Talanta 106 (2013) 229–236

Contents lists available at SciVerse ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Prediction of liquid chromatographic retention behavior based on quantum chemical parameters using supervised self organizing maps

Sila Kittiwachana [a,*], Sunanta Wangkarn [a], Kate Grudpan [a,b], Richard G. Brereton [c]

[a] Department of Chemistry, Faculty of Science, Chiang Mai University, Chiang Mai 50200, Thailand
[b] Center of Excellence for Innovation in Analytical Science and Technology, Chiang Mai University, Chiang Mai 50200, Thailand
[c] Centre for Chemometrics, School of Chemistry, University of Bristol, Cantocks Close, Bristol BS8 1TS, UK

ARTICLE INFO

Article history:
Received 15 September 2012
Received in revised form
5 December 2012
Accepted 5 December 2012
Available online 23 December 2012

Keywords:
Chemometrics
Self organizing map
Chromatography
Quantitative structure activity relationship
Non-linear multivariate calibration

ABSTRACT

Self organizing maps (SOMs) in a supervised mode were applied for prediction of liquid chromatographic retention behavior of chemical compounds based on their quantum chemical information. The proposed algorithm was simple and required only a small alteration of the standard SOM algorithm. The application was illustrated by the prediction of the retention indices of bifunctionally substituted *N*-benzylideneanilines (NBA) and the prediction of the retention factors of some pesticides. Although the predictive ability of the supervised SOM could not be significantly greater than that of some previously established neural network methods, such as a radial basis function (RBF) neural network and a back-propagation artificial neural network (ANN), the main advantage of the proposed method was the ability to reveal non-linear structure of the model. The complex relationships between samples could be visualized using U-matrix and the influence of each variable on the predictive model could be investigated using component planes—which can provide chemical insight.

1. Introduction

Prediction of chromatographic behaviors such as retention indices or retention factors from molecular structures of chemical compounds is one of the main interests of quantitative structure-activity relationship (QSAR) analysis. It is possible to relate chemical structure information of a group of chemical compounds to their chromatographic retention behavior and this could be done using some chemometric techniques such as artificial neural networks (ANNs) [1,2], partial least squares (PLS) [3,4] and multiple linear regression (MLR) [1,3]. Some studies reported the use of ANNs such as back-propagation neural network [3] and radial basis function (RBF) neural network [5,6] for predicting of the chromatographic retention. However, the disadvantage of using such learning models is that a complicated model optimization is necessary in order to guarantee a satisfactory solution. Moreover, the learning models generally conduct themselves as a 'black box', in which the chemical data are used as an input for the modelling. After an intensive calculation, the desired information, such as chromatographic retention, is given back from some output units. In most cases, the input data are linked successfully to the output information. However, it is rather difficult to investigate the structure inside the models. For example, what is the relationship between samples and which variables/parameters are most associated with the studied behavior. Calibration methods such as PLS and MLR are generally capable of revealing the model structure, but they may not be able to handle effectively with non-linearity.

A self organizing map (SOM) [7,8] is another ANN-based method that can be used to reduce the dimensionality of a data set by producing a low-dimensional map in which similar data are grouped together. SOMs can be used in an unsupervised mode, where a data set is presented to its system, and the models learn to recognize patterns within the data set [9]. This strategy is similar to PCA-based methods, but SOMs have advantage when the data are non-linear. Identically to the loading in PCA, using the component planes in SOMs can visualize non-linear structure of the data. Also, SOMs can be modified for the use in a supervised mode, where an additional information is given while the map is trained iteratively to assist the data organization [10].

Several reports on the use of SOMs were for data exploration [8,10], classification [11,12], process monitoring [13,14], and calibration [15]. In this paper, we proposed an alternative approach to applying SOM for multivariate calibration. The concept is simple and requires only a small alteration of the SOM algorithm. Initially, a non-linear relationship model between a prediction data (quantum chemical information) and a response data (the experimental retention indices) of training samples is built. The SOM map is trained in a supervised manner where additional component plane containing information of the response is given during the training process. The

[*] Corresponding author. Tel.: +66 87 9166692; fax: +66 53 892277.
E-mail address: silacmu@gmail.com (S. Kittiwachana).

0039-9140/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.talanta.2012.12.005

prediction of an unknown sample can be achieved by identifying its best matching unit (BMU), which will be referred to as the predictive result, using the additional response plane. The benefits of the proposed method are that the model does not assume that the data follow multivariate normal distribution and it is possible to visualize the influence of variables on the SOM model, which can provide chemical insight, using component planes. Also, the complex relationships between training samples can be revealed using the U-matrix.

In this paper, the use of supervised SOMs for prediction of the liquid chromatographic retention based on their quantum chemical information is described. The proposed method was applied to two data sets to demonstrate its application. The first data set consisted of the retention indices of 70 bifunctionally substituted *N*-benzylideneanilines (NBA) and the second data set contained retention factors of 26 commonly used pesticide compounds.

2. Experimental data

2.1. Data set 1: liquid chromatographic retention indices of bifunctionally substituted N-benzylideneanilines (NBA)

The proposed SOM algorithms were applied to predict the liquid chromatographic retention indices of bifunctionally substituted *N*-benzylideneanilines (NBA) based on their quantum chemical information. The quantum information of the compounds and the liquid chromatographic retention were extracted from the literature [6]. Here, the general formula of the substituted NBA are expressed as X-C_6H_4–CH$=$N–C_6H_4-Y, where X and Y indicate the substitutions on positions 3 or 4 of the left and right aromatic rings, respectively. The quantum chemical parameters involved in the modelling are dipole moment (μ), energies of the highest occupied and lowest unoccupied molecular orbitals (E_{homo}, E_{lomo}), net charge of the most negative atom (Q_{min}), sum of absolute values of the charge of all atoms in two given functional groups (Δ), total energy of the molecule (E_T), weight of the molecule (W) and numerical coding (N). The retention indices used in this work were obtained using normal-phase liquid chromatography with amino-bonded stationary phase and the eluents composed of solvent mixtures based on heptane and tetrahydrofuran. A total of 70 compounds used in this study are shown in Table 1. The compounds were categorized into 11 groups according to their left substitutions. Notably, it is possible to group the compounds according to their right substitutions; however, for brevity, it is not discussed here. These data were divided optionally into three sets: a training set, a validation set, and a test set, using the criterion reported in the source of data [6]. Each of the sub-data set consists of two sets of variables: 'predictive' variables or '***X***' block, which in this case are the quantum chemical parameters; and 'response' variable or '***c***' block, which is the experimental retention indices.

2.2. Data set 2: retention factors of pesticides

The second data set contained retention factors of 26 compounds (as shown in Table 2) which are commonly used insecticides, herbicides and fungicides and some metabolites. The compounds were categorized into 10 groups based on their chemical structures. This data was extracted from the literature [16]. Four quantum chemical descriptors, dipole moment (μ), mean polarizability (α), the anisotropy of polarizability (β^2) and a descriptor of hydrogen bonding ability based on the atomic charges on hydrogen bond donor and acceptor chemical functionalities (q_{HB}), were used. In addition to the effects of the solute molecular structures on the retention factors, the water-octanol partition coefficient (K_{wo}) and the organic modifier concentration

Table 1
The studied compounds and their labelled groups for data set 1.

No.	Compounds X–Y	Groups	No.	Compounds X–Y	Groups
1	H–4Br	1	36	4F–3Cl	6
2	H–H	1	37[b]	4F–4CN	6
3[a]	H–4F	1	38	4F–3NO_2	6
4[b]	H–4Cl	1	39[a]	4F–4NO_2	6
5	H–3Cl	1	40	4Cl–4CH_3	7
6	H–3CN	1	41	4Cl–H	7
7	H–4CN	1	42	4Cl–4F	7
8	H–3NO_2	1	43	4Cl–4Cl	7
9	H–4NO_2	1	44	4Cl–3Cl	7
10[b]	4NO_2–4OCH_3	2	45	4Cl–3NO_2	7
11[a]	4NO_2–4CH_3	2	46[b]	4Cl–4NO_2	7
12	4NO_2–3CH_3	2	47	3Cl–3OCH_3	7
13	4NO_2–3F	2	48[a]	3Cl–H	7
14	4NO_2–4F	2	49	3Cl–4F	7
15[a]	4NO_2–4Cl	2	50	3Cl–4Cl	7
16	4NO_2–4Br	2	51	3Cl–3Cl	7
17	4NO_2–3Cl	2	52	3Cl–4CN	7
18	4OCH_3–H	3	53	4CF_3–4OCH_3	8
19	4OCH_3–4F	3	54	4CF_3–4CH_3	8
20[b]	4OCH_3–4Cl	3	55	4CF_3–H	8
21	4OCH_3–3Cl	3	56	4CF_3–4F	8
22	4OCH_3–3NO_2	3	57[b]	4CF_3–4Cl	8
23	4CH_3–H	4	58	4CF_3–4CN	8
24	4CH_3–4F	4	59	4CF_3–4NO_2	8
25[b]	4CH_3–4Cl	4	60[b]	4CN–4CH_3	9
26	4CH_3–3Cl	4	61	4CN–H	9
27[a]	4CH_3–4CN	4	62	4CN–4F	9
28	4CH_3–3NO_2	4	63	3NO_2–4OCH_3	10
29	4CH_3–4NO_2	4	64	3NO_2–4CH_3	10
30	3OCH_3–H	5	65[a]	3NO_2–3CH_3	10
31	3OCH_3–4F	5	66	3NO_2–H	10
32	4F–4CH_3	6	67[b]	3NO_2–4F	10
33	4F–H	6	68	3NO_2–4Cl	10
34	4F–4F	6	69	3NO_2–3Cl	10
35[a]	4F–4Cl	6	70[a]	4Br–H	11

[a] The compounds are in validation set.
[b] The compounds are in test set. The rests are in training set.

Table 2
The studies compounds and their labelled chemical groups for data set 2.

Compound no.	Compound names	Chemical groups[a]
1	Atrazine	1
2	Desethylatrazine	1
3	Desisopropylatrazine	1
4	Simazine	1
5	Terbutylazine	1
6	Desethylterbutylazine	1
7	Aldicarb	2
8	Carbaryl	2
9	Carbofuran	2
10	Phenmedipham	2
11	2,4-Dichlorophenol	3
12	Fenitrothion	4
13	Malathion	4
14	Linuron	5
15	Iprodione	6
16	Procymidone	6
17	Vinclozolin	6
18	3,4-Dichloroaniline	7
19	3,5-Dichloroaniline	7
20	Dicloran	7
21	Chloridazon	8
22	Metalaxil	9
23	Oxadixil	9
24	Alachlor	10
25	Metolachlor	10
26	Metazachlor	10

[a] Chemical groups: (1) triazine, (2) carbamate, (3) phenol, (4) organophosphate, (5) phenylurea, (6) dicarboximide, (7) aniline, (8) pyridazinone, (9) acylalanine and (10) chloroacetanilide.

which is the v/v percentage of acetonitrile (% ACN) in the mobile phase were also used as predictive variables. Therefore, there are in total six predictive variables in the $\boldsymbol{X}$ block for this data set. The % ACN in the mobile phase was varied from 40% to 65% by steps of 5% resulting in 156 data points which were split into 120, 18 and 18 samples for a training set, a validation set and a test set, respectively, using the same criterion reported in the source of data [16]. For both of the data sets, the quantum chemical parameters and the retention were exported to Matlab version 7 (Mathworks, Natick, MA, USA) for further analysis and all software was written in Matlab.

3. Methods

3.1. Self organizing maps (SOMs)

The SOM algorithm has several stages and parameters that need to be set and these were well described in detail in the literatures [8,17]. Therefore, only essential steps are described here.

The first step is initialization. In this step, a trained map consisting of a grid of units is generated. The shape of the units is not specific, although squares and hexagons are particularly favorable because they have neighbors that have the same distance apart in numerous directions. In this work, a trained map consisted of a total of K ($=P\times Q$) map units are represented as hexagons (Fig. 1), where P and Q are the number of rows and columns of the map, respectively. Each map unit k is characterized by a weight for each variable, resulting in a $1\times J$ weight vector $\boldsymbol{w}_k$, where J corresponds to the number of variables. Therefore, each variable j has K weight units, which, for this work, were generated from randomly selected values from a uniform distribution within the measured range of variable j.

The next step is the training process. In this step, a sample vector, which is randomly selected from the training samples and is newly generated for each iteration, is compared to each of the map unit weight vectors. The dissimilarity (s) between a randomly selected sample $\boldsymbol{x}_z$ and a weight vector $\boldsymbol{w}_k$ for map unit k is given by

$$s_{(\boldsymbol{x}_z,\boldsymbol{w}_k)} = \sqrt{\sum_{j=1}^{J}\left(x_{zj}-w_{kj}\right)^2}$$

Here, the map unit with the most similar weight (having the lowest dissimilarity) vector is declared as the 'winner' or the best matching unit (BMU) and becomes the center of learning for that iteration. Once, the BMU for the randomly selected sample $\boldsymbol{x}_z$ can be identified, this BMU and its neighboring map units are then updated to become more like the selected sample. The amount by which the units can 'learn' to represent the input sample is controlled by the learning rate and the neighborhood width [8]. As the learning proceeds, the samples gradually move towards a region of the map that they are most similar to, and so samples that are close together in the high-dimensional input space are mapped onto the SOM units that are close together in the map space. This entire process is repeated for $i=1, 2,\ldots, I$, where I is the number of training iterations.

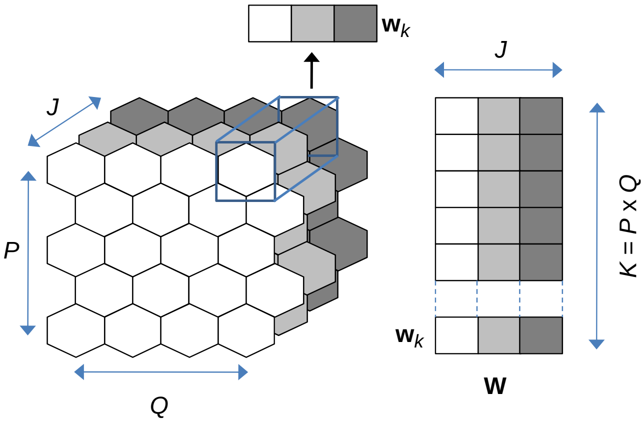


Fig. 1. $P\times Q$ map with J weights containing a total of K map units and the corresponding weight matrix $\boldsymbol{W}$.

For this work, the SOM maps were trained for 10,000 iterations, with an initial learning rate of 0.1 and an initial neighborhood width of half of the smallest dimension of the map [8]. If the map space is two-dimensional, there are a number of ways of visualizing the relationship between samples, such as the U-matrix [18] and hit histograms [19]. It is also possible to display component planes of the trained map for each variable separately, which is somewhat analogous to the loadings in PCA. These component planes can be used to investigate how each variable influences the map and which samples a variable is the most associated with.

3.2. SOMs for quantitative prediction

SOMs can be categorized into two different methods according to how the models are trained: unsupervised and supervised SOMs. For the unsupervised SOMs, only the information about the predictors (or measurements such as chromatographic peaks, quantum information etc.) is used; whereas, for the supervised SOMs, the information of the responses (in this case the retention indices for the data set 1 and the retention factors for the data set 2) is also included during the training process. The detail of how to construct an unsupervised SOM has been described in the previous section. The following section describes the supervised SOM, which will be applied for the prediction of the retention behavior of chemical compounds based on their quantum chemical information.

As mentioned previously, for unsupervised SOMs, the map is trained using only $\boldsymbol{X}$; the response $\boldsymbol{c}$ is not used. However, for supervised SOMs, the measurements $\boldsymbol{X}$ are used together with the response vector $\boldsymbol{c}$ in the training process. This can be done by adding an additional response vector to the predictor matrix resulting in a matrix $\boldsymbol{X}_s$ that is augmented by the additional column of the response vector. After that, the weight matrix can be trained in the same manner as for an unsupervised SOM. Here, it is recommended that $\boldsymbol{X}_s$ is standardised prior to the supervised training process to ensure that all variables, including the additional response vector, have equal influence on the model [20]. After the training process, the Euclidean distance [13] or the dissimilarity between the unknown sample $\boldsymbol{x}_i$ and the weight vector $\boldsymbol{w}_k$ can be calculated by

$$s_{(\boldsymbol{x}_i,\boldsymbol{w}_k)} = \sqrt{\sum_{j=1}^{J_s-1}\left(x_{ij}-w_{kj}\right)^2}$$

where J_s is the number of columns in the training matrix $\boldsymbol{W}$ for the supervised SOM, which is equal to the number of the supervised component planes ($J_s=J+1$). After that, the prediction for an unknown sample $\boldsymbol{x}_i$ can be determined by identifying the BMU on the trained map:

$$s_{(\boldsymbol{x}_i,\boldsymbol{w}_b)} = \min_k\left\{s_{(\boldsymbol{x}_i,\boldsymbol{w}_k)}\right\}$$

where b is the value of k with the most similar weight vector $\boldsymbol{w}_k$ to the unknown sample $\boldsymbol{x}_i$

$$b = \arg\min_k\left\{s_{(\boldsymbol{x}_i,\boldsymbol{w}_k)}\right\}$$

therefore, the prediction of the unknown sample $\boldsymbol{x}_i$ is

$$\hat{c}_i = w_{bJ_s}$$

where $\hat{c}_i$ is a predictive value for the unknown sample $\boldsymbol{x}_i$ and w_{bj_s} is a weight value in the component plane for the response variable. The diagram for the supervised SOM for prediction of liquid chromatographic retention is shown in Fig. 2.

Prior to the analysis, the training data were standardised as described previously and the test samples were pre-processed in the same way as the training set, where the means and the standard deviations of the training samples were used.

Since the initial weight vector is generated randomly and the input order of samples for the training process can be different, the trained map could be different each time it is generated. However, the sample organization in the trained maps should not be so different, and therefore, the variation in the predictive result is not significant. However, to cope with this variation, 10 supervised SOM models were generated and the results shown in this report were the mean values of these 10 prediction values.

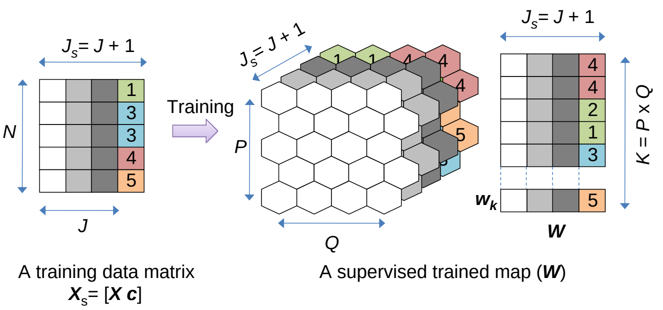


Fig. 2. Supervised SOM training from N samples and J variables using $P \times Q$ trained map resulting in weight matrix $\boldsymbol{W}$ with $J_s=J+1$ component planes where the last component plane corresponds to the response vector.

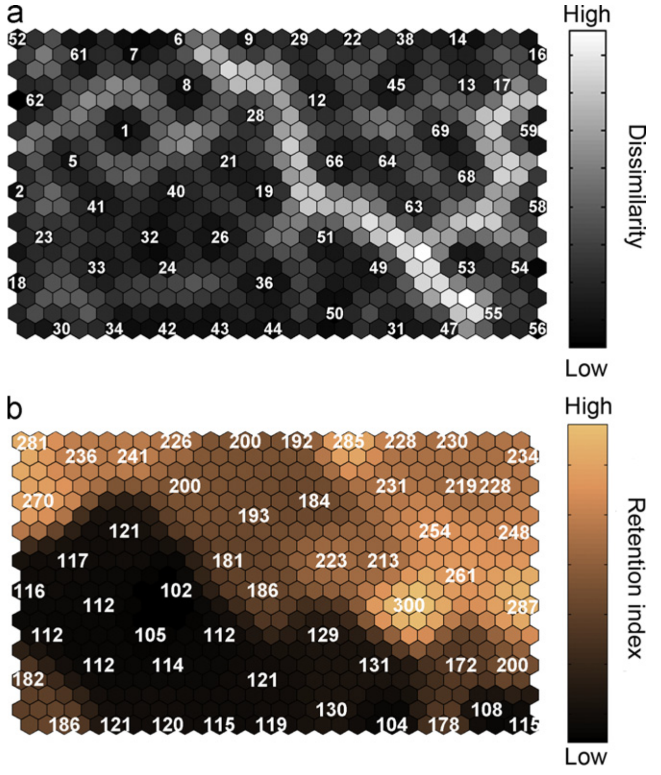


Fig. 3. (a) U-matrix visualization for the trained map of the training samples for data set 1 where the numbers represent the BMU of each sample and (b) The response plane, where the numbers represent the predictive retention values of the BMU of each training sample.

4. Results and discussion

4.1. Data set 1—bifunctionally substituted N-benzylideneanilines (NBA)

4.1.1. Interpretation of U-matrix

SOM of size 20 × 30 units was trained using the pre-processed data of the quantum chemical information. In this study, the trained map is visualized using a U-matrix; the similarity between a unit and its neighbors is represented. A low value in the U-matrix implies similar neighbors (e.g. in the middle of a cluster of map units), whereas a high value represents parts of the map where there are dissimilar neighbors (e.g. on the boundaries between different clusters or between an outlier and the main sample cluster). This visualization is shown in Fig. 3(a). The numbers labelled in the map represent the BMU of each sample. Notably, the use of higher resolution can describe the data in more detail, as there are more interpolation units (units that are not the BMU of any training sample and represent the transition between adjacent units), but it takes longer for training. In this

Table 3
Relative errors (%) between experimental and predicted retention indices of the test samples for data set 1 using RBF neural network and supervised SOM.

Sample no.	Retention indices			Relative errors (%)	
	Experimental	Predicted			
		RBF	SOM	RBF	SOM
5	117.3	114.6	116.8	2.30	0.43
11	287.2	251.1	272.9	12.57	4.98
21	186.5	174.4	183.0	6.49	1.88
26	110.2	110.6	106.0	−0.36	3.81
38	271.9	260.8	263.7	4.08	3.02
47	225.7	227.2	228.2	−0.66	−1.11
58	114.3	123.1	128.6	−7.70	−12.51
61	225.8	181.4	230.9	19.66	−2.26
67	255.7	255.9	261.6	−0.08	−2.31

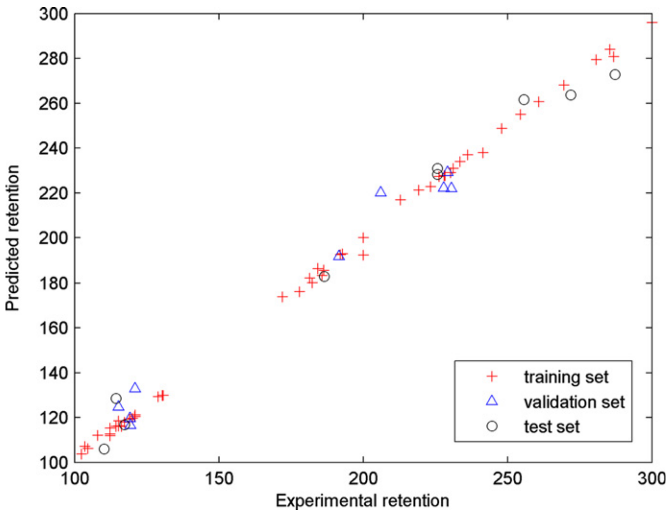


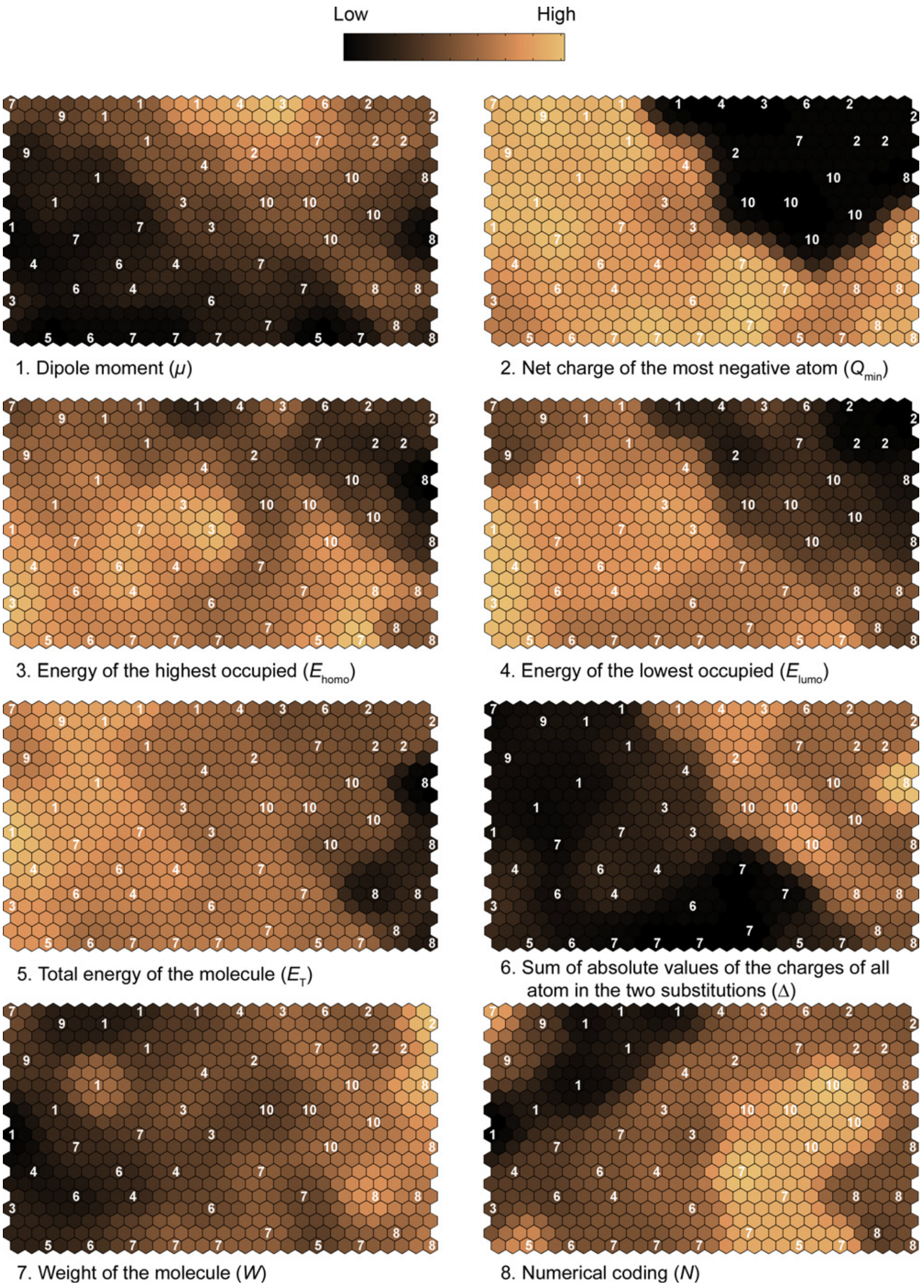
Fig. 4. Correlation plot between experimental and predicted retention indices using 20 × 30 supervised SOM model for data set 1.

work, the 20 × 30 trained map is chosen as a compromise between the training time and the map resolution.

From the U-matrix visualization, it can be seen that the training samples can be categorized into 3 clusters. The first cluster (cluster of the compounds in group no. 8 in Table 1) on the below corner of the right side of the map is the compounds that contain $–CF_3$ substitution on the left aromatic ring. The second cluster is the cluster of compounds no. 9, 12, 13, 14, 16, 17, 22, 29, 38, 45, 63, 64, 66, 68 and 69. It is possible that these samples are classified into this region because they have $–NO_2$ substitutions either on the left or the right of the aromatic ring. The rest of the samples are classified into the last cluster, which is the biggest one where most of the compounds in this cluster have neither $–CF_3$ nor $–NO_2$ substitutions except samples no. 8 and 28, which contain the $–NO_2$ substitution; however, they are located close to the boundary of the $–NO_2$ substitution region.

4.1.2. Prediction of retention time using supervised SOMs

Fig. 3(b) illustrates the 20 × 30 component plane of the response variable or, in this case, the retention index. This component plane is used as a predictor of the model. The prediction of an unknown sample can be achieved by identifying its best matching unit (BMU) using only the component planes corresponding to the prediction variables and then extracting the predictive value from the corresponding unit in a component plane of the response values. Based on the supervised SOM models, the predictive results of the test samples are shown in



1. Dipole moment (μ)

2. Net charge of the most negative atom (Q_{min})

3. Energy of the highest occupied (E_{homo})

4. Energy of the lowest occupied (E_{lumo})

5. Total energy of the molecule (E_T)

6. Sum of absolute values of the charges of all atom in the two substitutions (Δ)

7. Weight of the molecule (W)

8. Numerical coding (N)

Fig. 5. 20 × 30 map unit component planes of each parameter (quantum chemical information) for data set 1 where the BMU for each of the training samples is labelled, corresponding to the 20 × 30 map of Fig. 3.

Table 3. The results are also compared with the predicted retention indices using radial basis function (RBF) neural network [6]. A correlation plot between the predicted and the experimental data is shown in Fig. 4. The correlation coefficients (R) of the training, the validation and the test sets are 0.9989, 0.9798 and 0.9872; and the corresponding mean-absolute errors are 0.18, 1.53 and 1.47, respectively.

4.1.3. *Interpretation of component planes*

Besides the U-matrix discussed in the previous section, it is possible to visualize the trained map using component planes of each quantum chemical parameter. This is to evaluate how each parameter influences the map, and which samples a parameter is most associated with. The component planes of all of the quantum chemical parameters are shown in Fig. 5. Here, the BMU for each of the training samples is labelled, corresponding to the 20 × 30 map of Fig. 3(a). From the component planes, it can be seen that parameter no. 2, which is net charge of the most negative atom ($Q_{\min}$), is corresponded most to the samples in clusters 1 and 3 since this component plane is heavily weighted towards these samples. Parameter no. 6, which is the sum of absolute values of the charge of all atoms in two given functional groups (Δ), seems to correspond the most to the samples in clusters 1 and 2 for the same reason. However, it is not clear to conclude that which parameters are the most predictive parameters for this supervised SOM model as none of the pattern of the component planes of any quantum information is similar to that of the response plane, although a weak trend can be observed from the component plane for parameter no. 1. This could be due to the fact that the data were non-linear. The chemical compounds were different in terms of their structural complexities and each of the chemical parameters was generated from different chemical functionality. Therefore, there is no simple linear correlation between the retention indices of the studied compounds and the input parameters. If the data were assumed linear, the parameter functions would be monotonically increasing or decreasing, and so would the response values. Consequently, the intensities of the units in the component plan would have been higher for the larger retention values. However, in this case, when all of the parameters were used simultaneously for the modelling, the model could provide satisfactory predictive results. This has proved to be a success even when the model deals with non-linear situation.

4.2. *Data set 2—pesticides*

4.2.1. *Prediction of retention factors using supervised SOMs*

Based on the supervised SOM model trained using the selected training samples of the data set 2, the predictive retention factors of the test samples are shown in Table 4. For this data set, the predictive results were compared with those using back-propagation ANN and MLR [16]. A correlation plot between the predicted and the experimental retention factors is shown in Fig. 6. The correlation coefficients (R) of the training, the validation and the test sets are 0.9361, 0.9989 and 0.9606, respectively; and the corresponding mean-absolute errors are 0.79, 0.16 and 1.02. Compared with the mean-absolute errors of the test set using ANN and MLR which are 0.50 and 1.06, respectively, it appeared that, for this data set, the supervised SOM could not defeat the reference methods in term of its predictive ability. The mean-absolute error of the test samples using the supervised SOM model is slightly higher than that of the ANN and more or less the same as that of the MLR. The inferior predictive performance of the supervised SOM could be due to the fact that the

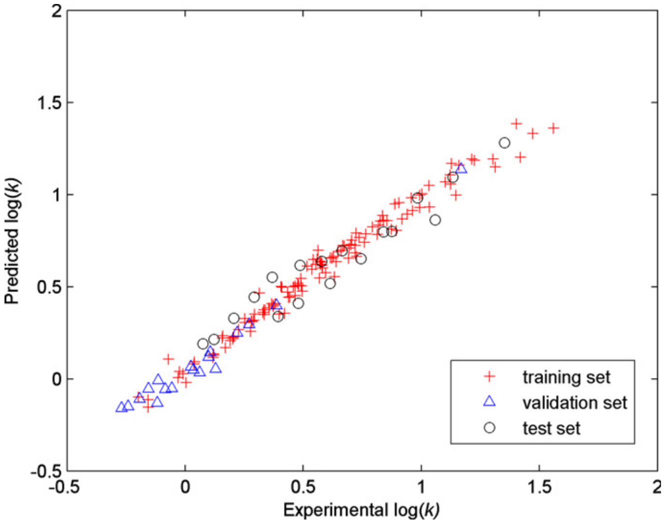


Fig. 6. Correlation plot between experimental and predicted retention factors for data set 2.

Table 4
Relative errors (%) between experimental and predicted retention factors of the test samples for data set 2 using ANN, MLR and supervised SOM.

Pesticides	% ACN	Retention factors (k)				Relative errors (%)		
		Experimental	ANN	MLR	SOM	ANN	MLR	SOM
Alachlor	65	3.79	4.23	4.60	4.33	11.6	21.4	14.3
	60	4.62	5.33	6.08	4.95	15.3	31.6	7.2
	55	6.90	7.08	8.04	6.29	2.7	16.4	−8.9
	50	9.65	9.99	10.62	9.60	3.6	10.1	−0.5
	45	13.62	15.07	14.04	12.40	10.6	3.1	−8.9
	40	22.51	24.48	18.56	19.11	8.7	−17.5	−15.1
3,5-Dichloroaniline	65	2.48	2.34	2.03	2.18	−5.2	−18.1	−12.0
	60	3.01	2.95	2.68	2.57	−2.2	11.0	−14.6
	55	4.10	3.79	3.54	3.29	−7.6	−13.6	−19.7
	50	5.54	4.99	4.68	4.48	−9.9	−15.5	−19.1
	45	7.52	6.83	6.19	6.33	−9.2	−17.6	−15.8
	40	11.43	9.82	8.18	7.29	−14.1	−28.4	−36.2
Simazine	65	1.19	1.04	1.10	1.55	−12.5	7.0	30.5
	60	1.33	1.24	1.46	1.64	−6.2	10.1	23.4
	55	1.61	1.52	1.93	2.13	−5.2	20.2	32.6
	50	1.96	1.90	2.55	2.79	−2.8	30.4	42.3
	45	2.33	2.43	3.37	3.57	4.2	44.7	52.9
	40	3.07	3.21	4.46	4.14	4.6	45.2	34.7

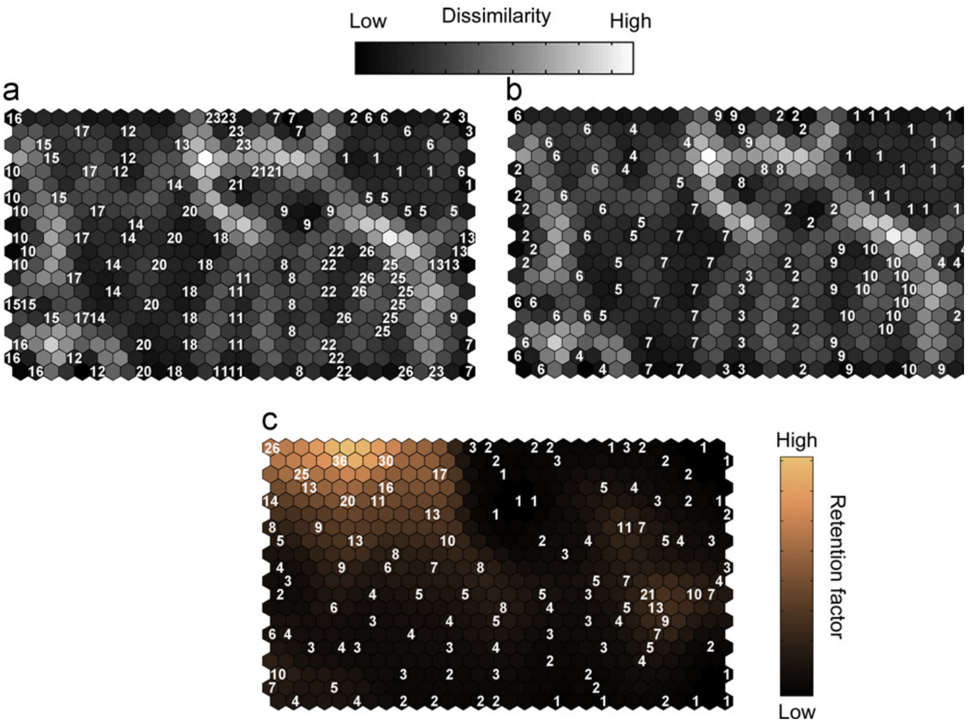


Fig. 7. U-matrix visualizations for the trained map of the training samples of data set 2 where (a) the numbers represent the BMU of each compound name and (b) the numbers represent of the BMU of each the chemical class listed in Table 2 and (c) The response plane, where the numbers represent the predictive retention factors of the BMU of each training sample.

data was originally based on the chemical parameters of the 23 training pesticides. The number of the training samples was increased to 120 samples as a result of the varying in the mobile phase composition (% ACN). If the parameter of the mobile phase composition was not included for the modelling, the SOM may have trained to recognize only the differences among the 23 training pesticide compounds.

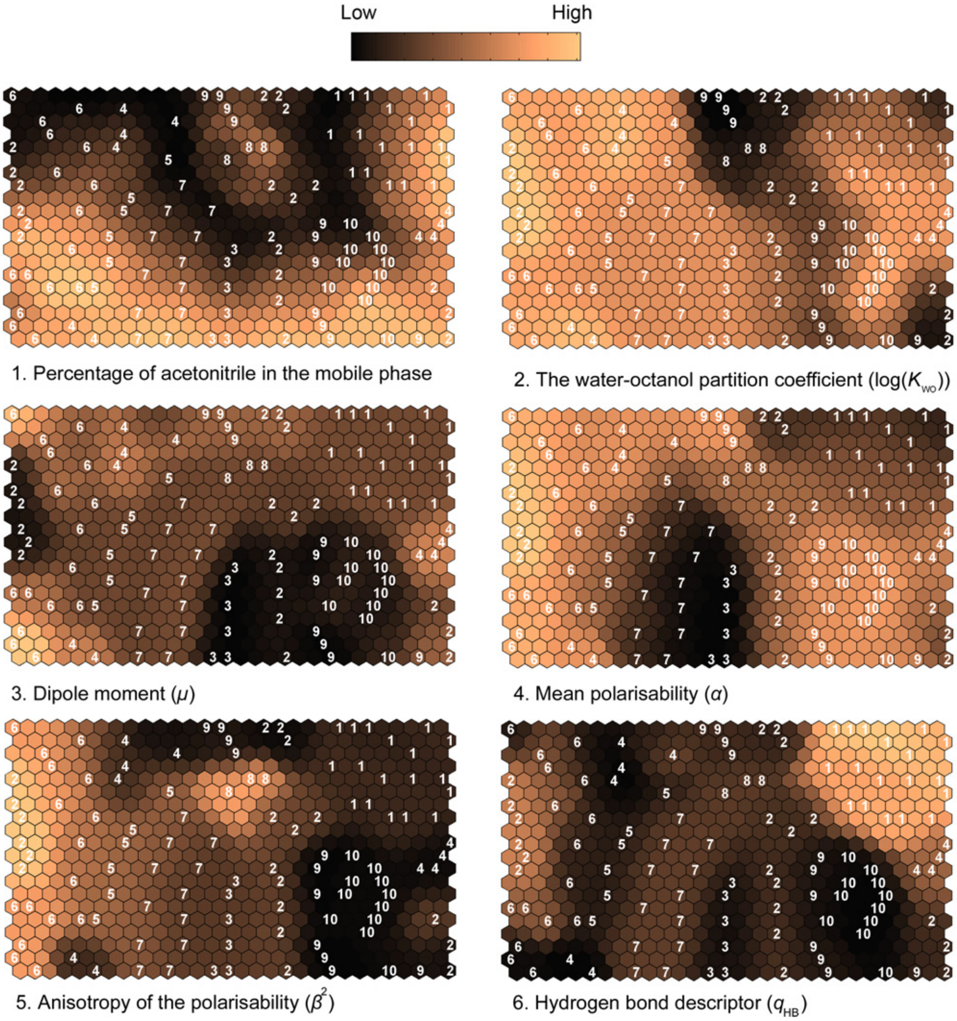
4.2.2. Interpretation of U-matrix and component planes

From the U-matrix visualization of the data set 2 shown in Fig. 7(a) where the numbers represent the compound numbers of training samples matched to the map units, it can be seen that mostly the samples which are derived from the same chemical compounds were organized into the same region. This is as expected because these training samples inherited the same chemical parameters and the variation within the cluster should be mainly because of the varying in the mobile phase composition. At the same time, when the same U-matrix was labelled using their chemical groups (Fig. 7(b)), the map revealed that the training compounds also could be organized in the same way as when they were labelled according to their chemical compounds. For example, on the top right corner, these low areas are positioned where the BMUs of the samples fall from the five different compounds from compound group 1 (triazine) and thus represent approximately where the compounds of this chemical group can be found. Although the distinction between some chemical groups is less clear, for example, between chemical groups 5 and 6, the distinctive trends in the sample organization of the compounds for each of the chemical groups still can be observed.

The component plane visualization shows where variables have been mapped onto the SOM, and labelling allows us to see which samples are best associated with that variable. In Fig. 8, parameter no. 6, which is the parameter of hydrogen bonding ability (q_{HB}), shows that the compounds of the chemical group 1 are all mapping to the region of the map where the units have large weights for this parameter. In fact, it was not expected that each of the chemical groups should be identified using component plants as the supervised SOM was established as a calibration model. Parameter no. 1 is the component plane responsible for the composition of acetonitrile in the mobile phase. This component plane could not clearly indicate which chemical groups the parameter most represented to. Nevertheless, it somehow showed the translation in the values of the map unit weights within the chemical group. This implies that this parameter could be quit important for the calibration model since it expressed the distinction among the pesticide compounds from the same chemical group. Although the pattern of this component plane was not really similar to that of the response plane as shown in Fig. 7(c), this indicated that several parameters were required to distinguish between the chemical samples as well as to establish a good calibration model as would be expected for multivariate data.

5. Conclusion

In this report, an extension of supervised SOM was applied to predict the retention indices of substituted *N*-benzylideneanilines and the retention factors of the studied pesticide compounds based on their quantum chemical information. The goal was to build a non-linear relationship model between a predictor (the quantum chemical information) and a response (the retention behavior of the training samples) using the supervised SOM model. The prediction of an unknown sample could be achieved by referring to the predictive values on the additional response plane, which was identified by the BMU of the unknown sample on the trained map. The supervised SOM could give satisfactory results. Although, in some cases, the predictive results using the supervised SOM was not significantly improved from some of the previous works using the established ANN methods, the supervised SOM has



1. Percentage of acetonitrile in the mobile phase

2. The water-octanol partition coefficient ($\log(K_{wo})$)

3. Dipole moment (μ)

4. Mean polarisability (α)

5. Anisotropy of the polarisability (β^2)

6. Hydrogen bond descriptor (q_{HB})

Fig. 8. 20 × 30 map unit component planes for the parameters used for the data set 2 where the BMU for each of the training samples is labelled, corresponding to the 20 × 30 map of Fig. 7(b).

advantage in that non-linear structure of the model can be visualized using U-matrix and component planes.

Acknowledgments

This work was supported by the Faculty of Science, Chiang Mai University, Chiang Mai, Thailand.

References

[1] W. Guo, Y. Lu, X.M. Zheng, Talanta 51 (2000) 479–488.
[2] B. Skrbic, A. Onjia, J. Chromatogr. A 1108 (2006) 279–284.
[3] V.K. Gupta, H. Khani, B. Ahmadi-Roudi, S. Mirakhorli, E. Fereyduni, S. Agarwal, Talanta, 83 (2011) 1014–1022.
[4] K. Bodzioch, A. Durand, R. Kaliszan, T. Baczek, Y. Vander Heyden, Talanta, 81 (2010) 1711–1718.
[5] X. Yao, X. Zhang, R. Zhang, M. Liu, Z. Hu, B. Fan, Talanta 57 (2002) 297–306.
[6] Y.H. Xiang, M.C. Liu, X.Y. Zhang, R.S. Zhang, Z.D. Hu, B.T. Fan, J.P. Doucet, A. Panaye, J. Chem. Inf. Comput. Sci. 42 (2002) 592–597.
[7] T. Kohonen, Biol. Cybern. 43 (1982) 59–69.
[8] G.R. Lloyd, R.G. Brereton, J.C. Duncan, Analyst 133 (2008) 1046–1059.
[9] K.M. Wolter, Introduction to Variance Estimation, Springer, New York, 2007.
[10] K. Wongravee, G.R. Lloyd, C.J. Silwood, M. Grootveld, R.G. Brereton, Anal. Chem. 82 (2010) 628–638.
[11] G.R. Lloyd, K. Wongravee, C.J.L. Silwood, M. Grootveld, R.G. Brereton, Chemom. Intell. Lab. Syst. 98 (2009) 149–161.
[12] F. Marini, J. Zupan, A.L. Magri, Anal. Chim. Acta 544 (2005) 306–314.
[13] S. Kittiwachana, D.L.S. Ferreira, L.A. Fido, D.R. Thompson, R.E.A. Escott, R.G. Brereton, Anal. Chem. 82 (2010) 5972–5982.
[14] I. Diaz, M. Dominguez, A.A. Cuadrado, J.J. Fuertes, Expert Syst. Appl. 34 (2008) 2953–2965.
[15] F. Marini, A.L. Magri, R. Bucci, A.D. Magri, Anal. Chim. Acta 599 (2007) 232–240.
[16] M. Aschi, A.A. D'Archivio, M.A. Maggi, P. Mazzeo, F. Ruggieri, Anal. Chim. Acta 582 (2007) 235–242.
[17] T. Kohonen, Self-Organizing Maps, Springer, New York, 2001.
[18] A. Ultsch, H.P. Siemon, Kohonen Self-Organizing Feature Maps for Exploratory Data-Analysis, Kluwer Academic Publ., Dordrecht, 1990.
[19] J. Vesanto, Intell. Data Anal. 3 (1999) 111–126.
[20] S. Kittiwachana, D.L.S. Ferreira, L.A. Fido, D.R. Thompson, R.E.A. Escott, R.G. Brereton, J. Chromatogr. A 1213 (2008) 130–144.