Monatsh Chem (2011) 142:1061–1068
DOI 10.1007/s00706-011-0527-1

ORIGINAL PAPER

In silico prediction of free-radical chain transfer constants for some organic agents in styrene polymerization

Mohammad H. Fatemi · Fereshte Dorostkar · Zahra Ghorbannezhad

Received: 20 November 2010 / Accepted: 3 May 2011 / Published online: 31 May 2011
© Springer-Verlag 2011

Abstract In the present work, quantitative structure–reactivity relationship (QSRR) approaches were used to predict the chain transfer constant log C_x of some organic compounds as chain transfer agents in free-radical polymerization of styrene. The energy of the lowest unoccupied molecular orbital, hydrogen-bonding-dependent hydrogen donor charged area, first-order Kier and Hall index, final heat of formation/number of atoms, count of H donor sites, and Min>(0.1) bond order of a C atom were selected as the most relevant variables from the pool of calculated descriptors by the stepwise multiple regression feature selection method. Then, an artificial neural network and multiple linear regressions were utilized to construct the nonlinear and linear QSRR models. The standard errors in the prediction of log C_x by the linear regression model were 0.641, 0.964, and 0.843 and by the neural network model were 0.049, 0.076, and 0.090 for training, internal, and external test sets, respectively. The predictivity of the artificial neural network model was further examined by cross-validation methods, which produce a Q^2 value of 0.85. The results of this study revealed the applicability of QSRR approaches in prediction of the chain transfer constant by using an artificial neural network.

Keywords Artificial neural network · Chain transfer constant · Molecular descriptor · Multiple linear regression · Quantitative structure–reactivity relationship

M. H. Fatemi (✉) · F. Dorostkar · Z. Ghorbannezhad
Laboratory of Chemometrics, Faculty of Chemistry,
University of Mazandaran, Babolsar, Iran
e-mail: mhfatemi@umz.ac.ir

Introduction

The chain transfer step in a free-radical polymerization process involves the reaction of a propagating chain ($P_n\cdot$) with a transfer agent (RX) to terminate one polymer chain and produce a new radical ($X\cdot$), which initiates a new chain ($P_1\cdot$) [1] as follows:

$$P_n\cdot + RX \xrightarrow{k_{tr,X}} P_nR + X\cdot \tag{1}$$

$$P_n\cdot + M \xrightarrow{k_p} P_{n+1}\cdot \tag{2}$$

$$X\cdot + M \xrightarrow{k_{t,x}} P_1\cdot \tag{3}$$

In the above equations, M represent the monomer molecule. Chain transfer agents have at least one weak chemical bond, and therefore can facilitate the chain transfer reaction. The efficiency of a chain transfer catalyst is expressed by the chain transfer constant (C_x), which is the ratio of the rate constant for the chain transfer reaction (k_{tr}) to propagation step (k_p) (Eq. 4). Chain transfer constants are generally determined from the Mayo equation (Eq. 5) [2]:

$$C_x = \frac{k_{tr,x}}{k_p} \tag{4}$$

$$\frac{1}{DP_n} = \frac{1}{DP_{n0}} + \frac{k_{tr}[X]}{k_p[M]} \tag{5}$$

In the above equation, DP_n is the number-average degree of polymerization, which is determined for a series of free radical reactions with different ratios of the chain transfer agent concentration [X] to the monomer concentration, and DP_{n0} is the number-average degree of polymerization in the absence of the chain transfer agent. Since understanding the chain transfer mechanism not only clarifies the micro-kinetic processes of a free-radical polymerization process but can also be used to measure the relative reactivity of the growing radical toward the transfer substance, theoretical calculations of the kinetic chain transfer constants play an important role in polymer chemistry [3].

We imagined that the value of chain transfer constants in free-radical polymerization processes is affected by the structural characteristics of the chain transfer agents, aND therefore it should be possible to predict it by quantitative structure–property relationship (QSPR) approaches. In QSPR methods, the chemical properties of molecules are quantitatively correlated to molecular structural features, which HAVE BEEN named molecular descriptors. The results of QSPR analyses were not only used in prediction of properties of new compounds but can also be used to further investigate the mechanism of processes of interest. A quantitative structure–reactivity relationship (QSRR) can be considered as a variant of QSPR, where the chemical reactivity of reactants or catalysts in a specified chemical reaction is related to their chemical structure [4]. The history of structure–reactivity relationship modeling goes back to the end of the 1970s when Carpenter and coworkers qualitatively studied the effects of substituents and of benzannelation on the rates of pericyclic reactions [5, 6]. They used semi-empirical quantum chemical descriptors to develop quantitative models for some unsaturated hydrocarbons. Today, there are a number of reports concerning QSRR investigation of chemical reactions. For example, Hemmateenejad et al. used multiple linear regressions (MLR) and partial least squares (PLS) for QSRR modeling of a Michael addition reactivity index of different substrates using different catalysts (SDS, silica gel, and $ZrOCl_2$) [4]. The results of their investigations revealed that the reactivity of different enones and substrates in Michael addition reactions is controlled by Coulombic interactions (dipole and charge) as well as the orbital energetic parameters. They concluded that different catalysts probably act in different mechanisms. In other work, Katritzky et al. [7] used QSRR approaches to investigate the solvent effect on the decarboxylation rate constants of 6-nitrobenzisoxazole-3-carboxylates. Moreover, Szentpaly et al. [8] developed a quantum chemical structure–reactivity relationship model to predict the chemical reactivity of polycyclic aromatic hydrocarbons. Also, Masunaga et al. developed some QSRR models to predict the transformation rate constants of *p*-substituted benzonitriles in raw sediments and in sediment extract fractions. They concluded that the electronic substituent constant (Hammett σ_p) controlled the transformation rate constants of these compounds [9]. In other work, Ignatz-Hoover et al. [10] used QSRR approaches to predict the kinetic chain transfer constants of some organic agents in polymerization of styrene at 60 °C. They obtained three- and five-parameter MLR models with correlation coefficients of $R^2 = 0.725$ and $R^2 = 0.818$, respectively. Descriptors involved in these equations were consistent with the proposed mechanism of the chain transfer reactions. In the present work, we try to improve these QSRR models by using an artificial neural network (ANN) as a nonlinear feature mapping technique.

Results and discussion

Molecular diversity analysis

In this study, diversity analysis was performed on the selected dataset to make sure that molecules in the structures of the training or test sets can represent those of the whole dataset [11]. We considered a database of n compounds generated from m highly correlated chemical descriptors $\{x_j\}_{j=1}^{m}$. Each compound, X_i, is represented as a vector of:

$$X_i = (x_{i1}, x_{i2}, x_{i3}, \ldots, x_{im}) \quad \text{for } i = 1, 2, \ldots, n \tag{6}$$

where x_{ij} denotes the value of descriptor j of compound X_i. The collective database ($X = \{X_i\}_{i=1}^{N}$) is represented by a $n \times m$ matrix (X) as follows:

$$X = (x_1, x_2, \ldots, X_N)^{\mathrm{T}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} & \\ x_{21} & x_{22} & \cdots & x_{2m} & \\ \vdots & \vdots & \ddots & & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{nm} & \end{bmatrix} \tag{7}$$

Here, the superscript T denotes the vector/matrix transpose. A distance score for two different compounds X_i and X_j can be measured by the Euclidean distance norm d_{ij}, based on the compound descriptors:

$$d_{ij} = \|X_i - X_j\| = \sqrt{\sum_{k=1}^{m} \left(x_{ik} - x_{jk}\right)^2} \tag{8}$$

Then, the mean distances $(\bar{d}_i)$ of one sample to the remaining ones were computed by means of Eq. 9.

$$\bar{d}_i = \frac{\sum_{j=1}^{n} d_{ij}}{n-1} \quad i = 1, 2, \ldots, n \tag{9}$$

The mean distances of samples were normalized within the interval 0–1 and plotted versus the experimental log C_x. The results of this test illustrate the good distribution of test sets among whole of the dataset. The training set with a broad representation of the chemistry space was adequate to ensure model stability and the diversity of the test set can prove the predictive capability of the model.

Model development

Descriptors, which were selected by stepwise variable subset selection, were used as independent variables and log C_x was considered as dependent variable to develop a MLR model. The correlation matrix between these descriptors is shown in Table 1. As can be seen in this table, the linear correlation between each two descriptor is lower than 0.60, which reveals that there is no significant correlation between selected descriptors. The obtained MLR model was used to calculate the log C_x values for test sets as well as the training set.

The MLR calculated values of log C_x are shown in Table 3 (in "Methodology"). Then, an ANN model was developed considering a nonlinear relationship among selected molecular descriptors and log C_x. The statistical parameters of the developed ANN model are shown in Table 2.

The values of standard errors in prediction of log C_x by the ANN model are 0.049, 0.076, and 0.091 for training, internal, and external test sets, respectively, while these values for MLR model are 0.641, 0.964, and 0.843 for training, internal, and external test sets, respectively. Comparison among these values and other statistics in Table 4 (in "Methodology") reveals the superiority of the ANN over the MLR model. These observations show that there are some nonlinear relationships among selected molecular descriptors and log C_x. The ANN predicted values of log C_x are plotted against their experimental values for training, internal, and external test sets in Fig. 1, which shows good correlation.

The residuals of this calculation are plotted against their experimental log C_x in Fig. 2. The random distribution of residuals around the zero line indicates that there is no systematic error contained in the developed ANN model.

Also, it can be seen in this figure that the residual of the predicted value of log C_x for 1,2-dibromoethylbenzene is 0.2, which is three times greater than the standard error of the ANN model and can be considered as an outlier. By removing of this outlier, the R^2 value was improved from 0.896 to 0.944, and the value of standard error became 0.558 for the external test set.

Table 1 Correlation matrix among selected descriptors

	$^1\chi$	ΔH_f/NA	E_{LUMO}	CHD	HDCA-1	$\mathrm{BO}_{\mathrm{C,min}}$
$^1\chi$	1	0.037	−0.314	−0.309	−0.497	−0.164
ΔH_f/NA		1	−0.08	0.033	−0.184	−0.457
E_{LUMO}			1	0.300	0.213	0.076
CHD				1	0.583	0.107
HDCA-1					1	0.230
$\mathrm{BO}_{\mathrm{C,min}}$						1

Model validation

Since the real utility of a QSRR model lies in its ability to accurately predict the modeled property for new molecules, a reliable assessment of the predictive power is necessary for a confident application. This is achieved by validating the model, which is performed in this research in three different ways: external validation test by dividing the dataset into training, internal, and external test sets, internal validation test by applying the cross-validation (CV) test, and the *Y*-randomization procedure. In the present study, we used the leave-many-out (LMO) cross-validation to evaluate the robustness of developed models, the results being indicated by Q^2_{LMO}, which can be calculated from the following equation [12]:

$$Q^2_{\mathrm{LMO}} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \tag{10}$$

where $y_i, \hat{y}_i$, and, $\bar{y}$ were the experimental, predicted, and mean values of log C_x. In LMO, *M* represents a group of randomly selected data points, which one would leave out at the beginning and that would be predicted by the model that was developed by using the remaining data points. So, M molecules are considered as prediction set. The result of the leave-nine-out CV test on the ANN model indicates a Q^2_{LMO} value of 0.85, which showed the robustness of the model. Randomization tests were also carried out to prove the possible existence of chance correlation [13]. The result of 30 times randomization of log C_x vectors gives $\bar{R}^2 = 0.125$, which reveals that there is no chance correlation in the dataset. As mentioned in the previous section, Ignatz-Hoover et al. used the QSRR approaches on the same dataset, and reported the statistics of SE = 0.825, $R^2 = 0.8181$ for their best five-parameter model, without further validation of their model. By comparison between

Table 2 Statistical results of ANN and MLR models

Models	Training set			Internal test set			External test set		
	R^2	SE	*F*	R^2	SE	*F*	R^2	SE	*F*
ANN	0.957	0.049	1417.605	0.931	0.076	94.943	0.896	0.090	60.114
MLR	0.886	0.641	496.273	0.824	0.964	32.708	0.857	0.843	41.897

Table 3 Data set and the observed and MLR and ANN predicted values of log C_x

No.	Name	log C_x exp	log C_x MLR	Residual	log C_x ANN	Residual
1	Benzene	−1.56	−0.66	−0.90	−1.07	−0.49
2	1-Chlorobutane	−1.40	−0.79	−0.61	−0.89	−0.51
3	*tert*-Butylbenzene	−1.30	−0.09	−1.21	−0.24	−1.06
4[a]	1-Bromobutane	−1.22	0.43	−1.65	−0.56	−0.66
5	Methyl 2-chloroacetate	−0.52	−0.72	0.20	−1.08	0.56
6	Acetonitrile	−0.36	0.37	−0.73	−0.87	0.51
7	Diethyl malonate	−0.33	−0.04	−0.29	−0.22	−0.11
8[a]	1-Ethylbenzene	−0.16	−0.34	0.18	−0.67	0.51
9	1-Chlorobenzene	−0.10	0.00	−0.10	−0.12	0.02
10[b]	2-Chlorobutane	0.08	−0.77	0.85	−0.89	0.97
11	Trichloromethane	0.08	0.92	−0.84	0.65	−0.57
12[a]	1-Chloro−2-methylpropane	0.15	−0.90	1.05	−0.89	1.04
13	2-Propen-1-ol	0.18	1.21	−1.03	0.76	−0.58
14	1-Bromobenzene	0.25	0.41	−0.16	0.28	−0.03
15	Phenylamine	0.30	0.67	−0.37	0.28	0.02
16	1,4-Diisopropylbenzene	0.52	0.36	0.16	0.47	0.05
17[a]	1,4-Hydroxybenzene	0.56	1.83	−1.27	0.93	−0.37
18	Dimethyl ketone	0.61	0.82	−0.21	0.30	0.31
19[b]	Benzaldehyde	0.66	1.15	−0.49	1.10	−0.44
20	*N,N*-Dimethylacetamide	0.66	1.46	−0.80	0.35	0.31
21	2-Butanone	0.70	1.24	−0.54	0.85	−0.15
22	Di(2-propenyl) propanedioate	0.72	0.46	0.26	0.63	0.09
23[b]	2-Phenylacetic acid	0.78	1.39	−0.61	0.88	−0.10
24	*sec*-Butylbenzene	0.79	0.02	0.77	−0.05	0.84
25	1,4-Dibutylbenzene	0.85	0.89	−0.04	0.92	−0.07
26	Acetaldehyde	0.93	0.57	0.36	0.67	0.26
27[b]	4-Chlorobenzaldehyde	0.94	1.69	−0.75	1.28	−0.34
28	1,4-Di(*sec*-butyl)benzene	1.03	0.83	0.20	0.91	0.12
29	4-Bromobenzaldehyde	1.08	2.06	−0.98	1.28	−0.20
30	3-Chlorobenzaldehyde	1.14	1.61	−0.47	1.19	−0.05
31	Phenol	1.15	0.63	0.52	0.62	0.53
32	Chloroacetic acid	1.46	1.45	0.01	1.83	−0.37
33	Diethyl 2,2-dichloropropanedioate	1.48	1.93	−0.45	1.93	−0.45
34	Dichloroacetic acid	1.54	1.92	−0.38	1.36	0.18
35[b]	4-Methylphenol	1.59	0.98	0.61	0.89	0.70
36[b]	2-Methylphenol	1.63	0.70	0.93	0.66	0.97
37	(4-Methoxyphenyl)acetonitrile	1.71	1.37	0.34	1.72	−0.01
38	(4-Chlorophenyl)acetonitrile	1.82	1.75	0.07	1.63	0.19
39	Trichloroacetic acid	1.82	2.19	−0.37	2.20	−0.38
40[a]	(3-Bromophenyl)acetonitrile	1.84	2.06	−0.22	2.52	−0.68
41	4-Formylbenzonitrile	1.88	2.86	−0.98	1.98	−0.10
42	Tetrachloromethane	2.03	2.15	−0.12	2.14	−0.11
43	*N,N*-Diethenylphenylamine	2.11	1.55	0.56	2.38	−0.27
44	1-Chloro-4-ethynylbenzene	2.21	1.18	1.03	1.56	0.65
45	1-Bromo-4-ethynylbenzene	2.28	1.55	0.73	2.62	−0.34
46	2,6-Di(2-propyl)phenol	2.49	2.48	0.01	3.16	−0.67
47	2-Bromoacetic acid	2.63	1.81	0.82	3.14	−0.51
48	2,3,5,6-Tetramethylphenol	2.76	2.17	0.59	3.05	−0.29

Table 3 continued

No.	Name	log C_x exp	log C_x MLR	Residual	log C_x ANN	Residual
49	Diethyl 2-bromopropanedioate	2.85	2.36	0.49	2.32	0.53
50	3-Methyl-3-buten-2-one oxime	3.04	3.53	−0.49	3.46	−0.42
51	Methanesulfonyl chloride	3.07	3.50	−0.43	2.94	0.13
52	(*E*)-2-Butenal oxime	3.18	3.63	−0.45	3.62	−0.44
53[b]	(1,2-Dibromoethyl)benzene	3.29	1.66	1.63	1.67	1.62
54	4-Methyl-2-pentanone oxime	3.36	3.22	0.14	3.03	0.33
55	Triphenylgermane	3.36	3.46	−0.10	3.16	0.20
56	Triethylgermane	3.38	2.77	0.61	3.20	0.18
57	3-Buten-2-one oxime	3.43	3.50	−0.07	3.47	−0.04
58[a]	4-Methoxybenzenesulfonyl chloride	3.49	4.19	−0.70	3.75	−0.26
59	1-Ethynyl-4-nitrobenzene	3.50	2.47	1.03	3.52	−0.02
60	4-Methylbenzenesulfonyl chloride	3.50	3.53	−0.03	3.32	0.18
61	Phenylmethanesulfonyl chloride	3.50	4.16	−0.66	3.90	−0.40
62	2-Chloroacetyl chloride	3.52	1.73	1.79	2.55	0.97
63	4-*tert*-Butyl-1,2-benzenediol	3.56	3.05	0.51	3.25	0.31
64	2-Methyl-1-penten-3-one oxime	3.63	4.03	−0.40	3.79	−0.16
65	Benzenesulfonyl chloride	3.64	3.43	0.21	3.57	0.07
66[a]	4-Chlorobenzenesulfonyl chloride	3.88	3.78	0.10	3.75	0.13
67	Iodoacetic acid	3.90	3.01	0.89	3.90	0.00
68	Acetyl bromide	3.93	2.02	1.91	3.29	0.64
69	1,2,3-Benzenetriol	4.02	4.15	−0.13	4.45	−0.43
70	1-Propenaldoxime	4.03	3.62	0.41	3.73	0.30
71[a]	Diethyl 2,2-dibromopropanedioate	4.08	3.16	0.92	3.71	0.37
72	2-Methyl-2-propenal oxime	4.11	3.52	0.59	3.63	0.48
73	Tetrabromomethane	4.33	3.93	0.40	4.29	0.04
74	Chlorodiethylgermane	4.50	4.51	−0.01	4.23	0.27
75[a]	Chlorodimethylgermane	4.52	4.70	−0.18	4.36	0.16
76	Dichloroethylgermane	4.76	5.21	−0.45	4.97	−0.21
77	2,4,6-Trinitrophenylamine	5.07	5.66	−0.59	5.52	−0.45
78	2-Methoxy-1,3,5-trinitrobenzene	5.31	6.19	−0.88	5.13	0.18
79[b]	2,4,6-Trinitrophenol	5.32	5.07	0.25	5.51	−0.19
80	1,3,5-Trinitrobenzene	5.55	5.11	0.44	5.59	−0.04
81	2,5-Dimethyl-2,5-cyclohexadiene-1,4-dione	5.63	5.76	−0.13	5.52	0.11
82[b]	Ethyl 2,4,6-trinitrobenzoate	5.76	5.74	0.02	5.80	−0.04
83	2-Bromo-1,3,5-trinitrobenzene	5.76	5.94	−0.18	5.84	−0.08
84	2,5-Cyclohexadiene-1,4-dione	6.36	5.99	0.37	5.78	0.58

[a, b] The internal and external test sets, respectively

of the results of ANN model (Table 2) and those obtained by Ignatz-Hoover, it was concluded that the developed ANN model was superior over previous work.

Sensitivity analysis and descriptor interpretation

Sensitivity analysis of inputs was used to identify the significance of individual molecular descriptors and to find the order of importance of descriptors. The sequential zeroing of weights (SZW) presented by Nord and Jacobsson was used to find the contribution of each variable in log C_x [14]. This method is based on SZW of the connection between the input variables and the first hidden layer of the generated ANN model. According to this method, the parameter that is calculated in order to show the importance of the *i*th input variable is the difference between the root mean square error (RMSE) of the complete network's predictions and RMSE obtained when the *i*th variable is excluded from the trained network ($RMSE_i$), both being calculated on the same dataset. Then, the differences

Table 4 Specifications of the multiple linear regression model

Descriptor	Notation	Coefficient	Standard error	*t* value	Sig
Kier and Hall index (order 1)	$^1\chi$	0.451	±0.050	8.989	0.000
Final heat of formation/# of atoms	ΔH_f/NA	0.100	±0.022	4.487	0.000
Energy of the lowest unoccupied molecular orbital	E_{LUMO}	−1.264	±0.084	−15.041	0.000
Count of H-donor sites (quantum-chemical PC)	CHD	0.115	±0.028	4.176	0.000
HA dependent HDCA-1/TMSA (quantum-chemical PC)	HDCA-1	73.071	±9.841	7.425	0.000
Min>(0.1) bond order of a C atom	$\mathrm{BO}_{\mathrm{C,min}}$	1.541	±0.250	6.166	0.000
Constant		−1.215	±0.268	−4.532	0.000

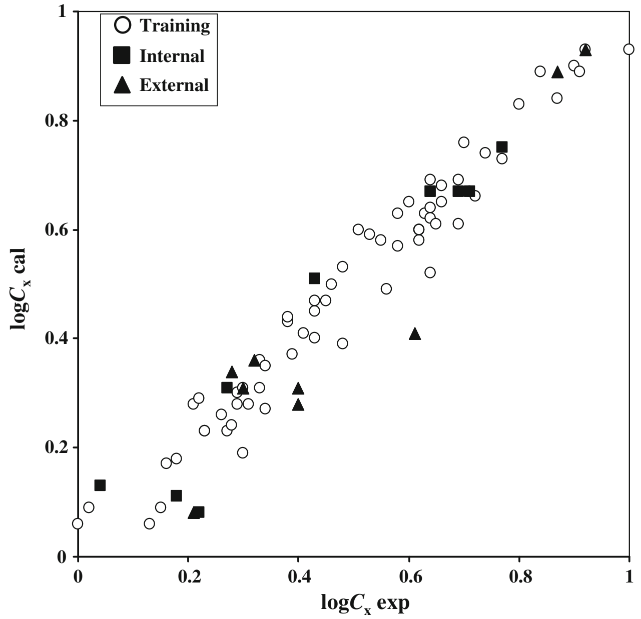

Fig. 1 Plot of ANN calculated versus experimental value of log C_x

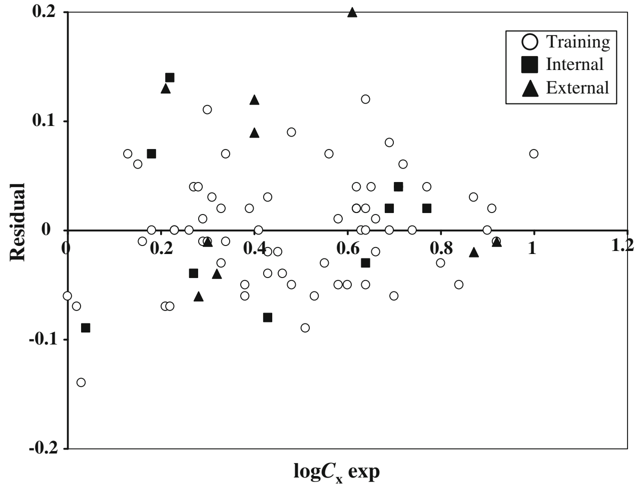

Fig. 2 Plot of ANN calculated residuals versus experimental value of log C_x

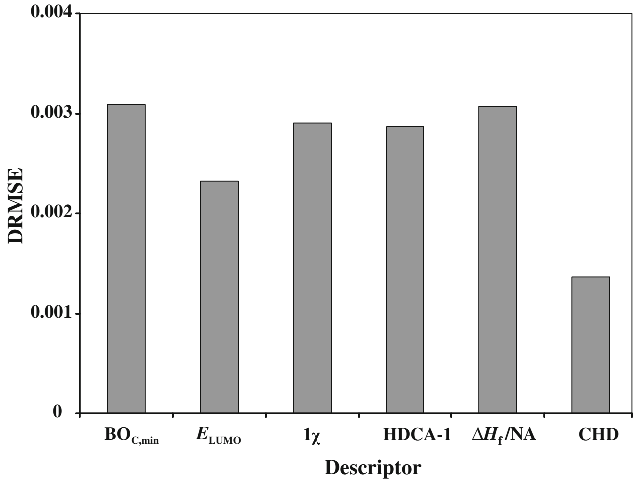

Fig. 3 Results of sensitivity analysis

between RMSE_i and RMSE were calculated (DRMSE) for all inputs. Each variable that produces a greater value of DRMSE is more important. The result of this procedure on the ANN model is shown in Fig. 3.

As can be seen in this figure, the order of importance of selected molecular descriptors is $\mathrm{BO}_{\mathrm{C,min}} > \Delta H_f/\mathrm{NA} > {}^1\chi >$ HDCA-1 $> E_{\mathrm{LUMO}} >$ CHD. According to this analysis, the most important descriptor in the model is the Min>(0.1) bond order of a C atom that is a quantum mechanical valency-related descriptor. This descriptor relates to the strength of intramolecular bonding interactions and characterizes the stability, conformational flexibility, and other valency-related properties of molecules and can account for the probability of formation of the carbon-centered radical (X·) that is the presumed intermediate in the chain transfer reaction [10, 15]. The second descriptor in the model is the final heat of formation/number of atoms, which is a quantum chemical descriptor that quantifies reactive bonds and stability of the molecule [16, 17]. The next descriptor is the first-order Kier and Hall index. This topological descriptor describes the atomic connectivity in the molecule [18, 19]. The first-order Kier and Hall index reflects the influence of the steric factor on the reactivity of the transfer agents [10]. Another descriptor is the hydrogen-bonding dependent hydrogen donor charged area, HDCA-1, which is a quantum chemical descriptor [15]. This descriptor is hydrogen bonding acceptor-dependent hydrogen bonding donor surface area, and encodes the hydrogen bonding acceptor properties of the compounds. Furthermore, the HA-dependent HDCA-1 descriptor reflects the importance of the polarity and hydrogen-bonding ability of the transfer agents, which can facilitate the chain transfer reaction. E_{LUMO} is the fifth descriptor in the model that denotes the energy of the lowest unoccupied molecular orbital. Eventually, as expected, molecules with lower LUMO energies are more reactive [10]. The last descriptor is the count of the H-donor sites in the transfer agents, which distinguishes the molecules according to the number of hydrogen donor sites that are capable of donating a hydrogen [10, 15].

Methodology

Dataset

The dataset consists of the numerical values of kinetic chain transfer constants (log C_{x}) of 84 different organic agents, which were used as the chain transfer agent in free-radical polymerization of styrene at 60 °C taken from [10]. A complete list of the compound's names and corresponding experimental log C_{x} values is given in Table 3. The maximum value of log C_{x} is 6.36 for 2,5-cyclohexadiene-1,4-dione and the minimum value is -1.56 for benzene.

The dataset was divided by using a diversity plot to training, internal, and external test sets, containing 66, 9, and 9 memberas, respectively. The training set was used for adjusting the model parameters, the internal test set was used for preventing of overtraining during ANN model development, and the external test set was used to evaluate the reliability of the obtained models.

Molecular descriptors

Since molecular structures and chemical characteristics of molecules affect their activities/properties, it was necessary to calculate the molecular structural features (molecular descriptors) of transfer agents to predict their chain transfer constant values. Molecular descriptors encode quantitatively the structural and physicochemical features of molecules [20]. In order to compute the molecular descriptors, the structure of each compound was entered via the drawing capabilities of HyperChem (v.7.0) [21]. Then the three-dimensional structures of molecules were pre-optimized with the MM+ molecular mechanics method and then submitted to the semi-empirical quantum chemical AM1 method for geometry optimization to generate three-dimensional structures of molecules [22]. The final optimization of molecular geometry was done using the MOPAC (v.6.0) package [23]. Then, the output of MOPAC and HyperChem files were used for producing constitutional, geometrical, topological, electrostatic, and quantum-chemical descriptors by using the CODESSA (comprehensive descriptors for structural and statistical analysis) package [24]. The CODESSA software, developed by the Katritzky group, enables the calculation of many quantitative descriptors based on the molecular structural information, and codes this chemical information into mathematical form [25–27].

Modeling

Multiple linear regression and artificial neural network methods were employed as linear and nonlinear modeling techniques in this study [28]. An important stage in multilinear regression is searching the best multi-linear equation among a given descriptor set, especially when using a large number of descriptors. Many of these descriptors contain very little information for response or are highly correlated with other descriptors. Thus, the next step is to reduce the number of descriptors by using statistical methods that ignore the dependent variable. Firstly, descriptors with constant values were discarded. Then, one of any two descriptors with a correlation greater than 0.90 was removed to reduce redundant and non-useful information. The stepwise multiple linear regression method was used to select the structural descriptors that are correlated with the chain transfer constant. Then, the variations of R^2 against the number of descriptors in models were used to select the number of descriptors in the model (break-point procedure).

As shown in Fig. 4, no improvement in R^2 was observed after the addition of six parameters to the MLR model. These parameters were: energy of the lowest unoccupied molecular orbital (E_{LUMO}), hydrogen-bonding dependent hydrogen donor charged area (HDCA-1), first-order Kier and Hall index (${}^{1}\chi$), final heat of formation/number of atoms (ΔH_{f}/NA), count of H donors sites (CHD), and Min>(0.1) bond order of a C atom ($\mathrm{BO}_{\mathrm{C,min}}$). These descriptors were selected to develop the MLR model. The specifications of this six-parameter MLR model are presented in Table 4.

Since linear regression analysis ignores nonlinear relationships that may exist between property and descriptors, these descriptors were fed to a three-layered ANN as input vectors to develop a nonlinear QSRR model. Detailed descriptions of the theory behind artificial neural networks have been adequately described elsewhere [29–32]. In

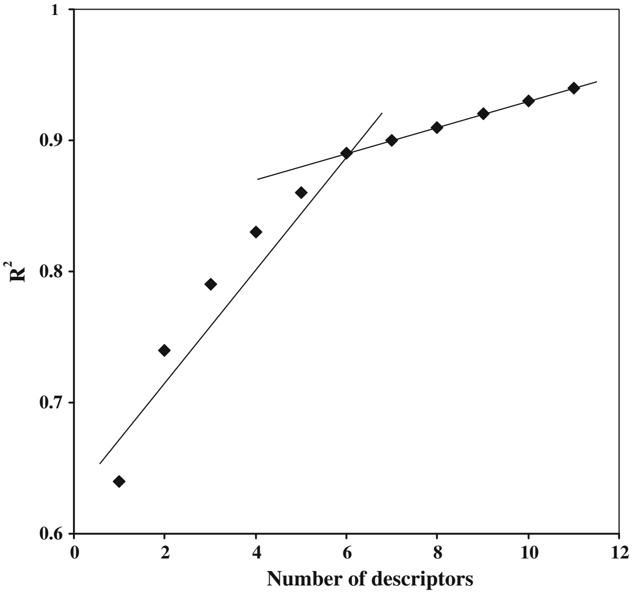

Fig. 4 Influence of the number of descriptors on R^2

addition, we reported some relevant principles of the ANNs in earlier papers [33–35]. In this work, the ANN program was coded in MATLAB 7 for windows [36]. In the first step of development of the ANN model, a three-layer network with a sigmoid transfer function was designed both for the hidden and output nodes. The values of log C_x were normalized between 0 and 1 and were used as target vectors. The number of nodes in the input layer of the network was equal to the number of descriptors, which appears in the MLR model. The adjustable weights among neurons have the random distribution between −0.3 and 0.3. The ANN parameters number of nodes in the hidden layer, weights learning rate, biases learning rate, and momentum optimized were set to 4, 0.1, 0.9, and 0.5, respectively. The procedure of optimizing these parameters was given in our previous work [37, 38]. The developed network was then trained by using the training set by back propagation strategy for optimization of the weights and bias values. The goal of training the network is to change the weights among the layers for minimizing the output errors. It should be noted that to prevent overfitting the training of the network must be stopped when the RMSE in the prediction of log C_x of the internal test set starts to increase. Then, the trained network was used to calculate the log C_x values of the external test set as well as the internal and training sets.

References

1. Moad G, Moad CL (1996) Macromolecules 29:7727
2. Odian G (1981) Principles of polymerization, 2nd edn. Wiley, New York
3. Braun D, Hempler P (1993) Polym Bull 30:55
4. Hemmateenejad B, Sanchooli M, Mehdipour A (2009) J Phys Org Chem 22:613
5. Carpenter BK (1978) Tetrahedron 34:1877
6. Wilcox CF, Carpenter BK, Dolbier WR (1979) Tetrahedron 35:707
7. Katritzky AR, Perumal S, Petrukhin R (2001) J Org Chem 66:4036
8. Szentpaly LV, Herndon WC (1988) Adv Chem Ser 217:287
9. Masunaga S, Wolfe NL, Carriera LH (1995) Environ Toxicol Chem 14:1827
10. Ignatz-Hoover F, Petrukhin R, Karelson M, Katritzky AR (2001) J Chem Inf Comput Sci 41:295
11. Maldonado AG, Doucet JP, Petitjean M, Fan BT (2006) Mol Divers 10:39
12. Osten DW (1988) J Chemom 2:39
13. Kraim K, Khatmi D, Saihi Y, Ferkous F, Brahimi M (2009) Chemom Intell Lab Syst 97:118
14. Nord LI, Jacobsson SP (1998) Chemom Intell Lab Syst 44:153
15. Luan F, Ma W, Zhang X, Zhang H, Liu M, Hu Z, Fan BT (2006) Chemosphere 63:1142
16. Dashtbozorgi Z, Golmohammadi H (2010) Eur J Med Chem 45:2182
17. Katritzky AR, Tatham DB (2001) J Chem Inf Comput Sci 41:1162
18. Kier LB, Hall LH (1986) Molecular connectivity in structure-activity analysis. Wiley, New York
19. Fatemi MH, Karimian F (2007) J Colloid Interface Sci 314:665
20. Li H, Ung CY, Yap CW, Xue Y, Li ZR, Chen YZ (2006) J Mol Graph Modell 25:313
21. HyperChem Release 7.0 for windows (2002) Hypercube, Saint-Laurent
22. Li Q, Chen X, Hu Z (2004) Chemom Intell Lab Syst 72:93
23. Stewart JPP (1989) MOPAC ver. 6.0, Quantum chemistry program exchange, QCPE, No. 455, India University
24. Katritsky A, Karelson M, Lobanov VS, Dennington R, Keith T (2004) CODESSA 2.7.2. Semichem, Shawnee
25. Katritzky AR, Lobanov VS, Karelson M (2002) Comprehensive descriptors for structural and statistical analysis, Reference Manual, Version 2.0. Semichem and University of Florida, Florida
26. Katritzky AR, Lobanov VS, Karelson M (1995) Chem Soc Rev 24:279
27. Katritzky AR, Lobanov VS, Karelson M (1997) Pure Appl Chem 69:245
28. Consonni V, Todeschini R (2002) Handbook of molecular descriptors. Wiley, Weinheim
29. Zupan J, Gasteiger J (1999) Neural networks in chemistry and drug design. Wiley, Weinheim
30. Bose NK, Liang P (1996) Neural network, fundamentals. McGraw-Hill, New York
31. Beal MT, Hagan HB, Demuth M (1996) Neural network design. PWS, Boston
32. Zupan J, Gasteiger J (1993) Neural networks for chemists: an introduction. VCH, Weinheim
33. Fatemi MH (2003) J Chromatogr A 1002:221
34. Fatemi MH (2002) J Chromatogr A 955:273
35. Jalali-Heravi M, Fatemi MH (2001) J Chromatogr A 915:177
36. MATLAB version 7.0 (2004) The MathWorks http://www.mathworks.com
37. Jalali-Heravi M, Fatemi MH (2000) Anal Chim Acta 415:95
38. Jalali-Heravi M, Fatemi MH (2000) J Chromatogr A 897:227