J Mol Model (2003) 9:365–371
DOI 10.1007/s00894-003-0137-x

ORIGINAL PAPER

Mohamed Nohair · Driss Zakarya

Prediction of solubility of aliphatic alcohols using the restricted components of autocorrelation method (RCAM)

Received: 19 December 2002 / Accepted: 7 April 2003 / Published online: 23 August 2003

Abstract Structure–water solubility modeling of aliphatic alcohols was performed using the multifunctional autocorrelation method. The molecule is represented by using a set of parameters describing global molecules, and others that take the structural environment of the edge O–C into account. Multiple linear regression (MLR) and multilayer feed-forward artificial neural network architectures are utilized to construct linear and nonlinear QSPR models, respectively. The optimal QSPR model was developed based on a 4–4–1 neural network architecture. The efficiency of the approach is demonstrated through the predictive ability of the ANN and MLR models by the leave-20%-out (L20%O) cross-validation method, demonstrating that the neural model is more reliable than that obtained using MLR. The root mean square errors in the solubility prediction (ln *SOL*) for the calibration and predictive models were 0.13 and 0.18 respectively. On the other hand, we tested four activation functions: the hyperbolic tangent, sigmoid function or Gaussian functions for the hidden layer and a linear, sigmoid, hyperbolic tangent or Gaussian function for the output layer. The influence and the contribution of each type of descriptor in the model is examined. After omission of a set of descriptors, we calculate the error for the solubility and classify them into discrete categories. The standard error and the percentage of the prediction in the precision interval considered have been estimated. The results imply that the solubility of aliphatic alcohols is dominated by the shape and branching of the molecule. The hydrogen-bonding interactions caused by the C–OH group seem to be a less important factor influencing the solubility. The model was compared with other models; especially that using weighted path numbers, which is considered to be the most accurate QSPR model for predicting the water solubility of aliphatic alcohols.

M. Nohair (✉) · D. Zakarya
Department of Chemistry, Faculty of Sciences and Techniques,
UFR of Applied Chemistry,
B.P. 146, 20450 Mohammadia, Morocco
e-mail: nohairm@hotmail.com

Keywords Water solubility · Aliphatic alcohols · Hydrogen-bonding interactions · Multifunctional autocorrelation method · Multiple linear regression · Neural network

Introduction

Many factors influence the physical properties of a molecule, among them are the molecular size, shape and the electronic structure. These factors are associated with various aspects of intermolecular interactions such as van der Waals forces. Numerous methods have been introduced in order to describe the chemical structure for a given set of molecules and several of them are based on the molecular graph in conjunction with the structures and properties of molecules. Among the large variety of descriptions, many are based on topological indices (TI) [1, 2, 3, 4, 5] contained in molecular graphs, usually hydrogen-depleted graphs. The topology of a chemical structure can be coded by the adjoining matrix **A**=(*aij*), where *aij* is the weight of the edge (*i*,*j*), *aij*=0 if the vertices *i* and *j* are not connected by an edge. The weight of *aij* is chosen in order to take into account the differences between the types of atoms and bonds. Another matrix **Dij**, called the distance matrix, will be defined, whose entries, *dij*, are equal to the number of edges connecting vertices *i* and *j* on the shortest path between them.

Different numbers characterizing the chemical structure of the molecule are calculated from its graph. Such numbers are called topological indices (TI). Topological indices have found a wide variety of applications in structural chemistry [6]. In particular, they can be used to code chemical information, in designing a chemical experiment, in the theory of the atomic structure and reactivity of molecules, and for the quantitative description of chemical structures in the analysis of the relation between the structure of a molecule and its properties. However, these conventional indices do not take into account the contributions of each of the individual atom

types or groups to properties and tend to obscure this fact. In this case, most of models correlating physical properties of complex compounds are unsuccessful.

The aim of this paper is to investigate the ability of the autocorrelation method to describe aliphatic alcohols in a QSPR model for the simulation of their water solubility. For molecules such as alcohols, the polarity and the ability of the molecule to participate in hydrogen bonding caused by heteroatoms may be a very important factor influencing physical properties that depend directly on the strength of intermolecular forces. We must then take into account the contribution of the individual group C–O to the physical property by adding new descriptors derived from the modified autocorrelation method.

The vector of the autocorrelation method used as a structural descriptor consists of only four components: the first two describe the global molecule; the second two encode the environment of the hydroxyl group.

For the statistical method used for deriving the model, we use a classical three-layer feedforward neural network (FFNN) trained by the back-propagation algorithm and multilinear regression (MLR). The linear model was essentially employed to investigate the behaviour of each type of component. A first model took into consideration only the usual components describing global molecules in order to see how much is gained by adding components that take the hydroxyl group of the molecule into account.

There are several models for predicting the solubility of aliphatic alcohols. The first was based on topological indices. Kier and Hall [7] constructed a simple model by introducing the vertex-connectivity index χ in Randic's original formulation [3]. The model (1) obtained was good enough for practical purposes because only one descriptor was considered. It leads to the following equation:

$$\ln\ Sol = 6.702 - 2.666\chi, n = 51, r = 0.978 \text{ and } s = 0.455$$

The vertex-connectivity index for aliphatic alcohols considered in this equation was computed for the related alkanes.

However, to derive more significant models to estimate water solubility of alcohols, it seemed logical to consider parameters that take the hydroxyl group in the molecule into account. Another parameter was added that defines whether we are dealing with primary, secondary or tertiary alcohols. Monoparametric correlation with the water solubility of alcohols improves significantly (r=0.991 and s=0.289)

Another model based on the surface of the molecule was recently proposed by Amidon et al. [8] who considered 51 aliphatic alcohols. The first model uses one descriptor representing the total surface area (TSA) in Å^2. The model produced a coefficient r close to 0.974, while the standard deviation s was close to 0.706. However, a model relating ln *Sol* to two parameters, the contribution of the hydrocarbon surface area and the hydroxyl group, has better statistical characteristics (r=0.978, s=0.462).

Table 1 The count of paths and weighted paths representing the hydrogen-depleted skeleton of 3,3-dimethyl-1-butanol

Vertex	P1	P2	P3	P4
1	1+x	1	3	
2	2	3+x		
3	4	1	x	
4	1	3	1	x
5	1	3	1	x
6	1	3	1	x
7	x	x	x	3x
Weighted path numbers	5+x	7+x	3+x	3x

The most accurate model uses weighted path numbers [9] combined with multilinear regression. A weighted path contains weighted edges. In the case of aliphatic alcohols, the weight of the edge representing the C–O bond is taken to be x, while the weights of the edges representing C–C bonds are equal to one. The length of a path containing the edge with weight x is simply denoted by x. For example, the weighted path numbers for 3,3-dimethyl-1-butanol (Table 1) are equal to the sum of weighted paths over all vertices in a graph divided by 2, x is left undefined and it will be adjusted by the regression analysis to obtain the smallest standard error. The optimum standard error of estimate, s for various values of x for four weighted paths (P_1, P_2, P_3, P_4) is found to be 0.216.

The examples listed below show that models that use descriptors that take not only molecular size but also the hydroxyl group into account give good statistical characteristics. Based on this observation, we use the autocorrelation method to develop two types of descriptors. In the first, we compute the autocorrelation vector based on the van der Waals volume of the alkane molecule corresponding to aliphatic alcohol considered. This gives us a global description of the molecular environment in space (molecular size, shape and branching). Then we compute an autocorrelation vector for the hydroxyl group within the molecule.

Method

The modified autocorrelation method (MAM) derived from the multifunctional autocorrelation method of Moreau et al. [10] was used in the structure–property relationships. The autocorrelation vector generated was used as a possible descriptor in QSARs and can be useful for database characterization and encoding various physicochemical properties. [11, 12, 13, 14]

The general relation used to calculate the autocorrelation component is defined below:

$$P_k = \sum_{k=0}^{n} (f(i)f(j))^x \quad (1)$$

where P_k is the autocorrelation component corresponding to the topological distance of k bonds (smallest number of bonds between i and j) to the specific property $f(i)$. The atomic contribution $f(i)$ depends on the property under study. These properties can be based, as an example, on the van der Waals volume (V) and surface (S) to account for the size and the shape of the molecule,

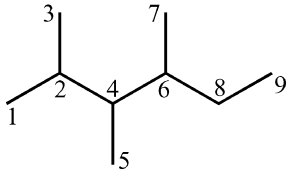

Fig. 1 A hydrogen-depleted molecular graph corresponding to the skeleton of 2,3,4-trimethylhexane

connectivity (number of non-hydrogen neighbors or vertex degree of the atom *i*) or electronegativity and charge to account for the electronic aspect.

To describe the local environments of atoms in a molecule, we define, similar to P_k, a new component P_{ik} by means of the following formula:

$$P_{ik} = \sum_{k=0}^{n} (f(i)f(j))^x \quad (2)$$

We compute P_{ik} by fixing the atom *i*, P_{ik} is defined as the sum of *f(i)*f(j)* of all chemical bonds existing between all pairs of carbon atoms *i* (fixed atom) and *j* separated by a topological distance equal to *k*.

Example of computation of P_k and P_{ik} components

2,3,4-Trimethylhexane (Fig. 1) was chosen to illustrate the calculation of components of autocorrelation vectors. The properties of different atoms present in the molecule (CH_3, CH_2, CH, and C) are given in Table 2.

Taking 2,3,4-trimethylhexane, shown in Fig. 1, for example, the procedure of computing the components of autocorrelation vectors, using the MAM method, is illustrated as follows. The properties of different atoms present in the molecule (CH_3, CH_2, CH, C) are given in Table 2.

f(i)=van der Waals volume

There are nine carbon atoms. Based on connecting C–C bond numbers between two atoms, they can be classified as four basic types: types 1, 2, 3 and 4 for primary (CH_3–), secondary (CH_2<), tertiary (–CH<), and quaternary (>C<) carbons, then we get for *k*=0:

$$V_0(k=0) = 5 \times (13.67) + 1 \times (10.23) + 3 \times (6.78) = 98.92$$

For a topological distance equal to 1 (*k*=1), there are eight pairs (*i*, *j*) of atoms [(1, 2); (2, 3); (2, 4); (4, 5); (4, 6); (6, 7); (6, 8); (8, 9)] and we compute V_1 as follows:

$$V_1(k=1) = [4(13.67 \times 6.78)^{1/2} + 2(6.78 \times 6.78)^{1/2} + (6.78 \times 10.23)^{1/2} + (10.23 \times 13.78)^{1/2}] = 72.22$$

For a local atom environment, we consider as an example only carbon atom number 4 (Fig. 1) to compute its components by using descriptors based on van der Waals volume as atomic properties

f(i)=van der Waals volume

P_{ik} is defined as V_{ik}

– For *k*=1

$$V_{41} = \left[2(6.76 \times 6.76)^{1/2} + (6.76 \times 13.67)^{1/2}\right] = 23.18$$

This corresponds to three pairs of atoms (4, 2), (4, 6) and (4, 5) with a topological distance equal to 1.

– For *k*=2

$$V_{42} = \left[3(6.76 * 13.67)^{1/2} + (6.76 * 10.23)^{1/2}\right] = 37.15$$

– Etc.

Experimental section

Data set

The data set used in this paper was reported by Amic et al [9]. They showed the chemical structure of 50 aliphatic alcohols and developed a model based on weighted path numbers. The 50 aliphatic alcohols (Table 3) consist of different types of structures: aliphatic (linear and branched), primary, secondary and tertiary carbon. The water solubility used in this work were expressed in ln *Sol*. Their values ranges from −8.2208 for decanol up to 0.0953 for the 1-butanol.

In order to simplify the computation of the components, molecules were coded by means of the SMILES system [15] and stored as input files. The computer program used to compute the components represents an algorithm for the construction of the connectivity matrix of any molecule from its SMILES code. The linear regression and the neural network were performed using Excel statistical procedures from the Microsoft office and Quiknet packages, respectively.

Linear regression

Both $P_k(V_k)$ and $P_{ik}(V_{ik})$ (the letter *P* is replaced by the letter *V*) components are computed for all molecules. The first takes the global description of the molecule's environment in space into account, the second that of the hydroxyl group within the molecule. In the first, a linear modeling method was used to investigate the behaviour of the V_k components. They were based on van der Waals volumes (V_0–V_5). Generally, the model offers

Table 2 Contributions of atoms to some molecular properties

No. of atoms in molecule (Fig. 1)									
Property	1	2	3	4	5	6	7	8	9
van der Waals volume *V* (cm^3 mol^{-1})	13.67	6.78	13.67	6.78	13.67	6.78	13.67	10.23	13.67

Table 3 The autocorrelation vector, experimental, calculated (calibration and cross-validation) water solubility of 50 aliphatic alcohols

No.	SMILES code	Autocorrelation vector				ln *Sol*				
		V_1	V_2	V_{11}	V_{12}	Exp	MLR_{cal}	MLR_{cv}	ANN_{cal}	ANN_{cv}
1	occ(c)c	39.41	46.95	11.83	9.63	0.0227	0.1422	0.1162	0.1217	–0.1180
2	oc(c)cc	39.41	46.95	9.63	25.50	0.0658	0.5987	0.6310	0.2185	0.1538
3	occc(c)c	49.64	57.47	11.83	11.83	–1.1680	–1.2708	–1.2710	–1.1648	–1.1410
4	occ(c)cc	49.93	53.14	11.83	9.63	–1.0584	–1.1531	–1.1470	–1.0955	–1.1210
5	oc(c)ccc	49.64	57.47	9.63	25.50	–0.6349	–0.7320	–0.7494	–0.6130	–0.6490
6	oc(cc)cc	49.93	53.14	9.63	23.65	–0.4861	–0.6049	–0.6085	–0.4651	–0.4144
7	oc(c)c(c)c	45.29	65.85	9.63	23.30	–0.4050	–0.5101	–0.5036	–0.3518	–0.3648
8	oc(c)(c)cc	37.90	83.23	6.75	39.17	0.3386	0.2524	0.2206	0.0294	–0.1188
9	oc(cc)ccc	60.16	63.66	9.63	23.65	–1.8326	–1.9364	–1.9377	–1.9013	–1.8535
10	oc(c)(cc)cc	48.82	84.70	6.75	37.32	–0.8301	–0.9044	–0.8676	–0.8569	–0.8336
11	oc(c)(c)ccc	48.13	94.15	6.75	39.17	–1.1178	–1.0752	–1.1000	–1.0186	–0.9385
12	oc(c(c)c)cc	55.82	72.33	9.63	21.45	–1.6094	–1.7106	–1.7475	–1.6125	–1.6072
13	oc(c)c(c)cc	55.82	72.33	9.63	23.30	–1.6399	–1.8023	–1.7848	–1.6844	–1.6854
14	oc(c)(c)c(c)c	44.25	97.06	6.75	36.97	–0.8510	–0.8072	–0.7902	–0.7627	–0.6018
15	occc(c)(c)c	48.13	94.15	11.83	11.83	–2.5903	–2.1598	–2.0913	–2.5039	–2.5648
16	oc(c)c(c)(c)c	44.25	97.06	9.63	20.42	–1.4106	–1.2893	–1.2965	–1.3408	–1.3197
17	occcc(c)c	59.87	67.70	11.83	11.83	–2.2828	–2.5651	–2.5576	–2.5862	–2.7410
18	oc(c)cc(c)c	55.17	81.42	9.63	25.50	–1.8140	–1.9589	–1.9855	–1.7927	–1.9794
19	occ(cc)cc	60.46	59.57	11.83	9.63	–2.7871	–2.4452	–2.3871	–2.6856	–2.5300
20	oc(c)(c)cccc	58.36	104.38	6.75	39.17	–2.4734	–2.3936	–2.4204	–2.3896	–2.4615
21	oc(c)(cc)ccc	59.05	95.61	6.75	37.32	–2.2634	–2.2325	–2.2335	–2.2211	–2.1468
22	oc(cc)(cc)cc	59.73	86.41	6.75	35.48	–1.9173	–2.0555	–2.0088	–1.9828	–2.0515
23	oc(c)(c)c(c)cc	54.77	103.93	6.75	36.97	–2.0025	–2.0950	–2.0535	–2.0087	–2.0489
24	oc(c)(cc)c(c)c	55.16	98.81	6.75	35.12	–1.9379	–1.9586	–1.9138	–1.9073	–1.9639
25	oc(c)(c)cc(c)c	53.66	118.56	6.75	39.17	–2.1456	–2.2996	–2.3296	–2.2908	–2.3955
26	oc(c(c)c)c(c)c	61.70	91.88	9.63	19.25	–2.8018	–2.8121	–2.8825	–2.7552	–2.8267
27	oc(cc)c(c)(c)c	54.77	103.93	9.63	18.57	–2.6437	–2.4855	–2.5619	–2.7061	–2.6232
28	oc(cc)cccc	70.39	73.89	9.63	23.65	–3.1942	–3.2536	–3.2731	–3.2174	–3.1674
29	oc(ccc)ccc	70.39	74.19	9.63	23.65	–3.1966	–3.2682	–3.2639	–3.2317	–3.1445
30	oc(c)(cc)c(c)(c)c	54.73	124.65	6.75	32.24	–2.9318	–2.6806	–2.5343	–2.9313	–2.9964
31	oc(c)cccccc	80.33	88.16	9.63	25.50	–4.7560	–4.6820	–4.7117	–4.7057	–4.8111
32	occ(cc)cccc	80.92	80.33	11.83	9.63	–4.9967	–5.0757	–5.0658	–4.7980	–5.0530
33	oc(c)ccccccc	90.56	98.39	9.63	25.50	–6.3200	–5.9983	–5.9851	–6.2780	–6.1767
34	oc(cc)cccccc	90.85	94.35	9.63	23.65	–6.1193	–5.8863	–5.8581	–5.9609	–5.9287
35	oc(ccc)ccccc	90.85	94.65	9.63	23.65	–5.9522	–5.9018	–5.9785	–5.9915	–6.0228
36	oc(cccc)cccc	90.85	94.65	9.63	23.65	–5.7446	–5.9026	–5.9574	–5.9915	–6.1090
37	oc(cc(c)c)cc(c)c	81.45	122.08	9.63	23.65	–5.7764	–5.7235	–5.7879	–5.7317	–5.4416
38	oc(c(c)cc)c(c)cc	82.75	104.85	9.63	19.25	–5.2983	–5.3991	–5.3810	–5.2631	–5.3916
39	occ(cc)(cc)ccc	80.88	99.28	11.83	6.75	–5.5728	–5.7285	–5.5789	–5.6834	–5.6023
40	occccccc(c)c	90.56	98.39	11.83	11.83	–5.7446	–6.5142	–6.5477	–5.8201	–6.1776
41	occc(c)cc(c)(c)c	74.42	135.27	11.83	11.83	–5.7699	–5.9234	–5.9521	–5.7514	–5.4932
42	occcc	44.11	33.88	11.83	11.83	0.0953	–0.0407	–0.0066	0.0384	–0.0519
43	occccc	54.34	44.11	11.83	11.83	–1.3471	–1.3570	–1.3569	–1.3068	–1.1813
44	occ(c)(c)c	37.90	83.23	11.83	6.75	–0.6463	–0.6863	–0.8173	–0.6972	–0.8952
45	occcccc	64.57	54.34	11.83	11.83	–2.7181	–2.6734	–2.6886	–2.7287	–2.8620
46	oc(c)cccc	59.87	67.70	9.63	25.50	–1.9951	–2.0492	–2.0572	–1.9423	–1.8259
47	occccccc	74.80	64.57	11.83	11.83	–4.0745	–3.9897	–3.9847	–3.9607	–4.0885
48	occcccccc	85.03	74.80	11.83	11.83	–5.4015	–5.3061	–5.2806	–5.4628	–5.3653
49	occcccccc	95.26	85.03	11.83	11.83	–6.9078	–6.6224	–6.5557	–7.0191	–6.9940
50	occcccccccc	105.49	95.26	11.83	11.83	–8.2208	–7.9388	–8.0047	–7.8950	–7.7613

us an adequate estimation of the water solubility of all the aliphatic alcohols (n=50; r=0.97; s=0.38).

The intercorrelation matrix of the components (V_0–V_5) for the 50 alcohols shows a high intercorrelation coefficient between V_0 and V_1. This indicates that these components express approximately the same type of structural information. V_0 represents only the size of the molecule, while the component V_1, defined as the sum of $f(i)*f(j)$ of all chemical bonds existing between all pairs of adjoining carbon atoms in the molecule under consideration, takes the molecular size and the branching of the molecule into account.V_0, V_3, V_4, and V_5 contribute poorly, so we define a new model in which components V_0, V_3, V_{4i} and V_{i5} are not considered and confirmed that the quality of regression did not change (r=0.97, s=0.38)

The model can be considered good enough. However, to derive a more significant model to estimate the water solubility, it seemed logical to consider the contribution of the edge O–C. When we add components related to the description of the hydroxyl group V_{ik} (k=1, 2) (V_{i0} was the same for all molecules), the model is much improved over the above model. The standard error estimated decreases

Table 4 Percentage of prediction in the interval [0–0.3] and >0.3 by considering respectively components (V_1, V_2,) and (V_1, V_2, V_{i1}, V_{i2})

Descriptor considered	Standard deviation (r^2, rms)	Percentage of deviation	
		[0–0.3]	[0–0.2]
(V_1, V_2,)	(0.99; 0.38)	27/50	17/50
(V_1, V_2, V_{i1}, V_{i2})	(0.99; 0.21)	47/50	44/50

dramatically from s=0.38 to s=0.21. For components P_{ik} (k>=3), there is not much difference, so we selected, according to the principle of Occam's razor [16], a model with the least number of parameters, that is, the four-descriptor model (V_1, V_2, V_{i1}, V_{i2}) for computing the water solubility for aliphatic alcohols. We calculate the error for each of the 50 values of the water solubility and classify them into a discrete category (here in [0–0.2], and [0–0.3]). The standard error and the percentage of the prediction in the precision interval considered are shown in Table 4. For the model based on components belonging to the second type, regressions were very much worse. In the above results, the V_k components up to k=2 represent factors of prime importance in the modeling of the water solubility of aliphatic alcohols.

We also used the leave-20%-out cross validation technique as a criterion for checking the quality of the model. In this procedure 20% of the whole set data were selected out one after another. For every selection the model was built with the remaining 80% examples. Next, the model was used to predict the water solubility for the selected molecules. This procedure is repeated five times until all patterns are selected in a prediction set once and only once. The combined results of the solubility values estimated gave information on the prediction ability. A linear regression between experimental and predicted values leads to the following results. The cross-validation standard coefficient r between ln Sol_{exp} and ln Sol_{calc} is 0.98, while the cross-validation standard s and the mean error are equal to 0.22 and 0.15, respectively. Fitted values by using cross validation are shown in column 7 of Table 3.

Neural network

Artificial neural networks (ANNs) appear to be very promising in obtaining models that convert structural features into different properties of chemical compounds. In the field of QSPR, the application of ANNs has demonstrated their efficiency compared to traditional multilinear regression, particularly in cases where it is difficult to specify an exact mathematical model that describes a structure–property relationship. In this paper, we show strong evidence that the mutifunctional autocorrelation method is useful for ANN modeling of the water solubility of aliphatic alcohols.

The main advantage of using neural networks in QSPR models is their capacity to offer a non-linear mapping of the structural descriptors to the physicochemical property. It is interesting to see how much is gained with ANNs in the prediction power of water solubility of aliphatic alcohols in comparison to the linear regression method.

The computational neural network used in this study was a three-layer (input-hidden-output), fully connected, feed-forward network. The input layer contains one node for each structural descriptor (autocorrelation component). The size of the input layer of the network is determined by the length of the code used to describe the environment of the molecule. The output layer has one node generating the scaled estimated value of the water solubility of the molecule considered. Although there are neither theoretical nor empirical rules to determine the number of hidden layers or the number of neurons in this layer, one layer seems to be sufficient in most chemical applications of ANNs. The number of hidden neurons needs to be sufficient to ensure that the information contained in the description data is adequately represented.

Four types of activation function were used: the hyperbolic tangent or a sigmoid or Gaussian function for the hidden layer and a hyperbolic tangent, sigmoïd or linear function for the output layer. A bias neuron was used in the input layer connecting to all neurons of the hidden layer and to the output layer.

The weights of connections between the neurons were initially assigned random values uniformly by using the standard back-propagation method [17]. The training was initiated and followed by examining the RMS error (RMS stands for root mean square, that is the square root of the average residual) for the total set and also for both the training and the test sets. Training was stopped when there was no further improvement in the test set RMS error. We also computed the correlation coefficient between the observed and predicted values.

Recent work, [18] based on empirical observations suggest that only networks with a ρ parameter greater than 2 (ρ is the ratio of the number of patterns in the training to the number of connections) should be used for QSPR in order to ensure that the network can give reliable predictions. The few data and the number of components (V_1, V_2, V_{i1}, V_{i2}) representing the vector of the autocorrelation method force us to search for the optimal network architecture according to the above limits by varying the number of hidden neurons from 1 to 4. In order to establish the optimal size of the hidden layer (the number of the neurons), a first application uses all the components (V_1, V_2, V_{i1}, V_{i2}) as input values by applying a cross-validation technique. This was carried out iteratively. We started from one neuron in the hidden layer, the statistical indices of the correlation between experimental and predicted water solubility for the whole data set (50

Table 5 Statistical results, outliers and their residuals for ANNs with 4–4–1 architecture and the MLR model: model calibration and prediction (leave-10%-out cross-validation method)

	MLR		ANN	
	MLR_{cal}	MLR_{cv}	ANN_{cal}	ANN_{cv}
(*r*, *s*, rms)	(0.99; 0.21; 0.12)	(0.98; 0.22; 0.15)	(0.99; 0.11; 0.08)	(0.99; 0.18; 0.13)
Percentage of prediction				
[0–0.1]	27	28	37	23
[0.1–0.2]	19	10	9	15
[0.2–0.3]	1	5	4	6
>0.3	3	7	0	6 (Ln_{Sol}<0.5)

Table 6 Statistical results, outliers and their residuals obtained using autocorrelation method and model based on weighted path numbers method for the calibration model

	Autocorrelation method		Weighted path numbers method	
	MLR_{cal}	ANN_{cal}	MLR_{cal}	ANN_{cal}
(*r*, *s*, rms)	(0.99; 0.21; 0.12)	(0.99; 0.11; 0.08)	(0.99; 0.21; 0.17)	(0.99; 0.14; 0.09)
Percentage of prediction				
[0–0.1]	27	37	25	36
]0.1–0.2]	19	9	12	11
]0.2–0.3]	1	4	8	1
>0.3	3	0	5	2 (Ln_{Sol}>0.4)

values) improves with increasing number of neurons. The optimal number of neurons in the hidden layer was found to be four with 5,000 iterations for a calibration model. The standard error of the estimate (RMS) was used as a criterion for the selection of the optimum number of hidden neurons.

Except for the linear output function, all four combinations of activation functions listed above gave close results for the calibration model, but we have observed that the hyperbolic function performs better. Thus, each neuron of a given layer (except the input one) takes a value Y calculated by using the following transfer function:

$$O_i = 1/\left(1 + \exp\left(-\left(\sum W_{ij}O_j + \theta_j\right)\right)\right) \quad (3)$$

where O_i and O_j are the outputs of neurons i and j, respectively, W_{ij} is the weight connecting neurons i and j, and θ_j is the bias of neuron j.

In order to compare the performance of the ANN model with the statistical results of the MLR equation, we have used the correlation coefficient r, the standard deviation s and the mean residual $mres$ of the linear correlation between experimental and predicted water solubility. On the other hand, we have used the same set of structural descriptors consisting of four components of the autocorrelation method. The comparison shown in Table 5 shows the neural network provides better calibration (4–4–1, number of epochs=5,000, r=0.99, s=0.11). For resitual values, we divided them into different discrete categories [0–0.1],]0.1–0.2],]0.2–0.3] and >0.3 as absolute residual. For the power predictive, the L20% cross-validation was performed for the ANN model by conserving the same partitioning of the patterns. The statistical indices were again superior to those obtained for the MLR model. We conclude that there is a nonlinear relation between the structural descriptor derived from the autocorrelation and the water solubility of the aliphatic alcohols. Consequently, the use of the ANN is justified.

To compare our approach to other models, we have selected the work based on the weighted path numbers. The size of the code used to describe the environment of the aliphatic alcohols consists of four parameters [9]. Only a model calibration for the multilinear regression and the neural network is considered. Our comparison is limited to the set considered in this work. On the whole, results obtained by using the autocorrelation method are better. In the MLR model, both models give almost the same statistical parameters, but the distribution of the fitted values is different. If we consider the number of cases having an error in prediction lower than 0.2, we find 46 cases in the autocorrelation method compared to 37 cases for the weighted path numbers model. On the other hand, the latter method gives two fitted values for which the error deviation exceeds three times the standard deviation in the ANN model. Results are reported in the Table 6. The success of the autocorrelation method depends on the correct description brought by each component. Two types of components are considered separately, the first ones describe the global molecule, and the second ones encode the local environment of the hydroxyl group.

Conclusion

In conclusion, the ANN approach using the multifunctional autocorrelation method with restricted components gives both a useful and simple mathematical model for the prediction of the water solubility of aliphatic alcohols. We are interested in this property because the toxic action of these compounds depends mainly on their solubility in water. A small number of topological parameters (four

components of the autocorrelation method) is efficient to take into account all topological considerations of the molecule. It has been demonstrated that water solubility appears to be largely determined by the first components (V_1 and V_2), which represent the size and the branching of the molecule disturbed by the hydrogen-bonding interactions caused by C–OH.

References

1. Rouvary DH (1997) J Comput Chem 4:470–480
2. Labanowski JK, Motoc I, Dammkoehler RA (1991) Comput Chem 1:47–53
3. Randic M (1975) J Am Chem Soc 97:6609–6615
4. Kier LB, Hall LH (1986) Molecular-connectivity in structure–activity analysis. Wiley, New York
5. Devillers J, Balaban AT (1999) Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach
6. Todeschini R, Consonni V (2000) The handbook of molecular descriptors. Wiley-VCH, New York
7. Kier LB, Hall LH (1976) J Pharm Sci 65:1806–1809
8. Amidon GL, Yalkowsky SH, Leung S (1974) J Pharm Sci 63:1858–1866
9. Amic D, Basak SC, Lucic B, Nikolic S, Trinajstic N (2002) SAR QSAR Environ Res 2:281–296
10. Moreau G, Broto P (1980) Nouv J Chim 4:359–360
11. Moreau G, Broto P (1980) Nouv J Chim 4:757–764
12. Chastrette M, Tiyal F, Peyraud JF (1990) C R Acad Sci Paris Ser II 310:514–515
13. Zakarya D, Tiyal F, Chastrette M (1993) J Phys Org 6:574–582
14. Nohair M, Zakarya D (2002) J Comput Chem 42:586–591
15. Weininger D (1988) J Chem Inf Comput Sci 28:31–36
16. Hoffman R, MinKin VI, Carpenter BK (1996) Bull Soc Chim Fr 133:117–130
17. Rumelhart DE, Hinton GE, Williams RJ (1986) 323:533–536
18. Livingstone DJ, Manallack DT (1993) J Med Chem 36:1295–1297