ELSEVIER

Journal of Chromatography A, 945 (2002) 173–184

JOURNAL OF CHROMATOGRAPHY A

www.elsevier.com/locate/chroma

Use of self-training artificial neural networks in modeling of gas chromatographic relative retention times of a variety of organic compounds

M. Jalali-Heravi*, Z. Garkani-Nejad

Department of Chemistry, Sharif University of Technology, PO Box 11365-9516, Tehran, Iran

Received 8 August 2001; received in revised form 7 November 2001; accepted 13 November 2001

Abstract

A quantitative structure–activity relationship study based on multiple linear regression (MLR), artificial neural network (ANN), and self-training artificial neural network (STANN) techniques was carried out for the prediction of gas chromatographic relative retention times of 13 different classes of organic compounds. The five descriptors appearing in the selected MLR model are molecular density, Winer number, boiling point, polarizability and square of polarizability. A 5-6-1 ANN and a 5-4-1 STANN were generated using the five descriptors appearing in the MLR model as inputs. Comparison of the standard errors and correlation coefficients shows the superiority of ANN and STANN over the MLR model. This is due to the fact that the retention behaviors of molecules show non-linear characteristics. Inspection of the results of STANN and ANN shows there are few differences between these methods. However, optimization of STANN is much faster and the number of adjustable parameters for this technique is much less compared with those of the conventional ANN. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Neural networks, artificial, self-training; Retention times, relative; Regression analysis; Multiple linear regression analysis; Quantitative structure–activity relationships

1. Introduction

The most important aim of mathematical and statistical methods in chemistry is to provide the maximum information about the selected molecular property by analyzing chemical data. In chromatography, retention is a phenomenon that depends on the solute–solute, solute–stationary phase and solute–mobile phase interactions. If the mobile and stationary phases are the same for the solutes, then only the differences in the structures of the solute molecules need to be investigated. Using the quantitative structure–activity relationship (QSAR) approach, structural parameters such as topological, geometric, electronic, and physicochemical descriptors can be generated for molecules and a subset can be selected that best describes the gas chromatographic retention parameters.

QSAR has been used to obtain models for predicting the chromatographic behavior of different groups of compounds [1]. Jurs and co-workers demonstrated the prediction of the retention indices for diverse sets of substituted pyrazines [2,3], polycyclic aromatic compounds [4], narcotics [5] and anabolic steroids

*Corresponding author. Tel.: +98-21-600-5718; fax: +98-21-601-2983.

E-mail address: jalali@sina.sharif.ac.ir (M. Jalali-Heravi).

0021-9673/02/$ – see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0021-9673(01)01513-8

[6]. Katritzky and coworkers have used the QSAR techniques for prediction of retention times of different organic compounds [7,8]. Collantes et al. have studied the chromatographic data for polycyclic aromatic hydrocarbons using the QSAR methods [9]. Some other works in this area are listed in Refs. [10–15].

In the present study, an artificial neural network (ANN) and, for the first time, a self-training artificial neural network (STANN) were employed to generate QSAR models between the relative retention times (RRTs) and the structural parameters (descriptors) of 13 different classes of organic compounds. As the first step, a multiple linear regression (MLR) model was developed and the descriptors appearing in this model were considered as inputs for the ANN and STANN. Then, the generated ANN and STANN were applied for the prediction of relative retention time of organic compounds with diverse structures. The main aim of this work was investigation of the use of a STANN in predicting the RRT and comparison of its results and ease of optimization with a conventional ANN.

2. Methods

There are many types of network architectures, but the type that has been most useful for QSAR studies is the multilayer feed-forward network with back-propagation (BP) learning [16]. The back-propagation learning method can be applied to any multilayer network that uses differentiable activation functions and supervised training. An ANN consists of a number of hidden units (nodes) that receive data from the outside, process the data, and output a signal. The back-propagation network receives a set of inputs, which are multiplied by each node's weights. These products are summed for each node and then a non-linear transfer function is applied. In order to train the network using the back-propagation algorithm, the differences between the ANN output and its desired value are calculated after each iteration. The changes in the values of the weights can be obtained using the following equation:

$$\Delta W_{ij}(n) = \eta \delta_i O_j + \alpha \Delta W_{ij}(n-1) \qquad (1)$$

where ΔW_{ij} is the change in the weight factor for each network node, δ_i is the actual error of node i, and O_j is the output of node j. The coefficients η and α are the learning rate and the momentum factor, respectively. These parameters are be optimized before training the network. Since we have used the self-trained artificial neural network for the first time, the optimization procedure of this method is therefore described in detail in the next section.

2.1. Optimization procedure of self-training artificial neural network

A self-training artificial neural network [17] is a new procedure for updating the node's weights and training of the networks in parallel fashion. An important aspect of the STANN is a network which trains another network. The architecture of a STANN is shown in Fig. 1. The structure of network 2 in this figure is the same as a BP-ANN. However, during

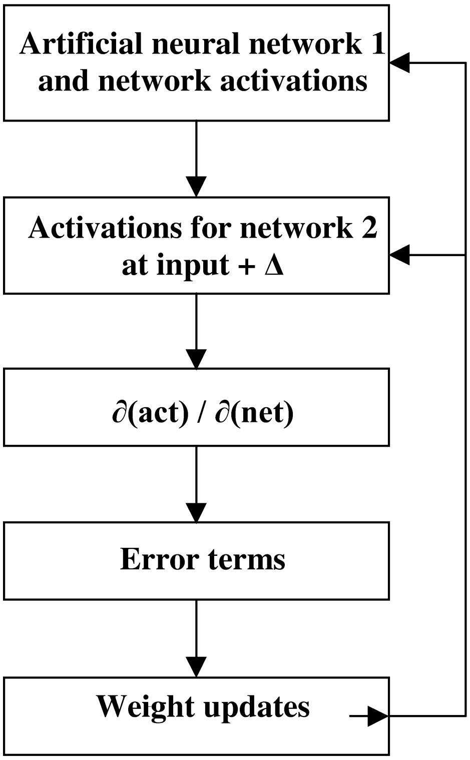
Artificial neural network 1 and network activations
Activations for network 2 at input + Δ
∂(act) / ∂(net)
Error terms
Weight updates

Fig. 1. The architecture of a self-training artificial neural network.

the training, the normalized inputs are increased by some infinitesimal amount, delta (Δ). In this regard, because the transfer function being utilized, a sigmoid, has a linear region around the value of 0.5, it is desirous when adding the delta value to the normalized input to adjust the input towards the linear region. Thus, the positive delta value should be added to normalize inputs which are less than 0.5 and the negative delta values should be added to normalize inputs which are greater than 0.5. For the hidden layers a similar manner are used. Network 1 uses from weight updates produced by the training network 2. Thus, training of artificial neural network 1 is not carried out with algorithmic code, but rather by a network training a network.

As mentioned before, the STANN was used for the first time for predicting the RRT of various organic compounds. The results obtained using this technique were compared with those of a conventional ANN.

3. Experimental

3.1. Data set

Data of 122 organic compounds taken from Ref. [18] were used as the data set. The compounds consist of 13 different classes of organic compounds containing various functional groups, i.e. alcohols, ketones, aldehydes, esters, alkenes, alkynes, alkanes, halides, thiols, nitro, ethers, cyanides, and sulfides. In Ref. [18], the RRTs of different compounds were determined using a Hewlett-Packard HP 5880 gas chromatograph. The carrier gas was helium and the chromatograms were obtained using a 30 m×0.25 mm I.D., 0.25 μm Rtx-5 column from Restek (Bellefonte, PA, USA). In the present work, these compounds were randomly divided into two groups: a training set and a prediction set (Table 1). The training and prediction sets consist of 105 and 17 compounds, respectively. The values of retention time relative to benzene (RRT) were used as the dependent variable. The training set was used for model generation and the prediction set was used for the evaluation of the models.

3.2. Descriptor generation

The next step in developing the model was generation of the numerical description of the molecular structures. The generated numerical descriptors were responsible for encoding important features of the structures and could be categorized as topological, geometric, electronic, and physicochemical properties. A total of 77 descriptors were calculated for each compound in the data set. Topological descriptors were calculated using two-dimensional representation of the molecules. In order to calculate the electronic and geometric descriptors, the molecular structures must first be optimized. Therefore, the three-dimensional structure of each molecule was optimized using the semi-empirical molecular orbital method of AM1 implemented in the MOPAC package (version 6) [19].

3.3. Regression analysis

The stepwise multiple linear regression procedure was used for model generation. The procedure for screening the descriptors and choosing the best model is given elsewhere [11]. A total of 19 out of 77 descriptors were removed using the mentioned procedure [11]. Then the stepwise addition method implemented in the software package of SPSS/PC was used for choosing the descriptors contributing to the RRT [20]. The best selected MLR model is presented in Table 2. The five parameters obtained using the stepwise method and appearing in this model are molecular density (MD), boiling point (b.p.), Winer number (WN), polarizability (α) and square of polarizability (α^2). The main goal of generating the MLR model was to choose a set of suitable descriptors as inputs for developing the ANN and STANN models.

3.4. Self-training artificial neural network and artificial neural network generation

The STANN and ANN programs were written in Fortran 77 in our laboratory. The networks were generated using the descriptors appearing in the MLR model as inputs. Therefore, the number of inputs in the STANN and ANN was five and the

Table 1
Experimental and calculated values of the RRT for the training and prediction sets

No.	Compound	RRT$_{\mathrm{EXP}}$	RRT$_{\mathrm{STANN}}$	RRT$_{\mathrm{ANN}}$	RRT$_{\mathrm{MLR}}$
Training set					
1	Dibromomethane	1.2668	1.2696	1.3334	1.6996
2	$CHCl_2CH_2Cl$	2.0898	1.8887	1.7737	1.7588
3	CCl_3CH_3	0.9191	0.8234	0.8802	1.1542
4	1,3-Dibromopropane	3.8305	3.8140	3.8523	3.1055
5	$CFCl_2CF_2Cl$	0.5878	0.6148	0.4398	0.4618
6	Ethyl disulfide	3.7395	3.7935	3.8705	3.8620
7	3-Bromopentane	2.3527	2.5003	2.4678	2.2350
8	Idomethane	0.5803	0.6502	0.6495	0.7308
9	2-Bromopentane	2.2239	2.3900	2.3646	2.1949
10	1-Pentanethiol	2.7509	2.7244	2.6258	2.1838
11	Ethyl iodide	0.7939	0.8003	0.8821	1.2659
12	1-Butanol	1.0562	1.0111	1.1071	1.3502
13	CH_2ClCH_2Cl	0.9423	0.8195	0.8453	1.0868
14	1-Bromobutane	1.5094	1.5757	1.6147	1.7260
15	1-Bromopentane	2.9502	2.9915	2.953	2.3854
16	2-Hexanone	2.4268	2.3007	2.1993	1.7985
17	Cyclopentylchloride	2.2720	1.9994	1.8885	1.7536
18	2-Methylheptane	1.9660	2.1251	2.0993	2.0778
19	2-Nitropropane	1.2814	1.5678	1.5986	1.5926
20	Butyl formate	1.5044	1.5676	1.4989	1.3835
21	2-Ethylbutyraldehyde	1.9289	2.0007	1.8791	1.6594
22	2,3,4-Trimethylpentane	1.7248	1.9725	1.9264	2.1144
23	*cis*-CHCl=CHCl	0.7732	0.6896	0.6719	0.9357
24	Propyl sulfide	3.4429	3.5283	3.6138	2.9332
25	Cycloheptane	2.4215	2.3209	2.2083	1.9883
26	Propyl acetate	1.4114	1.3918	1.3549	1.3157
27	1,3-Dichlorobenzene	4.2362	4.3229	4.4153	3.6739
28	$CHCl_2CH_3$	0.6875	0.6800	0.6824	0.6613
29	Ethylcyclohexane	2.9354	2.9560	2.9180	2.4984
30	Dipropyl ether	1.1711	1.0409	1.0686	1.2472
31	Isopropyl acetate	1.0243	1.0474	1.0618	1.1249
32	2,3-Butanedione	0.7187	0.8203	0.8432	0.9357
33	Nitromethane	0.6669	0.6678	0.7582	1.2068
34	2-Methyl-1-propanol	0.8618	0.7990	0.9489	1.1935
35	Cyclopropylcyanide	1.5194	1.4973	1.4477	1.6869
36	Allylsulfide	3.1745	3.2746	3.2018	3.3008
37	1,2-Dichlorobenzene	4.3976	4.4708	4.5224	3.7876
38	1-Methylcyclohexene	2.0159	1.9908	1.9310	2.0775
39	4-Methylcyclohexene	1.6085	1.6363	1.6159	1.8361
40	Methanol	0.4742	0.4303	0.3968	0.3066
41	Methyl propionate	0.8665	0.7821	0.7965	0.8272
42	1-Ethylcyclopentene	1.8348	1.7736	1.7368	1.9139
43	3-Pentanone	1.2444	1.1146	1.1491	1.2302
44	1-Bromopropane	0.8358	0.7936	0.8506	1.0958
45	*trans*-CHCl=CHCl	0.6553	0.6486	0.6379	0.5429
46	Allyl acetate	1.2808	1.4752	1.4516	1.4141
47	Valeraldehyde	1.2435	1.0941	1.1348	1.2058
48	Ethyl acetate	0.8026	0.7622	0.7727	0.7819
49	3,3-Dimethyl-2-butanone	1.3367	1.6112	1.5462	1.5162
50	2,2,4-Trimethylpentane	1.1403	1.3345	1.3343	1.8425
51	3-Ethylpentane	1.1358	1.1933	1.2097	1.4468
52	*trans*-2-Heptene	1.3048	1.3443	1.3519	1.6828
53	*sec.*-Butanol	0.7565	0.6474	0.8253	1.0423

Table 1. Continued

No.	Compound	RRT_{EXP}	RRT_{STANN}	RRT_{ANN}	RRT_{MLR}
54	2-Pentanone	1.1689	1.1568	1.2032	1.2766
55	3-Ethyl-2-pentene	1.3151	1.3204	1.3194	1.7577
56	Acetaldehyde	0.4794	0.5536	0.5228	−0.4215
57	4-Bromo-*m*-xylene	4.9433	5.1067	4.8638	5.4243
58	1-Heptyne	1.4709	1.3762	1.3695	1.4531
59	2-Methyl-2-butanol	0.9019	1.0821	1.1555	1.2529
60	2-Bromo-*p*-xylene	4.9562	4.9068	4.7300	5.1808
61	2-Bromopropane	0.6951	0.6973	0.7143	0.9146
62	3,3-Dimethylpentane	0.9475	1.0242	1.0498	1.3649
63	1-Heptene	1.1771	1.1782	1.1992	1.5007
64	Toluene	1.9918	2.0238	1.9647	2.1144
65	2,4-Dimethylpentane	0.8326	0.9053	0.9252	1.2143
66	Diisopropyl ether	0.7539	0.7382	0.7180	0.9238
67	*p*-Xylene	3.3243	3.2524	3.2315	3.1011
68	Trimethylacetonitrile	0.9480	1.0569	1.1647	1.2734
69	Ethyl benzene	3.2463	3.1611	3.1435	2.9006
70	*o*-Xylene	3.5036	3.4672	3.5216	3.1497
71	Cumene	3.7404	3.6233	3.6702	3.4842
72	*n*-Butylbenzene	4.4965	4.4954	4.5414	4.3501
73	3-Hexyne	1.0247	0.9067	0.9434	1.0975
74	3-Ethyl-1-pentene	0.9522	0.9731	0.9849	1.4133
75	Cyclohexene	1.1043	0.9869	1.0335	1.2499
76	Butyraldehyde	0.7281	0.5286	0.6295	0.6339
77	*sec*.-Butylbenzene	4.2867	4.2250	4.2885	4.2478
78	Methyl *tert*.-butyl ether	0.6647	0.6491	0.6470	0.4936
79	Isopropanol	0.5505	0.4015	0.5261	0.6425
80	Propionitrile	0.6943	0.6200	0.6132	0.8659
81	1-Chloropropane	0.6223	0.6206	0.6096	0.2653
82	Propionaldehyde	0.5513	0.4414	0.4930	0.0903
83	2-Butanone	0.7413	0.5378	0.6598	0.7184
84	Ethanol	0.5115	0.4388	0.4483	0.5236
85	Bicyclo[2,1]hepta-2,5-diene	1.3051	1.1173	1.1537	1.4008
86	Methylcyclopentane	0.8379	0.7649	0.8069	0.8602
87	Crotonaldehyde	0.9543	1.0524	1.1379	1.2827
88	Hexane	0.7363	0.7198	0.7463	0.7404
89	2-Chloropropane	0.5565	0.6255	0.6122	0.1043
90	Trifluoromethyl-benzene	1.3526	1.1501	1.3135	1.4211
91	1-Hexyne	0.8176	0.7234	0.7485	0.7261
92	2-Methyl-1-pentene	0.7097	0.6971	0.7033	0.7596
93	1-Hexene	0.7123	0.6981	0.7021	0.7365
94	2,2-Dimethylbutane	0.5941	0.6327	0.6272	0.4454
95	*m*-Xylene	3.3081	3.2883	3.2836	3.0839
96	2-Methyl-2-propanol	0.5852	0.5048	0.6642	0.7509
97	Acetonitrile	0.5439	0.4994	0.4548	0.5611
98	Methacrylonitrile	0.7548	0.6885	0.8264	0.9768
99	Cyclopentane	0.6568	0.5902	0.6170	0.2531
100	Pentane	0.5498	0.5786	0.5906	−0.0407
101	2-Methyl-2-butene	0.5762	0.6095	0.6180	0.1893
102	1,4-Difluorobenzene	1.1153	1.0085	1.0462	1.3913
103	Acrylonitrile	0.5884	0.4012	0.5457	0.6210
104	1-Pentene	0.5392	0.5959	0.5902	−0.0639
105	Hexafluorobenzene	0.7762	0.7299	0.8103	0.6996

Table 1. Continued

No.	Compound	RRT_{EXP}	RRT_{STANN}	RRT_{ANN}	RRT_{MLR}
Prediction set					
106	CH_2Cl_2	0.6001	0.6639	0.6278	0.3579
107	$CH_2ClCHClCH_3$	1.2406	1.2400	1.2340	1.3801
108	Ethyl sulfide	1.2571	1.2480	1.2819	1.4594
109	1,1-Dimethylcyclohexane	2.2450	2.3964	2.3248	2.3131
110	3,3-Diethylpentane	3.3426	3.4279	3.4891	3.0936
111	Heptane	1.2393	1.3299	1.3315	1.4810
112	Butyronitrile	1.0892	1.0090	1.0480	1.3127
113	1-Octyne	2.8834	2.5190	2.4834	2.1098
114	Propyl formate	0.8293	0.8069	0.8152	0.8399
115	1-Propanol	0.6508	0.6036	0.6430	0.9005
116	Tetrahydrofuran	0.8477	0.5892	0.6367	0.5502
117	Isobutyronitrile	0.8441	0.8682	0.9700	1.1553
118	2-Hexyne	1.1062	0.9619	0.9980	1.1539
119	Propyl benzene	3.9369	3.8704	3.9762	3.6283
120	Diethyl ether	0.5592	0.5946	0.5818	−0.0410
121	Acetone	0.5460	0.4066	0.4889	0.2288
122	Cyclopentene	0.6371	0.6208	0.6318	0.2761

number of nodes in the output layer was set to be one. A three-layer network with a sigmoidal transfer function was designed for both STANN and ANN. Before training the STANN and ANN, the input and output values of the networks were normalized between 0.1 and 0.9. The number of nodes in the hidden layer, learning rate and momentum were optimized. The initial weights were selected randomly between −1 and +1. In order to evaluate the performance of the STANN and ANN, the standard errors of training (SET) and prediction (SEP) were used.

4. Results and discussion

Table 1 shows that the data set consists of a very diverse set of molecules. The experimental values of the relative retention time of these compounds on an Rtx-5 column are also given in this table. Table 2 demonstrates the specifications of the selected MLR model. Also, the mean effect of each parameter is included in this table. The calculated values of RRT using this model for the training and prediction sets are presented in Table 1. The variables appearing in the selected MLR model encode different aspects of the molecular structure. These parameters mainly show topological and physicochemical characteristics indicating that these properties of molecules affect the RRT. MD is defined as the ratio of molecular mass to the Van der Waals volume of the molecules. This parameter with negative coefficient and mean effect indicates that as the ratio of mass to volume of the molecules increases the RRT decreases and these properties play different roles in the retention behavior. The presence of the WN as a topological descriptor in the model indicates that the RRT depends on the degree of branching and compactness

Table 2
Specifications of the selected multiple linear regression model

Descriptor	Notation	Coefficient	Mean effect
Molecular density	MD	−1.020 (±0.159)	−1.064
Boiling point	b.p.	+0.017 (±0.001)	+1.697
Winer number	WN	−0.013 (±0.002)	−0.513
Polarizability	α	−0.141 (±0.076)	−0.988
Square of polarizability	α^2	+0.035 (±0.006)	+1.884
Constant		+0.509 (±0.303)	

Table 3
The values of the descriptors appearing in the models studied in this work[a]

No.[b]	MD	b.p.	WN	α	α^2
Training set					
1	0.3733	97.0	4	3.5253	12.4274
2	0.6516	113.8	18	5.0319	25.3200
3	0.6531	74.1	16	5.3066	28.1603
4	0.4892	167.0	20	6.1559	37.8951
5	0.5564	47.6	58	6.2045	38.4954
6	0.9503	153.0	35	10.0648	101.2998
7	0.7561	119.0	31	7.3177	53.5483
8	0.3678	42.4	1	2.6754	7.1577
9	0.7563	117.0	32	7.3403	53.8797
10	1.0923	127.0	35	7.8459	61.5575
11	0.4433	72.0	4	4.1017	16.8243
12	1.1800	117.6	20	5.4352	29.5413
13	0.7384	83.5	10	4.1194	16.9692
14	0.7104	101.0	20	6.1372	37.665
15	0.7547	130.0	35	7.3484	53.9988
16	1.1479	127.0	52	7.5974	57.7198
17	0.9516	114.0	26	6.5587	43.0169
18	1.2827	118.0	79	9.6217	92.5772
19	0.9573	120.3	29	5.7899	33.5235
20	1.0487	107.0	56	7.2974	53.2528
21	1.1488	117.0	48	7.5509	57.0163
22	1.2728	113.0	65	9.4798	89.8668
23	0.6119	60.0	18	5.3571	28.6990
24	1.1059	146.0	56	9.3897	88.1667
25	1.1988	118.4	42	8.2177	67.5311
26	1.0449	102.0	52	7.1926	51.7333
27	0.7797	173.0	61	9.4005	88.3693
28	0.7403	57.3	9	4.2436	18.0082
29	1.2019	132.0	64	9.4177	88.6940
30	1.1946	89.0	56	8.1131	65.8228
31	1.0488	90.0	48	7.1037	50.4626
32	0.9681	87.5	29	5.4944	30.1889
33	0.8410	101.2	9	3.3072	10.9378
34	1.1808	107.9	18	5.3761	28.9027
35	1.1041	134.0	17	5.1462	26.4833
36	1.0443	138.0	56	10.2026	104.0925
37	0.7787	180.5	60	9.3398	87.2323
38	1.1695	110.0	42	8.6729	75.2194
39	1.1697	103.0	42	8.4124	70.7684
40	1.1487	64.6	1	1.8252	3.3315
41	1.0212	79.7	31	5.9131	34.9651
42	1.1761	106.0	43	8.5131	72.4723
43	1.1353	101.0	31	6.4230	41.2551
44	0.6542	71.0	10	4.9194	24.2001
45	0.6896	47.5	10	4.3309	18.7568
46	1.0097	103.0	52	7.3152	53.5122
47	1.1426	103.0	35	6.4302	41.3480
48	1.0233	77.1	32	5.9694	35.6337
49	1.1448	106.0	42	7.4558	55.5883
50	1.2771	99.2	66	9.4530	89.3600

Table 3. Continued

No.[b]	MD	b.p.	WN	α	α^2
51	1.2914	93.0	48	8.3892	70.3783
52	1.2606	98.0	56	8.8745	78.7560
53	1.1814	99.5	18	5.3577	28.7046
54	1.1391	105.0	32	6.4043	41.0149
55	1.2485	96.0	48	8.8527	78.3700
56	1.0834	20.1	4	2.7934	7.8032
57	0.7493	214.0	84	11.6086	134.7599
58	1.2394	100.0	56	8.2518	68.0928
59	1.1808	102.0	28	6.4597	41.7281
60	0.7493	200.0	84	11.6087	134.7623
61	0.6568	59.0	9	5.0002	25.0019
62	1.2868	86.0	44	8.3476	69.6820
63	1.2608	93.3	56	8.6632	75.0502
64	1.1366	110.6	42	8.6578	74.9581
65	1.2935	81.0	48	8.3409	69.5712
66	1.1939	68.5	48	7.9349	62.9624
67	1.1393	138.3	62	10.1550	103.1244
68	1.1800	105.5	28	6.3253	40.0097
69	1.1423	136.2	64	9.9172	98.3508
70	1.1390	144.0	60	10.0179	100.3583
71	1.1492	151.0	88	11.0232	121.5099
72	1.1550	183.0	133	12.3781	153.2175
73	1.2543	81.0	35	7.5431	56.8980
74	1.2624	84.0	48	8.5951	73.8749
75	1.1634	83.0	27	7.3255	53.6630
76	1.1258	76.0	20	5.2301	27.3544
77	1.1553	173.0	121	12.2542	150.1654
78	1.1875	55.2	28	6.6538	44.2729
79	1.1777	82.4	9	4.1821	17.4898
80	1.1756	97.2	10	4.0266	16.2138
81	0.9688	46.0	10	4.5265	20.4896
82	1.1136	49.0	10	4.0254	16.2042
83	1.1264	79.6	18	5.2105	27.1491
84	1.1654	78.3	4	3.0454	9.2747
85	1.0606	89.0	36	7.4913	56.1189
86	1.2164	71.8	26	6.8941	47.5290
87	1.0745	102.0	20	5.8154	33.8193
88	1.3116	69.0	35	7.3066	53.3869
89	0.9708	35.7	9	4.5609	20.8021
90	0.8263	101.0	114	8.9249	79.6535
91	1.2654	71.0	35	7.0407	49.5712
92	1.2656	62.0	32	7.4492	55.4901
93	1.2685	63.0	35	7.4567	55.6023
94	1.3046	49.7	28	7.1408	50.9903
95	1.1420	139.1	61	10.0817	101.6410
96	1.1800	82.2	16	5.2765	27.8411
97	1.1696	81.6	4	2.8005	7.8427
98	1.1284	90.3	18	5.5259	30.5352
99	1.2192	49.0	15	5.7120	32.6267
100	1.3296	36.1	20	6.1162	37.4080
101	1.2810	39.0	18	6.4599	41.7303
102	0.8665	88.5	62	7.8724	61.9739
103	1.1135	77.3	10	4.2805	18.3223
104	1.2850	30.0	20	6.2443	38.9912
105	0.6434	81.5	174	9.4022	88.4013

Table 3. Continued

No.[b]	MD	b.p.	WN	α	α^2
Prediction set					
106	0.6597	39.8	4	2.8711	8.2431
107	0.7963	96.8	18	5.3236	28.3411
108	1.0773	92.0	20	6.9664	48.5309
109	1.2030	120.0	59	9.3349	87.1409
110	1.2601	146.0	88	10.7168	114.8489
111	1.2932	98.4	56	8.4999	72.2476
112	1.1824	118.0	20	5.2555	27.6200
113	1.2363	126.0	84	9.4566	89.4268
114	1.0229	81.0	35	6.0765	36.9236
115	1.1731	97.2	10	4.2418	17.9931
116	1.0785	66.0	15	5.1164	26.1775
117	0.9779	107.0	32	5.1982	27.0214
118	1.2597	84.0	35	7.5684	57.2801
119	1.1482	159.0	94	11.1562	124.4602
120	1.1872	34.6	20	5.6781	32.2406
121	1.1109	56.2	9	4.0039	16.0312
122	1.1712	44.0	15	5.9404	35.2881

[a] The definition of the descriptors are given in Table 2.

[b] The numbers refer to the numbers of the molecules given in Table 1.

of the molecules. The b.p. of the molecules with a high positive mean effect indicates that as the boiling point of the molecules increases, the RRT increases and this parameter plays a major role in the gas chromatographic retention behavior of organic molecules. In order to improve the statistics of the model, different types of combination of descriptors, such as square and cubic function of descriptors, were examined. It can be seen from Table 2 that the square of polarizability (α^2) also appeared in the MLR model. The presence of this parameter improves the ability of the model compared with that of a simple MLR. This could be due to a non-linear relationship between the RRT and the descriptors appearing in the MLR model. It is noteworthy that α shows a negative contribution to the RRT, while α^2 shows a large positive contribution. Therefore, one may conclude that the polarizability shows an overall positive contribution to the RRT of organic compounds. This is in agreement with the experiment because dispersion interactions play some roles in the mechanism of the RRTs obtained by using polar or non-polar columns. It should be noted that for most columns the compounds in a homologous series will elute according to the chain length. This is caused by the additional Van der Waals attractive forces resulting from the additional carbon chain length. However, the non-polar Rtx-5 column with 5% diphenyl phase is extremely versatile, permitting the analysis of non-polar to polar compounds and therefore, dispersion interactions play some roles in the mechanism of the RRTs obtained using this column. The values of the five descriptors appearing in the MLR model are shown in Table 3 for all molecules included in the training and the prediction sets.

The next step was the generation of the STANN and ANN. Before the training of these networks, the parameters of the number of nodes in the hidden layer, learning rate and momentum were optimized. The procedure for the optimization of these parameters is reported in Refs. [21,22]. Table 4 shows the architecture and specifications of the optimized STANN and ANN. In order to control the overfitting of the networks during the training procedure, the values of the SET and SEP were recorded after each 500 iterations. Fig. 2a and b shows the learning curves for the STANN and ANN, respectively. As can be seen from these figures, for the STANN, after 10 500 iterations the values of SEP started to increase and overtraining began, but for the ANN, after 128 000 iterations overtraining began. Therefore, the training of the networks was stopped at these points. Comparison of the number of iterations for the

Table 4
Architecture of the STANN and ANN and their specifications

	STANN	ANN
Number of nodes in the input layer	5	5
Number of nodes in the hidden layer	4	6
Number of nodes in the output layer	1	1
Number of iterations in the beginning of overtraining	10 500	128 000
Learning rate	0.9	0.1
Momentum	0.9	0.1
Transfer function	Sigmoid	Sigmoid

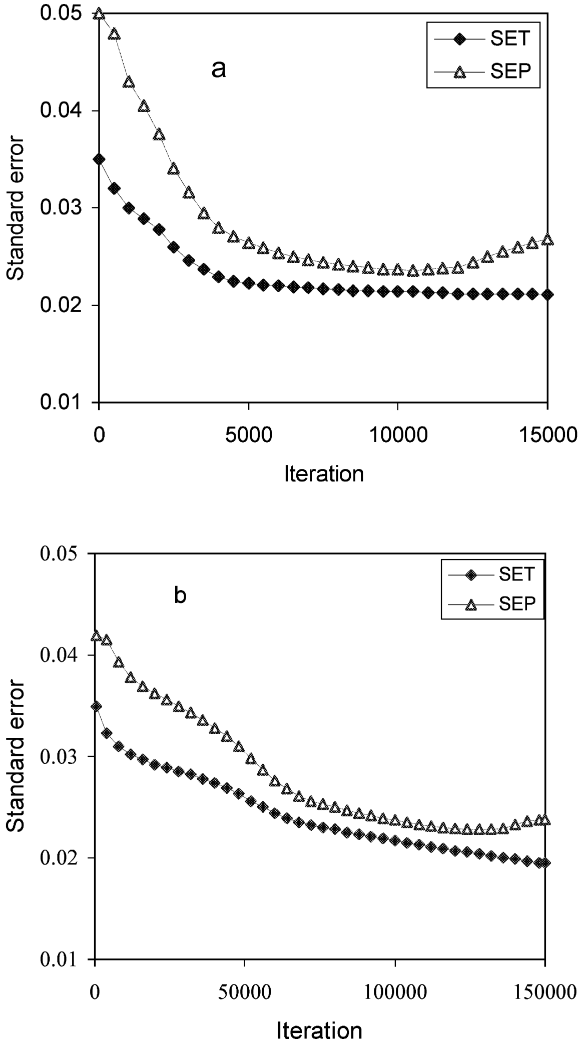

Fig. 2. A typical learning curve for (a) STANN, (b) ANN.

STANN and ANN indicate that updating of weights and optimization of the network for STANN is much faster for the ANN. Besides this advantage, in the case of the STANN, training of some networks may be performed in parallel. In addition, the topology of the STANN and ANN are 5-4-1 and 5-6-1, respectively. This means that the 29 adjustable parameters for the STANN should be compared with 43 adjustable parameters for the conventional ANN. The smaller number of adjustable parameters for the STANN reveals the validity of this model in predicting the RRT of organic compounds. For the evaluation of the prediction power of the STANN and ANN, the trained STANN and ANN were used to predict the RRT of the molecules included in the prediction set. The calculated values of the RRT using these models for the training and the prediction sets are presented in Table 1. For comparison purposes, the statistics for the STANN, ANN and MLR models are shown in Table 5. Correlation coefficients (*R*) and SEPs of these models indicate that the obtained results using STANN and ANN are much better than those obtained using the MLR model. This is believed to be due to the non-linear capabilities of the STANN and ANN.

In order to investigate the predictive ability of the generated networks, we have randomly chosen four different test sets, each consisting of 17 molecules,

Table 5
Statistical parameters obtained using the STANN and ANN and MLR models

Model	SET (%)	SEP (%)	R_{training}	$R_{\text{prediction}}$
STANN	2.135	2.364	0.996	0.992
ANN	2.036	2.279	0.995	0.992
MLR	32.027	33.326	0.960	0.951

Table 6
Comparison of the SET and SEP of the STANN and ANN models for the four different test sets with the prediction set

Method	Model	SET (%)	SEP (%)
STANN	Prediction set	2.135	2.364
	Test set I	2.182	1.873
	Test set II	2.462	1.624
	Test set III	2.117	2.453
	Test set IV	2.306	2.508
ANN	Prediction set	2.036	2.279
	Test set I	2.193	2.316
	Test set II	1.946	1.689
	Test set III	1.808	2.826
	Test set IV	2.042	1.889

and the networks were trained using the remaining molecules. The results for these test sets are given in Table 6. As can be seen from this table, the SET and SEP values for the prediction set and the test sets are similar for both STANN and ANN methods. This confirms the predictive ability of these models.

Fig. 3a and b shows the plot of the calculated RRT values versus the experimental values for the STANN and ANN models, respectively. These plots with correlation coefficients of 0.992 demonstrate the ability of these models in predicting the RRT of the molecules.

Fig. 4a and b shows the plot of the residuals against the experimental values of the RRT, for the STANN and ANN models, respectively. The propagation of the residuals on both sides of zero indicates that no systematic error exists in the development of the STANN and ANN.

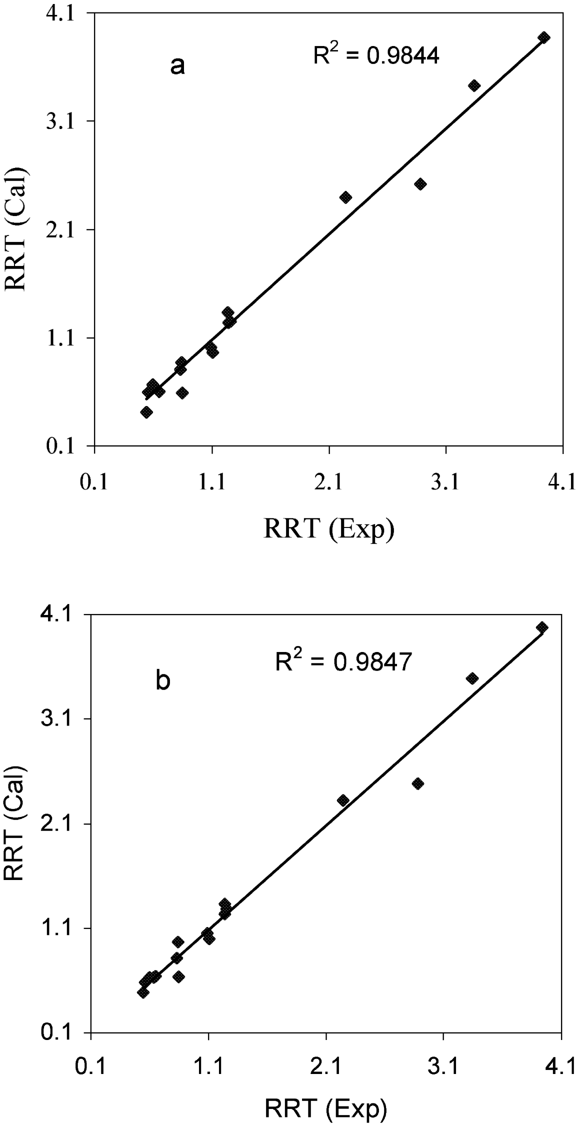

Fig. 3. Plot of the calculated RRT versus the experimental values: (a) STANN, (b) ANN.

5. Conclusions

The three methods of MLR, ANN and STANN were used for prediction of the gas chromatographic relative retention times of 13 different classes of organic compounds. As one may expect the retention behavior of organic molecules shows some non-linear characteristics and therefore applying a linear regression method cannot be justified. However, the aim of including the MLR model in this study was to choose a suitable set of numerical descriptors among the vast number of parameters available and to use them as inputs for neural network generation. As the results obtained using STANN and ANN are much better than those using the MLR method, one may conclude that the non-linear characteristics of the RRT are definite and the MLR model is a very useful method for screening the descriptors and choosing inputs for the networks. Inspection of the results obtained using STANN and ANN (Table 5) reveals that there are few differences between these methods in predicting the RRT. However, the only advantages

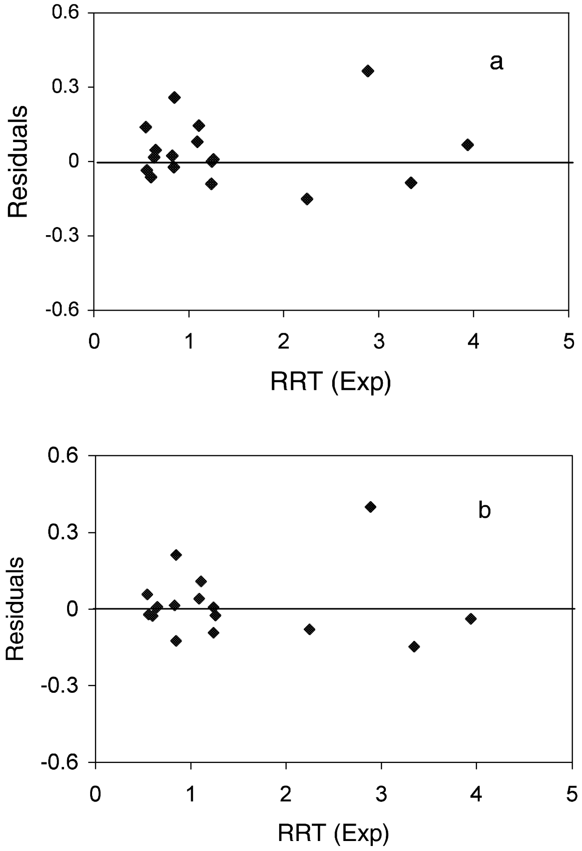

Fig. 4. Plot of the residuals versus the experimental values of RRT: (a) STANN, (b) ANN.

of the STANN over the conventional ANN are that the former can be optimized faster than the later, and also the number of adjustable parameters in the STANN is much less than in the ANN. This indicates that STANN is a reliable method for predicting the RRT of organic molecules.

References

[1] R. Kaliszman, Structure and Retention in Chromatography, Harwood, Amsterdam, 1997.

[2] D.T. Stanton, P.C. Jurs, Anal. Chem. 61 (1989) 1328.

[3] D.T. Stanton, P.C. Jurs, Anal. Chem. 62 (1990) 2323.

[4] E.K. Whalen-Pedersen, P.C. Jurs, Anal. Chem. 53 (1981) 2184.

[5] C.G. Georgakopoulos, J.C. Kiburis, P.C. Jurs, Anal. Chem. 63 (1991) 2021.

[6] J.M. Sutter, T.A. Peterson, P.C. Jurs, Anal. Chim. Acta 342 (1997) 113.

[7] A.R. Katritzky, E.S. Ignatchenko, R.A. Barcock, V.S. Lobanov, Anal. Chem. 66 (1994) 1799.

[8] B. Lucic, N. Trinajstic, S. Sild, M. Karelson, A.R. Katritzky, J. Chem. Inf. Comput. Sci. 39 (1999) 610.

[9] E.R. Collantes, W. Tong, W.J. Welsh, W.L. Zielinski, Anal. Chem. 68 (1996) 2038.

[10] T.F. Woloszyn, P.C. Jurs, Anal. Chem. 64 (1992) 3059.

[11] M. Jalali-Heravi, Z. Garkani-Nejad, J. Chromatogr. 648 (1993) 389.

[12] J. Kang, C. Cao, Z. Li, J. Chromatogr. A 799 (1998) 361.

[13] A.R. Katritzky, V. Lobanov, M. Karelson, R. Murugon, P.M. Grendze, J.E. Toomey, Rev. Roum. Chim. 41 (1996) 851.

[14] P. Payares, D. Diaz, J. Olivero, R. Vivas, I. Gomez, J. Chromatogr. A 771 (1997) 213.

[15] M. Pompe, M. Novic, J. Chem. Inf. Comput. Sci. 39 (1999) 59.

[16] D.W. Patterson, Artificial Neural Networks: Theory and Applications, Simon and Schuster, New York, 1996, Part III, Ch. 6.

[17] http://www.imagination-engines.com/adpcstanno.htm

[18] W.E. Wentworth, N. Helias, A. Zlatkis, E.C.M. Chen, S.D. Stearns, J. Chromatogr. A 795 (1998) 319.

[19] MOPAC Package, Version 6; US Air Force Academy, Colorado Springs, CO, 80840.

[20] SPSS/PC, The Statistical Package For IBMPC, Quiad Software, Ontario, 1986.

[21] M. Jalali-Heravi, M.H. Fatemi, Anal. Chim. Acta 415 (2000) 95.

[22] M. Jalali-Heravi, Z. Garkani-Nejad, J. Chromatogr. A 927 (2001) 211.