Classification and Regression Tree Analysis for Molecular Descriptor Selection and Binding Affinities Prediction of Imidazobenzodiazepines in Quantitative Structure-Activity Relationship Studies

Morteza Atabati, Kobra Zarei,* and Esmaeil Abdinasab

*School of Chemistry, Damghan University of Basic Sciences, Damghan, Iran. *E-mail: zarei@dubs.ac.ir*
Received May 17, 2009, Accepted October 4, 2009

The use of the classification and regression tree (CART) methodology was studied in a quantitative structure-activity relationship (QSAR) context on a data set consisting of the binding affinities of 39 imidazobenzodiazepines for the α_1 benzodiazepine receptor. The 3-D structures of these compounds were optimized using HyperChem software with semiempirical AM1 optimization method. After optimization a set of 1481 zero-to three-dimentional descriptors was calculated for each molecule in the data set. The response (dependent variable) in the tree model consisted of the binding affinities of drugs. Three descriptors (two topological and one 3D-Morse descriptors) were applied in the final tree structure to describe the binding affinities. The mean relative error percent for the data set is 3.20%, compared with a previous model with mean relative error percent of 6.63%. To evaluate the predictive power of CART cross validation method was also performed.

Key Words: CART, QSAR, Imidazobenzodiazepines, Binding affinity, Benzodiazepine receptor

Introduction

Benzodiazepines (BDZs) are the drugs of choice in the pharmacotherapy of anxiety and related emotional disorders, sleep disorders, status epilepticus, and other convulsive states; they are used as centrally acting muscles relaxants, for premedication, and as inducing agents in anesthesiology. They act *via* the benzodiazepine receptor site (BzR) on the γ-aminobutyric acid receptor ($GABA_A$) family.[1] These drugs have been subjected to extensive QSAR studies.[2-8] $GABA_A$ receptors are the major inhibitory neurotransmitter receptors in the brain, in the site of action of many clinically important drugs, and are important drug targets representing the sites of action of benzodiazepines, barbiturates, and neurosteroids. These receptors are ligand-gated chloride channels composed of five subunits that can belong to eight different subunit classes. Most $GABA_A$ receptor subtypes *in vivo* are believed to be composed of α-, β-, and γ-subunits. When BDZs bind to their receptors; they appear to induce a conformational change leading to an increase in the availability of $GABA_A$ receptors for $GABA_A$, leading to higher chloride influx and hyperpolarization. Receptors containing the $\alpha_{1\text{-}5}$-subunits in combination with any of the β-subunits and the γ_2-subunit are most prevalent in the brain. These receptors are sensitive to benzodiazepine modulation. The major receptor subtype is assembled from the subunits $\alpha_1\beta_2\gamma_2$ (diazepam-sensitive $GABA_A$ receptors).

Imidazobenzodiazepines are described novel pharmaceutically active substances which have a pronounced affinity to the central benzodiazepine receptors and which have only a low toxicity. There are a few QSAR studies on the imidazobenzodiazepines. Cook *et al*. in 1998, carried out a QSAR study on a number of imidazobenzodiazepines exhibiting affinities at recombinant $\alpha_1\beta_3\gamma_2$, $\alpha_1\beta_2\gamma_2$, $\alpha_1\beta_2\gamma_2$, $\alpha_1\beta_2\gamma_2$, and $\alpha_1\beta_2\gamma_2$ $GABA_A$/benzodiazepine receptor subtypes (α_1, α_2, α_3, α_4, α_5, α_6), by means of COMFA.[9] Hadjipavlou-litina and coworkers in 2004, derived different equations for above mentioned compounds with two descriptors overall molar refractivity and Taft's electronic effect.[1] They obtained r^2 value of 0.825 with three outliers and a mean relative error percent of 6.63% for the 38 investigated compounds.

In this study, another approach, classification and regression tree (CART) analysis was investigated. CART is a statistical method that explains the variation of a response variable using a set of explanatory variables, so called predictors. The method is based on a recursive binary splitting of the data into mutually exclusive subgroups containing objects with similar properties.[10] CART is extensively used for modeling and classification in several areas, such as medical diagnosis and prognosis,[11-13] ecology,[14] agriculture[15] and chemistry.[10,16-17] A very interesting advantage of CART is the possibility to deal with large numbers of both categorical and numerical variables. Another advantage is that no assumption about the underlying distribution of the predictor variables is required (even categorical variables can be used). Eventually, CART provides a graphical representation, which makes the interpretation of the results easy. Therefore, we felt that CART could be a very effective method to select and relate molecular descriptors with the medical properties of molecules.

Theory

In 1984, Breiman *et al*. have introduced a methodology for classification and regression, called "classification and regression tree analysis".[11] The goal of this statistical method was to explain the variation of a dependent variable, using a set of independent predictors, *via* a binary partitioning procedure. CART works by splitting the parent node in two nodes, called child nodes. The process is repeated by treating each child node as a parent node. Each split is defined by a simple rule, usually based on a single explanatory variable. For numerical explana-

tory variables, a splitting value (cut point) is selected to form two groups, which contain objects with values smaller and larger, respectively than the selected cut point. For categorical variables, a split is defined by relating one or more levels of the variable to a specific node. Trees are grown by selecting the splits in such a way that the impurity of the response variable within each node is minimized. The splitting procedure is continued until no further split can be performed, i.e., all child nodes are homogeneous, or contain one or a user-defined minimal number of observations. The tree thus obtained is called the maximal tree and the terminal nodes, the so-called leaves, represent the final groups formed by the tree. This maximal tree will usually contain too many leaves and will overfit the learning data set, which will cause poor predictive abilities for new sample. [10] Therefore, the selection of an optimal tree with a good compromise between model fit and predictive properties is required. Thus, in general, CART analysis consists of three steps: (i) the maximal-tree building, (ii) the tree "pruning", which consists of the cutting-off of nodes to generate a sequence of simpler trees, and (iii) the optimal tree selection.

Maximal tree building. CART is looking for the best possible variable, so called splitter, to divide the root node into two child nodes. To achieve this, the program looks at all possible variables, as well as at all possible values of the variable that can be used to split the data. The best splitter is defined as the variable (and associated splitting value) that will minimize the impurity, I, of the two child nodes. The goodness of a split is then defined as the impurity decrease between the parent node and its children:

$$\Delta i(s,t_p) = i_p(t_p) - p_L i(t_L) - p_R i(t_R) \quad (1)$$

where s is a candidate split, P_L and P_R are the fractions of observations of the parent node t_P that go into the child nodes t_L and t_R, respectively. The best splitter is the one that will maximize $\Delta i(s, t_P)$.

Different criteria to measure the impurity of a node have been proposed.[11] For regression trees, the total sum of squares of the response values about the mean of the node is the most popular measure of impurity:[10-11,17]

$$i(t) = \sum_{x_n \in t} (y_n - \bar{y}(t))^2 \quad (2)$$

where $i(t)$ is the impurity of node t; y_n, is the response value of observation x_n belonging to node t; $\bar{y}(t)$, the mean of all observations in node t. Absolute deviation about the node medians is another criterion which is used to build (robust) trees.[10]

Tree pruning. The resulting maximal trees are usually oversized and they describe the training set perfectly. It means that the model has been overfitted.[10,18] Such trees are often difficult to interpret and their predictive ability for new observations is poor in general, since they tend to fit also the noise in the data. The selection of a smaller tree, derived from the maximal is then necessary for predictive purposes. The procedure of pruning generates a sequence of smaller trees, obtained by removing successively branches of the maximal tree.

Optimal tree selection. Finally, the optimal tree is selected from the generated sequence of subtrees by evaluating the predictive error of the trees. The predictive error is often estimated using cross validation technique, especially for small data sets. In cross validation, some samples are randomly drawn from the data set, to test the tree, which is built with the rest of the data.[10,17] For a ten-fold cross validation, the original data set is divided into ten equal pairs (test sets), each containing a similar distribution for the response variable. A tree is then built using 90% of the observations (learning set), while the remaining 10% (test set) is used to test the tree. This step is repeated ten times using each time a different test set and the remaining observations as the learning set. The optimal tree is the one having the minimal cross validation error (most accurate tree). In practice, the optimal tree is chosen as the simplest tree with a predictive error estimate within one standard error of minimum. In this way, the chosen tree is the simplest with an error estimate comparable to that of the most accurate one.

Experimental

The binding affinities of 39 imidazobenzodiazepines were obtained from the paper by Hadjipavlou-Litina *et al.*[1] and were shown in Table 1.

Molecular modeling and geometry optimization were performed by HyperChem[19] (version 7.0, Hyper Cube, Inc.). Dragon software was used for calculation of descriptors.[20] SPSS software (version 13.0, SPSS, Inc.) was used for running CART.

The 3-D structures of these compounds were optimized using HyperChem software with semiempirical AM1 optimization method. After optimization a total of 1481 0-, 1-, 2-, and 3-D descriptors including constitutional, topological, molecular walk counts, BCUT-descriptors, GALVEZ topological charge indices, 2-D autocorrelations, charge, aromaticity indices, Randic molecular profiles, geometrical, RDF, 3D-MoRSE, WHIM descriptors, GETAWAY, functional group counts, atom-centered fragments, empirical and molecular properties were generated using Dragon software.

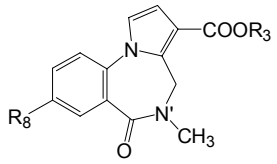
Results and Discussion

Maximal tree was grown using the binding affinities of 39 imidazobenzodiazepines (log $1/K_i$). A total of 1481 descriptors were used as explanatory variables. The regression tree was grown using Eq. (2) as impurity measure. The plot of maximal regression tree is shown in Fig. 1.

To select the optimal tree, ten fold cross-validation was used. The optimal tree was selected from the maximal tree, which was pruned back with no change in the split limit. Fig. 2 shows a plot of the prediction error, calculated as the root mean squared error of cross validation (RMSECV), as a function of the size of the tree (the tree size is defined as the number of leaves in a given tree). A horizontal line indicates the selection limit, situated one standard error above the minimal RMSECV. Applying this selection limit suggests a four-leaf tree size as optimal.

Fig. 3 shows the selected tree, indicating the splitting rules, the average response value and the numbers of objects of the

Table 1. K_i binding affinities of imidazobenzodiazepines for the benzodiazepine receptor isoform



no	Substituents R_3R_8	Observed log ($1/K_i$)
1	$R_3 = COOC_2H_5, R_8 = F$	9.097
2	$R_3 = COOC_2H_5, R_8 = Cl$	8.167
3	$R_3 = COOC_2H_5, R_8 = Br$	7.585
4	$R_3 = COOC_2H_5, R_8 = CN$	8.000
5	$R_3 = COOC_2H_5, R_8 = CH = CH_2$	8.081
6	$R_3 = COOC_2H_5, R_8 = C_2H_5$	7.690
7	$R_3 = COOC_2H_5, R_8 = OC_2H_5$	7.951
8	$R_3 = COOC_2H_5, R_8 = N_3$	8.481
9	$R_3 = COOC_2H_5, R_8 = CH = C = CH_2$	8.426
10	$R_3 = COOC_2H_5, R_8 = C \equiv CH$	7.547
11	$R_3 = COOC_2H_5, R_8 = C \equiv C(CH_3)$	7.996
12	$R_3 = COOC_2H_5, R_8 = C \equiv C[Si(CH_3)_3]$	6.917
13	$R_3 = COOC_2H_5, R_8 = C \equiv CCH_2[Si(CH_3)_3]$	6.523
14	$R_3 = COOC(CH_3)_3, R_8 = Cl$	7.762
15	$R_3 = COOC(CH_3)_3, R_8 = Br$	7.943
16	$R_3 = COOC(CH_3)_3, R_8 = I$	8.013
17	$R_3 = COOC(CH_3)_3, R_8 = OH$	8.824
18	$R_3 = COOC(CH_3)_3, R_8 = OCH_3$	8.171
19	$R_3 = COOC(CH_3)_3, R_8 = N(CH_3)_2$	7.883
20	$R_3 = COOC(CH_3)_3, R_8 = N - tetrahydropyrrole$	8.237
21	$R_3 = COOC(CH_3)_3, R_8 = N - hexahydropyridine$	8.191
22	$R_3 = COOC(CH_3)_3, R_8 = N_3$	8.140
23	$R_3 = COOC(CH_3)_3, R_8 = NCS$	7.767
24	$R_3 = COOC(CH_3)_3, R_8 = NO_2$	7.893
25	$R_3 = COOC(CH_3)_3, R_8 = C_2H_5$	7.830
26	$R_3 = COOC(CH_3)_3, R_8 = C \equiv CH$	7.570
27	$R_3 = COOC(CH_3)_3, R_8 = C \equiv C[Si(CH_3)_3]$	6.706
28	$R_3 = COOC(CH_3)_3, R_8 = C \equiv CCH_2[Si(CH_3)_3]$	6.561
29	$R_3 = COOCH_2 - cy - C_3H_5, R_8 = Cl$	7.785
30	$R_3 = COCH_3, R_8 = Cl$	4.756
31	$R_3 = COC_4H_9, R_8 = Cl$	5.801
32	$R_3 = CH_2OH, R_8 = Cl$	6.523
33	$R_3 = CH_2OCH_3, R_8 = Cl$	6.523
34	$R_3 = CH_2Cl, R_8 = Cl$	6.523
35	$R_3 = CH_2OC_2H_5, R_8 = Cl$	6.523
36	$R_3 = CH_2N(C_2H_5)_2, R_8 = Cl$	5.023
37	$R_3 = CH_2N[CH(CH_3)]_2, R_8 = Cl$	5.377
38	$R_3 = C_2H_5, R_8 = Cl$	6.389
39	$R_3 = C_5H_{11}, R_8 = Cl$	5.588

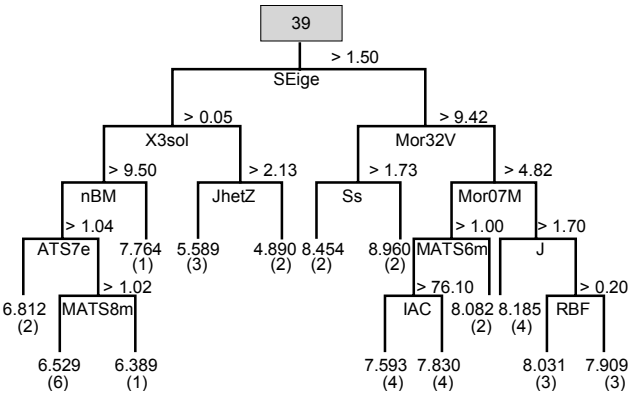


Figure 1. Maximal regression tree, grown for the log ($1/K_i$) values of 39 drugs using 1481 descriptors. For each leaf the mean log ($1/K_i$) value is given, as well as the number of objects (molecules), between the brackets. For each split the criterion is indicated.

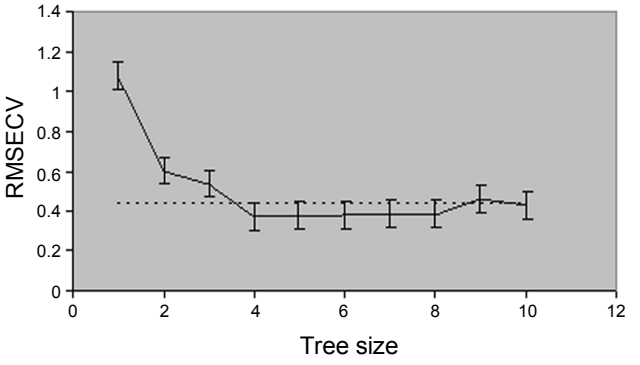


Figure 2. RMSECV versus tree size. The dotted line represents the selection limit.

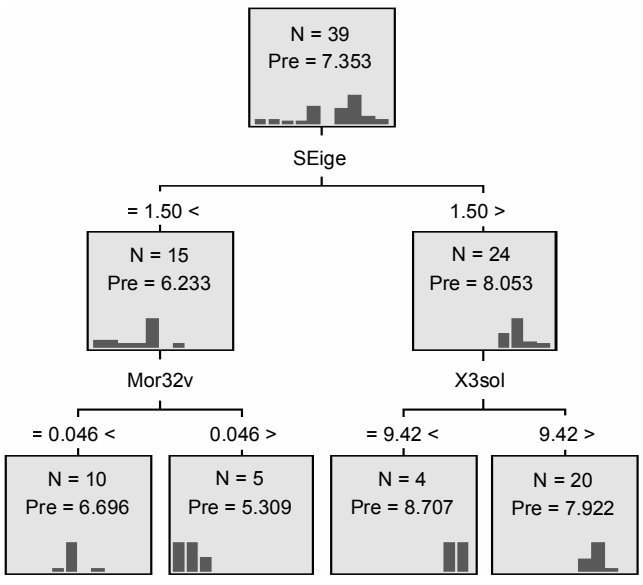


Figure 3. The optimal tree.

Table 2. The amounts of selected descriptors

no	SEige	X3sol	Mor32v
1	1.84	8.68	0.02
2	1.73	9.47	-0.01
3	1.67	9.73	-0.02
4	1.66	9.63	-0.01
5	1.52	9.63	0.01
6	1.52	9.63	-0.05
7	1.77	9.75	-0.04
8	1.70	9.00	0.04
9	1.52	9.22	-0.02
10	1.52	9.63	-0.03
11	1.52	9.75	-0.06
12	0.99	10.05	0.02
13	1.24	11.05	-0.02
14	1.73	9.65	-0.04
15	1.67	9.91	-0.07
16	1.54	10.17	-0.05
17	1.77	9.38	-0.05
18	1.77	9.81	0.11
19	1.66	10.10	-0.01
20	1.66	11.23	-0.01
21	1.66	11.48	0.01
22	1.94	9.93	-0.04
23	1.45	10.07	-0.04
24	2.16	10.10	0.00
25	1.52	9.81	-0.10
26	1.52	9.81	-0.03
27	1.24	10.98	-0.01
28	1.24	11.23	-0.06
29	1.73	10.16	0.04
30	1.48	8.75	0.06
31	1.48	9.74	0.05
32	1.48	8.65	0.04
33	1.48	8.92	0.02
34	1.45	8.80	0.01
35	1.48	9.17	-0.01
36	1.38	10.10	0.13
37	1.38	10.71	0.12
38	1.24	8.65	0.04
39	1.24	9.42	0.05

Table 3. The predicted values of cross validation method

No of Predicted data	Observed log ($1/K_i$)	Predicted log ($1/K_i$)	Absolute relative error
1	9.097	8.577	0.057
2	8.167	8.707	0.066
3	7.585	7.939	0.047
4	8.00	7.918	0.010
5	8.081	7.913	0.021
6	7.690	7.934	0.033
7	7.951	7.920	0.004
8	8.481	7.782	0.035
9	8.426	8.801	0.045
10	7.547	7.941	0.052
11	7.996	7.918	0.010
12	6.917	6.671	0.036
13	6.523	6.715	0.029
14	7.762	7.930	0.022
15	7.943	7.921	0.003
16	8.013	7.917	0.012
17	8.824	7.922	0.102
18	8.171	7.909	0.032
19	7.883	7.924	0.055
20	8.237	7.905	0.040
21	8.191	7.908	0.035
22	8.140	7.910	0.028
23	7.767	6.576	0.153
24	7.893	7.923	0.004
25	7.830	7.927	0.012
26	7.570	7.940	0.049
27	6.706	6.694	0.002
28	6.561	6.710	0.023
29	7.785	7.929	0.018
30	4.756	5.447	0.145
31	5.801	5.186	0.106
32	6.523	6.715	0.029
33	6.523	6.715	0.029
34	6.523	6.715	0.029
35	6.523	6.715	0.029
36	5.023	5.380	0.071
37	5.377	5.292	0.016
38	6.389	6.730	0.053
39	5.588	5.309	0.050

leaves. Additionally, histograms are plotted that representing the distribution of the response for the objects within each node.

For the optimal subtree with four terminal nodes, three molecular descriptors were selected to describe the binding affinities data. The amounts of these descriptors are shown in Table 2. The first selected molecular descriptor is Eigenvalue sum from electronegativity weighted distance matrix (SEige), which is a topological descriptor. Eigenvalue descriptors are independent of any molecular alignment, giving information about molecular size, shape and electronic properties.[21] As can be seen in Table 2, in presence of electronegativity groups this descriptor amount increases. The other descriptor is also a topological one, the solvation connectivity index chi-3 (X3sol). This descriptor is defined in order to model solvation entropy and describe dispersion interactions in solution.[21] Taking into account the characteristic dimension of the molecules by atomic parameters, they are defined as:

$$ {}^{m}\chi_{q}^{s} = \frac{1}{2^{m+1}} \sum_{k=1}^{K} \frac{\left(\prod_{a=1}^{n} L_a \right)_k}{\left(\prod_{a=1}^{n} \delta_a \right)_k^{1/2}} \tag{3} $$

where L_a is the principal quantum number (2 for C, N, O atoms, 3 for Si, S, Cl) of the ath atom in the kth subgraph and δ_a the corresponding vertex degree; K is the total number of mth order subgraphs; n is the number of vertices in the subgraph.[21] As Table 2 shows, the amount of this descriptor is high for the molecules containing atoms with big principal quantum number such as Si and Cl. The third selected descriptor is 3D-Morse-signal32 (Mor32v) from 3D Morse descriptors, 3D-molecule representation of structures based on electron diffraction.[21] These

Table 4. Verification of statistical validity of the model.

No of predicted data	The mean square error of calibration set	The mean square error of prediction set	R^2 (calibration set)	R^2 (prediction set)	The mean relative error of prediction set	The mean relative error of calibration set	Prediction set
1, 5, 10, 15, 20, 25, 30, 35, 39	0.8070	0.1133	0.9429	0.9493	4.19	2.94	Ser 1
2, 6, 12, 16, 22, 26, 29, 32, 36, 38	0.9461	0.8770	0.9169	0.9369	3.84	3.20	Ser 2
3, 7, 11, 17, 19, 21, 27, 31, 34, 37	0.9460	0.1571	0.9113	0.8931	3.93	3.39	Ser 3
4, 8, 9, 13, 14, 18, 23, 24, 28, 33	0.5850	0.1993	0.9529	0.7570	3.63	2.96	Ser 4

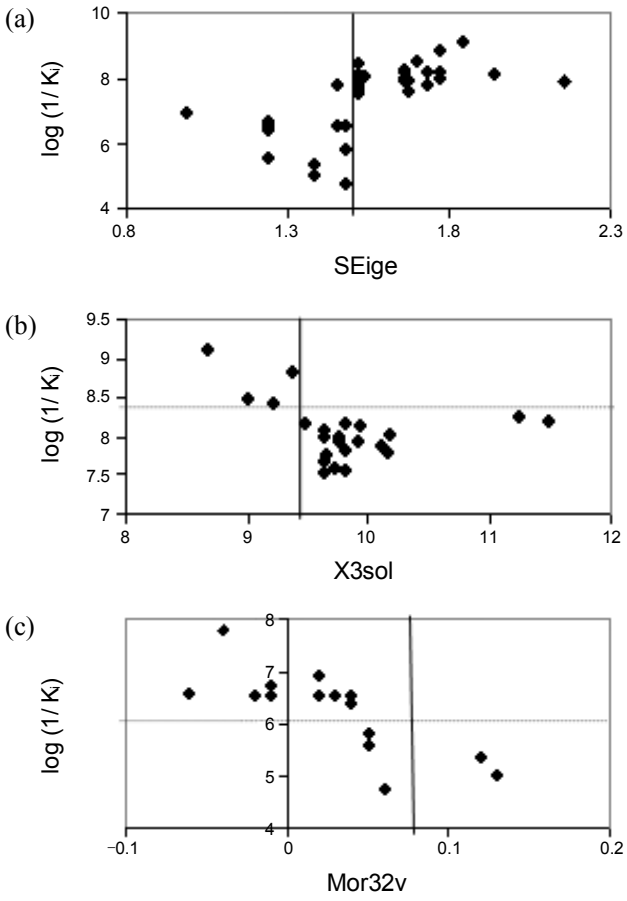


Figure 4. Log ($1/K_i$) versus the explanatory variables causing the splits in Fig. 3, (a) SEige, (b) X3sol, (c) Mor32V. The vertical line represents the limit value to divide into two child nodes.

descriptors are based on the idea of obtaining information from the 3D atomic coordinates by the transform used in electron diffraction studies for preparing theoretical scattering curves.

The relationships between the selected variables (from Fig. 3) and log ($1/K_i$) are shown in Fig. 4. The limit values defining the splits are indicated by a vertical line. Only the molecules relevant for a specific node are plotted. In Fig. 4a, for instance, all 39 molecules are plotted, whereas only 24 molecules are represented in Fig. 4b. The first split divides the data into two groups, which contain molecules with SEige values below and above 1.5, respectively. The second and third split divide the data into two groups, with molecules with X3sol values below and above 9.42, and Mor32v values below and above 0.046, respectively. The log ($1/K_i$) values are divided into two groups by these splits, very well.

The optimal tree was applied for the prediction of the whole data set. The mean relative error and R^2 were obtained as 3.20% and 0.9211, respectively. It has better prediction power rather than MLR model[1] with the mean relative error of 6.63% and R^2 of 0.8240.

In addition to the previous MLR model,[1] another MLR model was constructed with three descriptors which were selected between 1481 by stepwise selection method in SPSS software. These descriptors were C-041 (atom-centered fragments), BEHP7 (highest eigenvalue n.7 of Burden matrix/weighted by atomic polarizabilities) and SHP2 (average shape profile index of order 2 Randic molecular profiles). Then one MLR equation was derived for these descriptors as:

$$\log(1/K_i) = 54.118 \pm 7.555 + 2.433 \pm 0.176 \times (C - 041) - 14.833 \pm 2.056(BEHP7) - 21.104 \pm 5.866(SHP2) \quad (4)$$

The mean relative error and R^2 for this model were obtained as 3.85% and 0.8914, respectively.

To evaluate the predictive power of CART, leave one out cross validation method was also performed. The minimal tree built from the training sets always contained four leaves. The results were shown in Table 3. The mean relative error and Q^2 were 3.95% and 0.8736, respectively.

To make sure the demonstration of the absence of a chance correlation, the whole data set was divided into four subsets, and each subset was predicted by using the other three subsets as the training set. The results were shown in Table 4.

Conclusion

The main aim of the present work was the development of a QSAR method using classification and regression tree methodology for binding affinities of 39 imidazobenzodiazepines for the α_1 benzodiazepine receptor. The generated tree was evaluated and applied for the prediction of binding affinities of imidazobenzodiazepines. The results have shown that this methodology has good prediction power for this purpose. The application of CART to this data set has demonstrated that the CART analysis is able to perform a better prediction than MLR method in terms of prediction accuracy. Moreover, the output of rules sets from the CART analysis can provide useful insight into the relationships between the response and the predictor variables and the relative importance of predictor variables. The

statistical results were compared with MLR method results.

Acknowledgments. The authors acknowledge to the Research Council of Damghan University of Basic Sciences for the support of this work.

References

1. Hadjipavlou-Litina, D.; Garg, R.; Hansch, C. *Chem. Rev*. **2004**, *104*, 3751.
2. Verli, H.; Albuquerque, M. G.; de Alencastro, R. B.; Barreiro, E. J. *Eur. J. Med. Chem.* **2002**, *37*, 219.
3. Thakur, A.; Thakur, M.; Khadikar, P. *Bioorgan. Med. Chem*. **2003**, *11*, 5203.
4. Savini, L.; Massarelli, P.; Nencini, C.; Pellerano, C.; Biggio, G.; Maciocco, A.; Tuligi, G.; Carrieri, A.; Cinone, N.; Carotti, A. *Bioorgan. Med. Chem*. **1998**, *6*, 389.
5. Terletskay, A.; Shvets, N.; Dimoglo, A.; Chumakov, Y. *J. Mol. Struct.* (Theochem) **1999**, *463*, 99.
6. Blair, T.; Webb, G. A. *J. Med. Chem*. **1977**, *20*, 1206.
7. Greco, G.; Novellino, E.; Silipo, C.; Vittoria, A. *Quant. Struct.-Act. Relat*. **1992**, *11*, 461.
8. Gupta, S. P.; Paleti, A. *Quant. Struct.-Act. Relat*. **1996**, *15*, 12.
9. Huang, Q.; Liu, R.; Zhang, P.; He, X.; McKernan, R.; Gan, T.; Bennett, D. W.; Cook, J. M. *J. Med. Chem.* **1998**, *41*, 4130.
10. Put, R.; Perrin, C.; Questier, F.; Coomans, D.; Massart, L.; Vander Heyden, Y. *J. Chrom. A* **2003**, *988*, 261.
11. Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Wadsworth: Monterey, 1984.
12. Lavrac, N. *Artif. Intell. Med*. **1999**, *16*, 3.
13. Marshall, R. J. *J. Clin. Epidemiol*. **2001**, *54*, 603.
14. De'Ath, G.; Fabricius, K. E. *Ecology* **2000**, *81*, 3178.
15. Tittonell, P. A.; Shepherd, K.; Vanlauwe, B.; Giller, K. E. *Agr. Ecosyst. Environ*. **2008**, *123*, 137.
16. Questier, F.; Put, R.; Coomans, D.; Walczak, B.; Vander Heyden, Y. *Chemometr. Intell. Lab.* **2005**, *76*, 45.
17. Jalali-Heravi, M.; Shahbazikhah, P. *Electrophoresis* **2008**, *29*, 363.
18. Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; De Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics: Part A*; Elsevier: Amsterdam, 1997.
19. Hypercube, http://www.hyper.com.
20. Todeschini, R. Milano Chemometrics and QSAR Group, http://www.disat.unimib.it/vhm/.
21. Todeschini, R.; Consunni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.