Asian Journal of Chemistry; Vol. 25, No. 1 (2013), 349-352

ASIAN JOURNAL OF CHEMISTRY

http://dx.doi.org/10.14233/ajchem.2013.13061

Quantitative Structure-Electrochemistry Relationship Study for Prediction of Half-Wave Reduction Potentials of Some Chlorinated Organic Compounds by Genetic Algorithm-Multiple Linear Regression

MAJID MOHAMMADHOSSEIN* and MEHDI NEKOEI

Young Researchers Club, Shahrood Branch, Islamic Azad University, Shahrood, Iran

*Corresponding author: Fax: +98 273 3394537; Tel: +98 273 3394530; E-mail: majidmohammadhosseini@yahoo.com; m_mhosseini@iau-shahrood.ac.ir

(*Received*: 14 November 2011; *Accepted*: 26 July 2012) AJC-11876

Quantitative structure-electrochemistry relationship model has been used to predict and explain half-wave reduction potentials ($E_{1/2}$). This method allows for the prediction of $E_{1/2}$s in a variety of organic compounds based on their structures alone. Genetic algorithm-multiple linear regression (GA-MLR) was performed to build the model. The proposed methodology was validated using leave-one-out and leave-group-out cross validation using division of the available data set into training and test sets. The results illustrated that the linear techniques such as multiple linear regression combined with a successful variable selection procedure like GA are capable to generate an efficient quantitative structure-electrochemistry relationship model for predicting the $E_{1/2}$s of different compounds. A model with low prediction error and good correlation coefficient was obtained ($R^2_{calibration}$ = 0.923, $R^2_{prediction}$ = 0.940, Q^2_{LOO} = 0.810, Q^2_{LGO} = 0.803, R^2_{adj} = 0.889, RMSEP = 0.203). This model was used for the prediction of the $E_{1/2}$ values of some organic compounds, which were not used in the modeling procedure.

Key Words: QSER, Half-wave potentials, Genetic algorithm-multiple linear regression, Chlorinated organic compounds.

INTRODUCTION

Half-wave potential ($E_{1/2}$) is an important electrochemical property of diverse organic compounds. This property which is a constant characteristic for a reversible oxidation-reduction system can be useful for predicting electrochemical properties of other organic compounds. There are some different electrochemical methods, which permit the determination of the half-wave potentials of a wide variety of organic and organometallic compounds[1,2].

Nowadays much interest is devoted to the prediction of physicochemical properties of molecules, such as their biological activity, chemical property, their retention on chromatographic systems, or electrochemical property, *etc*. This is usually accomplished by implementing so-called quantitative structure-property relationship (QSPR) models, which relate the property of interest, with a set of molecular descriptors. These descriptors encode the chemical information and are related to certain physicochemical properties of the molecule[3]. In such studies, numerous physical properties of molecular systems have been successfully modeled, including boiling points, aqueous solubility and polymer properties, *etc*.[4-10]. Quantitative structure-electrochemistry relationships have been used to construct simple and reliable models to explain and predict the electrochemical behaviour of various classes of compounds[2,11-13]. In such studies, quantitative structure-electrochemistry relationships (QSERs) for half-wave potentials ($E_{1/2}$) have been reported for different types of organic compounds[14-17]. Quantitative structure-electrochemistry relationship has been demonstrated to be a powerful tool in electrochemistry.

Application of quantitative structure-electrochemistry relationships techniques usually requires selection of variables to build well-fitting models. In this work we used the genetic algorithm (GA) method for variable selection combined with multiple linear regression (MLR).

The main aim of this work is to search for an efficient method to build an accurate quantitative relationship between the molecular structure and the $E_{1/2}$ of the some organic compounds by genetic algorithm feature election strategy and multiple linear regression analysis.

EXPERIMENTAL

Computer hardware and software: A Pentium IV personal computer (CPU at 3.06 GHz) with the Windows XP operating system was used. The geometry optimization was performed with HyperChem. (Version 8.0 Hypercube, Inc). For the calculation of the molecular descriptors, the Dragon 2.1 software was used. The SPSS software (Version 14, SPSS, Inc.) was employed for the multiple linear regression analysis, while other calculations were performed in the MATLAB (Version 7, Math Works, Inc.) environment.

Data set: Experimental half-wave potentials ($E_{1/2}$) data of some chlorinated organic compounds were taken from reference[18] as our chosen dataset. The names of these compounds and their experimental half-wave potentials are listed in Table-1. The ranges of $E_{1/2}$ values of these compounds are -1.10 to -2.15. Moreover, the calculated half-wave potentials for these compounds by GA-MLR method are tabulated in Table-1.

TABLE-1
THE DATA SET AND THE CORRESPONDING OBSERVED AND PREDICTED $E_{1/2}$ VALUES BY GA-MLR FOR THE TRAINING AND TEST SETS

No.	Compound	Exp.[a]	Predicted[b]
Training set			
1	2-Chlorotoluene	-2.15	-2.10
2	3-Chlorotoluene	-2.09	-2.10
3	4-Chlorotoluene	-2.10	-2.02
4	2,3-Dichlorotoluene	-1.89	-1.77
5	2,4-Dichlorotoluene	-1.88	-1.84
6	3,4-Dichlorotoluene	-1.53	-1.72
7	2,3,4-Trichlorotoluene	-1.62	-1.58
8	2,4,5-Trichlorotoluene	-1.60	-1.56
9	2,4,6-Trichlorotoluene	-1.80	-1.86
10	2-Chloroanisol	-2.04	-2.02
11	3-Chloroanisol	-2.05	-2.10
12	4-Chloroanisol	-1.81	-1.92
13	2,3-Dichloroanisol	-1.77	-1.76
14	2,4-Dichloroanisol	-1.78	-1.74
15	3,4-Dichloroanisol	-1.8	-1.74
16	3,5-Dichloroanisol	-1.49	-1.45
17	2,3,4-Trichloroanisol	-1.10	-1.13
Test set			
1	2,4,6-Trichloroanisol	-2.10	-1.80
2	2,3,4,5-Tetrachloroanisol	-1.84	-1.63
3	2,3,4,6-Tetrachloroanisol	-2.03	-1.86
4	2,3,5,6-Tetrachloroanisol	-1.51	-1.42

[a]Experimental values; [b]GA-MLR

Data handling: The chemical structure of each component in our chosen data set was drawn using Hyperchem 8.0 (Hypercube, Inc) software package. The semi-empirical Austin Model 1 (AM1) Hamiltonian method was applied to optimize the corresponding 3D molecular structures. The geometry optimization was done using Polak-Ribiere algorithm until the root mean square gradient was 0.001 Kcal/mol. Geometry optimization was run multiple times over a variety of starting points for each molecule and the lowest energy conformation was utilized for the calculation of electronic properties. Regardless of any symmetry constraint, full optimization of all bond lengths and angles was performed. All calculations were accomplished at the restricted Hartree-Fock level without any configuration interaction.

Descriptor generation: Molecular descriptors are defined as numerical characteristics associated with chemical structures. The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number applied to correlate physical properties. The Dragon software was used to calculate the descriptors in this research and a total of 1481 molecular descriptors, from 18 different types of theoretical descriptors, were calculated for each molecule. Since the values of many descriptors are related to the bonds lengths and bonds angles *etc.*, the chemical structure of every molecule must be optimized before calculating its molecular descriptors. For this reason, chemical structure of the 21 studied molecules were drawn with the Hyperchem software and saved with the HIN extension. To optimize the geometry of these molecules, the AM1 geometrical optimization was applied. After optimizing the chemical structures of all compounds, the molecular descriptors were calculated using Dragon.

Data pretreatment: The calculated descriptors were first analyzed for the existence of constant or near-constant variables and those detected were removed. In addition, to reduce redundancy in the descriptor data matrix, probable correlations of the descriptors with each other and with the $E_{1/2s}$ of the molecules were examined and the collinear descriptors (*i.e.* r > 0.9) were detected. Among the collinear descriptors, that with the highest correlation with $E_{1/2}$ was retained while another variable was discarded from the data matrix. Then, the remaining descriptors were collected in a pool involving an n × m data matrix (D), where n = 21 and m = 284 are the numbers of the compounds and the descriptors, respectively.

Genetic algorithm: Nowadays, genetic algorithm is well-known as an interesting and the most widely employed variable selection method that is used to solve the optimization problems defined by fitness criteria, applying the evolution hypothesis of Darwin and different genetic functions, *i.e.* cross-over and mutation.

To select the most relevant descriptors, the evolution of the population was simulated. The population of the first generation was randomly selected. Each individual member in the population was defined by a chromosome of binary values, representing a subset of descriptors. Besides, number of the genes at each chromosome was equal to the number of the descriptors. A gene was given the value of 1, if its corresponding descriptor was included in the subset; otherwise, it was given the zero value. The number of the genes with the value of 1 was kept relatively low to have a small subset of descriptors. Consequently, the probability of generating 0 for a gene was set greater (at least 60 %) than the value of 1. The operators used here were cross-over and mutation. The application probability of these operators was varied linearly with a generation renewal (0-0.1 % for mutation and 60-90 % for cross-over). The population size was varied between 50 and 250 for different genetic algorithm runs. For a typical run, the evolution of the generation was stopped when 90 % of the generations took the same fitness.

RESULTS AND DISCUSSION

For the selection of the most important descriptors, we applied genetic algorithm as the variable selection procedure to select only the best combinations (most relevant) for obtaining the models with the highest predictive power by using the training set. The five most significant descriptors according to the GA-MLR algorithm are: Zagreb M2 index (ZM2), radial distribution function-70.0/weighted by atomic van der Waals (RDF070v), radial distribution function-80.0/weighted by

TABLE-2
SELECTED DESCRIPTORS OF GENETIC ALGORITHM MULTIPLE LINEAR REGRESSION

Descriptor	Type of descriptor	Notation	Coefficient	MF
Zagreb M2 index	Topological	ZM2	-330.33	0.7434
Radial Distribution Function-70.0 / weighted by atomic van der waals	RDF descriptors	RDF070v	-0.0894	0.4196
Radial Distribution Function-80.0 / weighted by atomic van der waals	RDF descriptors	RDF080v	1.6201	0.0091
3D-MoRSE signal 02/weighted by atomic masses	3D-MoRSE	Mor02m	-146.40	0.0033
3rd component accessibility directional WHIM index/weighted by atomic van der waals volume	WHIM	E3v	0.1937	0.1573
Constant			8.3733	

atomic van der Waals (RDF080v), 3D-MoRSE signal 02/ weighted by atomic masses (Mor02m) and 3rd component accessibility directional WHIM index/weighted by atomic van der Waals volume (E3v).

To examine the relative importance as well as the contribution of each descriptor in the model, the value of the mean effect (MF) was calculated for each descriptor. This calculation was performed with the equation given below:

$$MF_j = \frac{\beta_j \sum_{i=1}^{i=n} d_{ij}}{\sum_j^m \beta_j \sum_i^n d_{ij}} \qquad (2)$$

where, MF_j represents the mean effect for the considered descriptor j, β_j is the coefficient of the descriptor j, d_{ij} stands for the value of the target descriptors for each molecule and lastly m is the descriptors number for the model. The mean effect (MF) value indicates the relative importance of a descriptor, compared with the other descriptors implemented in the model. Its sign indicates the variation direction in the values of the activities as a result of the enhancement or decrease of the descriptor values. The mean effect values associated with selected variables are shown in Table-2.

After the selection of the most important descriptors by genetic algorithm, multiple linear regression was performed to build the linear model. Good correlations with the experimental $E_{1/2}$ data were selected based on the squared correlation coefficient (R^2), Fisher criterion (F), squared cross-validated correlation coefficient (Q^2) and standard error (SE) of the regression.

The following equation obtained by GA-MLR method: $E_{1/2}$ = 8.3733-330.33(ZM2)-0.0894(RDF070v) + 1.6201 (RFD080v) - 146.40 (Mor02m) + 0.1937 (E3v), N_{train} = 17, R^2_{train} = 0.9237, Q^2_{Loo} = 0.8100, Q^2_{LGO} = 0.8030, F = 26.66, N_{test} = 4, R^2_{test} = 0.9404, R^2_{adj} = 0.8891.

With the test set, the prediction results were also obtained. The predicted *versus* observed values based on GA-MLR are shown in Table-1. Fig. 1 shows the predicted *versus* observed RI for all of the 21 compounds studied encompassing the training set and the test set.

The results illustrated once more that the linear multiple linear regression technique combined with a successful variable selection procedure such as GA is adequate to generate an efficient QSER model for predicting the $E_{1/2}$ of compounds. Furthermore, the residuals (experimental RI- predicted RI) *versus* experimental RI value, obtained by the GA-MLR modeling is shown in Fig. 2. The distribution of the residuals on both sides of the zero line indicates there is no systematic error in both model.

For a more exhaustive testing of the predictive power of the model, validation of the model was also carried out using the LOO and the LGO cross-validation techniques on the training set of compounds. For LOO cross-validation, a data point is removed from the set and the model is recalculated. The predicted RI for that point is then compared with its actual value. This is repeated until each data point has been omitted once. For LGO, 20 % of the data points are removed from the dataset and the model was refitted; the predicted values for those points were then compared with the experimental values. Again, this is repeated until each data point has been omitted once. The results produced by the LOO (Q^2 = 0.8100) and the LGO (Q^2 = 0.8030) cross-validation tests illustrated the quality of the obtained model.

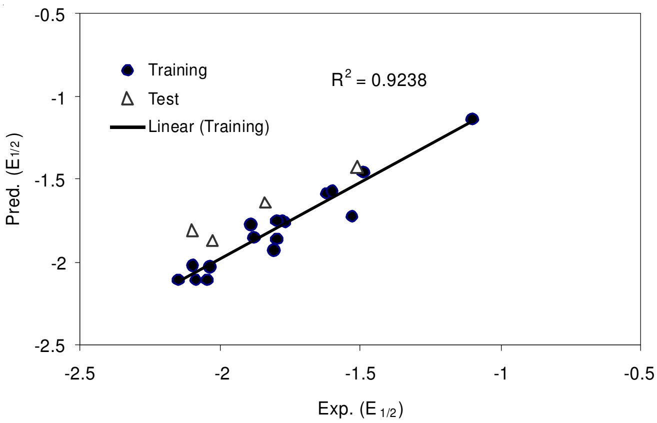

Fig. 1. Predicted $E_{1/2}$ values by the MLR modeling *vs*. the experimental $E_{1/2}$ values

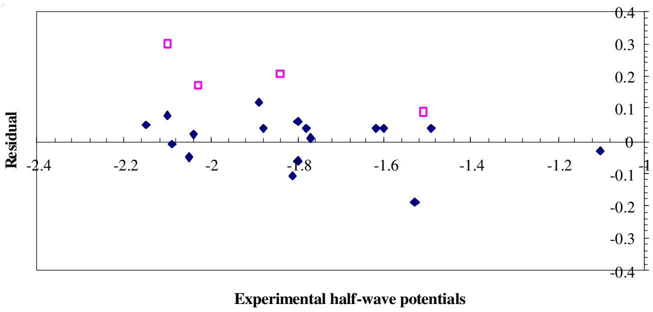

Fig. 2. Plot of the residuals against the experimental values of $E_{1/2}$ in the proposed GA-MLR model

The model was further validated by applying Y-randomization. Several random shuffles of the Y vector ($E_{1/2}$) were performed and the low R^2 and Q^2 resulted values clarify that the good results in the original model are not due to a chance correlation or structural dependency of the training set. The results of the Y-randomization test are presented in Table-3. High predictive ability and simplicity of the proposed method denote it could be a powerful aid as well as a proper alternative approach to the costly and time consuming experiments for determining the $E_{1/2}$ of other compounds.

TABLE-3
r^2 AND Q^2 VALUES AFTER SEVERAL *Y*-RANDOMIZATION TESTS

Iteration	R^2	Q^2
1	0.041	0.214
2	0.266	0.022
3	0.168	0.121
4	0.026	0.002
5	0.399	0.001
6	0.204	0.054
7	0.144	0.086
8	0.180	0.109
9	0.136	0.236
10	0.207	0.273

ACKNOWLEDGEMENTS

The authors express their special thanks both the office for Researchers Affair and Young Research Club of Islamic Azad University of Shahrood for financial supports of the research proposal entitled "Prediction of Half-Wave Potential of Some Chloroarenes by Using Quantitative Structure Property Relationship (QSPR)" and the time dedicated for performance the calculations of the study.

REFERENCES

1. A.G. Krivenko, A.S. Kotkin and V.A. Kurmaz, *Russ. J. Electrochem.*, **41**, 122 (2005).
2. M. Shamsipur, A. Siroueinejad, B. Hemmateenejad, A. Abbaspour, H. Sharghi, K. Alizadeh and S. Arshadi, *J. Electroanal. Chem.*, **600**, 345 (2007).
3. R. Todeschini and V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH: Weinheim (2000).
4. F.A.L. Ribeiro and M.M.C. Ferreira, *J. Mol. Struct.: Theochem.*, **663**, 109 (2003).
5. A. Jain and S.H. Yalkowsky, *Ind. Eng. Chem. Res.*, **46**, 2589 (2007).
6. A.R. Katritzky, L. Mu and M. Karelson, *J. Chem. Inf. Comput. Sci.*, **36**, 1162 (1996).
7. J. Wang, G. Krudy, T. Hou, W. Zhang, G. Holland and X. Xu, *J. Chem. Inf. Model.*, **47**, 1395 (2007).
8. F. Ignatz-Hoover, R. Petrukhin, M. Karelson and A.R. Katritzky, *J. Chem. Inf. Comput. Sci.*, **41**, 295 (2001).
9. A.R. Katritzky, U. Maran, V.S. Lobanov and M. Karelson, *J. Chem. Inf. Comput. Sci.*, **1**, 1 (2000).
10. E.H.P. Wolff and A.N.R. Bos, *Ind. Eng. Chem. Res.*, **36**, 1163 (1997).
11. R.L. McNaughton, A.A. Tipton, N.D. Rubie, R.R. Conry and M.L. Kirk, *Inorg. Chem.*, **39**, 5697 (2000).
12. S. Niu, X.B. Wang, J.A. Nichols, L.S. Wang and T. Ichiye, *J. Phys. Chem. A*, **107**, 2898 (2003).
13. A. Beheshti, S. Riahi and M.R. Ganjali, *Electrochim. Acta*, **54**, 5368 (2009).
14. M.H. Fatemi, M.R. Hadjmohammadi, K. Kamel and P. Biparva, *Bull. Chem. Soc. (Japan)*, **80**, 303 (2007).
15. B. Hemmateenejad and M. Yazdani, *Anal. Chim. Acta*, **634**, 27 (2009).
16. K. Nesmerak, I. Nemece, M. Sticha, K. Waisser and K. Palat, *Electrochim. Acta*, **50**, 1431 (2005).
17. B. Hemmateenejad and M. Shamsipur, *Intern. Electr. J. Mol. Design*, **3**, 316 (2004).
18. O. Krang and J. Voss, *Z. Naturforsch.*, **58b**, 1187 (2003).