Informatics aided QSRR study of retention index of some volatile compounds

Abbas Farmany

Dental Research Center, Hamadan University of Medical Sciences, Hamadan, Iran

Abstract

In the present work, an artificial neural network (ANN) model was used to study the quantitative structure retention relationship (QSRR) of retention index (RI) of some volatile compounds in natural cocoa and conched chocolate powder. Molecular structural descriptors are selected using genetic algorithm to construct the nonlinear QSRR models. Besides, kernel partial least squares PLS (KPLS) and Levenberg-Marquardt artificial neural network (L-M ANN) were employed.

Keywords: Volatile compounds; QSRR; Levenberg-Marquardt artificial neural network.

Introduction

Chocolate and cocoa are amongst the most popular flavors worldwide. Unlike many other natural flavors that have one principal compound responsible for their flavor [*e.g.* vanilla (vanillin), banana (amyl acetate), peach (γ-undecalactone) or almond (benzaldehyde)], chocolate and cocoa flavors reside in their volatile fraction, which is composed of a complex mixture of up to 500 compounds [1] with new research continuously increasing this number [2,3]. The complexity of this flavor is obvious and thus it has not been possible for the chemist to duplicate the flavor. It is obvious that the compounds that give cocoa and chocolate their flavors are N and O-containing molecules such as pyrazines, pyrroles, furans, aldehydes and amides, which are mainly generated during the Maillard reaction.

The analysis of volatile fractions of cocoa and chocolate is essential to study their aromatic profile. In general, the volatile fractions have been recovered by steam distillation or solvent extraction and analyzed by GC [4]. Considering the limitations of these sampling techniques, headspace solid-phase micro-extraction (HS-SPME) emerges as an attractive alternative [5]. The technique has been reported to be relatively cheap, solventless, fast and reproducible. SPME seems particularly appealing since it also eliminates problems associated with chemically and thermally unstable samples where generation of artifacts can be problematic as in the case of chocolate and cocoa. SPME, and HS-SPME in particular, and in combination with GC and GC–MS, have been largely used in food, environmental and biomedical analyses [6,7].

*Corresponding author: Abbas Farmany

Tel: +98 (83) 38394677, Fax: +98 (83) 38394677
E-mail: a.farmany@ut.ac.ir

Counet *et al.* [8] used HS-SPME-GC–MS to analyze the flavor of 8 cocoa liquors and tentatively identified 43 compounds, of which 5 had never been associated with chocolate flavor. In each case, 2 g of the cocoa liquor was adsorbed onto a CARPDMS fiber for 20 min at 25 °C and desorbed for 5 min at 250 °C in the GC injector. Pini *et al.* [9] assessed HS-SPME-GC-FID for the analysis of the alkylpyrazine content in single-roasted cocoa liquor. The optimum recovery was obtained using the DVB/CAR-PDMS fiber when the extraction was performed at 60 °C for 45 min, following a 15 min equilibration. In related studies, they had also analyzed the pyrazine content of cocoa mass and cocoa nibs using a CAR-PDMS fiber under similar conditions [10].

Using Chemometrics tools to predict compounds and chemical tissue distribution, membrane permeability or biphasic system partition is of major importance in physicochemical, environmental, and life sciences. Chemical distribution phenomena depend not only on molecular structure but also on the properties of the system in question [12]. Quantitative structure–retention relationship (QSRR) techniques based on different molecular descriptors have been successfully used to model organic chemicals properties [13-19,30-36].

The objective of this study was to investigate the relationship between retention index of some cocoa products obtained by HSSPME-GC–MS and their chemical structures.

Materials and Mathematical Methods

Data set
Retention index (RI) of 42 volatile compounds in cocoa products (natural cocoa powder and conched chocolate powder) which was obtained by solid-phase micro-extraction coupled to gas chromatography–mass spectrometry by the headspace analysis (HS-SPME-GC–MS) was taken from the literature [20] as it is presented in Table 1.

Computer hardware and software
A laptop computer with an AMD Turion64X2 processor and a Windows XP operating system was used. The molecular structures optimizations were done using HyperChem 7.0 and Dragon Version 3.0 software was used to calculate the descriptors. MATLAB (Version 7, Mathworks, Inc.) environment was used to perform LM ANN, GA KPLS and cross validation calculations.

Molecular descriptors
The molecular structure of the compounds was used to proceed the derivation of molecular descriptors (theoretical) using HyperChem version 7.0. Semi-empirical AM1 method was used to obtain the final geometries. The molecular structure optimizations were made using the Fletcher- Reeves algorithm [21].

Genetic algorithm for descriptor selection
Descriptor selection was made by simulating the population [22–24]. Subset of descriptors named as binary values chromosomes were from population. We set the population number as chromosome. At each chromosome, the number of genes was the same as the number of descriptors. First generation population was chosen as random. If the corresponding descriptor of each gene was given as one, otherwise it was zero. In the case that 90% of generations take the same, the evaluation process was ended.

Nonlinear ANN model
Three-layer networks are used as feed-forward artificial neural networks. In

each layer, neurons are connected to the succeeding layer. For artificial neural network training, the Levenberg–Marquardt back propagation algorithm was used. In the output and hidden layers, linear functions were employed.

Table 1. The data set, structure and the corresponding observed RI values

| No	Name	RI
	Calibration Set	
1	Acetone	495
2	Methyl acetate	521
3	2,3-Butanedione	581
4	2-Butanone	586
5	3-Methylbutanal	653
6	2-Methylbutanal	662
7	3-Hydroxy-2-butanone	707
8	35-Dimethyl-dihydro-furan-2-one	766
9	2,3-Butanediol	792
10	Hexanal	802
11	3-Methyl-butanoic acid	837
12	2-Methyl-butanoic acid	847
13	2,5-Dimethylpyrazine	921
14	Ethylpyrazine	926
15	α-Pinene	949
16	β-Pinene	999
17	35-Dimethyl-octane	1007
18	1H-Pyrrole-2-carboxaldehyde	1018
19	Benzyl alcohol	1048
20	3-Ethyl-2,5-dimethylpyrazine	1086
21	Tetramethylpyrazine	1094
22	3,5-Dimethyl-benzaldehyde	1250
23	Vanillin	1419
24	Caffeine	1783
	Prediction Set	
25	2-Methylpropanal	550
26	Pentanal	696
27	Methylpyrazine	832
28	2,3-Dimethylpyrazine	928
29	Trimethylpyrazine	1011
30	2-Acetylpyrrole	1072
31	Phenylethyl alcohol	1133
32	Isopentyl benzoate	1404
	Test Set	
33	Acetic acid	630
34	Dimethylpropanedioic acid	755
35	2,3-Butanediol	782
36	2-Furanmethanol	854
37	Benzaldehyde	981
38	2-Ethyl-6-methylpyrazine	1013
39	Benzeneacetaldehyde	1062
40	Benzoic acid	1157
41	3,5-Dihydroxy-6-methyl-2,3-dihydro-pyran-4-one	1164
42	8-Hydroxy-3-methyl-iso-chroman-1-one	1565

Result and discussion
GA KPLS
In this work, GA KPLS model was used to select the descriptors in the latent variables. For the training and the test, R^2 and RE were (0.894, 0.826) and (8.38, 15.70), respectively. To evaluate the RI prediction ability of the model, some parameters as R^2 as highest square correlation coefficient and RE as relative error were followed. R^2 and RE as main statistical parameters of the model were used to evaluate the statistical justification of the model. Note that KPLS model employs higher number of descriptors which extract better structural data resulted a model with a lower prediction error.

LM ANN
In this work LM ANN model was performed to improve the prediction performance of the QSRR nonlinear model. In this study, 15 descriptors were used as inputs of the GA-KPLS

model to generate the network output as retention index (RI). In the ANN modeling, three groups were separated from dataset as calibration, prediction and test. For each network, a three-layer sigmoid transfer function was constructed to optimize the weights and bias, the back propagation method was employed. For calibration, prediction and test set, the number of R^2 and RE were obtained as (0.976, 0.949, 0.926) and (4.58, 5.49, 8.71), respectively. The predicted and experimental values of RI were obtained by LM ANN method as presented in Fig. 1. As shown in Fig. 1, there is a negligible scattering near to the straight line with respective slope. At all, a good relationship between the predicted and experimental RI was obtained. The results of this study show a significant QSRR improvement to nonlinear one.
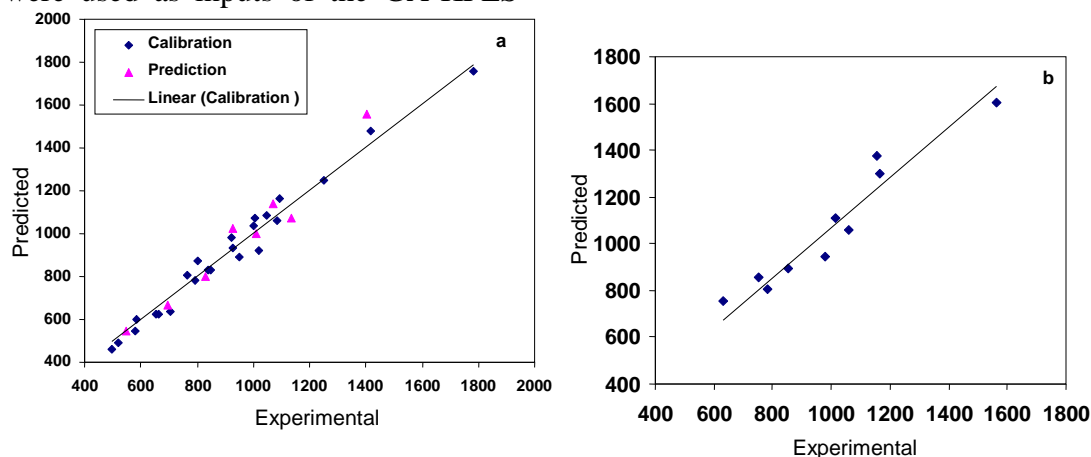
Figure 1. Predicted RI obtained by L-M ANN against the experimental values (a) calibration and prediction set of molecules and (b) for test se

Model validation and statistical parameters
To access the prediction ability of the obtained GA KPLS and LM ANN models, both internal and external validation methods were employed. LGO cross validation method was used as internal validation methods. The

process was applied to all dataset compounds. In this manner, the dataset was divided into some datasets as calibration, training and test sets. Model generation was based on the calibration set. The predictive ability of the generated models was evaluated using test set. Then, the prediction

power of it was checked. In this process, the test set was used as an external data set. We used 24, 8 and 10 compounds as calibration, training and test sets, respectively. During the analysis, the descriptive power of the generated model was checked by probing it's predict partition ability of unknown samples.

Conclusion

In this paper, nonlinear LM ANN and GA KPLS models were used to obtain a quantitative relation between descriptors and RI of some cocoa and chocolate powders. The results of this study show that LM ANN could generate a better model to predict the retention index of the above mentioned compounds.

Acknowledgements

References

[1] P.S. Dimick, J.C. Hoskin, in: S.T. Beckett (Ed.), Industrial Chocolate Manufacture and Use, Blackwell Science, **2003**, pp. 137–152.

[2] C. Counet, D. Callemien, C. Ouwerx, S. Collin, *J. Agric. Food Chem.,* **2002**, *50*, 2385-2391.

[3] F. Luna, D. Crouzillat, L. Cirou, P. Bucheli, *J. Agric. Food Chem.*, **2002**, *50*, 3527-3532.

[4] F. Augusto, A.L.E. Lopes, C.A. Zini, *Trends Anal. Chem*., **2003**, *22*, 160-169.

[5] H. Lord, J. Pawliszyn, *J. Chromatogr. A.,* **2000**, *885*, 153-193.

[6] B. Zygmunt, A. Jastrzebska, J. Namiesnik, *Crit. Rev. Anal. Chem.,* **2001**, *31*, 1-18.

[7] T. Kumazawa, X.P. Lee, K. Sato, O. Suzuki, *Anal. Chim. Acta.,* **2003**, *492*, 249-252.

[8] C. Counet, C. Ouwerx, D. Rosoux, S. Collin, *J. Agric. Food Chem*., **2004**, *52*, 6243-6249.

[9] G.F. Pini, E.S. de Brito, N.H.P. Garcia, A.L.P. Valente, F. Augusto, *J. Brazil. Chem. Soc.,* **2004**, *15*, 267-271.

[10] N.N.E.S de Brito, N.H. Pezoa Garc´ıa, A.L.P.Valente, G.F. Pini, *J. Sci. Food Agric*., **2002**, *82*, 534-538.

[11] A. Zumbe, N. Sanders, UK Patent US6296891, **2001**.

[12] G. Klopman, H. Zhu, *Mini Rev. Med. Chem.,* **2005**, *5,* 127-133.

[13] G. Schuurmann, R.U. Ebert, R. Kuhne, *Environ. Sci. Technol.,* **2006**, *40,* 7005-7011.

[14] A. Talevi, M. Goodarzi, E.V. Ortiz, P.R. Duchowicz, C.L. Bellera, G. Pesce, E.A. Castro, L.E. Bruno-Blanch, *Eur. J. Med. Chem.,* **2011**, *46*, 218-228.

[15] R. Heidarimoghadam, S.S. Mortazavi, A. Farmany, *Iran. Chem. Commun*, **2018**, *4*, 437-449.

[16] G. Carlucci, A.A. D'Archivio, M.A. Maggi, P. Mazzeo, F. Ruggieri, *Anal. Chim. Acta.,* **2007**, *601*, 68–76

[17] S.H. Woo, C.O. Jeon, Y.S. Yun, H. Choi, C.S. Lee, D.S. Lee, *J. Hazard. Mater.*, **2009**, *161*, 538–544

[18] N. Krämer, A. L. Boulesteix, G. Tutz, *Chemom. Intell. Lab. Syst.,* **2008**, *94*, 60–69.

[19] H. Noorizadeh, A. Farmany, *Chromatogr*, **2010**, *72*, 563-569.

[20] S. Ducki, J. Miralles-Garcia, A. Zumbe, A. Tornero, D.M. Storey, *Talanta*., **2008**, *74*, 1166–1174.

[21] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, DRAGON-Software for the calculation of molecular descriptors. Version 3.0 for Windows, **2003**.

[22] E. Konoz, H. Golmohammadi, *Anal. Chim. Acta*., **2008**, *619*, 157–164.

[23] S. Ahmad, M.M. Gromiha, *J. Comput. Chem*., **2003**, *24*, 1313-1320.

[24] H. Noorizadeh, A. Farmany, *Drug Test Anal.*, **2012**, *4*, 151-157.

[25] J. Aires-de-Sousa, M.C. Hemmer, J. Casteiger, *Anal. Chem.*, **2002**, *74*, 80-90.

[26] S. Kara, A.S. Güven, M. Okandan, F. Dirgenali, *Comput. Biol. Med.,* **2006**, *36*, 473–483.

[27] R. Put, Y. Vander Heyden, *Anal. Chim. Acta.,* **2007**, *602*, 164-172.

[28] S. Kara, M. Okandan, *Pattern Recognit.,* **2007**, *40*, 2967 – 2973.

[29] K. Kim, J.M. Lee, I.B. Lee, *Chemom. Intell. Lab. Syst.,* **2005**, *79*, 22-30.

[30] Y. Boukarai, F. Khalil, M. Bouachrine, *Chem. Method*., **2017**, *1*, 173-193.

[31] M. Shahpar, S. Esmaeilpoor, *Asian J. Green Chem*., **2017**, *1*, 116-129.

[32] M. Nekoei, *Iran. Chem. Commun.,* **2017**, *5*, 79-98.

[33] M. Shahpar, S. Esmaeilpoor, *Asian J. Green Chem.*, **2017**, *1*, 116-129. DOI: 10.22631/ajgc.2017.94413.1010

[34] M. Shahpar, S. Esmaeilpoor, *Chem. Method.*, **2017**, *1*, 98-120. DOI: 10.22631/chemm.2017.96397.1008

[35] H. Noorizadeh, S. Esmaeilpoor, Z. Moghadam, S. Nosratolahy, *Iran. Chem. Commun.,* **2014**, *2*, 283-299.

[36] S. Sajjadifar, *Chem. Method.*, **2017**, *1*, 1-11. DOI: 10.22034/chemm.2017.49740

[37] I. Amini, K. Pal, S. Esmaeilpoor, A. Abdelkarim, *Adv. J. Chem. A,* **2018**, *3*, 12-31.

[38] M. Shahpar, S. Esmaeilpoor, *Chemical Methodologies*, **2018**, *2*, 1-22. doi:10.22631/chemm.2017.100307.1012, http://www.chemmethod.com/article_53806.html

[39] M. Shahpar, S. Esmaeilpoor, *Chemical Methodologies*, **2017**, *1*, 98-120. doi:10.22631/chemm.2017.96397.1008, http://www.chemmethod.com/article_50557.html