Asian Journal of Chemistry Vol. 18, No. 4 (2006), 2623–2636

Artificial Neural Network and Topological Indices to Predict Retention Indices in Gas Chromatography

S. BATOUCHE, N. REBBANI and A. GHEID*
Departement de Chimie, Laboratoire de Chimie Physique et des Matériaux Inorganiques
Universite Badji Mokhtar , B.P. No. 12, Annaba, Algeria
E-mail: hakgheid@yahoo.fr

A comparative study was undertaken to test the ability of different methods to predict the retention indices of a series of acrylates using statistical treatment as criteria of fit. In this paper, a three-layer back-propagation neural network was applied to analyze the QSAR of acrylates in gas chromatography on five different stationary phases. Nine topological indices, Wiener, Balaban, Harary, Shultz, Zagreb, and Randic of first, second, third and fourth order were calculated using a computer program in comparison with the multilinear regression and stepwise methods. The results showed that the ANN model outperformed the MLR predictions. The training phase of the ANN model was extremely short owing to the high performance of the Levenberg-Marquardt algorithm.

Key Words: **Gas chromatography, Retention indices, Topological indices, Multiple regression, Stepwise, Neural network.**

INTRODUCTION

Gas chromatography is a very important analytical tool for simple and complex compounds. However, for complex systems, it can be difficult to find optimum conditions for relatively speedy separations with a satisfactory resolution. Most of the time, these conditions are obtained by means of experimental trials. It is therefore obvious that this approach is not only tedious but can be costly too.

One of the most important parameters in gas chromatography is the retention index. The latter is a useful tool for the comparison of retention data obtained by various authors in different conditions, as it is nearly independant on many of the parameters and conditions of the gas chromatography analysis[1]. It depends on temperature, the stationary phase and the solute. Hence for isothermal operations and for a fixed stationary phase, the retention index becomes only a function of the structure of the solute, in other words, on the topological indices of the solute. The topological indices reflect the molecular shape, branching and composition. Topological indices can be advantageously exploited to predict certain physicochemical properties[2–10]. In recent years, there has been an increased interest in methods that can predict the retention index by means of models based on the

quantitative structure-activity relationships (QSAR)[11–13]. There are advantages in this approach: there are no costly experiments involved and results are obtained faster. For example, Sutter *et al.*[14] calculated six molecular descriptors of 150 alkyl benzene compounds and related them to their respective retention indices using a statistics based model. Soler *et al.*[15] established two statistical models, one of which related the retention properties of a series of benzodiazepines to their molecular connectivity indices. They discriminated between the models using the correlation coefficient, the standard error and Fisher's test. F. Vilma *et al.*[16] correlated the retention indices of linear alkyl benzene isomers with connectivity indices. They employed a multiple linear regression method and discriminated between rival models by means of the correlation coefficient and Fischer's test.

Yan *et al.*[17] are among the few workers who employed neural networks to predict retention indices. They worked on a series of alkyl benzenes on carbowax-20M and employed an extended delta-bar-delta back propagation learning algorithm. The bulk of publications involving neural networks attempted to predict properties other than retention indices.

The objective of this work is to apply rigorous model discrimination criteria on models and apply the most recent neural network algorithm to predict retention indices of a series of acrylates and corresponding propionates and haloproprionates. The merits and shortcomings of each method will then be discussed.

EXPERIMENTAL

Data set

In this study the retention indices of 63 acrylates selected from literature[18] served as an example to build QSAR models using multi-linear regression, stepwise, and neural network methods. In the original paper[18], the retention behaviour of 86 acrylates with various substituents at different positions have been examined isothermally on 5 different capillary columns. Among the 86 acrylates only 63 analogs which had the values of their retention indices for all the columns have been selected for this study (Table-1). Because no structural variables were available in the original paper, the 9 structural descriptors used in this work should be determined above all (Table-2). This work was performed on a Pentium-III personnel computer using programs written by ourselves. The basic operation of the back-propagation neural network program was performed using Matlab software.

The 63 groups of data set were randomly divided into three sets: a training set (40 members), validation set (10 members) and a testing set (13 members).

TABLE-1
RETENTION INDICES OF C_1–C_6 *n*-ALKYL AND C_3-C_6 ISOALKYL ACRYLATES AND THE CORRESPONDING PROPIONATES AND HALOPROPIONATES ON SQUALANE, OV-101, SE-54, UCON LB-550-X, AND SP-1000 PHASES AT 100°C

Esters	Squalane	OV-101	SE-54	UCON-LB-550-X	SP-1000
	I_1	I_2	I_3	I_4	I_5
Acrylates:					
Methyl*	596.5	596.7	603.9	719.5	940.0
Ethyl	654.0	675.7	700.0	793.2	992.1
Propyl	733.2	775.0	797.2	884.6	1078.3
Butyl†	835.2	875.2	895.1	983.3	1175.2
Pentyl	934.8	974.8	994.9	1082.7	1272.9
Hexyl	1033.6	1075.2	1095.3	1181.6	1370.3
Isopropyl*	684.4	722.5	739.3	820.9	995.9
Isobutyl	794.5	835.5	854.7	936.0	1113.8
Isopentyl	899.8	939.0	958.8	1041.7	1223.2
Isohexyl†	992.7	1035.4	1054.5	1132.8	1311.4
Propionates, X = H_2:					
Methyl	611.0	615.0	628.7	715.3	904.6
Ethyl	669.0	694.9	708.3	786.2	954.4
Propyl*	748.1	794.1	807.2	880.1	1041.7
Butyl	841.9	892.2	905.5	979.0	1138.3
Pentyl	944.3	989.5	1005.4	1078.2	1236.1
Hexyl	1042.3	1086.5	1105.1	1176.9	1332.6
Isopropyl†	688.1	739.4	750.2	814.0	956.0
Isobutyl	807.5	854.3	866.4	932.8	1079.2
Isopentyl*	910.8	954.6	968.2	1037.2	1186.4
2-*Chloropropionates*, X = 2-Cl-3H:					
Methyl	711.3	766.1	793.2	922.2	1197.8
Ethyl	789.3	847.8	873.7	978.8	1226.5
Propyl*	878.2	932.1	958.5	1067.7	1303.9
Butyl	975.3	1029.0	1055.2	1161.9	1390.6
Pentyl	1072.7	1126.4	1152.9	1257.5	1483.5
Hexyl†	1172.2	1225.1	1252.1	1356.2	1577.4
Isopropyl	820.5	875.6	899.0	997.1	1216.5
Isobutyl*	937.2	989.1	1014.2	1114.0	1335.9
Isopentyl	1039.5	1090.0	1116.9	1218.1	1440.6
3-*Chloropropionates*, X = 3-Cl-2H:					
Methyl†	770.0	823.0	855.2	1007.7	1331.7
Ethyl	846.7	901.4	930.9	1071.5	1371.6
Propyl	944.4	1000.1	1029.6	1164.0	1454.2
Butyl	1042.7	1098.3	1127.8	1260.3	1544.5
Pentyl*	1140.7	1195.8	1226.6	1358.3	1638.7
Hexyl	1239.3	1294.3	1326.1	1456.7	1734.7
Isopropyl†	888.2	945.0	971.5	1095.9	1366.2
Isobutyl	1004.7	1059.6	1088.0	1213.0	1488.7
Isopentyl*	1105.6	1160.1	1189.6	1316.8	1591.6

Esters	Squalane	OV-101	SE-54	UCON-LB-550-X	SP-1000
	I_1	I_2	I_3	I_4	I_5
2,3-*Dichloropropionates*, X = Cl_2:					
Methyl	869.1	923.6	960.1	1130.8	1490.7
Ethyl*	940.8	994.5	1028.0	1182.5	1512.1
Propyl	1036.1	1088.7	1122.6	1269.6	1585.4
Butyl	1131.5	1183.7	1218.4	1361.4	1669.0
Pentyl†	1227.5	1279.8	1314.4	1457.3	1761.4
Isopropyl	977.2	1032.0	1055.4	1198.6	1494.1
Isobutyl	1095.9	1145.3	1176.4	1315.5	1614.8
Isopentyl*	1193.0	1243.0	1275.3	1414.6	1713.7
2-*Bromopropionates*, X = 2-Br-3-H:					
Methyl	788.3	837.8	868.3	1000.0	1299.7
Ethyl	858.1	909.1	937.4	1056.1	1326.9
Propyl	953.5	1004.6	1032.6	1145.8	1405.3
Butyl	1050.4	1101.7	1129.2	1239.7	1493.5
Pentyl†	1146.6	1198.5	1226.0	1335.7	1583.5
Hexyl*	1245.8	1298.1	1325.7	1435.2	1681.1
Isopropyl	896.4	947.4	972.2	1075.3	1316.4
Isobutyl*	1016.3	1062.8	1088.5	1192.6	1438.3
Isopentyl†	1115.1	1162.9	1188.7	1292.2	1539.6
3-*Bromopropionates*, X=3-Br-2-H:					
Methyl*	852.4	901.7	937.1	1099.7	1435.2
Ethyl	928.3	978.5	1012.7	1160.0	1474.2
Propyl	1026.9	1076.9	1110.9	1251.3	1555.8
Butyl	1126.0	1175.3	1209.2	1346.9	1645.5
Pentyl	1223.7	1273.0	1307.3	1444.9	1739.7
Isopropyl	970.0	1021.1	1051.8	1183.6	1465.9
Isobutyl	1087.4	1135.5	1168.0	1300.0	1586.0
Isopentyl	1188.9	1235.9	1269.1	1396.7	1687.1
Isohexyl†	1281.6	1330.4	1363.4	1493.0	1776.0

*Test compounds, †Validation compounds.

Topological indices

Nine topological indices for each component were computed using a computer program developed by our team of workers. These topological indices are as follows: Wiener, Zagreb, Balaban, Shultz, Harary, Randic (of orders 1 to 4). Table-2 shows these indices.

Statitical models

Multiple linear regression with a variant method called stepwise regression have been employed to establish predictive models for retention indices for the compounds cited above.

TABLE-2
TOPOLOGICAL INDICES OF C_1–C_6 *n*-ALKYL AND C_3–C_6 ISOALKYL ACRYLATES AND THE CORRESPONDING PROPIONATES AND HALOPROPIONATES

Compds.	Topologic indices								
	Wiener W	Balaban J	Harary H	Shultz MTI	Zagreb Z	Randic1 χ^1	Randic2 χ^2	Randic3 χ^3	Randic4 χ^4
Acrylates:									
Methyl	22.375	3.746	35.239	60.719	11.031	3.974	3.179	2.847	0.691
Ethyl	38.000	3.703	36.990	108.641	14.531	4.428	3.543	2.809	1.269
Propyl	60.625	3.581	38.610	180.313	18.531	4.914	3.911	3.098	1.307
Butyl	91.250	3.463	40.195	279.234	22.531	5.414	4.254	3.358	1.511
Pentyl	130.875	3.367	41.769	409.406	26.531	5.914	4.608	3.600	1.695
Hexyl	180.500	3.292	43.340	574.828	30.531	6.414	4.961	3.850	1.867
Isopropyl	55.625	3.953	38.992	164.563	20.031	4.763	4.407	2.877	1.663
Isobutyl	85.250	3.727	40.480	259.984	24.531	5.264	4.752	3.303	1.376
Isopentyl	123.875	3.564	42.029	386.156	28.531	5.770	5.078	3.544	1.656
Isohexyl	172.500	3.445	43.593	547.578	32.531	6.270	5.436	3.775	1.827
Propionates, X = H_2:									
Methyl	24.875	3.425	31.842	77.469	13.531	3.451	2.720	2.161	0.423
Ethyl	41.000	3.477	33.581	130.641	17.031	3.906	3.085	2.145	1.050
Propyl	64.125	3.422	35.194	208.563	21.031	4.392	3.452	2.435	1.100
Butyl	95.250	3.347	36.774	314.734	25.031	4.892	3.795	2.694	1.305
Pentyl	135.375	3.279	38.346	453.156	29.031	5.392	4.149	2.937	1.489
Hexyl	185.500	3.223	39.916	627.828	33.031	5.892	4.502	3.187	1.661
Isopropyl	59.125	3.763	35.569	191.813	22.531	4.240	3.949	2.222	1.454
Isobutyl	89.250	3.594	37.057	249.484	27.031	4.741	4.293	2.639	1.176
Isopentyl	128.375	3.466	38.605	428.906	31.031	5.247	4.620	2.880	1.450
2-Chloropropionates, X = 2-Cl-3H:									
Methyl	34.941	4.084	42.840	105.044	16.068	4.010	3.705	2.955	0.781
Ethyl	54.919	4.026	44.647	170.010	19.568	4.465	4.069	2.951	1.262
Propyl	82.897	3.885	46.302	263.079	23.568	4.950	4.437	3.240	1.320
Butyl	119.875	3.741	47.912	387.751	27.568	5.450	4.780	3.500	1.524
Pentyl	166.853	3.615	49.504	548.026	31.568	5.950	5.134	3.743	1.708
Hexyl	224.831	3.531	51.091	747.904	35.568	6.450	5.487	3.993	1.879
Isopropyl	76.897	4.221	46.703	242.976	25.068	4.799	4.933	3.032	1.606
Isobutyl	112.875	3.983	48.207	346.148	29.568	5.300	5.278	3.445	1.398
Isopentyl	158.853	3.798	49.771	520.423	33.568	5.806	5.605	3.687	1.669
3-Chloropropionates, X = 3-Cl-2H:									
Methyl	38.941	3.565	42.097	116.000	15.362	4.212	3.260	2.441	1.207
Ethyl	59.919	3.624	43.878	184.319	18.862	4.666	3.625	2.425	1.854
Propyl	88.897	3.581	45.520	280.741	22.862	5.152	3.991	2.714	1.905
Butyl	126.875	3.508	47.122	408.766	26.862	5.652	4.335	2.974	2.109
Pentyl	174.853	3.433	48.709	572.393	30.862	6.152	4.689	3.217	2.293
Hexyl	233.831	3.367	50.292	775.624	34.862	6.652	5.042	3.467	2.465
Isopropyl	82.897	3.870	45.909	260.638	24.362	5.000	4.489	2.502	2.266
Isobutyl	119.875	3.724	47.412	385.163	28.862	5.502	4.833	2.919	1.980
Isopentyl	166.853	3.600	48.973	544.790	32.862	6.008	5.160	3.160	2.254

Compds.	Topologic indices								
	Wiener W	Balaban J	Harary H	Shultz MTI	Zagreb Z	Randic1 χ^1	Randic2 χ^2	Randic3 χ^3	Randic4 χ^4
2,3-Dichloropropionates, X = Cl_2:									
Methyl	50.713	4.308	53.439	147.942	17.898	4.778	4.111	3.773	1.503
Ethyl	75.544	4.253	55.287	228.055	21.398	5.233	4.476	3.769	2.004
Propyl	109.375	4.114	56.972	339.624	25.398	5.719	4.843	4.058	2.061
Butyl	153.206	3.962	58.603	486.149	29.398	6.219	5.186	4.318	2.265
Pentyl	208.037	3.823	60.212	671.630	33.398	6.719	5.540	4.561	2.449
Isopropyl	102.375	4.413	57.386	316.168	26.898	5.568	5.340	3.850	2.355
Isobutyl	145.206	4.183	58.906	459.193	31.398	6.068	5.684	4.263	2.139
Isopentyl	199.037	3.992	60.484	640.674	35.398	6.575	6.011	4.505	2.410
2-Bromopropionates, X = 2-Br-3-H:									
Methyl	33.943	4.198	68.885	100.339	15.246	4.082	3.805	3.045	0.813
Ethyl	53.739	4.108	70.698	164.125	18.746	4.536	4.170	3.035	0.294
Propyl	81.536	3.944	72.356	255.832	22.746	5.022	4.537	3.325	0.348
Butyl	118.332	3.784	73.680	378.961	26.746	5.522	4.880	3.584	0.552
Pentyl	165.129	3.648	75.562	537.512	30.746	6.022	5.234	3.827	0.736
Hexyl	222.925	3.538	77.149	735.484	34.746	6.522	5.588	4.077	0.908
Isopropyl	75.536	4.290	72.760	235.911	24.246	4.871	5.034	3.115	.0640
Isobutyl	111.332	4.032	74.265	355.540	28.746	5.372	5.379	3.530	1.425
Isopentyl	157.129	3.834	75.829	510.090	32.746	5.878	5.705	3.771	1.697
3-Bromopropionates, X = 3-Br-2-H:									
Methyl	37.943	3.645	67.885	110.932	14.903	4.321	3.336	2.535	1.297
Ethyl	58.739	3.685	69.669	177.889	18.403	4.776	3.701	2.519	1.935
Propyl	87.536	3.627	71.313	272.768	22.403	5.262	4.068	2.808	1.986
Butyl	125.332	3.543	72.916	399.069	26.403	5.762	4.412	3.068	2.190
Pentyl	173.129	3.460	74.504	560.791	30.403	6.262	4.765	3.311	2.374
Isopropyl	81.536	3.923	71.703	252.847	23.903	5.111	4.565	2.596	2.343
Isobutyl	118.332	3.763	73.207	375.647	28.403	5.611	4.910	3.013	2.061
Isopentyl	165.129	3.630	74.769	533.369	32.403	6.118	5.236	3.255	2.335
Ilsohexyl	222.925	3.523	76.343	730.512	36.403	6.618	5.594	3.485	2.506

The general form of the model is:

$$I_i = b_0 + b_i * \text{Wiener} + b_2 * \text{Zagreb} + b_3 * \text{Harary} + b_4 * \text{Shultz} + b_5 * \text{Balaban} + b_6 * \chi^1 + b_7 * \chi^2 + b_8 * \chi^3 + b_9 * \chi^4 \quad (1)$$

where i refers to the retention index corresponding to each column in Table-1. The b's are the coefficients estimated by linear regression and the χ's are the Randic indices (of orders 1 to 4). The model coefficients and the associated statistics were obtained by means of the statistical software STUDENT SYSTAT[19].

Artificial Neural Network Model

There has been an "explosion" of application of neural networks to areas relevant to chemical engineers. Neural networks have been used for a wide variety of purposes[20–27]. Most of the articles published on the subject concentrate on

applying neural networks in novel ways to solve important problems. Their application in QSAR analysis has been explored widely since 1990[28–34]. Our goal in this paper is to concentrate more on the application of neural net models (NNM) to predict retention indices in gas chromatography and less on their properties.

Excellent reviews of artificial networks are available[35]. Artificial neural networks are composed of many simple computational elements (nodes) locally interacting across very low bandwidth channels (connections). The architecture of these models is specified by the node characteristics, network topology and learning algorithm. Nodes in artificial neural networks are very simple processors inspired by their biological counterparts.

An artificial neural network provides a nonlinear mapping between some input (or independent) and output (or dependent) variables. The mapping is performed by use of processing elements (PE) and connection weights.

The development of the neural network involves the construction of the network of PEs (number of layers, number of elements in each layer), the connectivity between the layers and the strength of each connection. Due to the uncertainty concerning the true number of independent components, the number of hidden elements is usually adjusted to give the best model fit to the database.

Several ANN topologies[36–38] have been proposed. Each differs in the number and character of the processing nodes, the connections, the training procedures and whether the input-output values are continuous or discrete. The network topology chosen for this work is specified by a multi-layer feed-forward neural network. This choice was based on the relatively well-established training behaviour of this type of neural network. Feed-forward neural networks are networks that transfer information in the forward direction from input to output without feedback. For such networks, the connection weights are determined by training the neural network model with process data by adjusting the weights in an order fashion to minimize the deviation of the predicted outputs from data outputs. This process is called back propagation.

RESULTS AND DISCUSSION

The statistical treatment of the multi-linear regression models incorporating all topological indices[9] showed that the models were not significant at the 95% level. Although the Fisher's coefficient, F, is very meaningful for the 5 columns, due to the no significance of the majority of the coefficients of all the models, multi-linear regression models have been rejected. This was based on the Student's test which showed that some independent variables in the models have no significance. The no significance of coefficients of obtained models has also been confirmed by the values of tolerance that is the parameter that indicates the redundancy of each variable relative to the other independent variables. The registered value is one less than the square of the multiple correlation between one variable and the other ones. It is the fraction of its variability that is not explained by the other variables. A tolerance less than 0.1 is problematic. When it is less than 0.01 it indicates that variables are identical and that a variable can be predicted by the other ones which means that data are collinear. This evidence

was already foreseeable to the seen of correlation matrixes that showed that some independent variables are greatly correlated (r = 0.999). Even though the F report is very meaningful, in this case the coefficients of regression can have high standard errors that can make them become no significant from a statistical point of view. In other terms, there is a paradox: the F report indicates that coefficients are not all equal to zero, but none of these coefficients is meaningfully different from zero. Because of the colinearity of predictors, their coefficients are not meaningful and therefore one can suppress any redundant predictor without taking off a lot of explanatory power to the model. However, when a term is suppressed, it is necessary to know that the behaviour of variables changes and the other terms can become meaningful[39].

Since multi-linear regression models had been rejected essentially because of the number of independent variables in the models[39], we used the strategy of reduction of the number of predictors step by step by the stepwise method; the algorithm (SYSTAT) chooses the predictor that has the most elevated correlation with the dependent variable (retention index), then it examines every other predictor to see the one that, combined with the first predictor, gives the smallest value of the residual sum-of-squares. This process is reiterated; however, to every stage, it tests the significance of the predictor newly introduced and does not add it to the model if it is not meaningful to the level 5%. It also refuses to integrate a variable to the model if tolerance is too weak even though this one is meaningful to the level 5%.

The mathematical forms and the statistical parameters of the models obtained with the stepwise method are:

(a) For Squalane:

$$I = 389.310 + 86.345\ \chi^4 + 0.599\ MTI + 3.916\ H$$

$$n = 40,\quad R = 0.983,\quad s = 29.693,\quad F = 351.592$$

(b) For OV-101:

$$I = 430.468 + 94.995\ \chi^4 + 0.581\ MTI + 3.829\ H$$

$$n = 40,\quad R = 0.972,\quad s = 39.262,\quad F = 201.922$$

(c) For SE-54:

$$I = 418.502 + 101.690\ \chi^4 + 0.589\ MTI + 4.333\ H$$

$$n = 40,\quad R = 0.978,\quad s = 36.577,\quad F = 257.781$$

(d) For UCON-LB-550-X:

$$I = 418.502 + 101.690\ \chi^4 + 0.589\ MTI + 4.333\ H$$

$$n = 40,\quad R = 0.963,\quad s = 50.601,\quad F = 154.423$$

(e) For SP-1000:

$$I = 547.990 + 210.006\ \chi^4 + 0.352\ MTI + 6.754\ H$$

$$n = 40,\quad R = 0.930,\quad s = 83.791,\quad F = 76.664$$

The models obtained are statistically significant above the 5% level. The stepwise methods gives us meaningful models to a level. In order to assess the predictive power of the above models, the deviation between the experimental and the predicted retention indices has been calculated and is presented in Table-3.

TABLE-3
CALCULATED VALUES OF RETENTION INDICES (I_C) OF TEST COMPOUNDS AND THEIR RESIDUES $\delta = | I_C - I_{exp} |$= FOR THE DIFFERENT COLUMNS USING THE STEPWISE METHOD

Squalane		OV-101		SE-54		UCON-LB-550-X		SP-1000	
I_c	δ	I_c	δ	I_c	δ	I_c	δ	I_c	δ
623.341	26.841	660.317	69.617	677.224	73.324	762.530	43.030	952.481	12.481
784.168	99.768	833.356	110.856	853.492	114.192	970.308	149.408	1218.506	222.606
834.523	7.377	878.104	14.096	895.927	9.573	985.514	6.514	1181.000	42.700
922.602	11.802	965.224	10.624	985.854	17.654	1073.084	35.884	1264.211	77.811
842.188	36.012	886.000	46.100	908.313	50.187	1008.887	58.813	1230.526	73.374
906.142	31.058	948.967	40.133	973.427	40.773	1071.606	42.392	1289.012	46.888
1120.906	19.194	1167.359	28.441	1199.873	26.727	1314.671	43.629	1559.927	78.773
1102.039	3.561	1148.627	11.473	1180.792	8.808	1296.474	20.326	1543.874	47.726
915.454	25.346	965.032	29.468	996.172	31.828	1131.748	50.752	1567.767	55.667
1218.020	25.020	1263.231	20.231	1303.009	27.709	1424.599	9.999	1688.129	25.571
1296.726	50.926	1334.438	36.338	1380.013	54.313	1485.714	50.514	1705.099	24.000
1016.142	0.158	1056.765	6.035	1094.610	6.110	1210.252	17.652	1473.983	35.683
837.501	14.899	878.060	23.640	909.878	27.222	1035.964	63.736	1317.910	117.290

$\delta = | I_C - I_{exp} |$

Inspecting the deviation values in Table-3, one can see that the values can exceed 99.768 i.u. for squalane, 110.876 i.u. for OV-101, 114.192 i.u. for SE-54, 149.408 i.u. for UCON-LB-550-X and 222.060 i.u. for SP-1000. Therefore these models are not precise for the prediction of retention indices. Besides, this has been confirmed by the statistical study of the simple linear regression between experimental retention indices, I_{exp}, and calculated ones, I_c, for reference compounds as well as test compounds. The results of the test compounds are presented in Table-4.

TABLE-4
LINEAR REGRESSION $I_c = b_0 + b_1 I_{exp}$ FOR TEST COMPOUNDS USING STEPWISE METHOD

Column	b_0	b_1	r	F	S_r
Squalane	36.709	0.968	0.980	268.123	38.978
OV-101	93.471	0.910	0.975	207.517	43.911
SE-54	87.008	0.922	0.972	188.652	48.096
UCON-LB-550-X	120.415	0.896	0.962	137.659	57.567
SP-1000	248.188	0.824	0.941	85.531	78.884

Each of the models leads to correlation coefficients lower than 0.99 and to elevated standard errors. We conclude that models obtained by the stepwise method are meaningful, but are not precise enough for the prediction of the retention indices of our compounds.

In order to refine our precision, we used a new method used to solve complex cases of optimization, neural networks. It uses techniques based on the working of biological neurons for data prediction and classification.

The neural network used in our work can be presented as follows:

The database using the input parameters given in Table-1 was randomly split into three sets: a training set including 40 compounds of the data, the validation set including 10 compounds and the test set including the remaining 13 compounds.

The neural network software was used to scale the input data over the interval [−1 − +1], and to initialize the network weight using a tansigmoide function.

In order to determine the optimal number of hidden layer nodes, neural networks with different numbers of hidden layer nodes were trained. The number of hidden layer nodes was varied from three to fifteen. According to its generalization ability on the validation sets, we calculated the root-mean-square (RMS) error on different numbers of the hidden layer nodes and the lowest was picked as the optimal neural network model.

It is a multi-layer back propagation network with 3 layers (input–hidden layer–output). The number of neurons in the hidden layer is: 5 for squalane, 7 for OV-101, 3 for SE-54, 8 for UCON-LB-550-X and 3 for SP-1000. The network includes:

- An input layer that contains the nine descriptors.
- A hidden layer of several neurons.
- An output layer that contains the variables (retention indices).

A schematic diagram of the ANN is shown in Fig. 1, where the circles are nodes and the connections represent weights that describe the importance of the signal being transmitted along a given path. The neural network models were obtained using MATLAB.

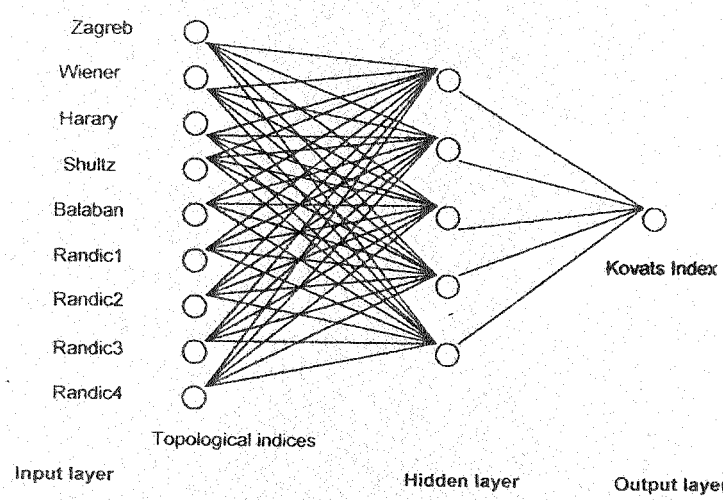

Fig. 1. Neural network diagram employed for the prediction of retention indices

To train the network we used the LEVENBERG MARQUARDT algorithm (Trainlm)[40]. The number of iterations has been fixed to 10.000 and the precision to 10^{-12}.

The choice of reference and test compounds is the same that for the two previous methods.

To test the validity of the model found, we compared the experimental results of the training compounds in the first place to those obtained by the models. Values of calculated retention indices for reference compounds are identical to their observed values ($\delta = 10^{-12}$), indicating that the training was successful. We did the same comparisons for test compounds (Table-5) and the found differences were very low.

Even the study of the linear regression between calculated and experimental retention indices for test compounds leads to values of correlation coefficient higher than 0.998 and to low standard errors. Table-6 and Fig. 2 show the goodness-of-fit results for the optimal neural network models (NNM).

TABLE-5
CALCULATED VALUES OF RETENTION INDICES (I_c) OF TEST COMPOUNDS AND THEIR RESIDUES ($\delta = I_c - I_{exp}$) FOR THE DIFFERENT COLUMNS USING NEURAL NETWORK

Squalane		OV-101		SE-54		UCON-LB-550-X		SP-1000	
I_c	δ	I_c	δ	I_c	δ	I_c	δ	I_c	δ
597.400	0.900	609.500	12.800	627.600	23.700	725.700	6.200	931.300	8.700
682.100	2.300	705.700	16.800	742.900	3.600	823.300	2.400	997.700	1.800
839.000	2.900	891.300	0.900	909.700	4.200	977.700	1.300	1133.900	4.400
913.900	3.100	937.000	17.600	964.200	4.000	1046.400	9.200	1185.100	1.300
878.400	0.200	938.700	6.600	960.400	1.900	1070.400	2.700	1304.000	0.100
936.700	0.500	988.800	0.300	998.600	15.600	1119.000	5.000	1361.700	25.800
1157.400	16.700	1198.700	2.900	1232.500	5.900	1354.100	4.200	1642.000	3.300
1116.800	11.200	1161.100	1.000	1186.400	3 .200	1324.300	7.500	1595.500	3.900
940.600	0.200	994.900	0.400	1024.600	3.400	1183.400	0.900	1503.700	8.400
1194.600	1.600	1248.000	5.000	1264.800	10.500	1426.300	11.700	1716.000	2.300
1237.100	8.700	1294.800	3.300	1314.100	11.600	1409.700	25.500	1627.200	53.900
1018.000	1.700	1058.700	4.100	1081.700	6.800	1179.600	13.000	1431.000	7.300
829.200	23.200	875.600	26.100	926.400	10.700	1091.600	8.100	1441.500	6.300

TABLE-6
LINEAR REGRESSION $I_c = b_0 + b_1 I_{exp}$ FOR TEST COMPOUNDS USING NEURAL NETWORK

Column	b_0	b_1	r	F	S_r
Squalane	–12.397	1.013	0.999	5033.017	9.418
OV-101	–8.932	1.006	0.999	3833.719	11.293
SE-54	30.412	0.968	0.999	7113.637	8.224
UCON-LB-550-X	16.860	0.985	0.999	5397.270	10.103
SP-1000	17.585	0.985	0.998	2343.209	17.997

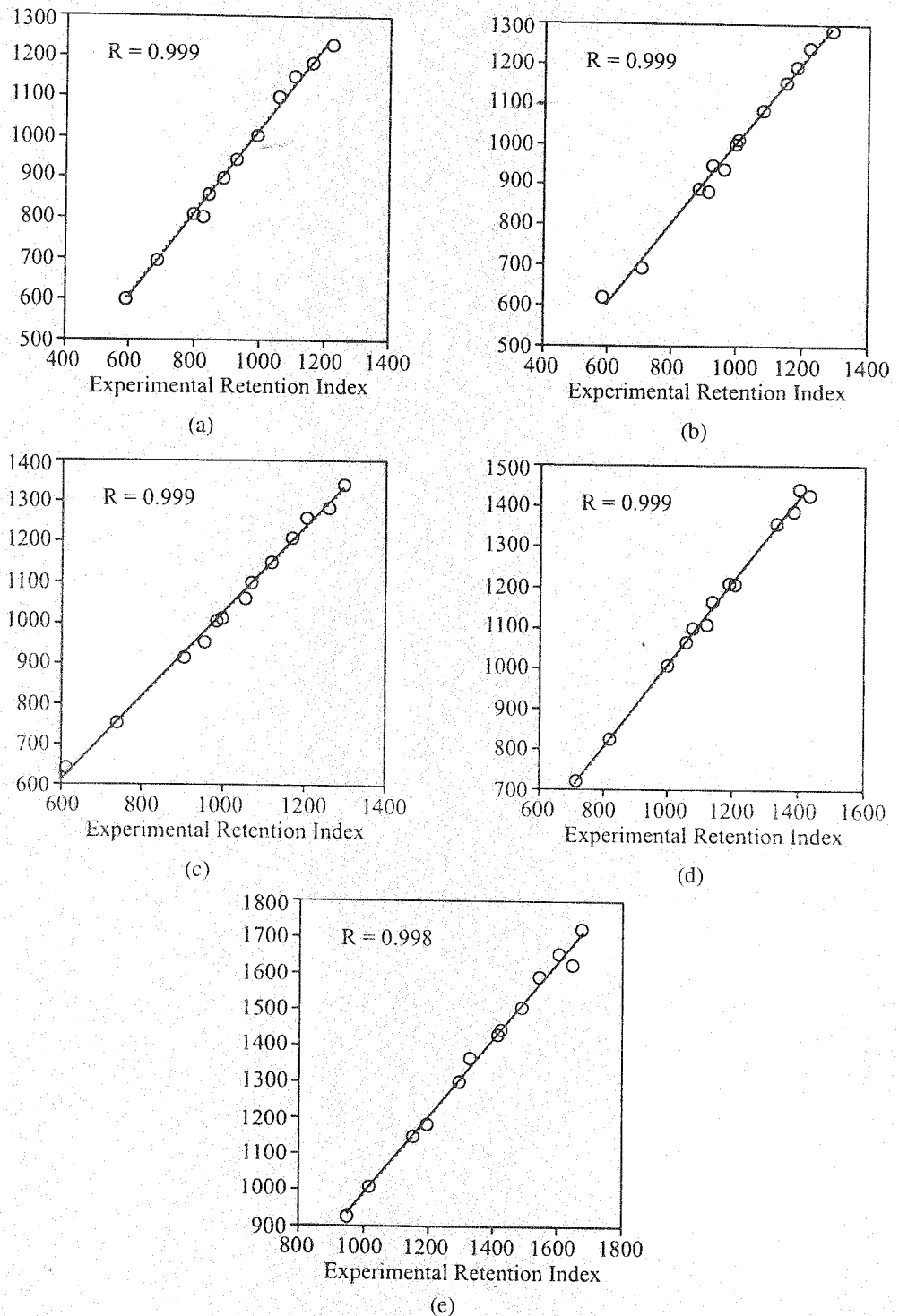

Fig. 2. Parity plots showing the goodness-of-fit for all the columns: (a) Squalane; (b) OV-101; (c) SE-54; (d) UCON-LB-550-X; (e) SP-1000

The retention indices clearly predicted by the ANN model are closer to the corresponding experimental values. This demonstrates the superiority of the ANN model in predicting the retention indices of our compounds.

Conclusion

The key results of this work suggest that a neural network model is capable of providing reasonable estimates of the retention indices. The NNM cannot provide the fundamental information needed to know which of the topological indices influencing the retention indices is the most important. However, it does provide a fast and accurate method for correlating retention indices to topological indices.

In this paper we have presented an example of QSAR model built by neural network method which is a powerful tool to predict chromatographic parameters. Comparing the results from stepwise method with those from neural network analysis, they are consistent in the relationships of retention indices with descriptors, but the results of neural network are much better. Rule-following behaviour occurred without any explicit representation of rules due to the spontaneous generalization, thereby allowing the network to classify similar input patterns not used to train the network. In conclusion, we have demonstrated that an artificial neural network (ANN) can predict successfully retention indices in gas chromatography.

REFERENCES

1. J. Castelle, *J. Chromatogr. A*, **842**, 51 (1999).
2. V. Makovskaya, J.R. Dean, W.R. Tomlinson and M. Comber, *Anal. Chim. Acta*, **315**, 193 (1995).
3. V.S. Raman and C.D. Maranas, *Comput. Chem. Sagag.*, **22**, 747 (1998).
4. X. Yao, X. Zhang, M. Liu, Z. Hu and B. Fan, *Comput. Chem.*, **25**, 475 (2001).
5. B. Ren, *Chemom. Intell Lab. Syst.*, **66**, 29 (2003).
6. E.A. Castro, M. Tucros and A.A. Toropov, *Comput. Chem.*, **24**, 571 (2000).
7. B.S. Junkes, R.C. Amboni, R.A. Yunes and V.F. Heinzen, *Anal. Chim. Acta*, **477**, 29 (2003).
8. R.C. Amboni, B.S. Junkes, V.F. Heinzen and R.A. Yunes, *J. Molec. Struct. (Theochem.)*, **579**, 53 (2002).
9. J. Galvez, R. Garcia-Domenech, J. Bernal and F. Garcia-March, *An. Real Acad. Farm.*, **57**, 533 (1991).
10. J. Huuskonen, *J. Chem. Inf. Comput. Sci.*, **40**, 773 (2000).
11. J. Beens, R. Tijssen and J. Blomberg, *J. Chromatogr. A*, **822**, 233 (1998).
12. I.G. Zenkevich, *J. Anal. Chem.*, **53**, 816 (1998).
13. R. Soler, R. Garcia and J. Galvez, *An. Real Acad. Farm.*, **57**, 563 (1991).
14. J.M. Sutler, T.A. Peterson and P.C. Jurs, *Anal. Chim. Acta*, **342**, 113 (1997).
15. R.M.S. Roca, F.J.G. March, G.M.A. Fos and R.G. Domenech, *J. Chromatogr.*, **607**, 91 (1992).
16. E.F. Vilma, Heinzen and R.A. Yunes, *J. Chromatogr. A*, **654**, 183, 189 (1993).
17. A. Yan, G. Jiao, Z. Hu and B.T. Fan, *Comput. Chem.*, **24**, 171 (2000).
18. A. Horna, J. Tabosky and J. Churacek, *J. Chromatogr.*, **348**, 141 (1985).
19. K.N. Berk and J.W. Steagall, Analyse Statistique de Donnees avec Student Systat, Thomson, Paris (1995).
20. Y. Tang, K.X. Chen, H.L. Jiang and R.Y. Ji, *Eur. J. Med. Chem.*, **33**, 647 (1998).
21. T.M. Leib, P.L. Mills, J.J. Lerou and J.R. Turner, *Trans. IchemE, Part A*, **73**, 690 (1995).
22. E. Hernandez and Y. Arkunt, *Comput. Chem. Eng.*, **16**, 227 (1992).

23. E.P. Nahas, M.A. Henson and D.E. Seborg, *Comput. Chem. Eng.*, **16**, 1039 (1992).
24. L. Hadjiiski, P. Geladi and P. Hopke, *Chemom. Intell Lab. Syst.*, **49**, 91 (1999).
25. A.P. Borosy, *Chemom. Intell. Lab. Syst.*, **47**, 227 (1999).
26. H.G. Bohr, P. Rogen and K.J. Jalkanen, *Comput. Chem.*, **26**, 65 (2001).
27. J.D. Hirst, R.D. King and M.J. Sternberg, *J. Computer-aided Molecular Design*, **8**, 421 (1994).
28. T. Aoyama, T. Suzuki and H. Ichikawa, *J. Med. Chem.*, **33**, 2583 (1990).
29. D.T. Manallack, D.D. Ellis and D.J. Livingstone, *J. Med. Chem.*, **37**, 3758 (1994).
30. S. So and M. Karplus, *J. Med. Chem.*, **39**, 5246 (1996).
31. T. Chiu and S. So, *J. Chem. Inf. Comput. Sci.*, **44**, 147 (2004).
32. J. Huuskonen, C. Living and I. Tetko, *J. Chem. Inf. Comput. Sci.*, **40**, 97 (2000).
33. T. Chiu and S. So, *J. Chem. Inf. Comput. Sci.*, **44**, 154 (2004).
34. I. Tetko, V. Tanchuk and A.E. Villa, *J. Chem. Inf. Comput. Sci.*, **41**, 1407 (2001).
35. A.J. Morris, G.A. Montague and M.J. Willis, *Trans. IchemE, Part A*, **72**, 3 (1994).
36. J.J. Hopfield, *Proc. Nat. Acad. Sci.*, **79**, 2254 (1982).
37. J.A. Feldman and D.H. Ballard, *Cognitive Sci.*, **6**, 205 (1982).
36. T. Kohonen, Self Organization and Associative Memory, Springer-Verlag, Berlin (1984).
37. P. Dagnelie, Statistique théorique et appliquée, Presses Agronomiques de Gembloux, Tome 1, Belgique (1992).
38. M.T. Hagan and M. Menhaj, *IEEE Trans. Neural Networks*, **5**, 989 (1994).

(*Received*: 28 May 2005; *Accepted*: 10 April 2006) AJC-4774

11th ASIAN PACIFIC CONFEDERATION OF CHEMICAL ENGINEERING CONGRESS

AUGUST 27–30, 2006

KUALA LUMPUR CONVENTION CENTER, MALAYSIA

Contact:
Congress Secretariat
The Institution of Engineers, Malaysia
Lot 60 & 62, Jalan 52/4
PO Box 223 (Jalan Sultan)
46720 Petaling Jaya, Selangor Malaysia
Fax: (603)79577678
Tel: (603)79684008, (603)79684015
Email: siti@iem.org.my, janet@iem.org.my