Journal of the Korean Chemical Society
2014, *Vol. 58, No. 6*
Printed in the Republic of Korea
http://dx.doi.org/10.5012/jkcs.2014.58.6.543

Quantitative Structure Activity Relationship Prediction of Oral Bioavailabilities Using Support Vector Machine

Mohammad Hossein Fatemi* and Fatemeh Fadaei

*Department of Chemistry, University of Mazandaran, Babolsar, Iran. *E-mail: mhfatemi@umz.ac.ir*

(Received July 16, 2014; Accepted September 18, 2014)

ABSTRACT. A quantitative structure activity relationship (QSAR) study is performed for modeling and prediction of oral bioavailabilities of 216 diverse set of drugs. After calculation and screening of molecular descriptors, linear and nonlinear models were developed by using multiple linear regression (MLR), artificial neural network (ANN), support vector machine (SVM) and random forest (RF) techniques. Comparison between statistical parameters of these models indicates the suitability of SVM over other models. The root mean square errors of SVM model were 5.933 and 4.934 for training and test sets, respectively. Robustness and reliability of the developed SVM model was evaluated by performing of leave many out cross validation test, which produces the statistic of Q^2_{SVM} = 0.603 and SPRESS = 7.902. Moreover, the chemical applicability domains of model were determined via leverage approach. The results of this study revealed the applicability of QSAR approach by using SVM in prediction of oral bioavailability of drugs.

Key words: Quantitative structure activity relationship, Support vector machine, Multiple linear regressions, Oral bioavailability, Molecular descriptors

INTRODUCTION

Oral bioavailability is an important pharmacokinetic property, which is defined as the fraction of an administered dose of drug that reaches the systemic circulation and is a critical property to be considered during the early stages of drug design. Among the absorption, distribution, metabolism and elimination (ADME) properties of a chemical, unfavorable oral bioavailability is indeed an important reason for stopping further development of the drug candidates.[1] The early evaluation of ADME properties in drug research has driven the need for large scale screening methods. In vitro and in vivo ADME assays are lengthy, complex and relatively expensive in terms of resources, reagents and detection techniques. Computational methods have emerged during the past decade as a powerful strategy for the prediction of human pharmacokinetics. In this regard, a variety of useful in silico models has been developed with different levels of complexity for the screening of large data sets of compounds, to creating tools that are faster, simpler, and more cost effective than traditional experimental procedures.[2] Among theoretical methods, quantitative structure activity relationships (QSAR) approaches have been successfully established to predict the properties/activities of chemicals from their structural features. The advantage of this approach over other methods lies in the fact that it requires only the knowledge of chemical structure. The main steps involved in QSAR are: data collection, molecular geometry optimization, molecular descriptor generation, descriptor selection, model development and finally evaluation of the model performance.[3]

There are some reports about QSAR prediction of oral bioavailabilities of chemicals. For example, Andrew et al. proposed a QSAR model that can achieve the coefficient of multiple correlation of R^2 = 0.71 by using 85 molecular descriptors.[2] In 2004, Tuner et al. developed QSAR model for prediction of oral bioavailabilities of some chemicals by using the artificial neural network (ANN).[4] The training and testing correlation coefficients given by the model were 0.736 and 0.897, respectively. In 2007 Moda and the coworkers reports the application of the hologram quantitative structure activity relationships (HQSAR) technique to construct a prediction model for the prediction of human oral bioavailability.[1] The correlation between experimental and calculated values of oral bioavailabilities by their model was 0.93 for training set. The main aim of the present work is developing of some QSAR models by using multiple linear regression (MLR), artificial neural network, support vector machine (SVM) and random forest (RF) as modeling techniques to developing better QSAR models.

EXPERIMENTAL

Data Set

Data set consist the experimental bioavailabilities of 302 structurally divers chemicals that were reported in ref.[1]

***Table* 1.** Experimental and predicted values of human oral bioavailability for both training and test set compounds for SVM mode

No.	Name	Experimental	Predicted	Residual	No.	Name	Experimental	Predicted	Residual
1	Abacavir*	86	76.6	9.4	49	Dihydroergosine	10	15.8	−5.8
2	Acetaminophen*	88	83.9	4.1	50	Diltiazem*	40	30.2	9.8
3	Acetylsalicylic*	68	81.1	−13.1	51	Diphenhydramine*	72	65.7	6.3
4	Albuterol	71	77	−6	52	Disopiramide*	83	93.1	−10.1
5	Almotriptan	70	72.6	−2.6	53	Dixyrazine	10	14.5	−4.5
6	Alosetron	55	59.8	−4.8	54	Domperidone	14	19.4	−5.4
7	Amiodarone	50	54.5	−4.5	55	Doxapram	61	55.8	5.2
8	Anastrozole	84	83.5	0.5	56	Doxazosin	65	64.9	0.1
9	Aprepitant	62.5	61.9	0.6	57	Dronabinol*	15	27	−12
10	Atomoxetine	63	68	−5	58	Drospirenone	76	69.9	6.1
11	Atorvastatin	14	19.3	−5.3	59	Dutasteride	60	54.3	5.7
12	Atovaquone	23	47	−24	60	Eletriptan	62.5	57.1	5.4
13	Bepridil	60	55.4	4.6	61	Emtricitabine	41	45.8	−4.8
14	Bosentan	50	55.1	−5.1	62	Endralazine*	75	79.1	−4.1
15	Budesonide	11	4.9	6.1	63	Eproxindine	70	65.3	4.7
16	Bufuralol*	46	46.4	−0.4	64	Escitalopram*	80	75.8	4.2
17	Bumetanide	81	76	5	65	Esomeprazole	90	85	5
18	Bupropion	70	64.5	5.5	66	Estradiol valerate*	3	4.5	−1.5
19	Busulfan	80	75.9	4.1	67	Ethambutol	77	81.7	−4.7
20	Caffeine*	100	95.5	4.5	68	Ethinyl	40	42.7	−2.7
21	Calcitriol	61	49.3	11.7	69	Famciclovir	77	72.3	4.7
22	Candesartan	15	13.1	1.9	70	Felodipine	15	20.4	−5.4
23	Carbamazepine	70	65.3	4.7	71	Fenflumizole*	50	47.2	2.8
24	Cefaclor	90	84.5	5.5	72	Fenfluramine	89	87.2	1.8
25	Cephalexin	90	85.7	4.3	73	Fenoprofen*	80	84.4	−4.4
26	Chlorambucil	87	81.2	5.8	74	Fenoximone*	53	82.3	−29.3
27	Chloramphenicol palmitate	80	74.6	5.4	75	Flecainide	95	89.3	5.7
28	Chloramphenicol	69	74.2	−5.2	76	Flucloxacillin	49	54.1	−5.1
29	Chloroquine	80	74.6	5.4	77	Flunisolide	20	25.3	−5.3
30	Chlorpromazine	32	37	−5	78	Fluocortolone	83.5	78.1	5.4
31	Chlorthalidone*	64	74.8	−10.8	79	Fluphenazine	2.7	8.7	−6
32	Cicloprolol	100	95.4	4.6	80	Flurbiprofen*	92	91.3	0.7
33	Cimetropium bromide	2	6.7	−4.7	81	Fluticasone	1	6.5	−5.5
34	Ciprofloxacin	70	75.7	−5.7	82	Fluvastatin	24	32.3	−8.3
35	Citalopram*	80	73.5	6.5	83	Gabapentin	60	64.8	−4.8
36	Clavulanate*	75	75.6	−0.6	84	Gatifloxacin*	96	88.3	7.7
37	Clofibrate	95	89.7	5.3	85	Gefitinib	60	65.1	−5.1
38	Clonazepam*	90	78.5	11.5	86	Gemifloxacin	71	76.9	−5.9
39	Clonidine*	60	70.3	−10.3	87	Glimepiride	100	94.5	5.5
40	Clorazepate	91	85.4	5.6	88	Glipizide	95	90.6	4.4
41	Cloxacillin	37	41.8	−4.8	89	Glyburide	95	100	−5
42	Clozapine	55	48.8	6.2	90	Granisetron	60	59.3	0.7
43	Codeine	55	49	6	91	Haloperidol	60	66.2	−6.2
44	Cyclophosphamide	74	68.5	5.5	92	Hydromorphone	24	18	6
45	Dapsone*	93	95.6	−2.6	93	Ibuprofen	80	85.6	−5.6
46	Delavirdine	96	90.5	5.5	94	Idarubicin	28	32.8	−4.8
47	Diazepam*	100	80.1	19.9	95	Imipramine	42	46.3	−4.3
48	Diflunisal	100	95	5	96	Irbesartan	70	66.5	3.5

***Table* 1.** Experimental and predicted values of human oral bioavailability for both training and test set compounds for SVM mode – *Continued*

No.	Name	Experimental	Predicted	Residual	No.	Name	Experimental	Predicted	Residual
97	Isradipine	19.5	24.3	–4.8	145	Ofloxacin	97.5	91.5	6
98	Ketoprofen*	85	96.2	–11.2	146	Olmesartan	26	31.6	–5.6
99	LAAM	47	51.6	–4.6	147	Ondansetron*	62	57.7	4.3
100	Lamivudine	86	80.4	5.6	148	Oseltamivir	75	78.8	–3.8
101	Lansoprazole	80	81.7	–1.7	149	Pantoprazole	77	83.2	–6.2
102	Levofloxacin	99	93.6	5.4	150	Penicillin	22.5	27.8	–5.3
103	Levonorgestrel	87	67.1	19.9	151	Pentazocine*	18	20.7	–2.7
104	Linezolid	100	94.4	5.6	152	Phenobarbital	96	90.2	5.8
105	Lorazepam*	93	77.7	15.3	153	Phenylethylmalonamide	91	95.4	–4.4
106	Lormetazepam*	75	68.7	6.3	154	Phenytoin*	90	88	2
107	Losartan	33	28.3	4.7	155	Physostigmine	6	11.9	–5.9
108	Lovastatin*	5	14.9	–9.9	156	Pimozide	50	56	–6
109	Melphalan	71	76.6	–5.6	157	Pinacidil	57	62.4	–5.4
110	Mepindolol	82	76.6	5.4	158	Pipotiazine*	26	9.2	16.8
111	Metergoline	23	18.7	4.3	159	Pirazolac	93.5	88.9	4.6
112	Methadone	92	89.7	2.3	160	Piroxicam	100	94.3	5.7
113	Methylphenobarbital	73	78.5	–5.5	161	Pramipexole	90	90.9	–0.9
114	Methylprednisolone	82	76.5	5.5	162	Pravastatin	17	22.5	–5.5
115	Methysergide	13	17.3	–4.3	163	Prazosin	68	65.2	2.8
116	Metoclopramide*	76	78.4	–2.4	164	Prednisone	80	75.2	4.8
117	Metopimazine	19	23.9	–4.9	165	Primaquine	96	90.3	5.7
118	Mexiletine	87	92.3	–5.3	166	Primidone	92	86.9	5.1
119	Mianserin	20	24.4	–4.4	167	Procainamide	83	89	–6
120	Midalcipran	84	88.7	–4.7	168	Procyclidine*	75	87.4	–12.4
121	Midazolam	40.5	45	–4.5	169	Promethazine	25	29.9	–4.9
122	Midodrine	93	89.6	3.4	170	Propiomazine	33	38.6	–5.6
123	Milrinone	92	86.9	5.1	171	Propylthiouracil	78	73.8	4.2
124	Minocycline	97.5	92.7	4.8	172	Proscillaridin*	7	10.1	–3.1
125	Moexipril	13	16.3	–3.3	173	Protriptyline	85	46.9	38.1
126	Montelukast	73	68.1	4.9	174	Proxyphylline	100	94.4	5.6
127	Morphine	40	44.7	–4.7	175	Quinidine	75	70.6	4.4
128	Moxifloxacin	90	95.4	–5.4	176	Raloxifene	2	6.7	–4.7
129	Nadolol	34	40.3	–6.3	177	Repaglinide	56	50.5	5.5
130	Nalbuphine	11	16.2	–5.2	178	Risperidone	70	60.9	9.1
131	Naloxone	2	7	-5	179	Rizatriptan	45	51.1	–6.1
132	Naltrexone*	20	7.1	12.9	180	Rofecoxib	93	87.3	5.7
133	Naratriptan	67.5	63.4	4.1	181	Ropinirole*	55	81.3	–26.3
134	Nateglinide*	73	87.8	–14.8	182	Sildenafil	40	34	6
135	Nevirapine	93	88.1	4.9	183	Simvastatin*	5	13.3	–8.3
136	Nifedipine	50	45.6	4.4	184	Sobrerol	72	66.4	5.6
137	Nimodipine*	13	19.6	–6.6	185	Sotalol*	60	71.9	–11.9
138	Nitrazepam	78	83.1	–5.1	186	Spironolactone	25	30.4	–5.4
139	Nitrendipine	16	20.9	–4.9	187	Stavudine	82	87.8	–5.8
140	Nizatidine	70	65.2	4.8	188	Suprofen	92	86.4	5.6
141	Nomifensine	27	35.7	–8.7	189	Telenzepine	54	53.6	0.4
142	Norethindrone	64	60.2	3.8	190	Testosterone	7	22.8	–15.8
143	Norfenfluramine	85	79.4	5.6	191	Theophylline	98.8	93.1	5.7
144	Norzimelidine	66	60.6	5.4	192	Tiagabine	90	85	5

***Table* 1.** Experimental and predicted values of human oral bioavailability for both training and test set compounds for SVM mode – *Continued*

No.	Name	Experimental	Predicted	Residual
193	Tiapamil	22	26.6	–4.6
194	Tocainide	89	83.5	5.5
195	Tolbutamide	85	90.2	–5.2
196	Toliprolol	90	94.9	–4.9
197	Topiramate*	70	72.7	–2.7
198	Topotecan*	32	50	–18
199	Torasemide	91	96.6	–5.6
200	Tramadol	75	69.9	5.1
201	Trandolapril	10	14.2	–4.2
202	Triamterene	51	55.5	–4.5
203	Triazolam	44	40.7	3.3
204	Trimethoprim	63	67.8	–4.8
204	Trimethoprim	63	67.8	–4.8
205	Trospium	9.6	15.6	–6
206	Valganciclovir*	59.4	60.1	–0.7
207	Venalafaxine*	45	57.8	–12.8
208	Verapamil	13.5	18.6	–5.1
209	Verdenafil	15	20.2	–5.2
210	Viloxazine	85	79.3	5.7
211	Warfarin	93	87.9	5.1
212	Zaleplon	30	52.6	–22.6
213	Ziprasidone	60	54.9	5.1
214	Zolmitriptan	40	46.1	–6.1
215	Zolpiclone	80	74.9	5.1
216	Zolpidem	72	66.9	5.1

In the above table the superscript of ((*)) indicate the test set compounds.

Some of these chemicals (118) are enantiomers that had the same oral bioavailabilities value so one of each pair of enantiomer was deleted from the data set. Then the principle component analysis (PCA) was performed on the remaining 243 molecules to detecting the dissimilar (outlier) chemical. After this step remaining molecules (216) were considered for further investigations, which their names are shown in *Table* 1. In other words, the final data is a 216×9 matrix, which corresponds to 216 chemicals and 9 descriptors. The distribution of the human oral bioavailability data for the complete data set is presented in form of a histogram in *Fig*. 1. As can be seen in this figure values of bioavailability (%) are acceptably distributed across the range of values.

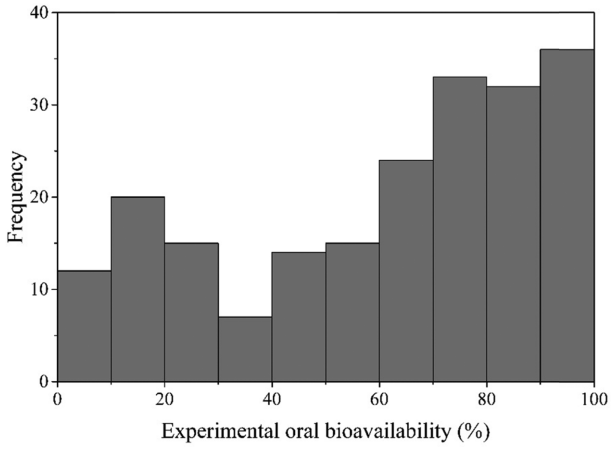

***Figure* 1.** Histogram representation of the distribution of human oral bioavailability for the 216 data set compounds.

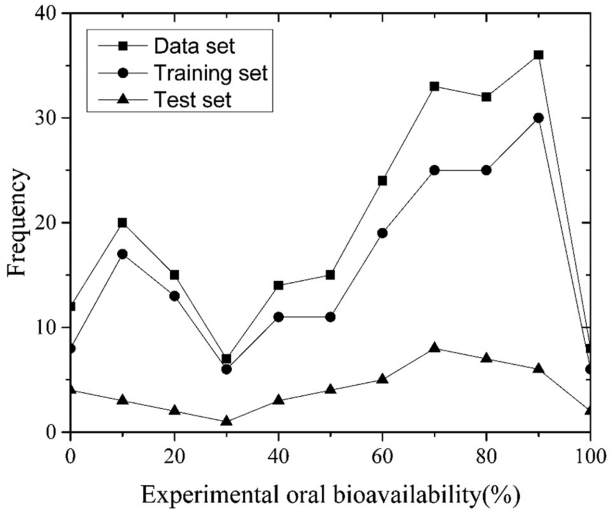

***Figure* 2.** Data set, training set, and test set composition.

Data sets were splitted randomly into two separate groups; the training and test sets, which containing, 171 and 45 members, respectively. The training set was used for the generation of the model and the test set was used for evaluation of the predictive power of generated model. The distribution of training and test set among data set space was shown in *Fig.* 2, which indicates the structurally diverse molecules possessing oral bioavailability values of a wide range were included in both sets.

Descriptor Calculation and Selection

In order to developing a QSAR model, structural features of molecules were converted to the numerical code, which were named molecular descriptors. Molecular descriptor is a result of standardized numerical calculation from logical and mathematical interpretation of chemical information, such as chemical formula, molecular structure, interaction and etc., from a molecule.[5] The molecular descriptors used to search for the best model were calculated by

the Dragon program[6] on the basis of the minimum energy molecular geometries optimized by the Hyperchem package[7] based on AM1 semi empirical method. In addition electronic descriptors were calculated by the MOPAC package (Ver.6).[8] During developing of models, great care was taken in order to avoid inclusion of highly collinear molecular descriptors. The collinear descriptors encoded similar molecular information, therefore, it was vital to test descriptors and eliminate those with low variation and those which encoded similar information (descriptors with the absolute value of Pearson correlation coefficient above 0.9). Then the most significant descriptors were selected from the pool of remaining molecular descriptors by stepwise multi linear variable selection method. These descriptors were used as inputs independent variables for developing of QSAR models.

Support Vector Machine

Support vector machine (SVM) introduced by Vapnik[9] to solve the classification and nonlinear regression problems. SVM, mapped input data into the higher dimensional feature space by the use of a kernel function then linear regression is performed in the feature space. The kernel functions that were used for nonlinear transformation of the input data can be linear, polynomial or radial basis function (RBF), which the later more commonly used in QSAR studies.The performance of SVM depends on the type of kernel function and parameters of C, ε and γ. The parameter of C is regularization constant that represent the tradeoff between minimizing the training set error and maximizing the margin.[10] The parameter of ε is sensitive loss function and the parameters of γ are used to control the width of radial basis function. The optimization of these parameters was done by simultaneous changing of these parameters and monitoring of the root mean squared error (RMS) of SVM. The values of RMS calculate according to the following equation:

$$RMS = \sqrt{\frac{\sum_{i=1}^{n_s}(y_k - \hat{y}_k)^2}{n_s}} \quad (1)$$

where y_k and $\hat{y}_k$ are the experimental and predicted response and n_s is the number of compounds in data set. SVM was used as feature mapping techniques in some QSAR studies such as, estimation of selectivity coefficients of univalent anions for anion selective electrode[11] and prediction of pharmacokinetic properties of QSAR Prediction of Oral Bioavailabilities Using SVM drugs,[12] modeling of blood brain partitioning behavior of chemicals[13] and prediction of the retention of peptides in immobilized metal affinity chromatography.[14]

RESULTS AND DISCUSSION

Descriptors

In this work, quantitative relationships between oral bioavailabilities of some drugs and their molecular structural descriptors were investigated by using linear and nonlinear feature mapping techniques. After calculations of descriptors and prescreening of them, the method of stepwise multiple linear regression was performed on the remaining descriptors to select the most important of them, which relate to the oral bioavailability of interested drugs. These descriptors are; number of circuts (nCIR), mean square distance index or Balaban index (MSD), number of Al–O–Ar or Ar–O–Ar or R–O–R or R–O–c=x substructure (where R is any group linked carbon, Al and Ar are aliphatic and aromatic groups, respectively and X is any electronegative atom) (O060), number of oxygen double bonds (O058), number of R–N–R or R–N–x groups (N075), 3D MORSE signal 10/weighted by atomic masses (Mor10m), Broto-Moreau autocorrelation of topological structure lage8 (ATS8m), leverage weighted autocorrelation of lage-6/unweighted (HATS6u), leverage weighted autocorrelation of lage-8/weighted by atomic van der Waals volumes (HATS8v).

***Table* 2.** Correlation matrix between selected descriptors.

Descriptors	O060	O058	nCIR	Mor10m	HATS6u	HATS8v	MSD	N075	ATS8m
O060	1								
O058	0.297	1							
nCIR	−0.04	−0.033	1						
Mor10m	0.087	−0.005	0.387	1					
HATS6u	−0.206	−0.091	0.15	−0.265	1				
HATS8v	0.2	−0.056	−0.071	−0.051	−0.295	1			
MSD	−0.12	−0.171	−0.58	−0.379	−0.089	−0.38	1		
N075	0.029	−0.052	−0.043	0.063	−0.182	0.01	−0.016	1	
ATS8m	0.224	0.333	0.1	0.386	−0.581	0.24	−0.305	0.119	1

The methods for calculations of these descriptors and the meaning of them are explained in the Handbook of Molecular Descriptors by Todeschini and Consonni.[5] The inter correlation among these descriptors are shown in *Table* 2. As can be seen in this table, there is not any high correlation between selected molecular descriptors. In this study sensitivity analysis approach is used to rank descriptors.[15] This method used to determine how different values of an independent variable will impact a particular dependent variable. According to the results of sensitivity analysis on SVM model, the importance order of descriptors was O058 > Mor10m > HATS8v > N075 > O060 > MSD > nCIR > HATS6u > ATS8m. Brief explanations of these descriptors that were utilized in developing of linear and nonlinear models are found in the following.

The descriptor of nCIR is a constitutional descriptor that indicates the number of circuits in a molecule. Constitutional descriptors reflecting the molecular composition of a compound without connectivity and geometry information. Descriptors of HATS8v and HATS6u are geometry topology and atomic weight assembly (GETAWAY) type descriptors.[16] These 3D descriptors encode geometrical information given by influence matrix, topological information given by molecular graph and chemical information from selected atomic properties. The next descriptor is Mor10m that is belonged to 3D MoRSE descriptors (3D molecule representation of structures based on electron diffraction). 3D MoRSE descriptors are derived from infrared spectra simulation using a generalized scattering function and mainly reflect the molecular size and 3D information. Three other descriptors in the model are N075, O060 and O058, which were belonged to atom centered fragment (ACF) descriptors. ACF descriptors represent of a single central atom surrounded by one or several atoms that separated from the central one by the same topological distance, which can describe some terms of element and bonding type in a molecule.The next molecular descriptor is mean square distance index or Balaban index. This descriptor introduce by Balaban in 1983 and encode the size and compactness of a molecule.[17] The final descriptor is ATSC8m, which is a topological 2D auto correlation indices and calculate by summing the products of atom weights of the terminal atoms of all paths in the considered path length.[18] All of these descriptors can encode topological and electronic aspects of molecules, which their variation effects significantly on the oral bioavailabilities of interested chemicals.

Modeling

Selected molecular descriptors were used as independent variable to developing QSAR models. The methods of MLR, ANN, SVM and RF were used as feature mapping methods. In order to comparison these models, standard errors (SE) and coefficient of multiple determination (R) calculated from below equation.

***Table* 3.** Statistical parameters of MLR, ANN, SVM and RF models

Methods	Parameters	Training	Test
MLR	R	0.72	0.83
	SE	21.07	12.26
ANN	R	0.73	0.86
	SE	15.96	14.76
SVM	R	0.97	0.93
	SE	5.70	10.46
RF	R	0.97	0.78
	SE	5.58	12.45

$$SE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{n-1}} \tag{2}$$

and

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

where y_i and $\hat{y}_i$ were the experimental, predicted and average experimental values of oral bioavailability, respectively and n is the number of data.

The statistical parameters of these models are shown in *Table* 3. As can be seen in this table the SVM model is superior over other models in terms of standard errors (SE) and coefficient of multiple determination (R) on training and test set. Therefore this model is further explained in the following.

In order to developing of a predictive QSAR based SVM model, SVM's parameters must be optimized. These parameters are the kernel parameter (γ), sensitive loss function (ε) and regularized constant (C). To finding the optimal value for γ, it was varied in the range of 0.1 to 300 and examine the performance of SVM model to get the minimum of RMS.The value of ε is depended on the type of noise in the data. To find an optimal value for ε, the SE for the SVM models with different ε values (in the range of 0.1 to 1) were calculated and its optimum value was selected based on minimizing of SE. The other parameter is the regularized constant, C. If C is too large, the SVM model will over fit to the training set; likewise, if C is too small, deficient stress will be placed on fitting the training set and it will cause an under fit state on the training set. In order to find an optimal value for C, SE for the SVM mod-

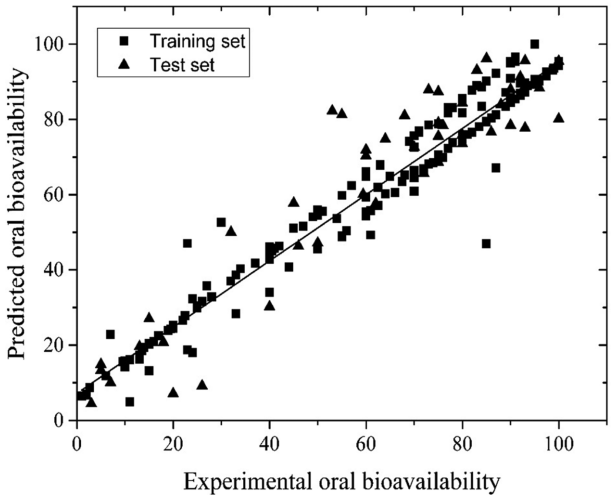

***Figure* 3.** Plot of predicted versus experimental oral bioavailability.

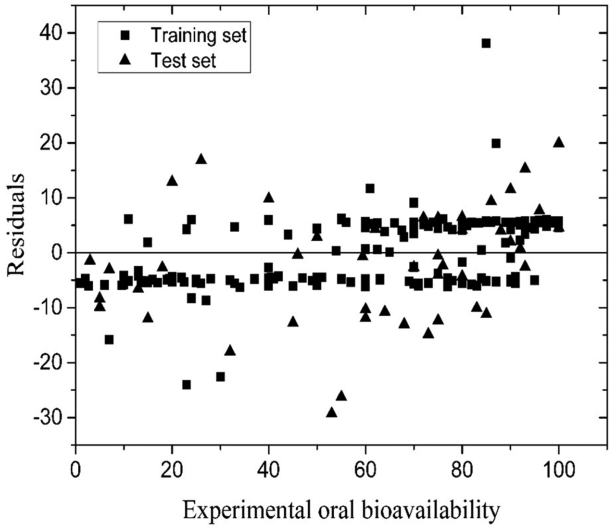

***Figure* 4.** Plot of residuals versus experimental oral bioavailability.

els with different C values were calculated and the best value for C was selected based on minimizing of SE. According to this procedure, the optimal values for these parameters were C = 10, γ = 7 and ε = 0.1. Then the developed SVM model was used for prediction of oral bioavailabilities of chemicals in test set as well as training set. These values are shown in *Table* 1. The standard error of this calculation are 5.70 and 10.46 for training and test set respectively. The predicted values of oral bioavailability were plotted versus their experimental values in *Fig.* 3 which indicate the good agreement between these values (R^2_{train} = 0.95, R^2_{test} = 0.86). The residuals of this calculation were plotted in *Fig.* 4. The random distribution of the residuals around the zero line indicated that there were not any systematic errors in this model.

In order to evaluate the robustness and predictive power of SVM model, the leave eleven out cross validation test was performed and the values of the cross validation correlation coefficient (Q^2) and standardized predicted error sum of square (SPRESS) were calculated,[19] according to the following formulas:

$$Q^2_{lmo} = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y})^2} \quad (4)$$

and

$$SPRESS = \sqrt{\frac{\sum(\hat{y}_i - y_i)^2}{n-k-1}} \quad (5)$$

In the above equations k is the number of independent variables in regression equation. The obtained Q^2 and SPRESS were 0.603 and 7.902, respectively, that indicated the robustness of developed SVM model. Y randomization test is another validation approach that must be used to investigate chance correlation among data matrix. In order to performing this test, dependent variable is randomly scrambled and a new model is developed using the original independent variables matrix. The average value of R after 30 times Y scrambling was 0.056, which revealed that the proposed model is well founded and not just the result of chance correlations.

Also the domain of applicability of developed QSAR model was investigated. The applicability domain (AD) indicate the scope of model and define the model limitations with respect to structural domain and response space.[20] Through the leverage approach, it is possible to verify whether a new chemical will lie within the structural model domain (in this case predicted data can be considered as interpolated and with reduced uncertainty, hence reliable) or outside the domain (so predicted data are extrapolated by the model and must be considered to have increased uncertainty, hence unreliable). Leverage, h_i, is defined as:

$$h_i = x_i^T (X^T X)^{-1} X_i \quad (i = 1,..., n) \quad (6)$$

where x_i is the descriptor row vector of the query compound and X is the n × k–1 matrix of k model descriptor values for n training set compounds. The superscript T refers to the transpose of the matrix/vector. To visualize the AD of QSAR models, the standardized residuals were plotted against leverages in *Fig.* 5 (William plot). In this plot, the horizontal and vertical lines delineate the limits of acceptable values; the former for the Y outliers (i.e. compounds

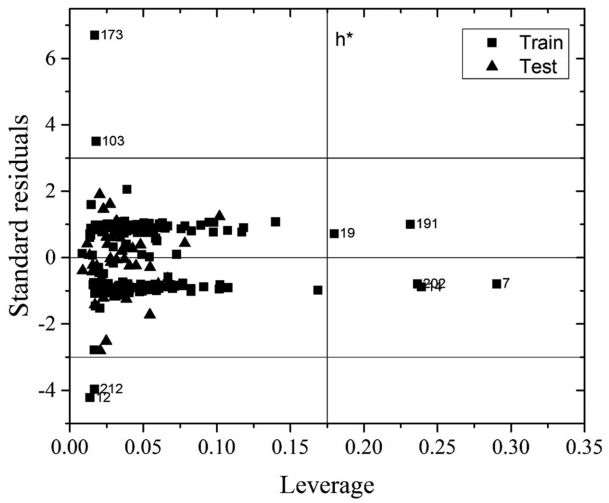

***Figure* 5.** The plot of standardized residuals versus leverage (William plot).

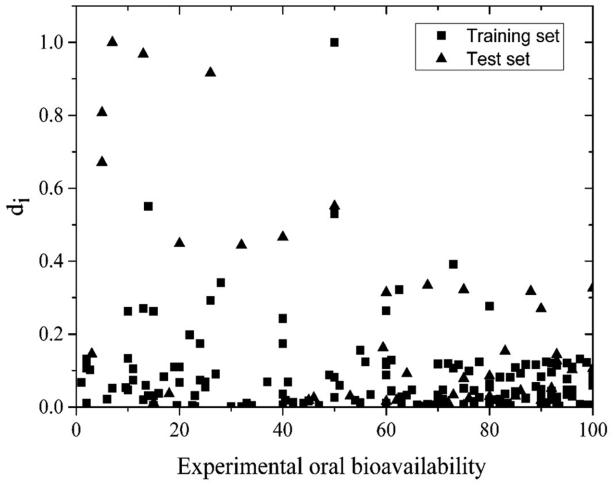

***Figure* 6.** The results of diversity test.

with standardized residuals greater than 3 standard deviation units) and the latter for X outliers, respectively. The limit of X outliers is determined by their warning hat values (h*) calculated by 3p/n, where p is the number of model variable plus one (k+1), and n is the number of training compounds. As can be seen from this figure, all predictions were reliable, except for number 7, 14, 19, 191 and 202 in training set which is not within the cut off value of h* = 0.175 and the number of 12, 103, 173 and 212 in William plot seems to be outlier.

Moreover, diversity analysis was done on the dataset to make sure the structures of the training or test sets could represent those of the whole ones.[21] In this way, a database of n compounds generated from m highly correlated chemical descriptors $\{x_j\}_{j=1}^{m}$ was considered. Each compound, X_i, is represented as following vector:

$$X_i = (x_{i1}, x_{i2}, x_{i3},..., x_{ij},...x_{im}) \text{ for } i = 1,2,..., n \quad (7)$$

where $x_{i,j}$ denotes the value of descriptor j of compound i. The collective data base $x = \{x_j\}_{i=1}^{n}$ is represented by an n × m matrix of X:

$$X = (X_1, X_2,..., X_n)^{\mathrm{T}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \quad (8)$$

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

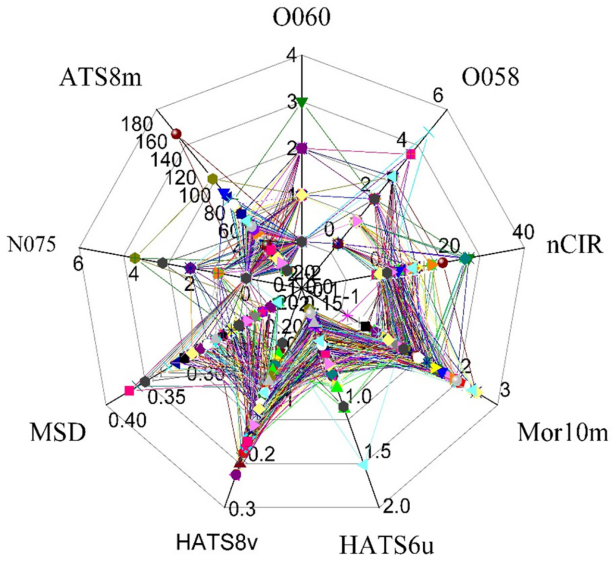

***Figure* 7.** Radar plot of the distribution of selected descriptors used in QSAR modeling.

$$d_{ij} = \|X_i - X_j\| = \sqrt{\sum_{k=1}^{m}(X_{ik} - X_{jk})^2} \quad (9)$$

The mean distances of one sample to the remaining ones were computed as follows:

$$\bar{d}_i = \frac{\sum_{j=1}^{n} d_{ij}}{n-1} \quad i = 1,2,..., n \quad (10)$$

Then the mean distances were normalized within the interval of zero to one and the resulting values were plotted against oral bioavailability (*Fig.* 6). As can be seen from this figure, the structures of the compounds are diverse for both training and test sets. Distribution of selected descriptors for SVM model are shown in the radar chart (*Fig.* 7). A

***Table* 4.** Different in silico models to predict oral bioavailability

Reference	Modeling method	Number of descriptors	Number of chemicals	Statistical parameters	
				RMS	R^2
Andrews[2]	MLR	85 fragment descriptors	591	RMS = 20.40	$R^2 = 0.71$
Yoshida[22]	ORMUCSa method	$LogD_{6.5}$, $(logD_{6.5})^2$, ΔlogD, and 15 fragment descriptors	272	–	$R_{training} = 71\%$ $R_{test} = 60\%$
Turner[23]	MLR	Eight molecular descriptors	169	$RMS_{train} = 21.87$ $RMS_{test} = 71.08$	$R^2_{training} = 0.35$ $R^2_{test} = 0.51$
Pintore[22]	Adaptive fuzzy partitioning	Ten molecular descriptors	Data1: 272 Data2: 432	– –	$R_{traning1} = 82\%$ $R_{test1} = 40\%$ $R_{traning2} = 70\%$ $R_{test2} = 64\%$
Turner[4]	ANN	Ten molecular descriptors	167	$RMS_{train} = 19.21$ $RMS_{test} = 16.15$ $RMS_{validation} = 20.74$	$R^2_{training} = 0.54$ $R^2_{test} = 0.80$
Wang[24]	GA[b]- QSAR	Eight molecular descriptors and 42 fragment descriptors	577	RMS = 21.9	$R^2 = 0.55$
Moda[*l*]	HQSAR	HQSAR derived molecular descriptors	302	$RMS_{train} = 7.46$ $RMS_{test} = 10.87$	$R^2_{training} = 0.93$ $R^2_{test} = 0.85$
Ma[22]	GA-CG[c]-SVM	Multiple molecular descriptors	766	–	$R_{training} = 0.80$ $R_{test} = 0.86$
Kumar[25]	SVM	Twelve molecular descriptors	511	–	–
Present work	SVM	Nine molecular descriptors	216	$RMS_{train} = 5.93$ $RMS_{test} = 4.93$	$R^2_{training} = 0.95$ $R^2_{test} = 0.86$

[a]bordered multi categorical classification method using the simplex technique; [b]genetic algorithm; [c]conjugate gradient.

radar chart is a graphical method that displays multivariate data in the form of a two dimensional chart with several quantitative variables represented on axis starting from the same point. The purpose of a radar chart is to compare m options across n parameters so that audience can be convinced that option A is better than say option B. Radar charts are often used when neighboring variables are unrelated, creating spurious connections so in this work instead of use a column chart for show independent variables, we illustrated them graphically because column chart might look cluttered. Each of the independent variables are in individual axes and each line is drawn connecting the data values for each drug.

Comparison between statistical parameter of developed SVM model with these obtained by other researchers are shown in *Table* 4. As can be seen in this table over developed SVM model is superior over other models.

CONCLUSION

Multiple linear regression, artificial neural network, support vector machine and random forest methods were used as feature mapping techniques for modeling and prediction of oral bioavailabilities of some drugs from their molecular structural descriptors. The results show that the SVM model exhibit reliable statistical and prediction performance over other models. The structural descriptors resulted from the stepwise multi linear variable selection method can be used to guide chemists on how to design new drugs. The results of this study revealed that quantitative structure-activity relationship approach has a good applicability for accurate prediction of oral bioavailability.

Acknowledgments. Publication cost of this paper was supported by the Korean Chemical Society.

REFERENCES

1. Moda, T. L.; Montanari, C. A.; Andricopulo, A. D. *J. Bioorg. Med. Chem*. **2007**, *15*, 7738.
2. Andrews, C. W.; Bennett, L.; Lawrence, X. Y. *J. Pharm. Res*. **2000**, *17*, 639.
3. Yasri, A.; Hartsough, D. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1218.
4. Turner, J. V.; Maddalena, D. J.; Agatonovic-Kustrin, S. *J. Pharm. Res*. **2004**, *21*, 68.
5. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; WILEY-VCH Verlag GmbH: 2000, Vol. 11, p 516.

6. The Dragon Website. http://www.disat.unimib.it/chem.
7. Hyper Chem Release 7.0 for windows; Hypercube, Inc., 2002.
8. Stewart, J. P. P. MOPAC 6.0, *Quantum Chemistry Program Exchange*, vol. 455; India University: Bloomington, 1989.
9. Cortes, C.; Vapnik, V. *J. M. l. R*. **1995**, *20*, 273.
10. Bennett, K. P.; Campbell, C. *J. ACM. SIGKDD*. **2000**, *2*, 1.
11. Fatemi, M. H.; Gharaghani, S.; Mohammadkhani, S.; Rezaie, Z. *J. Electrochim. Acta*. **2008**, *53*, 4276.
12. Yang, S. Y.; Huang, Q.; Li, L. L.; Ma, C. Y.; Zhang, H.; Bai, R.; Teng, Q. Z.; Xiang, M. L.; Wei, Y. Q. *J. Artif. Intell. Med*. **2009**, *46*, 155.
13. Golmohammadi, H.; Dashtbozorgi, Z.; Acree Jr., W. E. *J. Pharm. Sci*. **2012**, *47*, 421.
14. Kermani, B.; Kozlov, I.; Melnyk, P.; Zhao, C.; Hachmann, J.; Barker, D.; Lebl, M. *J. Sens. Actuators, B*. **2007**, *125*, 149.
15. Fatemi, M. H.; Dorostkar, F.; Ghorbannezhad, Z. *J. Monatsh. Chem. Chem. Mon*. **2011**, *142*, 1061.
16. Consonni, V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 682.
17. Balaban, A. T, 1983. *J. Pure & Appl. Chem*. **1983**, *55*, 199.
18. Broto, P.; Moreau, G.; Vandycke, C. *J. Med. Chem*. **1984**, *19*, 66.
19. Gramatica, P. *J. QSAR. Comb. Sci*. **2007**, *26*, 694.
20. Tropsha, A.; Gramatica, P.; Gombar, V. K. *J. QSAR. Comb. Sci*. **2003**, *22*, 69.
21. Maldonado, A. G.; Doucet, J.; Petitjean, M.; Fan, B.T. *J. Mol. Diversity.* **2006**, *10*, 39.
22. Zhu, J., Wang, J.; Yu, H.; Li, Y.; Hou, T. *Chem. High Throughput Screening*. **2011**, *14*, 362.
23. Turner, J. V.; Glass, B. D.; Agatonovic Kustrin, S. *J. Anal. Chim. Acta*. **2003**, *485*, 89.
24. Wang, J.; Krudy, G.; Xie, X. Q.; Wu, C.; Holland, G. *J. Chem. Inf. Model*. **2006**, *46*, 2674.
25. Kumar, R.; Sharma, A.; Varadwaj, P. K. *J. Nat. Sci. Bio. Med*. **2011**, *2*, 168.