Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(6):1557-1564

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Prediction of Henry's law constants for organic compounds using multilayer feedforward neural networks based on linear solvation energy relationship

Hao Li[1*], Xiaoting Wang[2*], Tianqi Yi[3], Zhihan Xu[4] and Xifeng Liu[5,6]

[1]College of Chemistry, Sichuan University, Chengdu, Sichuan, China
[2]School of Civil Engineering and Architecture, Nanchang University, Nanchang, Jiangxi, China
[3]College of Polymer Science and Engineering, Sichuan University, Chengdu, Sichuan, China
[4]College of Light Industry, Sichuan, Chengdu, Sichuan, China
[5]Department of Chemistry, Michigan State University, East Lansing, Michigan, United States
[6]College of Chemistry and Chemical Engineering, Hunan University, Changsha, China

ABSTRACT

Henry's law constants are crucial to correctly estimate the solubilities of different organic compounds in water. However, the precise values of these constants are difficult to obtain by traditional approaches in laboratory. In previous studies, a Linear Solvation Energy Relationship (LSER) method has been used to express a relationship between the Henry's law constant and the relative descriptors of organic compounds. Some of these studies have developed a linear regression model to calculate the Henry's law constants using a LSER method. In our study, instead of using linear prediction approach, we successfully established an Artificial Neural Network (ANN) model to predict Henry's law constants based on 72 typical organic compounds, using a LSER method to describe the independent variables of the ANN model. This research work indicates that the linear relationship provided by the LSER method can be calculated to be a non-linear relationship with a lower error using ANN models. Within a permissible error range (30% tolerance), results showed that the Multilayer Feedforward Neural Network (MLFN) model with two nodes (MLFN-2) is an effective model for predicting the Henry's law constants of organic compounds, whose average RMS error is 0.14 logH units.

Keywords: Henry's law constant, organic compounds, Artificial Neural Networks, Multilayer Feedforward Neural Networks, linear solvation energy relationship.

INTRODUCTION

Henry’s law describes the equilibrium liquid and vapor phase concentrations of a solute in the limit of low solute concentrations [1]. The Henry's law constant *H* is the partition coefficient of the two phases, representing the migratory direction and velocity of the organic compounds existing in the equilibrium liquid and vapor phases. As for the organic compounds, those with low value of *H* are easy to aggregate in the water phase, whereas those with high value of *H* are more concentrated in the gas phase.

In practical applications, we can estimate the aggregation tendency of the organic compounds in water environment by knowing the Henry's law constant, which is crucial to the environmental pollution control [2-4]. In the field of electrochemistry, Henry's law is also significant in the modeling of PEM fuel cells and Ballard Mark IV solid polymer electrolyte fuel cell [5-6], as well as the relative researches on the superoxide electrochemistry in ionic liquid [7].

However, the determination of Henry's law constants is complex and has a low reproducibility, which generates a great obstacle to practical applications.

A number of methods in analyzing the Henry's law constants were presented in previous studies [8-12]. English and Carroll [8] have developed two estimation models by quantitative structure property relationship and neural networks, using 10 and 12 descriptors respectively. Yao and his co-workers [9] utilized radial basis function network-based quantitative structure–property relationship to predict the Henry's law constant of organic compounds. Hine [10] and Gharagheizi [11] have developed group-contribution-based models to calculate the values of Henry's law constant. These methods can predict the Henry's law constant effectively with a small error. Nevertheless, the establishment of these models are still complex and difficult to be applied to the practical applications. In addition, Y.B. He and his co-workers [12] used multiple linear regression to calculate the constant, which is easy-understanding and convenient. But the average stand error of their conclusion is 0.25 log*H* units, which is generally higher than other prediction methods. In general, these studies are useful and can be taken as references, nevertheless, whatever the degree of accuracy or the maneuverability of applications, these models have disadvantages (see *4.3: Comparisons with other models*).

In our study, we aimed at developing a reasonable artificial neural network (ANN) model to predict the values of organic compounds' Henry's law constant, using a precise but uncomplicated description based on Linear Solvation Energy Relationship (LSER) [13-15].

EXPERIMENTAL SECTION

Fundamental of Artificial Neural Networks

Artificial Neural Networks (ANNs) [16-18] are computational models inspired by animals' central nervous systems that are capable of machine learning and pattern recognition [19]. They are usually presented as systems of interconnected "neurons" that can calculate different values from inputs by feeding information through the network. As the development of the algorithm, this method is mature and has been packed into a module of the software [20]. Represented by nonlinear functions, artificial neural network analysis is an artificial intelligence (AI) approach to modeling.

In natural conditions, elements form groups and connect each other as neurons within the discrete layer. Each connection of them has its identified weight coefficient. The multiple layer consisted of the structure of such network [21]. Usually, there are one or more than one layers of the elements followed by an output layer. Multiple layers of elements can drive the network to learn nonlinear and linear relationships between input and output vectors.

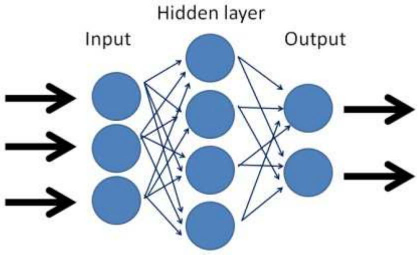

Figure 1. A schematic view of artificial neural network structure

Fig.1 shows the main structure of the ANN. It is chiefly made up of the input layer and the output layer [22]. The input layer introduces the input variables to the network. The output of the nodes in this layer represents the predictions made by the network for the response variables. In addition, it contains hidden layers [23]. The optimal number of neurons in the hidden layers depends on the type and complexity of the process or experimentation and it's usually iteratively determined.

As for the Henry's law constant, the relationships between constant *H* and other variables are so complex that can be considered to be described by the non-linear situation [8]. Therefore, we did a series of computational experiments on developing the ANN models that can describe these relationships accurately.

Linear Solvation Energy Relationship

Linear solvation energy relationship (LSER) is a method that is deemed to be one of the best patterns of Quantitative Structure-Activity Relationships (QSARs) [15]. Kamlet [24] pointed out that, a number of properties of chemicals depend on the interaction between solute and solvent. Hence we can predict a series of chemicals' properties by the equations of LSER.

According to the equations of LSER [25-26], the description of Henry's law constant was obtained as follows:

$$\log H = SP_0 + mV_1/100 + s\pi^* + a\alpha_m + b\beta_m \qquad (1)$$

where *H* represents the Henry's law constant; *V_1* represents the molar volume, a measurement of the energy effect of the void generated by dissolution; π^* represents the dipole term, a measurement of the energy effect of the dipole-dipole interaction; α_m represents the acidity of the hydrogen bond's donor; and β_m represents the basicity of the hydrogen bond's acceptor. Both α_m and β_m are measurements of the energy effect of the formation of hydrogen bonds, and *SP_0, m, s, a,*and *b* are constant terms.

According to Eq.1 proveded by a LSER method [25-26], the descriptors of Henry's law constants of 72 typical organic compounds are shown as follows :

Table 1: LSER descriptors and log*H* of 72 typical organic compounds.

Number	Compound	V_1/100	π^*	α_m	β_m	log*H*
1	benzene	0.491	0.59	0.10	0	-0.65
2	methylbenzene	0.591	0.55	0.11	0	-0.56
3	propylbenzene	0.768	0.51	0.12	0	-0.39
4	fluorobenzene	0.520	0.62	0.07	0	-0.59
5	fluorobenzene	0.581	0.71	0.07	0	-0.74
6	iodobenzene	0.671	0.81	0.05	0	-1.28
7	1, 2 – dichlorobenzene	0.671	0.80	0.03	0	-1.00
8	1,2, 3 - trichlorobenzene	0.761	0.85	0	0	-1.30
9	1,2,3,5 - four chlorobenzene	0.851	0.70	0	0	-0.63
10	dibromobenzene	0.758	0.89	0.02	0	-1.07
11	Chlorotoluene	0.679	0.67	0.08	0	-0.76
12	parabromotoluene	0.722	0.75	0.08	0	-1.02
13	phenol	0.536	0.72	0.33	0.61	-4.79
14	4 - bromophenol	0.669	0.79	0.23	0.69	-5.21
15	parachlorophenol	0.626	0.72	0.23	0.67	-4.77
16	2 - cresol	0.634	0.68	0.34	0.58	-4.30
17	4 – nitrophenol	0.676	1.15	0.32	0.82	-7.77
18	1 - chlorine naphthalene	0.843	0.80	0.11	0	-1.45
19	anisole	0.639	0.73	0.32	0	-1.68
20	cyanobenzene	0.590	0.90	0.37	0	-2.50
21	benzaldehyde	0.606	0.92	0.44	0	-2.95
22	hypnone	0.690	0.90	0.49	0.04	-3.36
23	paranitrotoluene	0.729	0.97	0.31	0	-2.55
24	2, 6 - dinitrotoluene	0.869	1.02	0.56	0	-3.61
25	1-bromo-4-nitrobenzene	0.764	1.01	0.26	0	-2.74
26	pyridine	0.470	0.87	0.64	0	-3.44
27	4 - methyl pyridine	0.568	0.84	0.67	0	-3.61
28	4 - ethyl pyridine	0.666	0.85	0.65	0	-3.46
29	2, 4 - dimethyl pyridine	0.666	0.79	0.67	0	-3.56
30	ethane	0.272	0	0	0	1.31
31	butane	0.455	0	0	0	1.58
32	octane	0.842	0.01	0	0	2.12
33	ethylene	0.243	0.08	0.07	0	0.94
34	n-butene	0.428	0.08	0.07	0	1.01
35	hexane	0.624	0.08	0.07	0	1.25
36	methane chloride	0.252	0.45	0.10	0	-0.39
37	tetrachloromethane	0.514	0.28	0.10	0	0.07
38	1, 2 - dichloroethane	0.442	0.81	0.10	0	-1.27
39	1,1,2 – trichloroethane	0.519	0.81	0.10	0	-1.43
40	pentachloroethane	0.700	0.62	0.10	0	-1.00
41	1 - chloropropane	0.450	0.39	0.10	0	-0.26
42	1, 2-2 vinyl chloride	0.541	0.70	0.10	0	-0.92
43	trichloro ethylene	0.492	0.53	0.05	0	-0.32
44	3 - allyl chloride	0.424	0.49	0.05	0.05	-0.42
45	acetic acid	0.323	0.60	0.45	0.56	-4.91

46	methyl alcohol	0.205	0.40	0.42	0.35	-3.72
47	allyl alcohol	0.372	0.40	0.43	0.33	-3.69
48	butanol	0.499	0.40	0.45	0.33	-3.46
49	hexyl alcohol	0.690	0.40	0.45	0.33	-3.20
50	octanol	0.882	0.40	0.45	0.33	-3.01
51	cyclohexanol	0.636	0.45	0.51	0.31	-3.61
52	methyl formate	0.326	0.62	0.37	0	-2.04
53	ethyl acetate	0.521	0.55	0.45	0	-2.26
54	ethyl propionate	0.622	0.53	0.45	0	-2.05
55	methyl butyrate	0.620	0.55	0.45	0	-2.08
56	methyl pentanoate	0.716	0.46	0.45	0	-1.86
57	methyl caproate	0.814	0.55	0.45	0	-1.82
58	methyl caprylate	1.010	0.55	0.45	0	-1.50
59	diethyl ether	0.505	0.27	0.47	0	-1.28
60	butyl ether	0.895	0.27	0.47	0	-0.61
61	acetaldehyde	0.284	0.63	0.41	0	-2.57
62	hexanal	0.674	0.63	0.41	0	-2.06
63	heptanal	0.772	0.63	0.41	0	-1.96
64	acetone	0.380	0.71	0.48	0.04	-2.79
65	2 – butanone	0.477	0.67	0.48	0.03	-2.72
66	2 - undecane ketone	1.159	0.61	0.48	0	-1.58
67	ethylamine	0.335	0.32	0.70	0.14	-3.38
68	butyl amine	0.535	0.31	0.69	0.14	-3.21
69	dimethylamine	0.339	0.25	0.70	0.14	-3.14
70	triethylamine	0.704	0.14	0.71	0	-2.22
71	acetonitrile	0.271	0.75	0.31	0.09	-2.85
72	nitroethane	0.445	0.80	0.25	0.12	-2.72

Training Process of The Neural Networks

The ANN prediction models were constructed by the NeuralTools® software (Trial Version, Palisade Corporation, NY, USA). We chose the linear regression (LR) module, General Regression Neural Networks (GRNN) [27-29] module and Multilayer Feedforward Neural Networks (MLFN) [30-32] module as the training modules. 85% data groups were used for training set, while the rest of groups were used for testing set. To ensure the accuracy of the results, models were trained repeatedly. Each process was generated randomly by the different composition of training sets. The average training results are shown as follows:

Table 2: The average training results of Henry's law constants in different models

Model	Trained Samples	Tested Samples	RMS Error	Training Time	Finishing Reason
Linear Prediction	61	11	0.18	0:00:01	Auto-Stopped
GRNN	61	11	0.36	0:00:00	Auto-Stopped
MLFN 2 Nodes	61	11	0.14	0:01:10	Auto-Stopped
MLFN 3 Nodes	61	11	0.16	0:01:15	Auto-Stopped
MLFN 4 Nodes	61	11	0.15	0:01:32	Auto-Stopped
MLFN 5 Nodes	61	11	0.20	0:01:40	Auto-Stopped
MLFN 6 Nodes	61	11	0.35	0:01:31	Auto-Stopped
MLFN 7 Nodes	61	11	0.45	0:02:14	Auto-Stopped
MLFN 8 Nodes	61	11	0.37	0:02:05	Auto-Stopped
MLFN 9 Nodes	61	11	0.40	0:02:48	Auto-Stopped
MLFN 10 Nodes	61	11	0.45	0:03:00	Auto-Stopped
MLFN 11 Nodes	61	11	0.60	0:03:00	Auto-Stopped
MLFN 12 Nodes	61	11	0.54	0:03:38	Auto-Stopped
MLFN 13 Nodes	61	11	0.57	0:03:40	Auto-Stopped
MLFN 14 Nodes	61	11	1.51	0:04:42	Auto-Stopped
MLFN 15 Nodes	61	11	0.60	0:04:50	Auto-Stopped
MLFN 16 Nodes	61	11	0.61	0:04:34	Auto-Stopped
MLFN 17 Nodes	61	11	0.62	0:05:32	Auto-Stopped
MLFN 18 Nodes	61	11	0.92	0:05:14	Auto-Stopped
MLFN 19 Nodes	61	11	1.16	0:04:53	Auto-Stopped
MLFN 20 Nodes	61	11	0.59	0:04:56	Auto-Stopped

Table 2 shows that the MLFN model with 2 nodes (MLFN-2) generates the lowest RMS error (0.14). With the increasing nodes, RMS errors of MLFN models are also higher than MLFN-2 model. Therefore, it is not necessary to establish more MLFN models with over 20 nodes, since the training time may increase rapidly with the increasing quantity of nodes. In addition, the RMS error of linear regression is 0.18 log*H* units, which is corresponded with the result of the reference [12].

RESULTS AND DISCUSSION

Training Results of MLFN-2 Model

Training results of MLFN-2 model in predicting the Henry's law constants are shown as follows:

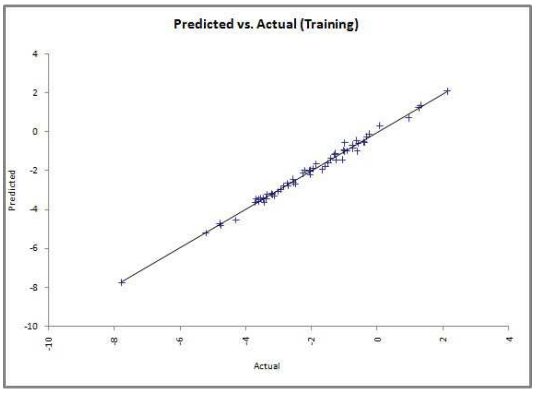

Figure 2. Comparison between predicted values and actual values of Henry's law constant of organic compounds using MLFN-2 model during training process

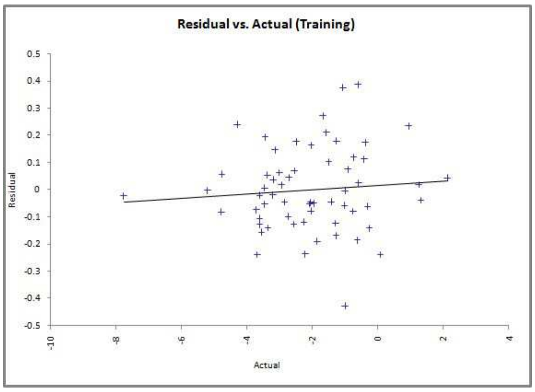

Figure 3. Comparison between residual values and actual values of Henry's law constant of organic compounds using MLFN-2 model during training process

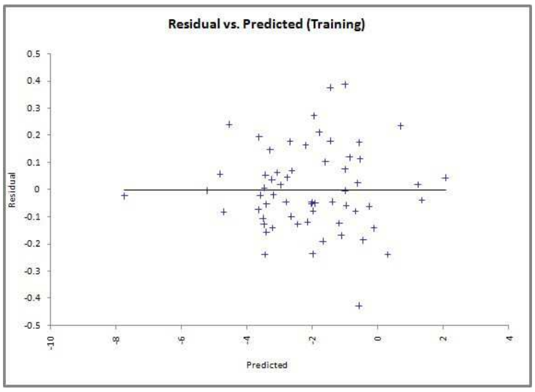

Figure 4. Comparison between residual values and predicted values of Henry's law constant of organic compounds using MLFN-2 model during training process

Fig.2 to 4 depict the training results of MLFN-2 model in predicting the Henry's law constants of organic compounds. There into Fig.2 presents the good fitting process of the neural networks, and residual values shown in Fig.3 and Fig.4 are concentrated nearby the regression line. Results show that the process is accurate and normal.

Testing Results of MLFN-2 Model

In order to test the robustness of MLFN-2 model in predicting Henry's law constants of organic compounds, in each process, we took different data components as the training set and testing set. Fig.5 to 7 are one of the examples of testing results of MLFN-2 model:

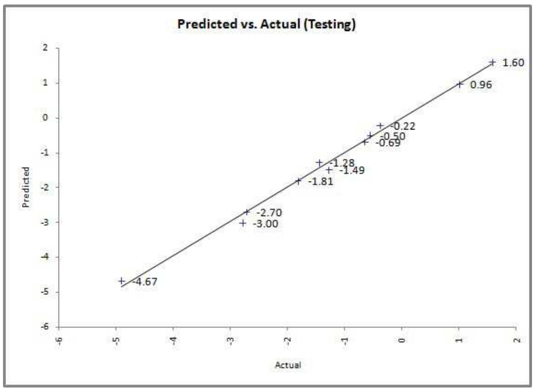

Figure 5. Comparison between predicted values and actual values of Henry's law constant of organic compounds using MLFN-2 model during testing process

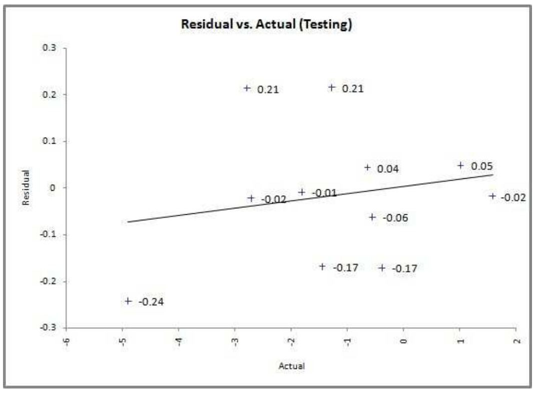

Figure 6. Comparison between residual values and actual values of Henry's law constant of organic compounds using MLFN-2 model during testing process

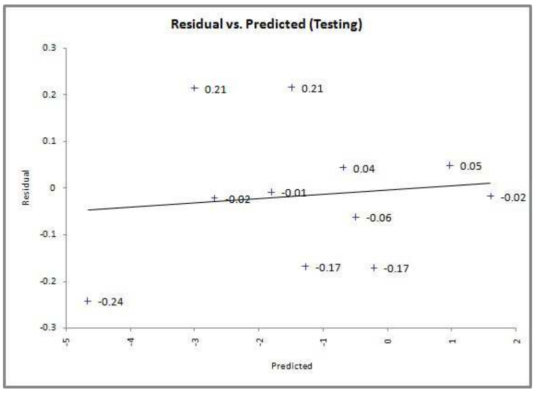

Figure 7. Comparison between residual values and predicted values of Henry's law constant of organic compounds using MLFN-2 model during testing process

Fig.5 to 7 depict the testing results of MLFN-2 model. It's obvious that the predicted values are very closed to the actual values. Among the three figures, Fig.5 presents a good testing results, all the predicted values are corresponded with the actual values within the permission error. Fig.6 and Fig.7 also show that the residual values are concentrated nearby the regression line. What is worth mentioning is that the results are averages, being concluded from a series of repeated experiments, proving that the development of MLFN-2 model is a robust process in predicting Henry's law constants of organic compounds.

Comparisons with Other Models

According to previous studies [8-12], similar researches on predicting the Henry's law constants were provided. In English and Carroll's research [8], the QSAR and ANN models were developed. Their models are based on a complex molecular recognition method, using more than 10 descriptors, which may be difficult to practical applications. Yao and his co-workers' prediction model was based on the radial basis function network-based quantitative structure–property relationship [9]. However, RMS error of their tests and the overall data sets are 0.3121 and 0.3038 log*H* units, which may be a little high compared with other previous studies. Group-contribution-based models developed by Hine [10] and Gharagheizi [11] are robust, but the processes of establishment are not easy to operate.

In our study, we aimed at using LSER method to find out the variables to develop the ANN model innovatively. Previous opinions considered that there exists a linear relationship between the independent variables of Eq.1 and log*H* [12,25-26]. Our research found that this linear relationship can transform to be a nonlinear situation with a lower error via ANN models.

Previous studies were successful that can be used for references to our study, having different advantages respectively by using various prediction or calculation methods. Nevertheless, by contrast, we considered that using Multilayer Feedforward Neural Networks combined with linear solvation energy relationship to predict the Henry's law constants of organic compounds is a easier and more precise method (RMS error: 0.14 log*H* units). The model can not only obtain the precise value of Henry's law constants, but also be applied to practical applications.

CONCLUSION

Instead of measuring the values of Henry's law constant of organic compounds from experiments in laboratory, it is now possible to use the artificial neural networks with known experimental data and linear solvation energy relationship to predict this property of organic compounds. The neural network can now be put to use with the actual data, which involves the values of Henry's law constant.

Acknowledgements

This work is an extended version of the article entitled "Prediction of Henry's law constant using a LSER method: a comparison between Artificial Neural Networks and linear prediction", which has been pre-accepted by *Asian Journal of Chemistry*.

REFERENCES

[1] D Mackay, WY Shiu, *J. Phys. Chem. Ref. Data*, **1981**, 10(4), 1175-1199.
[2] D Mackay, WY Shiu, RP Sutherlan. *Environ. Sci. Technol.*, **1979**, 13(3), 333-337.
[3] F Gharagheizi, A Eslamimanesh, AH Mohammadi, et al., *J. Chem. Thermodyn.*, **2012**, 47, 295-299.
[4] KN McPhedran, R Seth, KG Drouillard, *Chemosphere*, **2013**, 91(11), 1648-1652.
[5] RF Mann, JC Amphlett, BA Peppley, et al., *J. Power Sources*, **2006**, 161(2), 768-774.
[6] JC Amphlett, R M Baumert, RF Mann, et al., *J. Electrochem. Soc.*, **1995**, 142(1), 1-8.
[7] IM AlNashef, ML Leonard, MA Matthews, et al., *Ind. Eng. Chem. Res.*, **2002**, 41(18), 4475-4478.
[8] NJ English, DG Carroll, *J. Chem. Inf. Model*, **2001**, 41(5), 1150-1161.
[9] X Yao, M Liu, X Zhang, et al., *Anal. Chim. Acta*, **2002**, 462(1), 101-117.
[10] J Hine, PK Mookerjee, *J. Org. Chem.*, **1975**, 40(3), 292-298.
[11] F Gharagheizi, R Abbasi, B Tirandazi, *Ind. Eng. Chem. Res.*, **2010**, 49(20), 10149-10152.
[12] YB He, YY Wang, CH Wu, *Acta Scierntentiae Circumstantiae*, **1997**, 17(2), 227-231.
[13] MJ Kamlet, JLM Abboud, RW Taft, *Phys. Org. Chem.*, **1981**, 13, 485-630.
[14] DS Van Meter, OD Stuart, AB Carle, et al, *J. Chromatogr. A*, **2008**,1191(1), 67-71.
[15] DJW Blum, RE Speece, *Environ. Sci. Technol.*, **1990**, 24(3), 284-293.
[16] N Gupta, *Network and Complex Systems*, **2013**, 3(1), 24-28.
[17] Y Shen, A Bax, J. Biomol. *NMR*, **2013**, 56(3), 227-241.
[18] A Georgieva, SJ Payne, M Moulden, et al., *Neural Comput. Appl.*, **2013**, 22(1), 85-93.
[19] T Saba, A Rehman, *Int. J. Mach. Learn. & Cyber.*, **2013**, 4(2), 155-162.
[20] GR Finnie, GE Wittig, JM Desharnais, *J. Syst. Software*, **1997**, 39(3), 281-289.
[21] AW Minns, MJ Hall, *Hydrolog. Sci. J.*, **1996**, 41(3), 399-417.
[22] B Samanta, KR Al-Balushi, *Mech. Syst. Signal Pr.*, **2003**, 17(2), 317-328.
[23] GE Dahl, D Yu, L Deng, et al., *IEEE T. Audio Speech*, **2012**, 20(1), 30-42.

[24] MJ Kamlet, RM Doherty, MH Abraham, et al., *J. Phys. Chem*., **1987**, 91(7), 1996-2004.
[25] MJ Kamlet, RM Doherty, PW Carr, et al., *Environ. Sci. Technol.*, **1988**, 22(5), 503-509.
[26] M J Kamlet, RM Doherty, MH Abraham, et al., *J. Phys. Chem.*, **1988**, 92, 5244-5255.
[27] Ö Polat, T Yıldırım, *Digit. Signal Process.*, **2010**, 20(3), 881-886.
[28] O Er, F Temurtas, AÇ Tanrıkulu, *J. Med. Syst.*, **2010,** 34(3), 299-302.
[29] WT Pan, *Knowl-Based Syst.*, **2012**, 26, 69-74.
[30] CH Chen, TK Yao, CM Kuo, et al., *J. Vib. Control*, **2013**, 19(16), 2413-2420.
[31] SA Mirjalili, SZ Mohd Hashim, H Moradian Sardroudi, *Appl. Math. Comput*., **2012**, 218(22), 11125-11137.
[32] R Pahlavan, M Omid, A Akram, *Energy*, **2012**, 37(1), 171-176.