Available online at www.sciencedirect.com
SCIENCE DIRECT
Talanta 69 (2006) 736–740
Talanta
www.elsevier.com/locate/talanta

Prediction of selectivity coefficients of a theophylline-selective electrode using MLR and ANN

S. Riahi [a], M.F. Mousavi [a,*], M. Shamsipur [b]

[a] *Department of Chemistry, Tarbiat Modares University, P.O. Box 14115-175, Tehran, Islamic Republic of Iran*
[b] *Department of Chemistry, Razi University, Kermanshah, Islamic Republic of Iran*

Received 21 June 2005; received in revised form 3 November 2005; accepted 4 November 2005
Available online 13 December 2005

Abstract

The selectivity coefficient of 24 interfering compounds (drugs, amino acids and organic compounds) of a theophylline-selective electrode was predicted using an artificial neural network (ANN). The multiple linear regression (MLR) technique was used to select the descriptors as inputs for the artificial neural network. The neural network employed here is a connected back-propagation model with a 2-2-1 architecture. Two topological indices for the interfering compounds, namely, Narumi harmonic topological index, HNar, and sum of topological distances between nitrogen and oxygen, T(N···O), were taken as inputs for the ANN. Standard errors of training and prediction were 0.954 and 0.945, respectively, for the MLR model and 0.032 and 0.007, respectively, for the ANN model. Two topological indices for the interference of the electrode were taken as inputs for ANN.

Keywords: Selectivity coefficient; Theophylline-selective electrode; Artificial neural network; Regression analysis; Structure-activity relationships

1. Introduction

Ion-selective electrodes (ISEs) are electrochemical sensors that respond selectively to the activity of ionic species. Since, in practice, no electrode is sensitive to a particular ion while excluding all others, the term 'selective' is preferred over the term 'specific'. In fact, the presence of any species other than the ion being measured may affect the measured emf of the cell to some extent. Other possible sources of change in the emf include the chemical reactions with the membrane material, and reactions in the sample solution, which may cause precipitation, complexation, oxidation or reduction of the ion being measured. Selectivity coefficient is an expression of the extent to which an ion-selective electrode responds to an interfering ion relative to the measured ion. The selectivity coefficient depends on a number of factors including the ionic strength of solution, the concentration of both ions and temperature [1].

The selectivity is clearly one of the most important characteristics of a sensor, as it often determines whether a reliable measurement in the target sample is possible. It is especially critical in clinical applications of ion-selective electrodes, where for whole blood or serum measurements the allowed emf deviation (error) should not be larger than 0.1 mV [2].

Thorough theoretical description of selectivity allows researchers to identify the key parameters for optimizing the performance of potentiometric sensors, e.g. by adjusting weighing parameters (i.e. absolute membrane concentrations) or choosing different plasticizers or matrices [3,4].

The non-ideal behavior of a liquid membrane ion-selective electrode is usually described by potentiometric selectivity coefficients ($K^{\mathrm{Pot}}_{i,j}$), defined in semiempirical Nikolsky–Eisenman equation [5]. Generally, $K^{\mathrm{Pot}}_{i,j}$ is used to express the ability of an electrode to distinguish between the desired ion (i) and the interfering one (j). Different experimental methods have been proposed for the determination of selectivity coefficients [5–9]. In the IUPAC commission held in 1975 [8], the separate solution method (SSM) was recommended only if the electrode exhibits Nernstian responses to both the mother ion and the interfering ions. However, it was considered less desirable compared to the fixed interference method (FIM), because it does not represent the actual conditions under which the electrodes are used. In order to overcome the limitations and inconveniences of

* Corresponding author. Fax: +98 21 88006544.
E-mail address: mousavim@modares.ac.ir (M.F. Mousavi).

0039-9140/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2005.11.010

the selectivity coefficients based on Nikolsky–Eisenman equation including values found for ions of unequal charges, non-Nernstian behavior of interfering ions and activity dependence of $K_{i,j}^{Pot}$, in 1975, Umezawa et al. proposed the matched potential method (MPM), which is independent of Nikolsky–Eisenman equation [10]. However, due to its critical importance, the determination of selectivity coefficients has remained as a challenging task and, thus, a number of recent modifications have been proposed in this respect [11–18].

The use of artificial neural networks (ANNs) in chemistry has grown substantially [19,20]. There are several reports on the use of neural networks in the modeling of retention behavior and optimization of conditions in micellar liquid chromatography [21,22]. Artificial neural networks have also been applied to a wide variety of chemical problems such as quantitative structure-activity relationship (QSAR) studies [23–26], prediction of ^{13}C NMR chemical shift [27], selectivity coefficients of ion-selective electrodes [28], simulation of mass spectra [29] and modeling of ion-interaction chromatography [30,31].

The main goal of the present work was the development of an ANN for modeling of the selectivity coefficient of a series of drugs, amino acids and other compounds reported for a new theophylline-selective electrode [32]. A linear regression model was also developed and its results were compared with the calculated ANN selectivity coefficients. This comparison clarified the non-linear characteristics of the selectivity coefficient of different compounds studied in this work.

2. Experimental

2.1. Data set

We have recently reported the selectivity coefficients of a series of 24 drugs, amino acids and organic compounds for a new theophylline-selective electrode based on bis(phenyl)-4-(phenyl)-3H-thiopyran (PPT) (Fig. 1) [32], which were employed as the data set for this work. It should be noted that the selectivity coefficient of each compound was mean of five determinations with no systematic errors, as indicate by the *t*-test.

In the present work, this data set was randomly divided into two groups, a training set consisting of 18 compounds and a prediction set that includes 6 compounds (Table 1). The training set was used for the generation of the network and the prediction set was used to evaluate the generated network. As it is obvious from Table 1, the $K_{i,j}^{Pot}$ values of the studied compounds ranged from 1.32×10^{-5} for glycine to 7.92×10^{-2} for the caffeine.

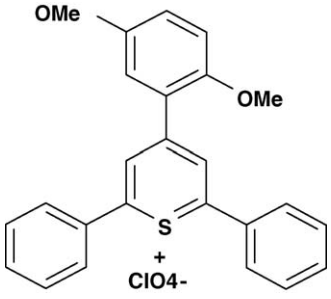

Fig. 1. Structure of PPT.

Table 1
Experimental, ANN and MLR calculated values of selectivity coefficients together with the values of the descriptors appearing in the model for the training and prediction sets

Number	Compound	Descriptors		−Logarithm selectivity coefficient		
		HNar	T(N···O)	K_{ANN}	K_{MLR}	K_{EXP}
Training set						
1	Proline	1.714	6	2.914	2.527	2.92
2	Serine	1.355	9	4.117	4.113	4.12
3	Cysteine	1.355	6	4.235	4.231	4.23
4	Aspartic acid	1.385	14	3.636	3.774	3.64
5	Arginine	1.5	50	1.907	1.814	1.90
6	Caffeine	1.68	20	1.116	1.721	1.10
7	Histamine	1.846	0	1.397	1.951	2.27
8	Threonine	1.333	9	3.652	4.217	3.66
9	Glycine	1.304	6	4.876	4.473	4.88
10	Tryptophan	1.837	18	2.495	1.923	2.49
11	Leucine	1.385	6	4.119	4.088	4.12
12	Valine	1.333	6	4.029	4.335	4.25
13	Phenylalanine	1.714	6	4.090	3.695	4.12
14	Tyrosine	1.66	13	3.308	2.508	3.22
15	Glutamic acid	1.429	16	3.418	3.487	3.41
16	Isoleucine	1.385	6	4.502	4.088	4.02
17	Ephedrine	1.714	3	2.586	2.645	2.65
18	Asparagine	1.385	20	2.929	3.538	3.00
Prediction set						
19	Methionine	1.459	6	4.012	3.737	4.01
20	Glutamine	1.429	23	3.028	3.211	3.02
21	Lysine	1.5	20	3.339	2.993	3.34
22	Histidine	1.692	28	1.338	1.767	1.34
23	Imidazole	2	0	1.254	1.406	1.25
24	Alanine	1.286	6	4.256	4.558	4.25

2.2. Descriptor generation

In the present work, the topological descriptors were employed as numerical parameters that relate selectivity coefficient of the only 18 molecules with their structures. A total of 256 descriptors were calculated for each molecule of the data set by DRAGON software version 3 [33].

2.3. MLR analysis

A stepwise multiple linear regression procedure was used for model generation. From pairs of variables with $R > 0.90$, only one of them is used in modeling and those variables that over the 90% of which were equal to zero are eliminated. By using these criteria, 192 out of 256 original descriptors were eliminated and the remaining descriptors were used to generate the models using the SPSS/PC software package [34]. A stepwise procedure was used for selection of descriptors. This method combines the forward and backward procedures. Due to the complexity of inter-correlations, the variance explained by certain variables will change when new variables enter the equation.

Table 2
Selected model of multiple linear regression[a]

Descriptor	Notation	Coefficient	Mean effect
Narumi harmonic topological index	HNar	−4.745 (±0.451)	−2.480
Sum of topological distances between nitrogen and oxygen	T(N···O)	−0.0393 (±0.008)	−0.158
Constant		10.896 (±0.66)	

[a] Statistics of the model: $n = 18$, $R = 0.919$, S.E. = 0.954 and $F = 45.0$.

Sometimes a variable that is qualified to enter loses some of its predictive validity when other variables enter. If this takes place, the stepwise method will remove the weakened variable.

A final set of selected equations was then tested for stability and validity through a variety of statistical methods. The choice of equation suitable for further consideration was made by using four criteria, namely, multiple correlation coefficients (R), standard error (S.E.), F-statistic and the number of descriptors in the model. The best multiple linear regression (MLR) model is one that has high R and F-values, low standard error, least number of descriptors and high ability for prediction. The best model selected in this work is presented in Table 2.

2.4. Neural network generation

The ANN program was written in MATLAB 6 in our laboratory. The network was generated using descriptors appearing in the MLR model as inputs. A three-layer network with a sigmoidal transfer function was designed. The initial weights were randomly selected from a uniform distribution between −1 and +1. Before training, the input and output values were normalized between 0 and 1. Number of neurons in the hidden layer, learning rate and momentum were optimized. A feed-forward neural network with back-propagation of an error algorithm was constructed to model the structure-activity relationship. Our network has one input layer, one hidden layer and one output layer. The number of nodes in the input layer is dependent on the number of descriptors introduced in the network. A bias unit with a constant activation of unity is connected to each unit in the hidden and output layers. The ANN models are confined to a single hidden layer because the network with more than one hidden layer is harder to train. The number of nodes in the hidden layer is optimized through a learning procedure. There is only one node in the output layer. For each descriptor subset, the best topology of the ANN's was searched by using the training and validation data sets. The validation set is used to monitor the overall performances of the trained network. Once the best topology of the network is obtained and the convergence criterion is reached, a leave-5-out cross-validation procedure is also employed to more validate the performances of the resulted networks. The root-mean-square errors of the training, validation, and cross-validation (RMSET, RMSEV and RMSEC-V, respectively) and the corresponding correlation coefficients (R^2_T, R^2_V and $R^2_{C\text{-}V}$, respectively) have been monitored during the training of the networks [35]. To evaluate the performance of the ANN, root-mean-square errors of the training and validation were used. The number of neurons of the hidden layer with the minimum value of RMSET was selected as the optimum number. Learning rate and momentum were optimized in a similar way. We have used the validation set to examine the validity of the ANN model.

The standard error (S.E.) for the training or prediction sets was calculated as follows:

$$\text{S.E.} = \sqrt{\frac{\sum_{i=1}^{n}(t_i - y_i)^2}{n}} \quad (1)$$

where t and y are the target and calculated output values, respectively, and n indicates the number of training or prediction patterns. It should be noted that the number of input nodes is 2, which is equal to the number of descriptors appearing in the MLR model, and the number of output nodes is 1. We applied every time the same number of epochs to training in the learning and testing set (number of epochs = 5500).

3. Results and discussion

The experimental and calculated values of the selectivity coefficients using both the MLR and ANN methods for the drugs, amino acids and organic compounds studied in this work together with the values of the two descriptors appearing in the selected MLR model are given in Table 1.

3.1. Multiple regression analysis

The MLR technique was performed on the molecules of the training set shown in Table 1. After regression analysis, a few suitable models were obtained among which the best model was selected and presented in Table 2. As seen, the two descriptors appeared in this model consist of Narumi harmonic topological index, HNar, and sum of topological distances between nitrogen and oxygen, T(N···O). In order to obtain the extent of contributions of each descriptor in the selectivity coefficient, the mean effect of each parameter was calculated and given in Table 2. The mean effect of a descriptor is a product of its mean value and the regression coefficient in the MLR model. For the calculation of these parameters, the algorithms given by Narumi [36], for HNar, and by Trinajstic [37,38], for T(N···O), were used.

Molecular descriptors represent important tools for predicting the chemical properties, for classifying the chemical structures, and for seeking similarities among them. The properties of the topological descriptors are discussed and it is shown that they can serve for modeling and predicting a wide range of properties. They are numerical quantifiers of molecular topology that reveal the rule of the size, shape, branching pattern, cyclicity, steric interactions, molecular complexity and symmetry of molecular graphs on the selectivity coefficient. Besides, these indices are

related to the number of atoms and how they are connected in a molecule [39–42].

HNar is a harmonic topological index related to the molecular branching and represents the number of non-hydrogen atoms divided by the reciprocal vertex degree summation. As it is obvious seen Table 2, HNar is the much more effective descriptor than T(N···O), meanwhile based on the data given in Table 2, the relationship between selectivity coefficient and the descriptors' coefficient is $-\log K_{sel} = -4.91\ \mathrm{HNar} - 0.047\ \mathrm{T(N\cdots O)} + 11.04$, which indicates a direct relationship between the selectivity coefficient and HNar. This means that, as it is expected, the increased branching of the interfering compounds will result not only in easier diffusion of the molecule inside the lipophilic membrane, but also in their enhanced tendency for interaction with the ionophore PPT. The net result would be an increase in the interfering effect of these compounds on the potentiometric behavior of the theophylline ion-selective electrode. Moreover, the positive influence of T(N···O) on the selectivity coefficient emphasizes the fact that the increased distance between nitrogen and oxygen atoms of the interfering molecules will result in their increased interfering effect on the electrode response.

In order to check the suitability of PLS method for non-linear modeling of the system, the inverse, square, cube and product of the two descriptors were calculated, and were used as new descriptors in the MLR method. The result revealed that only the two main descriptors (i.e. HNar and T(N···O)) were entered, and the rest of descriptors did not show any improvement in the system. Thus, it was concluded that ANN should be a much more efficient method than PLS for the modeling of such a complicated system.

3.2. Neural network analysis

In order to investigate the non-linear interactions between different parameters in the MLR model, an ANN was developed to predict the selectivity coefficients of the drugs, amino acids and organic compounds. The ANN was generated using the descriptors appearing in the MLR model as inputs. A 2-2-1 ANN was developed. The ANN calculated values of the selectivity coefficients for the training and prediction sets are also included in Table 1. Based on the data given in Tables 1 and 3, a comparison between the results obtained by the ANN and MLR methods clearly indicates the superiority of ANN over that of the MLR model. As it is seen from Table 3, the standards error of training (SET) and standards error of prediction (SEP) have been reduced from 0.354 and 0.22, for the MLR model, to 0.032 and 0.007, for the ANN model, respectively. Table 1 shows that the trend of variation in the ANN predicted values of selectivity coefficients are in agreement with the experiment values for different compounds. This confirms the validity of the ANN model and selection of the MLR descriptors as inputs for ANN modeling.

Table 3
Comparison between the results obtained using the MLR and ANN models

Data set	MLR		ANN	
	R	S.E.	*R*	S.E.
Training set	0.919	0.954	0.999	0.032
Prediction set	0.982	0.945	0.998	0.007

Table 4
Comparison of the SET and SEP of the selected model with the test models obtained using different molecules as prediction set

Model	SET	SEP	Molecules in the prediction set[a]
Selected model	0.032	0.007	19, 20, 21, 22, 23 and 24
Test model I	0.007	0.015	1, 5, 8, 11, 15 and 18
Test model II	0.008	0.019	2, 4, 7, 10, 12 and 13
Test model III	0.009	0.020	3, 6, 9, 14, 15 and 17

[a] Numbers refer to the number of the compounds given in Table 1. The remaining molecules for each set are due to the corresponding training set.

It is noteworthy that the ANN has a 2-2-1 architecture with 24 adjustable parameters, while the training set consists of 18 compounds. Since the data set was small, different prediction and training sets were chosen and the network was trained using these training sets. Each time, a set of six compounds out of 24 molecules was chosen randomly as a prediction set, and a network was developed using the remaining compounds and then the selectivity coefficients of these six compounds were predicted by using the ANN model. This procedure was repeated three times and the results for the three test sets are given in Table 4. Obviously, the remaining molecules of each set were the corresponding training sets. As can be seen from Table 4, the results obtained are not depend on the molecules employed in the prediction set.

A plot of the calculated against the experimental selectivity coefficients (Fig. 2) indicates an excellent correlation between the experimental and predicted values. In Fig. 3 are plotted the residuals of ANN predicted values of selectivity coefficients against the experimental values. As the calculated residuals are distributed on both sides of the zero line, one may conclude that there is no systematic error in the development of the neural network.

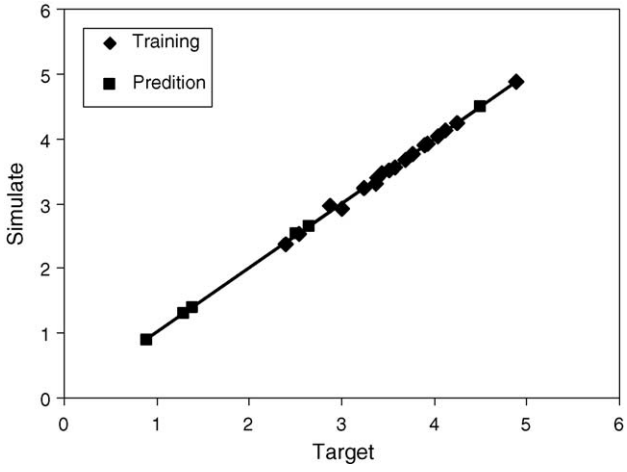

Fig. 2. Plot of the calculated selectivity coefficient against the experimental values.

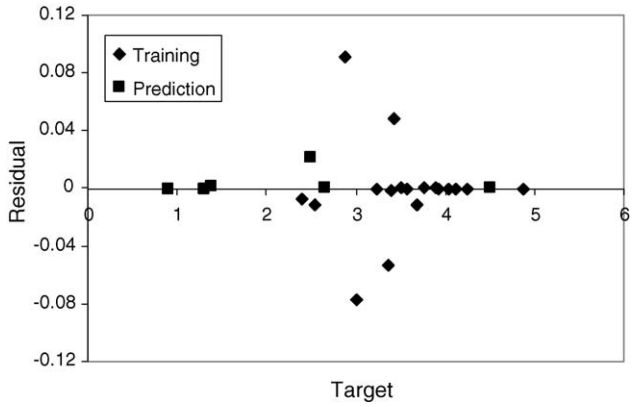

Fig. 3. Plot of the residuals vs. experimental selectivity coefficient.

4. Conclusion

The MLR and ANN modelings were applied to the prediction of the selectivity coefficients of a variety of interfering compounds for a theophylline ion-selective electrode. Both methods seem to be useful, although a comparison between the two methods revealed the superiority of the ANN over the MLR model. Moreover, the ANN is able to predict the trend of variation in the values of selectivity coefficients for different compounds, while the MLR has lower predictive ability in this respect. The superiority of ANN over MLR indicates that the selectivity coefficients of drugs, amino acids and organic compounds possess some non-linear characteristics. The results clearly showed that the MLR is a suitable technique for choosing the inputs for the ANN modeling.

Acknowledgment

We gratefully acknowledge the support of this work by the Tarbiat Modares University (TMU) Research Council.

References

[1] http://www.nico2000.net/Book/Guide1.html.
[2] U. Oesch, D. Ammann, W. Simon, Clin. Chem. 32 (1986) 1448.
[3] P.C. Meier, W.E. Morf, M. Laubli, W. Simon, Anal. Chim. Acta 156 (1984) 1.
[4] U. Schaller, E. Bakker, U.E. Spichiger, E. Pretsch, Anal. Chem. 66 (1994) 391.
[5] K. Srinivasan, G.A. Rechnitz, Anal. Chem. 41 (1969) 1203.
[6] G.J. Moody, J.D.R. Thomas, Lab. Pract. 20 (1971) 307.
[7] G.J. Moody, J.D.R. Thomas, Selective Sensitive Electrodes, Merrow, England, 1971.
[8] G.G. Guibault, R.A. Durst, M.S. Frant, H. Freiser, E.H. Hansen, T.S. Light, E. Pungor, G.A. Rechnitz, N.M. Rice, T.J. Rohm, W. Simon, J.D.R. Thomas, Pure Appl. Chem. 48 (1976) 127.
[9] Y. Umezawa, K. Umezawa, H. Sato, Pure Appl. Chem. 67 (1995) 507.
[10] V.P.Y. Gadzekpo, G.D. Christian, Anal. Chim. Acta 164 (1984) 279.
[11] E. Bakker, R.K. Meruva, E. Pretsch, E. Meyerhoff, Anal. Chem. 66 (1994) 3021.
[12] E. Bakker, J. Electrochem. Soc. 143 (1996) L83.
[13] E. Bakker, Anal. Chem. 69 (1996) 1061.
[14] R.J. Forster, D. Diamond, Anal. Chim. Acta 276 (1993) 75.
[15] F.J. Saez de Viteri, D. Diamond, Electroanalysis 6 (1994) 9.
[16] P. Kane, D. Diamond, Talanta 44 (1997) 1847.
[17] W. Zhang, A. Fakler, C. Demuth, U.E. Spichiger, Anal. Chim. Acta 375 (1998) 211.
[18] M. Nagele, E. Bakker, E. Pretsch, Anal. Chem. 71 (1999) 1048.
[19] M. Gevrey, I. Dimopoulos, S. Lek, Ecol. Modell. 160 (2003) 249–264.
[20] H. Chan, A. Butler, D.M. Falck, M.S. Freund, Anal. Chem. 69 (1997) 2373.
[21] O. Jimenz, I. Benito, M.L. Marina, Anal. Chim. Acta 353 (1997) 367.
[22] H.J. Metting, P.M.J. Coenegracht, J. Chromatogr. A 728 (1996) 47.
[23] K.L. Peterson, Anal. Chem. 64 (1992) 379.
[24] A.R. Katritzky, E.V. Gordeeva, J. Chem. Inf. Comput. Sci. 33 (1993) 835.
[25] A. Moosavi-Movahedi, D. Safarian, S. Riahi, M.F. Mousavi, Nucleot. Nucleos. Nucleic Acid 23 (2003) 115.
[26] M. Jalali-Heravi, M.H. Fatemi, J. Chromatogr. A 897 (2000) 227.
[27] S.L. Anker, P.C. Jurs, Anal. Chem. 64 (1992) 1157.
[28] W.L. Xing, X.W. He, Anal. Chim. Acta 349 (1997) 283.
[29] M. Jalali-Heravi, M.H. Fatemi, Anal. Chim. Acta 415 (2000) 95.
[30] E. Maremgo, M.C. Gennaro, S. Anglino, J. Chromatogr. A 799 (1998) 47.
[31] G. Sacchero, M.C. Bruzzonoti, C. Sarzamini, E. Mentasti, H.J. Metting, P.M.J. Coenegracht, J. Chromatogr. A 799 (1998) 35.
[32] S. Riahi, M.F. Mousavi, S.Z. Bathae, M. Shamsipur, Anal. Chim. Acta 548 (2005) 192.
[33] R. Todeschini, Milano Chemometrics and QSAR Group, http://www.disat.unimib.it/vhm/.
[34] SPSS/PC, Statistical Package for IBMPC, Quiad software, Ontario, 1986.
[35] B. Hemmateenejad, M. Akhond, R. Miri, M. Shamsipur, J. Chem. Inf. Comput. Sci. 43 (2003) 1328.
[36] H. Narumi, Comm. Math. Comp. Chem. 22 (1987) 195.
[37] D. Bonchev, N. Trinajstic, J. Chem. Phys. 67 (1977) 4517.
[38] Z. Mihalic, S. Nikolic, N. Trinajstic, J. Chem. Inf. Comput. Sci. 32 (1992) 28.
[39] K. Baumann, TRAC 18 (1999) 36.
[40] M. Jalali-Heravi, M.H. Fatemi, J. Chromatogr. A 915 (2001) 177.
[41] S. Agatonovic-Kustrin, L.H. Ling, S.Y. Tham, R.G. Alany, J. Pharm. Biomed. Anal. 29 (2002) 103.
[42] E.A. Castro, M. Tueros, A.A. Toropov, Comput. Chem. 24 (2000) 571.