of Artemisia Essential Oils by the Combination of Genetic Algorithms and PLS Analysis

Hadi Noorizadeh and Abbas Farmany

Department of Chemistry, Faculty of Sciences, Arak Branch, Islamic Azad University, Arak, Iran

Abstract: Genetic algorithm and partial least square (GA-PLS) and kernel PLS (GA-KPLS) techniques were used to investigate the correlation between retention time (RT) for *Artemisia* essential oils compounds which obtained by gas chromatography-mass spectrometry (GC-MS). The applied internal (leave-group-out cross validation (LGO-CV)) and external (test set) validation methods were used for the predictive power of models. The results indicate that GA-KPLS can be used as an alternative modeling tool for quantitative structure-retention relationship (QSRR) studies.

Key words: Artemisia essential oils · Gas chromatography-mass spectrometry · QSRR · Genetic algorithm-kernel partial least squares

INTRODUCTION

Aromatic plants are frequently used in traditional medicine as antimicrobial agents and their essential oils, mixtures of natural volatile compounds isolated by steam distillation, have been known since antiquity to possess antibacterial and antifungal properties. Previous works have suggested that several essential oils showed important antimicrobial activity against bacteria, yeasts, dermatophyte and Aspergillus strains [1, 2] and have therapeutic potential, mainly in diseases involving mucosal, cutaneous and respiratory tract infections. The major constituents of many of these oils are phenolic compounds (terpenoids and phenylpropanoids) like thymol, carvacrol or eugenol, of which antimicrobial and antioxidant activities are well documented [3].

Nevertheless, aromatic plants producing non-phenolic essential oils, like some *Artemisia* species, are also used as spices and in folk remedies as antiseptics. Powdered leaves of *A. absinthium*, *A. biennis*, *A. frigida* and *A. ludoviciana* have been applied externally in salves and washes by North American native people for treating sores and wounds and, internally to treat chest infections. Antioxidants retard oxidation and are sometimes added to meat and poultry products to prevent or slow oxidative degradation of fats. Antioxidant agents are effective due to different mechanisms such as free radical scavenging, chelating of pro-oxidant metal ions or quenching singlet-oxygen formation. The aromatic leaves of A. frigida and A. dracunculus (tarragon) have been also used as spice and to preserve meat [4]. These species might be a source of natural antioxidants and antimicrobials.

Artemisia dracunculus ethanolic extract significantly reduced hyperglycemia in mice with chemically induced insulin deficiency and diabetes and, an activity-guided fractionation revealed six active polyphenolic compounds [5]. The essential oil was screened in guinea pig and rat plasma in order to assess antiplatelet activity and inhibition of clot retraction. Gas chromatography (GC) and gas chromatography-mass spectrometry (GC-MS) are the main methods for identification of these volatile plant oils. To increase the reliability of the MS identification, comprehensive two-dimensional GC-MS can be used.

Quantitative structure-retention relationship (QSRR) is statistically derived relationships between chromatographic parameters and descriptors related to the molecular structure of the analytes. A number of reports, deals with QSRR retention calculation of several compounds have been published in the literature [6-8].

There is a trend to develop QSRR from a variety of methods. In particular, genetic algorithm (GA) is frequently used as search algorithms for variable selection in chemometrics and QSRR. GA is a stochastic method to solve the optimization problems defined by fitness criteria, applying the evolution hypothesis of Darwin and different genetic functions, i.e. crossover and mutation [9, 10].

Corresponding Author: Hadi Noorizadeh, Department of Chemistry, Faculty of Sciences, Arak Branch, Islamic Azad University, Arak, Iran. Tel: + 989183452507.

Partial least square (PLS) is the most commonly used multivariate calibration method [11, 12]. Moreover, non-linear statistical treatment of QSRR data is expected to provide models with better predictive quality as compared with related PLS models. In recent years, nonlinear kernel-based algorithm as kernel partial least squares (KPLS) has been proposed [13-15]. KPLS can efficiently compute latent variables in the feature space by means of nonlinear kernel functions. In the present work, a QSRR study has been carried out on the GC-MS system retention times (RT) for essential oils compounds.

MATERIALS AND METHODS

Data Set: The chemical composition of essential oils isolated from aerial parts of seven wild sages from Western Canada - *Artemisia absinthium* L., *Artemisia biennis* Willd., *Artemisia cana* Pursh, *Artemisia dracunculus* L., *Artemisia frigida* Willd., *Artemisia longifolia* Nutt and *Artemisia ludoviciana* Nutt., was investigated by GC-MS. The retention data of these compounds were taken from literature [16] is shown in Table 1. The data set was randomly divided into two groups including training set (calibration and prediction sets) and test set, which consists of 92 and 23 molecules, respectively. The calibration set was used for model generation. The prediction set was applied o deal with overfitting of the network, whereas test set which its molecules have no role in model building was used for the evaluation of the predictive ability of the models for external set.

Table 1: The data set and corresponding observed RT values

No	Name	RT_{Exp}
	Calibration Set	
1	3-Methyl-2-buten-1-ol	773
2	Hexanal	801
3	(2E)-Hexenal	854
4	Heptanal	903
5	Santolina triene	909
6	2,5-Diethenyl-2-methyl-tetrahydrofuran	918
7	alpha-Thujene	931
8	alpha-Pinene	938
9	Camphene	953
10	Thuja-2,4(10)-diene	959
11	Benzaldehyde	964
12	Artemiseole	978
13	beta-Pinene	981
14	6-Methyl-5-hepten-2-one	989
15	Myrcene	992
16	Mesitylene	996
17	n-Octanal	1004
18	delta-3-Carene	1013
19	alpha-Terpinene	1020
20	para-Cymene	1028
21	Limonene	1032
22	1,8-Cineole	1035
23	(Z)-beta-Ocimene	1041
24	(E)-beta-Ocimene	1051
25	gamma-Terpinene	1062
26	cis-Sabinene hydrate	1070
27	Artemisia alcohol	1085
28	trans-Sabinene hydrate	1099
29	Linalool	1100
30	n-Nonanal	1105
31	Filifolone	1105
32	1,3,8-para-Menthatriene	1114
33	trans-Thujone	1120
34	Isophorone	1123
35	Chrysanthemone	1128
36	Allo-ocimene	1132
37	(Z)-Myroxide	1136
38	trans-para-Menth-2-en-1-ol	1145
39	(E)-Myroxide	1146
40	Camphor	1149
41	Borneol	1168
42	Terpinen-4-ol	1180
43	trans-Isocarveol	1189
44	alpha-Terpineol	1191
45	Methyl chavicol	1199
46	n-Decanal	1207
47	trans-Carveol	1220
48	Nerol	1230
49	Nor-davanone	1233
50	Geraniol	1257
51	Piperitone	1257
52	cis-Verbenyl acetate	1287
53	Bornyl acetate	1287
54	trans-Sabinyl acetate	1295
55	Eugenol	1348
56	cis-Carvyl acetate	1366
57	alpha-Copaene	1377
58	(Z)-Jasmone	1398
59	Methyl eugenol	1407
60	cis, Threo-davanafuran	1417
61	para-Menth-1-en-9-ol acetate	1423
62	(Z)-beta-Farnesene	1445
63	(E)-beta-Farnesene	1461
64	ar-Curcumene	1484
65	Davana ether (isomer)	1494
66	Bicyclogermacrene	1496
67	(E,E)-alpha-Farnesene	1509
68	Davana ether (isomer)	1514
69	delta-Cadinene	1525
70	Artedouglasia oxide A	1536
71	Artedouglasia oxide D	1561
72	Davanone B	1566
73	(E)-Nerolidol	1566
74	Spathulenol	1578

Table 1: Continued

No	Name	RT_{Exp}
75	Artedouglasia oxide B	1583
76	Caryophyllene oxide	1583
77	Neryl isovalerate	1586
78	beta-Copaen-4-alpha-ol	1587
79	Davanone	1589
80	epi-alpha-Cadinol	1642
81	epi-alpha-Muurolol	1643
82	beta-Eudesmol	1650
83	alpha-Cadinol	1655
84	alpha-Bisabolol	1684
85	(2Z,6E)-Farnesyl acetate	1819
86	(Z)-en-yn-Dicycloether	1881
87	(E)-en-yn-Dicycloether	1894
88	alpha-Bisabolol	1684
89	Chamazulene	1728
90	(2Z,6E)-Farnesyl acetate	1819
91	(Z)-en-yn-Dicycloether	1881
92	(E)-en-yn-Dicycloether	1894
	Test Set	
93	n-Hexanol	868
94	Tricyclene	927
95	alpha-Fenchene	952
96	Sabinene 977	
97	alpha-Phellandrene	1006
98	beta-Phellandrene	1032
99	cis-Arbusculone	1055
100	Terpinolene	1091
101	cis-Thujone	1108
102	cis-para-Menth-2-en-1-ol	1124
103	trans-Pinocarveol	1143
104	Hexyl isobutanoate	1152
105	para-Cymen-8-ol	1187
106	Myrtenol	1197
107	trans-Piperitol	1208
108	Carvone 1247	
109	cis-Chrysanthemyl acetate	1264
110	Nor-chrysanthemic acid methyl ester	1348
111	beta-Elemene	1393
112	beta-Caryophyllene	1420
113	Germacrene D	1482
114	Cubebol 1516	
115	Chamazulene	1728

Computer Hardware and Software: All calculations were run on a HP Laptop computer with AMD Turion64X2 processor with windows XP operating system. The optimizations of molecular structures were done by the HyperChem 7.0 (AM1 method) and descriptors were calculated by Dragon Version 3.0 software's.. Cross validation, GA-PLS, GA-KPLS and other calculation were performed in the MATLAB (Version 7, Mathworks, Inc.) environment.

Cross Validation Technique: Cross validation is a popular technique used to explore the reliability of statistical models. Based on this technique, a number of modified data sets are created by deleting in each case one or a small group (leave-some-out) of objects. For each data set, an input-output model is developed, based on the utilized modeling technique. Each model is evaluated, by measuring its accuracy in predicting the responses of the remaining data (the ones or group data that have not been utilized in the development of the model) [17]. In particular, the LGO procedure was utilized in this study.

RESULTS AND DISCUSSION

Linear Model

Results of the GA-PLS Model: 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. The remained descriptors were employed to generate the models with the GA-PLS and GA-KPLS program. The best model is selected on the basis of the highest square correlation coefficient (R^2) and relative error (RE) of prediction and simplicity of the model. These parameters are probably the most popular measure of how well a model fits the data. The best GA-PLS model contains 17 selected descriptors in 9 latent variables space. The R^2 and RE for training and test sets were (0.938, 0.911) and (3.11, 5.37), respectively. For this in general, the number of components (latent variables) is less than number of independent variables in PLS analysis. 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.

Nonlinear Model

Results of the GA-KPLS Model: With the aim of improving the predictive performance of QSRR model, GA-KPLS modeling was performed. 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 [18]. It means that the value of c depends on the system under the study. The 10 descriptors in 6 latent variables space chosen by GA-KPLS feature selection methods were contained. The R^2 and RE for training and test sets were (0.941, 0.919) and (2.61, 4.07), respectively. The statistical parameters R^2 and RE were obtained for proposed models.

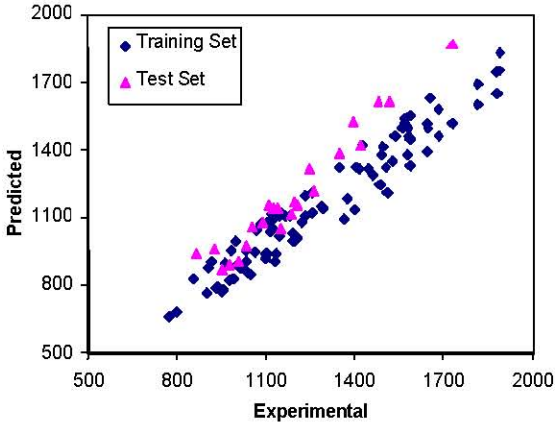

Fig. 1: Plot of predicted RT obtained by GA-KPLS against the experimental values

Each of the statistical parameters mentioned above were used for assessing the statistical significance of the QSRR model. Inspection of the results reveals a higher R^2 and lowers other values parameter for the training and test sets GA-KPLS compared with their counterparts for GA-PLS. The GA-PLS linear model has good statistical quality with low prediction error, while the corresponding errors obtained by the GA-KPLS model are lower. Plots of predicted RT versus experimental RT values by GA-KPLS for training and test set are shown Fig. 1. Obviously, there is a close agreement between the experimental and predicted RT and the data represent a very low scattering around a straight line with respective slope and intercept close to one and zero. This clearly shows the strength of GA-KPLS as a nonlinear feature selection method.

CONCLUSION

In this research, an accurate QSRR model for estimating the retention time of *Artemisia* essential oils compounds was developed by employing the GA-PLS and GA-KPLS techniques. These models have good predictive capacity and excellent statistical parameters. A comparison between these models revealed the superiority of the GA-KPLS to GA-PLS model. It is easy to notice that there was a good prospect for the GA-KPLS application in the QSRR modeling. It can also be used successfully to estimate the RT for new compounds or for other compounds whose experimental values are unknown. This indicates that RT of these compounds possesses some nonlinear characteristics.

REFERENCES

1. Griffin, S.G., G. Wyllie, J.L. Markham and D.N. Leach, 1999. The role of structure and molecular properties of terpenoids in determining their antimicrobial activity. Flav. Fragr. J., 14: 322-332.
2. Rios, J.L., M.C. Recio and A. Villar, 1988. Screening methods for natural products with antimicrobial activity: a review of the literature. J. Ethnopharmacol., 23: 127-149.
3. Lawrence, B.M., 2005. Antimicrobial/Biological Activity of Essential Oils. Allured Publishing Corporation, Illinois, United States.
4. Kershaw, L., 2000. Edible & Medicinal Plants of the Rockies. Lone Pine, Edmonton, Canada.
5. Schmidt, B.M., D.M. Ribnicky, P.E. Lipsky and I. Raskin, 2007. Revisiting the ancient concept of botanical therapeutics. Nat. Chem. Biol., 3: 360-366.
6. Chen, J., T. Yang and S.M. Cramer, 2008. J. Chromatogr. A., 1177: 207.
7. Noorizadeh, H. and A. Farmany, 2010. Chromatographia, 72: 563.
8. Hemmateenejad, B., K. Javadnia and M. Elyasi, 2007. Anal. Chim. Act., 592: 72.
9. Noorizadeh, H., A. Farmany and M. Noorizadeh, 2011. Quim. Nova., 34: 242-249.
10. Aires-de-Sousa, J., M.C. Hemmer and J. Casteiger, 2002. Prediction of H-1 NMR chemical shifts using neural networks, Anal. Chem., 74: 80-90.
11. Riahi, S., E. Pourbasheer, M.R. Ganjali and P. Norouzi, 2009. J. Haz. Mat., 166: 853.
12. Bodzioch, K., A. Durand, R. Kaliszan, T. Bączek and Y.V. Heyden, 2010. Talanta, 81: 1711.
13. Niazi, A., S. Jameh-Bozorghi and D. Nori-Shargh, 2008. J. Hazard. Mater., 151: 603.
14. Woo, S.H., O. Jeon Ch, Y.S. Yun, H. Choi, S. Lee Ch and D.S.J. Lee, 2009. Hazard. Mater., 161: 538.
15. Krämer, N., A.L. Boulesteix and G. Tutz, 2008. Chemom. Intell. Lab. Syst., 94: 60.
16. Lopes-Lutz, D., D.S. Alviano, C.S. Alviano and P.P. Kolodziejczyk, 2008. Phytochemistry, 69: 1732-1738.
17. Noorizadeh, H. and A. Farmany, 2011. Drug Test Anal.
18. Kim, K., J.M. Lee and I.B. Lee, 2005. Chemom. Intell. Lab. Syst., 79: 22.