Asian Journal of Chemistry; Vol. 25, No. 5 (2013), 2604-2610

ASIAN JOURNAL OF CHEMISTRY

http://dx.doi.org/10.14233/ajchem.2013.13515

Quantitative Structure-Property Relationship for Predicting Surface Tension of Organic Compounds Using Associative Neural Networks

P. NEELAMEGAM[1,*] and S. KRISHNARAJ[2]

[1]School of Electrical and Electronics Engineering, SASTRA University, Thirumalaisamudram, Thanjavur-613 401, India
[2]PRIST University, Thanjavur-613 403, India

*Corresponding author: E-mail: neelkeer@yahoo.com

(*Received*: 27 January 2012; *Accepted*: 15 November 2012) AJC-12412

This paper explains associative neural network based quantitative structure property relationship study for prediction of surface tension of organic compounds using molecular descriptors derived from molecular structures. A set of 116 organic compounds, which includes 48 alkanes, 31 alcohols, 20 amines, 14 alkenes and 3 aldehydes as data series are selected for the present study. Unsupervized forward selection strategy is used for descriptor selection from the large set of descriptors using E-DRAGON software and six descriptors are selected for model development for surface tension. Associative neural network method is used to construct the non-linear prediction model for surface tension. The selected descriptors are used as input data for training and testing the associative neural network. The predicted results are in good agreement with the experimental surface tension of organic compounds with squared correlation co-efficient (R^2) of 0.98 for training and 0.932 for testing. The results are cross-validated by leave-one-out procedure. The model is suitable to a large variety of compounds, which predicts better than other models reported in previous studies.

Key Words: Quantitative structure-property relationship model, Surface tension, Descriptors, Associative neural network.

INTRODUCTION

Surface tension is an important thermo-physical property in the chemical process industry. Surface tension of a liquid characterizes the free energy per unit area required for the formation of a liquid-air interface at constant temperature, pressure and composition. Surface tension data is used in many engineering applications such as mass transfer operations including distillation, liquid-liquid extraction, adsorption and absorption. Accurate and reliable values of surface tension are necessary for optimal design of the equipment, this leads to better operation and ultimately reduction in costs[1,2]. For the chemical engineer, surface tension determines the quality of many of the products resulting from the different industries such as those producing coatings, paints, detergents, cosmetics and agrochemicals, but also affects some important steps in the production process. Since experimental measurements of surface tensions are often unavailable, expensive and time-consuming, theoretical models are regularly used[3]. The theoretical prediction of the surface tension of organic compounds is required in many chemical engineering calculations.

Surface tension is closely related to the effects of intermolecular interaction in bulk liquid of organic compounds. It should be noted that London dispersion forces are responsible for intermolecular interaction. London force is an attractive force that results when electrons in two adjacent atoms form temporary dipoles; the London potential is given by:

$$U_{attr}(r) = -\frac{3\alpha^2 I}{4(4\pi\varepsilon_0)^2}\frac{1}{r^6} \quad (1)$$

where, α is the atomic polarizability ($\alpha = 4\pi\varepsilon_0 r^3_m$, where r_m is the radius of the molecule), I is the ionization potential, r is the distance between molecules[4].

Quantitative structure property/activity (QSPR/QSAR) study has become one of the most explored areas of research in computational chemistry in past couple of decades. A fundamental goal of QSPR/QSAR studies is to predict complex physical, biological, chemical and technological properties of chemical compounds from simpler descriptors, preferably those calculated solely from molecular structure. The general idea of QSAR and QSPR is that property/activity of a new untested molecule can be readily estimated from the molecular structure of similar compounds whose properties/activities have already been determined. The correlation between the properties/activities of the molecules and their structures are necessary to obtain reliable models. If a good correlation is found, then it would be easy to determine the properties/activities of various compounds, including those not yet

synthesized[5]. Quantitative structural property relationship studies are performed on the basis of the correlation between the experimental values of the property and molecular descriptors reflecting the molecular structure of the respective compounds. The descriptors provide quantitative information about the properties of models. Rigorous testing of the predictive power of the equations obtained is possible. Hence the QSPR approach is a general and reliable method to predict various physico-chemical properties[6,7]. To develop a QSPR model the following steps are involved *i.e.*, data collection, molecular geometry optimization, molecular descriptors generation, selecting the best descriptors, model development and finally model performance evaluation (validation)[8]. One of the important problems in QSPR is the description of molecular structures using molecular descriptors, which can include structural information as much as possible. Theoretical descriptors such as constitutional descriptors and topological indices have found the major popularity in QSPR studies for several reasons such as i) their calculation is simple and fast, ii) they do not need information about three dimensional structure of molecules, iii) They are exact number without uncertainty and iv) they represent high correlation with many physico-chemical properties[9]. Recently, various statistical methods such as multiple linear regression, cluster analysis, Principal component analysis and partial least square regression have been applied to the QSPR studies. For the prediction of physical properties, high-quality models are obtained based on predictive equations using linear regression techniques, are used to correlate structure related descriptors with observed properties. Currently, neural networks are used with encouraging success in development of various QSPR models. An artificial neural network represents non-linear methods are well suited to describe structure-property models. Moreover, artificial neural network is able to consider not only particular structure characteristics, but also interrelations and interdependencies between mutually influencing structural features. Therefore, they can be easily adapted for processing large data set formed by a set of descriptors[10,11].

Several papers are reported in the literature for prediction of surface tension using multivariate regression. Stanton and Jurs[12] have designed a multivariate regression based 10 descriptor model for a data set of 146 structures having a R^2 of 0.983. Kauffman and Jurs[13] published a paper which employs multiple linear regression as a tool for prediction of surface tension of 159 structures. Their model involves eight descriptors yielded a squared correlation coefficient of 0.83. Delgado and Diaz[14] report a six descriptor model for a data set of 320 chemicals with a R^2 of 0.96 using multiple linear regression (MLR) methods. Only few works are available in literature for prediction of surface tension using neural network Hence, this paper explains associative neural network based prediction of surface tension of organic compounds based on 6 molecular descriptors provided by E-DRAGON[15] having specific physical meaning corresponding to different molecular interactions occurring in the bulk solution are reported.

EXPERIMENTAL

Data set: The experimental surface tension data set for the 116 organic compounds considered in this study are compiled from the published literature[16]. The heterogeneous data set includes alkanes, alcohols, amines, alkenes and aldehydes. Tables 1 and 2 shows the compounds and their corresponding experimental surface tension as logarithmic values at 20 ºC. The data set is randomly divided into two subsets: the training set containing 90 compounds (78 %) and the test set containing 26 compounds (22 %). The training set is used to build models using associative neural network. The test set is used evaluate the predictive ability of the models obtained.

Molecular descriptors: The chemical structures of the 116 organic compounds are drawn with MarvinSketch[17] and exported as SMILES notation. Next, organic compounds represented by SMILES format are used as input for calculation of various types of descriptors with the online software, E-DRAGON[15], which converts the molecules from SMILES notation to 3-dimensional structures.

Selection of molecular descriptors: The challenge in developing the QSPR model is the selection of molecular descriptors from the pool of available descriptors that strongly correlate with desired physical property. As the number of descriptors increases, the capability of prediction analysis methods decreases because of the redundancy of information incorporated by the different descriptors. Techniques to minimize the problem, such as principal component analysis and Unsupervized Forward selection, have been used in QSPR model development.

Unsupervized forward selection is a data selection procedure by eliminating redundant variables, that selects from a data matrix a maximal linearly independent set of columns with a minimal amount of multiple correlations. Unsupervized forward selection was designed for use in the development of QSPR models, where the m by n data matrix contains the values of n variables (typically molecular properties) for m objects (typically compounds). QSPR data sets often contain redundancy (exact linear dependencies between subsets of the variables) and multicollinearity (high multiple correlations between subsets of the variables). Both of these features inhibit the development of QSPR models with the ability to generalize successfully to new objects. Continuum regression, an algorithm encompassing ordinary least squares regression, regression on principal components and partial least squares regression, is used to construct models from the selected variables. Unsupervized forward selection produces a reduced data set that contains no redundancy and a minimal amount of multicollinearity. The variable selection routine is shown to produce simple, robust and easily interpreted models for the chosen data sets. The developed algorithm of unsupervized forward selection is available online at[18].

Neural network: Artificial neural network is one of recently emerged directions in the field of information processing technology. It has a origin in efforts to produce a computer model of the information processing that takes place in the nervous system. In many applications, including the present work, the biological relevance of neural networks of nervous system function is unimportant. A neural network may simply be viewed as a highly parallel computational device and is found to be useful in a variety of tasks including solving certain optimization problems and pattern recognition. The

artificial neural network are trained to perform a particular function by adjusting the values of the connections, or weights, between elements until a particular input leads to a specific output. The artificial neural network consists of input layer, hidden layer and output layer. These three layers are connected with each other. The input layer receives the input data outside the network and sends them to the hidden layer. The hidden layer contains interconnected neurons for the pattern recognition and the relevant information interpretation for adjusting the weights on the connections. Afterwards, the results from the hidden layer are sent to the output layer for the outputs. The neurons contain several functions and variables including weights, non-linear transfer functions, methods to add up all inputs and bias values. The sum of all products of all the inputs multiplied with the weights and the bias values passes through a non-linear transfer function as the output of each neuron[19].

Associative neural network: The traditional artificial feed forward neural network is a memory-less approach. This means that after training is complete, all information about the input patterns is stored in the neural network weights and input data are no longer needed, *i.e.* there is no explicit storage of any presented example in the system. Contrary to that, associative neural network is a method with improved predictive abilities including combination of memory-based and memory-less method. It offers an elegant approach to incorporate "on the fly" the user's data[9]. The associative neural network is an extension of the committee of machines that goes beyond a simple/weighted average of different models. An associative neural network represents a combination of an ensemble of feed forward neural networks (memory-less) and the K-nearest neighbour technique (memory-based). It uses the correlation between ensemble responses as a measure of distance among the analyzed cases for the nearest neighbour techniques. An associative neural network has a memory that can coincide with training set. If new data is available the network improves its predictive ability and gives a good approximation of unknown function without a need to retrain the neural network ensemble. This method dramatically enhances its predictive ability over traditional neural network and K-nearest neighbour techniques[20].

The associative neural network models are selected based on selection processes that include the algorithm, the number of neurons and hidden layers and the iterations and number of ensembles. The early stopping over ensemble (ESE) method was used for training the neural networks). In ESE, initial training sets were randomly constructed with equal size learning and validation sets for each neural network in the ensemble. Thus, each neural network had its own learning and validation sets. The learning set was used for adjusting neural network weights. The training was stopped when a minimum error for the validation set was calculated ('early stopping point'). Following ensemble learning, a simple average of all networks was used for predicting the test patterns. Associative neural network is available online at the VCCLAB website[21].

RESULTS AND DISCUSSION

E-DRAGON software is used to capture all possible diverse structural information, Unsupervized forward selection method has been used for descriptor selection or model development in different systems. The descriptors selected for present study must not be highly correlated. Only those descriptors having intercorrelation co-efficient below 0.8 are considered for the present study.

The selected descriptors involved in the present QSPR model are: (i) Mp: Mean atomic polarizability (scaled on carbon atom); (ii) L1v: 1st component size directional WHIM index/ weighted by atomic vander Waals volumes; (iii) C001: Atom centered fragments/ CH3R; (iv) X2Av: Average valence connectivity index chi-2; (v) nN: Number of nitrogen atoms; (vi) nO: Number of oxygen atoms

Mp is a constitutional descriptor calculated by dividing sum of atomic polarizabilities by number of atoms. nN and nO are also constitutional descriptors. L1v is a WHIM (weighted holistic invariant molecular) descriptor based on the statistical indices calculated on the projections of atoms along principal axes. They are built in such a way as to capture relevant molecular 3D information regarding the molecular size, shape, symmetry and atom distribution with respect to invariant reference frames. X2Av is a topological descriptor encodes presence of heteroatom, double and triple bonds calculated from hydrogen suppressed graph. C-001 is a atom centered fragment descriptor defined by counting first neighbours of carbon atoms (CH3R), where R is the presence of heteroatoms.

The selected descriptors listed in Table-1 are applied to the associative neural network for training. During the training process, the network involves six neurons (six descriptors) in the input layer, six neurons in the hidden layer and one neuron in the output layer [log (ST)] for 90 compounds. The network is trained using the Leven Berg Marquardt algorithm. Number of hidden neuron is decided by training and predicting the 'training data' by varying the number of hidden neurons in the hidden layer. A suitable configuration has to be chosen for the best performance of the network. Out of the different configuration tested, a hidden layer with 6 hidden neurons produced the best result for prediction of surface tension of organic compounds. Seed number is used in to start sequence of random numbers for neural network weights initialization and partition of initial training set data on training/test sets. After the training process, predictive ability of the model is estimated from an external test set of chemicals not included in the training set. The computed surface tension for 26 compounds using associative neural network is given in Table-2. The validation set included 26 compounds with diverse set of chemical compounds. The quality of prediction is evaluated by using two parameters: squared correlation co-efficient (R^2) and root mean square error. The architecture of the final model is shown in Table-3. The statistical performance of the associative neural network QSPR model for surface tension estimation is summarized in Table-4. The root mean square errors of associative neural network model for training and testing are 0.035 log units and 0.0809 log units respectively. Figs. 1 and 2 show scatter plot of the associative neural network predicted *versus* experimental values of log (ST) for the training and test set respectively. Squared correlation co-efficient (R^2) of 0.988 for training and 0.932 for testing confirms the

TABLE-1
TRAINING SET WITH SELECTED DESCRIPTORS FOR SURFACE TENSION

S. No	Compound name	m.p.	L1v	C-001	X2Av	nN	nO	Experimental log st	Predicted log (ST)	Residual
1	Propane	0.55	1.432	2	0.707	0	0	2.024193	2.12	-0.1
2	2,2 Dimethyl butane	0.57	1.851	4	0.416	0	0	2.791778	2.85	-0.06
3	Pentane	0.56	3.703	2	0.451	0	0	2.775709	2.8	-0.02
4	2-Methyl Pentane	0.57	3.445	3	0.437	0	0	2.85532	2.87	-0.01
5	3,3-Dimethyl pentane	0.57	2.779	4	0.359	0	0	2.967847	3.03	-0.06
6	3-Ethyl 3-methyl pentane	0.57	2.483	4	0.319	0	0	3.090588	3.09	0
7	2,2,3-Trimethyl pentane	0.57	2.866	5	0.368	0	0	3.028683	3.04	-0.01
8	2,3,3-Trimethyl pentane	0.57	2.723	5	0.35	0	0	3.071303	3.08	-0.01
9	Hexane	0.57	5.256	2	0.427	0	0	2.912351	2.92	-0.01
10	2-Methyl hexane	0.57	5.023	3	0.423	0	0	2.959587	2.95	0.01
11	3-Ethyl hexane	0.57	3.764	3	0.353	0	0	3.073619	3.06	0.01
12	2,3 Dimethyl hexane	0.57	4.428	4	0.376	0	0	3.044522	3.04	0
13	2,4 Dimethyl hexane	0.57	4.551	4	0.393	0	0	2.998229	3.01	-0.01
14	2,5 Dimethyl hexane	0.57	5.088	4	0.421	0	0	2.988708	2.97	0.02
15	3,3 Dimethyl hexane	0.57	4.18	4	0.363	0	0	3.030134	3.06	-0.03
16	3,4-Dimethyl hexane	0.57	4.221	4	0.346	0	0	3.078233	3.09	-0.01
17	4-Ethyl-2-methyl hexane	0.57	3.796	4	0.368	0	0	3.08191	3.04	0.04
18	3-Ethyl-3-methyl hexane	0.57	3.792	4	0.327	0	0	3.146305	3.12	0.03
19	2,2,3-Trimethyl hexane	0.57	4.233	5	0.369	0	0	3.085116	3.07	0.02
20	2,2,4-Trimethyl hexane	0.57	4.337	5	0.389	0	0	3.020913	3.03	-0.01
21	2,2,5-Trimethyl hexane	0.57	4.893	5	0.408	0	0	2.997231	3	0
22	2,3,3-Trimethyl hexane	0.57	4.019	5	0.354	0	0	3.109061	3.1	0.01
23	2,3,4-Trimethyl hexane	0.57	3.609	5	0.349	0	0	3.126761	3.1	0.03
24	Heptane	0.57	7.017	2	0.412	0	0	3.002708	3	0
25	2-Methyl heptane	0.57	6.955	3	0.413	0	0	3.032546	3.02	0.01
26	3- Methyl heptane	0.57	6.391	3	0.379	0	0	3.059176	3.07	-0.01
27	4-Methyl heptane	0.57	6.204	3	0.383	0	0	3.05022	3.05	0
28	2,4-Dimethyl heptane	0.57	6.209	4	0.391	0	0	3.061052	3.06	0
29	2,5-Dimethyl heptane	0.57	6.516	4	0.387	0	0	3.061052	3.07	-0.01
30	2,6-Dimethyl heptane	0.57	7.075	4	0.413	0	0	3.044522	3.03	0.01
31	3,3-DImethyl heptane	0.57	5.862	4	0.362	0	0	3.093766	3.1	-0.01
32	3,4-Dimethyl heptane	0.57	5.732	4	0.35	0	0	3.128951	3.12	0.01
33	3-Ethyl heptane	0.57	5.477	3	0.353	0	0	3.128075	3.1	0.03
34	Octane	0.57	9.08	2	0.402	0	0	3.069912	3.06	0.01
35	Nonane	0.57	11.372	2	0.395	0	0	3.128951	3.11	0.02
36	Decane	0.57	13.95	2	0.39	0	0	3.171365	3.16	0.01
37	*n*-Tridecane	0.58	26.818	2	0.378	0	0	3.258097	3.26	0
38	Nonadecane	0.58	48.757	2	0.371	0	0	3.353057	3.34	0.01
39	Eicosane	0.58	53.929	2	0.37	0	0	3.362803	3.35	0.01
40	Methanol	0.5	0.718	0	0	0	1	3.113515	3.12	-0.01
41	Ethanol	0.53	1.292	1	0.316	0	1	3.108614	3.12	-0.01
42	Propanol	0.54	2.051	1	0.362	0	1	3.165897	3.17	0
43	Isopropanol	0.54	1.35	2	0.365	0	1	3.059646	3.08	-0.02
44	1-Butanol	0.55	3.162	1	0.359	0	1	3.234355	3.22	0.01
45	2-Methyl-1-butanol	0.56	2.666	2	0.339	0	1	3.246491	3.21	0.04
46	3-Methyl-1-butanol	0.56	2.775	2	0.381	0	1	3.165897	3.16	0.01
47	1-Pentanol	0.56	4.48	1	0.358	0	1	3.233173	3.24	-0.01
48	2-Pentanol	0.56	3.591	2	0.328	0	1	3.15487	3.19	-0.04
49	1-Hexanol	0.56	6.103	1	0.357	0	1	3.250762	3.26	-0.01
50	1-Heptanol	0.56	7.957	1	0.356	0	1	3.292126	3.29	0
51	2-Methyl 2-heptanol	0.57	6.835	3	0.327	0	1	3.209229	3.21	0
52	2-Octanol	0.57	9.205	2	0.337	0	1	3.270329	3.27	0
53	1-Nonanol	0.57	12.477	1	0.356	0	1	3.328268	3.34	-0.01
54	1-Decanol	0.57	15.134	1	0.355	0	1	3.363149	3.36	0
55	2-Dodecanol	0.57	20.158	2	0.343	0	1	3.372798	3.36	0.01
56	3-Dodecanol	0.57	19.838	2	0.331	0	1	3.353407	3.35	0
57	4-Dodecanol	0.57	19.583	2	0.333	0	1	3.326115	3.34	-0.01
58	Benzyl alcohol	0.58	2.955	0	0.299	0	1	3.663562	3.56	0.1
59	Allyl alcohol	0.57	2.003	0	0.236	0	1	3.247658	3.29	-0.04
60	Cyclopentanol	0.58	1.585	0	0.277	0	1	3.497113	3.48	0.02
61	Cycloheptanol	0.58	1.947	0	0.294	0	1	3.501646	3.5	0
62	1,2-Propanediol	0.53	1.576	1	0.258	0	2	3.819908	3.76	0.06
63	1,3-Propanediol	0.53	2.691	0	0.267	0	2	3.83298	3.78	0.05

S. No	Compound name	m.p.	L1v	C-001	X2Av	nN	nO	Experimental log st	Predicted log (ST)	Residual
64	1,5-pentanediol	0.55	5.283	0	0.302	0	2	3.768153	3.75	0.02
65	Methylamine	0.5	0.838	0	0	1	0	2.988708	3.02	-0.03
66	Dimethylamine	0.53	1.53	0	0.5	1	0	3.294725	3.2	0.09
67	Trimethylamine	0.54	1.221	0	0.447	1	0	2.636912	2.81	-0.17
68	Ethylamine	0.53	1.351	1	0.408	1	0	2.990217	3	0
69	Diethylamine	0.55	3.821	2	0.319	1	0	3.016515	3.06	-0.04
70	Propylamine	0.54	2.274	1	0.394	1	0	3.107721	3.09	0.02
71	Dipropylamine	0.56	7.196	2	0.35	1	0	3.127637	3.15	-0.02
72	Tripropylamine	0.57	4.211	3	0.316	1	0	3.127637	3.14	-0.01
73	Butylamine	0.55	3.423	1	0.381	1	0	3.178054	3.15	0.03
74	Allylamine	0.57	2.127	0	0.262	1	0	3.215671	3.25	-0.03
75	Hexylamine	0.56	6.54	1	0.37	1	0	3.268047	3.24	0.03
76	Isohexylamine	0.56	4.6	2	0.387	1	0	3.152309	3.14	0.01
77	Dihexylamine	0.57	23.488	2	0.352	1	0	3.312366	3.31	0
78	Ethylene diamine	0.52	2.096	0	0.289	2	0	3.736955	3.68	0.06
79	Cyclohexyl amine	0.58	2.168	0	0.3	1	0	3.459781	3.42	0.04
80	Dibenzyl amine	0.68	11.155	0	0.188	1	0	3.716008	3.7	0.02
81	Aniline	0.66	2.292	0	0.176	1	0	3.753496	3.71	0.04
82	1-Decene	0.59	13.086	1	0.356	0	0	3.178054	3.18	0
83	1-Heptene	0.59	6.492	1	0.357	0	0	3.010621	3.06	-0.05
84	1-Nonene	0.59	10.634	1	0.356	0	0	3.136363	3.14	0
85	1-Tridecene	0.59	22.066	1	0.355	0	0	3.267285	3.27	0
86	Cyclohexene	0.61	1.475	0	0.293	0	0	3.287655	3.25	0.04
87	Cyclopentene	0.62	1.187	0	0.28	0	0	3.196221	3.2	0
88	Benzaldehyde	0.7	2.645	0	0.17	0	1	3.651697	3.64	0.01
89	Butyraldehyde	0.58	2.615	1	0.318	0	1	3.21165	3.26	-0.05
90	2-Furaldehyde	0.68	2.481	0	0.142	0	2	3.77872	3.76	0.02

TABLE-2
TESTING SET FOR 26 COMPOUNDS WITH DESCRIPTORS

S. No	Compound name	m.p.	L1v	C-001	X2Av	nN	nO	Experimental log (ST)	Predict log (ST)	Residual
1	Butane	0.56	2.496	2	0.5	0	0	2.53	2.61	-0.08
2	2,3-Dimethyl butane	0.57	2.045	4	0.415	0	0	2.86	2.89	-0.03
3	3-Methyl pentane	0.57	3.163	3	0.384	0	0	2.9	2.98	-0.08
4	3-Ethyl-2-methylpentane	0.57	2.748	4	0.353	0	0	3.07	3.04	0.03
5	2,2,4 -Trimethyl pentane	0.57	3.115	5	0.416	0	0	2.93	2.93	0
6	3-Methyl hexane	0.57	4.65	3	0.384	0	0	2.99	3.03	-0.04
7	2,2 -Dimethyl hexane	0.57	4.705	4	0.407	0	0	2.98	3	-0.02
8	2,3,5-Trimethyl hexane	0.57	4.46	5	0.385	0	0	3.06	3.05	0.01
9	4-Ethyl heptane	0.57	4.51	3	0.356	0	0	3.13	3.07	0.06
10	Undecene	0.58	16.768	2	0.386	0	0	3.13	3.16	-0.03
11	2-Butanol	0.55	2.342	2	0.314	0	1	3.09	3.16	-0.07
12	2-Methyl-2-butanol	0.56	2.000	3	0.309	0	1	3.3	3.32	-0.02
13	1-Octanol	0.57	10.096	1	0.356	0	1	3.31	3.34	-0.03
14	5-Dodecanol	0.57	19.417	2	0.333	0	1	3.54	3.51	0.03
15	Cyclohexanol	0.58	1.866	0	0.286	0	1	3.85	3.76	0.09
16	1,3-butanediol	0.54	3.002	1	0.272	0	2	2.98	3.15	-0.17
17	Diisopropylamine	0.56	5.379	2	0.338	1	0	3.2	3.24	-0.04
18	Dibutylamine	0.57	11.573	2	0.351	1	0	3.69	3.41	0.28
19	Benzylamine	0.58	3.217	0	0.304	1	0	2.91	3.04	-0.13
20	1-Hexene	0.59	4.779	1	0.358	0	0	3.08	3.12	-0.04
21	1-Octene	0.59	8.408	1	0.356	0	0	3.21	3.19	0.02
22	1-Dodecene	0.59	18.807	1	0.355	0	0	3.24	3.24	0
23	1-Tetradecene	0.59	25.569	1	0.355	0	0	3.29	3.29	0
24	1-Pentadecene	0.59	29.346	1	0.355	0	0	3.3	3.3	0
25	1-Hexadecene	0.59	33.371	1	0.355	0	0	3.32	3.32	0
26	1-Heptadecene	0.59	37.668	1	0.355	0	0	3.34	3.33	0.01

suitability of the associative neural network model and shows a good agreement of associative neural network predicted values with experimental one. The residual of the associative neural network predicted values the log (ST) are plotted against their experimental values (Fig. 3). The propagation of residuals on both sides of zero indicates that no systematic error exists in the development of associative neural network. The number of compounds not correctly predicted by the model is very limited. Diisopropyl and benzylamine have high residual value. Cross-validation is a popular technique to explore the stability of developed models. In this validation technique, a number of modified data sets are created by deleting, one compound.

For each reduced data set, the model is calculated and responses for the deleted compounds are predicted from the model. In this study, the predictive power of the models is checked by leave one out cross-validation and the square of the cross-validated correlation coefficient (q^2) is used to measure the models predictivity. A good correlation is obtained with LOO correlation co-efficient $q^2 = 0.983$. So the predictive power of this model is very significant.

TABLE-3
ARCHITECTURE AND SPECIFICATION OF THE GENERATED ASNN

No. of nodes in the input layer	6
No. Of nodes in the hidden layer	6
No. of nodes in the output layer	1
Seed value	38
Number of KNN	10
Activation function	Logistic 1/(1+exp (-x))

TABLE-4
STATISTICAL ANALYSIS OF ASSOCIATIVE NEURAL NETWORK

Data set	R^2	q^2	RMSE	MAE
Training	0.988	0.983	0.035	0.023
Testing	0.932	0.912	0.0809	0.05

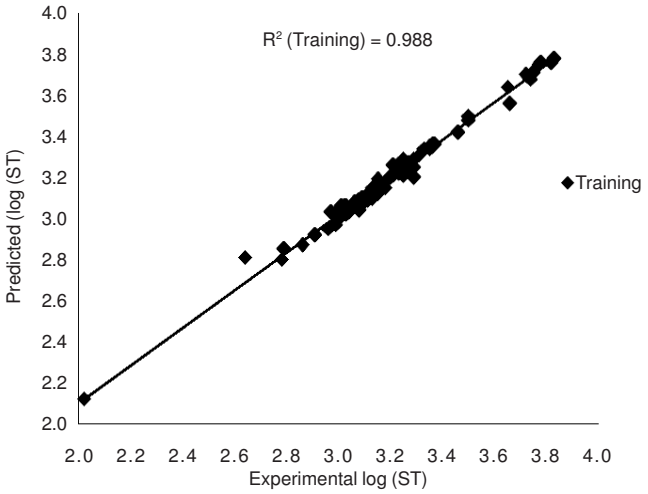

Fig. 1. Scatter plot of the experimental *vs*. predicted surface tension values of training set

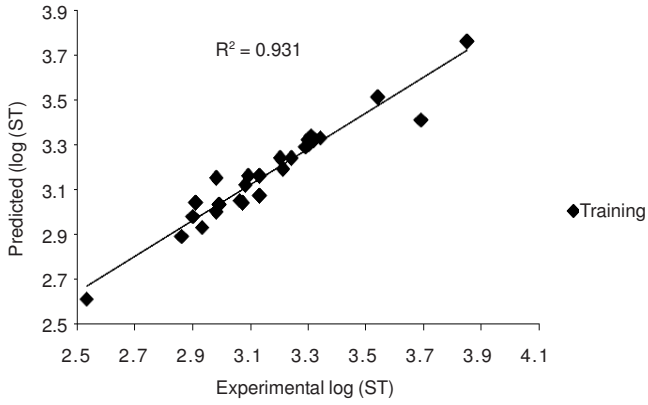

Fig. 2. Scatter plot of the experimental *vs*. predicted surface tension values of test set

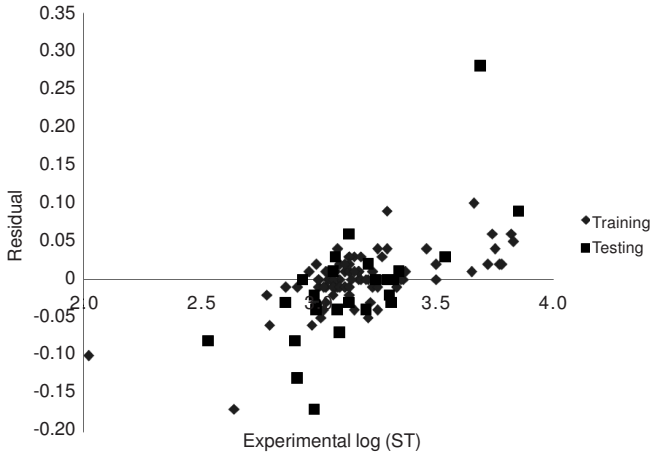

Fig. 3. Scatter plot of experimental *vs*. residual error for training and test set

Interpretation of descriptors: It is well accepted that surface tension is governed by London dispersion forces in organic compounds. From eq. 1, it should be concluded that molecular interaction force is directly proportional to polarizability. Therefore descriptor mean atomic polarizability (Mp) directly encodes information related to molecular polarizability. The atom-centered fragment descriptor (C-001) is used to differentiate the isomers within same group of compounds. The descriptors nN and nO in the molecules relates a polar interaction among the molecules in the bulk liquid. It is well known that molecular polarizability is directly proportional to number of electrons in the molecule. When the size (number of electrons) of the molecule increases, intermolecular interaction is stronger, therefore the higher the surface tension will be. Hence the WHIM descriptor L1v represents a measure of molecular polarizability. The descriptor X2Av is used for heteroatom differentiation. It can be concluded that the descriptors in the present QSPR model has definite chemical meaning and these can account the structural features that affect on the surface tension of the organic compounds.

Conclusion

The results of this study indicate that it is possible to estimate the surface tension of organic compounds from their theoretically derived molecular descriptors. The associative neural network with 6-6-1 architecture produces high statistical quality and low prediction error model. The six descriptors involved in the present work, which can be calculated from molecular structures, have definite physical meaning corresponding to the different intermolecular interactions, which takes place in bulk solution. The obtained results in this paper suggest that the associative neural network predicts surface tension of organic compounds very well compared with previous studies. The QSPR models developed in this study can provide a useful tool to predict the surface tension of new compounds.

REFERENCES

1. A.W. Adamson, Physical Chemistry of Surfaces, Interscience Publishers, New York, edn 5 (1990).
2. B.E. Poling, J.M. Prausnitz and J.P. O'Connell, The Properties of Gases and Liquids, McGraw-Hill, USA (2004).

3. A.J. Quiemada, I.M. Marrucho, J.A.P. Coutinho and E.H. Stenby, Proceedings of the 15 th Symposium on Thermophysical Properties, Boulder (USA), June 22-27 (2003).
4. H.Y. Erbil, Surface Chemistry of Solid and Liquid Interfaces, Oxford, Blackwell (2006).
5. R. Katritzky, U. Maran, US Lobanov and M. Karelson, *J. Chem. Inf. Comput. Sci.*, **40**, 1 (2000).
6. A.R. Katritzky, V.S. Lobanov and M. Karelson, *Chem. Soc. Rev.*, **24**, 279 (1995).
7. A.R. Katritzky, M. Karelson and V.S. Lobanov, *Pure. Appl. Chem.*, **69**, 245 (1997).
8. F. Ashrafi, R. Saadati and A.B. Amlashi, *Afr. J. Pure Appl. Chem.*, **2**, 116 (2008).
9. R. Todeschini and V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH Verlag, Weinheim (2000).
10. P.X. Liu and W. King, *Int. J. Mol. Sci. Rev.*, **10**, 1978 (2009).
11. J. Devillers, Neural Networks in QSAR Drug Design, Academic Press, London (1996).
12. D.T. Stanton and P.C. Jurs, *Anal. Chem.*, **62**, 2323 (1990).
13. G.W. Kaffman and P.C. Jurs, *J. Chem. Inf. Comput. Sci.*, **41**, 408 (2001).
14. E.J. Delgado and G.A. Diaz, *SAR and QSAR Environ. Res.*, **17**, 483 (2006).
15. E.-Dragon, 1.0 On-line software; Virtual Computational Chemistry Laboratory. http://146.107.217.178/lab/edragon/index.html.
16. J.J. Jasper, *J. Phys. Chem. Ref. Data*, **1**, 841 (1972).
17. http://www.chemaxon.com/Marvin/Sketch/index.php
18. D.C. Whitley and M.G. Ford, *J. Chem. Inf. Comput. Sci.*, **40**, 1160 (2000).
19. T.E. Quantrille and Y.A. Liu, Artificial Intelligence in Chemical Engineering, Academic Press, New York (1992).
20. I.V. Tetko, *J. Chem. Inf. Comput. Sci.*, **42**, 717 (2002).
21. http://www.vcclab.org/asnn/