QSAR Studies of Hallucinogenic Phenylalkylamines by Using Neural Network and Support Vector Machine Approaches

Zhuoyong Zhang*, Yuhong Xiang, and Liying An,
Department of Chemistry, Capital Normal University
Beijing 100048, P.R. China

***Abstract*—Quantitative structure-activity relationship (QSAR) studies based on a data set of 88 phenylalkylamines has been implemented. These chemicals used are among the most widely abused hallucinogens especially for young people. Because of the difficulty of assaying hallucinogenic activities, it is particularly important to develop predictive models. In this work, quantitative structure-activity relationships of phenylalkylamines were determined by using three methods, multiple liner regression (MLR), generalized regression neural network (GRNN), and support vector machine (SVM). Seven molecular descriptors, accounting for distributions of atomic charges, molecular orbital energy, molecular size and hydrophobic property were selected by stepwise regression method to build QSAR models. Comparison of the results obtained from the three models showed that the SVM and GRNN methods exhibited better performance than MLR method. SVM revealed better predictive performance comparing to the GRNN method. All the three methods should be useful to rapidly identify potential hallucinogenic phenylalkylamines.**

Keywords- QSAR; Generalized Regression Neural Network (GRNN); Support vector machine (SVM); Hallucinogen; Phenylalkylamines

I. INTRODUCTION

Hallucinogens are substances that can produce characteristic changes in human consciousness, including disorientation, derealization and depersonalization, giving rise to a variety of abnormal phenomena. The drugs are of great importance for the reason that they throw on normal and abnormal function, and they are used in pharmacological preparations for screening of new drugs for antipsychotic properties. The hallucinogens are also used as components of drugs in some countries and abused among some people, especially young people. Consequently, there is a need to develop screening and testing procedures for hallucinogens. Most reported determinations of hallucinogenic activity were done by human experiments, which seems immoral. Considering the difficulties in testing the hallucinogenic activity, it is critical to develop simple, efficient and harmless modeling approaches. Quantitative structure-activity relationship (QSAR) methods are the most promising and successful tools to provide rapid and useful means for predicting the biological activity and/or toxicity of chemical compounds [1].

* Corresponding author.

As phenylalkylamines include a relatively large amount of compounds and their hallucinogenic activity data were complete and available in literature, so relatively more studies on this category have been reported. To get better understanding on the activity of hallucinogens at molecular level, the QSAR studies of these compounds have been paid more attention and several QSAR models have been established [2-9]. Kang and Green [2] had studied LSD, 12 phenylalkylamines and 8 tryptamines, and found that their hallucinogenic activity correlated with the energy of the highest occupied molecular orbital (HOMO). However, this model was established based on a relatively small data set, and it could not represent a large diversity set of the hallucinogens. Clare introduced a new descriptor named atomic orbital phase angle to study the activities of hallucinogens [5], and got perspective results about the electronic effect of the hallucinogenic activities for a larger data set of the phenylalkylamines. In our previous work, the 3D-QSAR for 90 phenylalkylamines was investigated by using comparative molecular field analysis (CoMFA) approach [9]. In this paper, three QSAR models of 88 phenylalkylamines were established using some quantum chemical descriptors including atomic charges, orbital energy, dipole moments, and other descriptors such as hydrophobic constants, molecular volume, and molecular topological index. Results showed that the two nonlinear methods GRNN and SVM are better for the test of hallucinogenic activity.

II. DATA SET AND METHODOLOGY

A. Biological data

The chemical structures of 88 phenylalkylamines and their biological activities were cited from the reference [9]. The biological data were collected by Shulgin and co-workers. The reported work was about the hallucinogenic effect on human (oral activity data) and it has been considered as a benchmark in QSAR studies. The detail of the data can be found in the reference [9] and will not be given in the present paper. The structures and activity data of all the compounds were not listed here because of limited length of this paper. Two molecules in the reference [9] i.e. 2C-I and DOI were excluded in this paper, because the Gaussian 03 computation at HF/6-31G* is unable to deal with iodine atom.

B. Computation and descriptor generation

Molecular building was done using the molecular modeling package sybyl 7.1 (Tripos Inc., USA) with molecular sketch program. Since the crystal structure of DOET (4-ethyl-2, 5-dimethoxyamphetamine) has been reported [10], the rest of the molecules were constructed using DOET as template. Partial atomic charges were assigned to each atom using Gasteiger-Hückel method, and then the molecular geometry of each compound was minimized using Powell method and tripos standard force field with a distance-dependent dielectric function. The minimization was terminated when energy gradient convergence criterion of 0.005 kcal/mol was reached or when the 2000-step minimization cycle limit was exceeded. Then the resultant conformation of each compound was input into the Gaussian 03 package of programs for further optimization. Conformation optimization and frequency computations were done at HF/6-31G* level. Frequency analysis shows that there were no virtual frequencies, which suggested that all the molecules take stable conformations. Quantum descriptors such as atomic charge, atomic orbital energy E_{HOMO} (the highest occupied molecular orbital energy), E_{SHOMO} (the second highest occupied molecular orbital energy), E_{LUMO} (the lowest unoccupied molecular orbital energy), E_{SLUMO} (the second lowest unoccupied molecular orbital energy), and the energy gaps between them, dipole moment (μ_x, μ_y, μ_z, and μ) and polarizability α were extracted from the resultant files. Other descriptors such as ClogP, topological parameters Wiener Index and molar refractivity MR were acquired from the Chemoffice 2005 software package. The volumes and hydrophobic constants of the substituents on the phenyl ring were cited from the reference [6]. The volumes of the substituents were obtained by subtracting the volume of benzene from the volume of the benzene mono-substituted with the substituent, using the model builder geometries. And these calculations were performed using the QSAR module of Hyperchem. Similarly, the lipophilicity of a substituent was obtained from the lipophilicity of benzene substituted with the substituent, calculated with the program ClogP, minus the lipophilicity of benzene itself. The volume or lipophilicity of a ring system fused to the benzene ring was halved and the halves ascribed to both of the two points of attachment. Indicator variable I_{Me} was introduced to explain the difference between the phenylethylamines and amphetamines, if there is neither an H atom nor a methyl at the α position of the amino group, and then a zero would be set to the I_{Me}.

C. Selection of structural descriptors

To select the significant structural descriptors, the stepwise regression was performed using SPSS software (version of 14.0, SPSS Inc., Chicago, Illinois). After optimization, seven parameters, including the energy gap between the E_{SLUMO} and E_{LUMO}, labeled as $E_{\Delta L}$, the lipophilicity of the 4-substituent of the phenyl ring π_4, the volume of the 2- substituent of the phenyl ring V_2, the topological parameter Wiener index (WI), the indicator variable I_{Me}, the atomic charge of the nitrogen atom of the amino group σ_N, and the summation of atomic charges of the six carbon atoms on the benzene ring σ_P, were selected to build the QSAR models. Table 1 showed the coefficient correlation matrix of these variables. It can be seen in the table that their correlation coefficient are less than 0.581, which suggest that no obvious correlations exist between the selected descriptors.

D. Data Preparation

Data pre-processing step is necessary to ensure that each parameter is on the comparable scale. In this paper, all the variables were scaled to transform each variable to have a distribution with a mean value of zero and unity standard deviation. The autoscaling was performed by the following equations:

$$X_{ij} = \frac{x_{ij} - \overline{x}_j}{\sqrt{V_j}} \quad (1)$$

$$V_j = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \overline{x}_j\right)^2 \quad (2)$$

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \quad (3)$$

Where x_{ij} is the measured value in the *j*th column of training set, $\overline{x}_i$ is the average value of the *j*th column, V_j is the deviation of the *j*th column, and X_{ij} is the scaled valued to be used for the modeling.

Computation was performed on a Pentium IV PC with 512 Mb RAM.

TABLE 1. COEFFICIENT CORRELATION MATRIX OF VARIABLES ENTERED IN THE FINAL MODEL

	σ_N	σ_P	WI	V_2	π_4	I_{Me}	$E_{\Delta L}$
σ_N	1	-0.168	0.090	-0.050	-0.005	0.017	0.184
σ_P		1	0.203	-0.011	-0.236	0.098	-0.294
WI			1	0.026	0.581	-0.073	0.141
V_2				1	0.269	0.239	0.405
π_4					1	-0.288	0.293
I_{Me}						1	0.088
$E_{\Delta L}$							1

III. RESULTS AND DISCUSSION

A. Model validation

Validation of the models is required to test the predictive ability and generalization of the methods as well as to enable comparison between them. The whole data set was divided into a training set (71 compounds) and a test set (17 compounds). The training set was applied for model development. In the modeling process, leave-one-out (LOO) procedure was implemented. Each compound in the training set was omitted once, and the model developed from all the remaining compounds in the training set predicted its hallucinogenic activity. The test set was used for external

prediction of the obtained models. The goodness of the models was measured from RMSE, which is defined as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (y_k - \hat{y}_k)^2}{N}} \quad (4)$$

where y_k is the desired output and $\hat{y}_k$ is the output of the model, and N is the number of compounds in the analyzed set.

B. MLR model

A linear model was constructed based on the seven selected descriptors. Figure 1 shows the correlation between predicted values of logA and experimental ones. The linear function was built as follows, with parameters selected from the stepwise regression.

$$\log A = 0.516 * \sigma_P - 0.868 * \sigma_N - 0.002 * WI + 0.001 * V_2 + 0.532 * \pi_4 + 19.833 * E_{AL} + 0.383 * I_{ME}$$

R=0.789, s=0.302, F=14.889, N=71(Training set)
R=0.761, s=0.358, N=17 (Test set)

C. GRNN model

GRNN was used to develop a nonlinear model based on the same subset of descriptors. This network could be designed as 7-71-1 net to indicate the number of units in the input, the pattern layer and the output layer, respectively. The GRNN does not require an iterative training procedure. Instead, it is a single pass procedure. The smoothing parameter σ determines how tightly the underlying regression surface will be approximated, and the GRNN's success is dependent on this choice. A larger value of σ will result in smoother final curve and thus lead to greater generalization, whereas smaller values of σ will produce a more detailed final curve with a more accurate fit to the data points.

In this paper, the adjustment of the smoothing parameter σ was performed using the LOO cross-validation procedure, the best σ was acquired when the lowest RMSE reached. Each RMSE on LOO cross-validation was plotted versus the σ wa obtained. It can be concluded that the optimal condition was σ=0.15. And the optimal model give a RMSE of 0.140 logA unit for the training set, 0.301 for the test set, and 0.183 for the whole set. The corresponding correlation coefficients were 0.96, 0.83, and 0.93, respectively. The predictive values based on the best GRNN model were listed in Table 2. The plot of predicted logA versus experimental values of the compounds both in the training set and the test set is shown in Figure 2.

D. SVM model

The first step of SVM training is the selection of the type of the kernel function. The radial basis function (RBF) is commonly used because of its good general performance. So, in this paper, the RBF kernel function was chosen to build the model. Then, the training process included the selection of ε-insensitive loss function, regularized factor C and the corresponding parameter γ of the RBF kernel function. The parameters γ greatly affect the number of support vector, which has a close relation with the performances of the SVM and the training time. At the same time too many support vector could produce overfitting and increase the training time. In addition, γ controls the amplitude of the RBF function and therefore controls the generalization ability of SVM.

To obtain the optimal γ, the support vector learning machines with different γ were trained, the γ varying from 0.02 to 0.10. The RMSE was calculated on different γ, according to the generalization ability of the model based on LOO cross-validation for the training set in order to determine the optimal one. The curve of RMSE versus the gamma was obtained. The optimal γ was found as 0.04.

Parameter ε-insensitive loss function prevents the entire training set meeting boundary conditions and so allows for the possibility of sparsity in the dual formulation's solution. The optimal value for ε depends on the type of noise present in the data, which is usually unknown. The RMSE of LOO cross-validation on different epsilon is recorded and the optimal value was 0.13.

The regularization constant C controls the trade-off between maximizing the margin and minimizing the training error. If C is too small, then insufficient stress will be placed on fitting the training data. If C is too large then the algorithm will overfit the training data. To make the learning process stable, a relatively large value should be set up for C initially. The relationship of the RMSE versus C value was obtained for γ=0.04, ε=0.13. It can be seen from the figure that the selection of the C parameter has some influence on the performance. The RMSE value dramatically decreased as the C value increased and got the minimum at C=120, then, the curve began to ascend. So, the best choice for γ, ε, and C were 0.04, 0.13 and 120, with the support vector number of 31. The optimal model gave a RMSE for the LOO cross-validation of 0.274, a RMSE for the training set of 0.229, 0.282 for the test set, and 0.243 for the whole set. The corresponding correlation coefficients were 0.877, 0.838 and 0.875 respectively. The predictive values of all the compounds based on the best SVM model were obtained. Figure 3 showed the predicted versus experimental values of logA.

E. Comparison of the results

Comparison between the results obtained by MLR, GRNN and SVM based on the RMSE. The SVM and the GRNN models give higher correlation coefficient. This result indicates that the SVM and the GRNN performed better than the MLR method. For the whole data set, the GRNN method seems to outperform the SVM, but the GRNN brought a larger RMSE in the test set. This conforms to the shortcoming of the ANNs of overtraining and overfitting. While for the prediction of the holdout test set, the SVM method showed much superiority and also showed the better generalization ability. The reason may be that the SVM method embodies the structural risk minimization principle, which minimizes an upper bound of the generalization error rather than minimizes the training error. This eventually leads to better generalization than neural networks, which implement the empirical risk minimization principle and may not converge to global solutions.

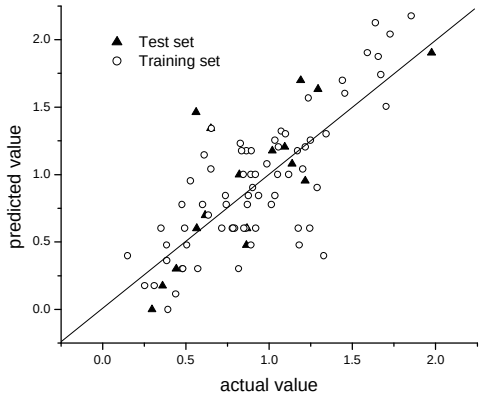

Figure 1. The predicted versus the actual value for the training and test set by the MLR method

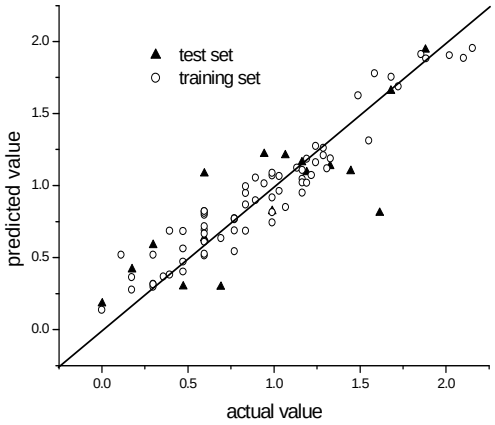

Figure 2. Predicted value versus actual value for compounds both in the training set and test set by the GRNN method

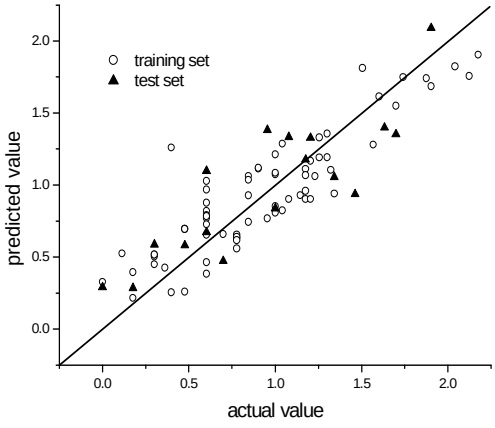

Figure 3. Predicted value versus actual value for compounds both in the training set and test set by the SVM

F. Discussion on the descriptors

From a practical point of view, interpreting the descriptors used in the models could provide some insight into factors that are likely to govern the hallucinogenic activities of these compounds and help us to understand which interactions may play an important role in the hallucinogenic effect.

TABLE 2. EXPERIMENTAL AND CALCULATED LOGA FOR THE TEST SET BY THE THREE QSAR MODELS

Designation	Exp. (logA)	MLR		GRNN		SVM	
		Pre.	Residue	Pre.	Residue	Pre.	Residue
2C-B	1.204	1.094	0.110	1.105	0.099	1.331	-0.127
2C-T	0.602	0.867	-0.265	1.091	-0.489	1.101	-0.499
2C-T-F	1.462	0.561	**0.901**	1.117	0.345	0.938	**0.524**
3-TE	0.602	0.565	0.037	0.640	-0.038	0.679	-0.077
4-Br-3, 5-DMA	1.633	1.292	0.341	0.827	**0.806**	1.409	0.224
4-MA	0.699	0.616	0.083	0.305	0.394	0.479	0.220
5-TASB	0.301	0.441	-0.140	0.591	-0.290	0.596	-0.295
ALEPH-2	1.699	1.190	0.509	1.678	0.021	1.331	0.368
BOD	1.176	1.020	0.156	1.175	0.001	1.170	0.006
DOPR	1.903	1.976	-0.073	1.966	-0.063	2.097	-0.194
G	1.079	1.138	-0.059	1.223	-0.144	1.34	-0.261
HOT-2	1.342	0.649	0.693	1.151	0.191	1.073	0.269
J	0.176	0.360	-0.184	0.420	-0.244	0.281	-0.105
MDA	0.477	0.864	-0.387	0.303	0.174	0.577	-0.100
ME	0.000	0.297	-0.297	0.181	-0.181	0.261	-0.261
MEM	0.954	1.216	-0.262	1.229	-0.275	1.387	-0.433
TMA-6	1.000	0.821	0.179	0.834	0.166	0.833	0.167

The most influential descriptor was the $E_{\Delta L,}$ which represents the electronic effect. The $E_{\Delta L}$ is the energy gap between the lowest unoccupied orbital and the second lowest orbital. The E_{LUMO} is related to the electron affinity energy, a small value of the E_{LUMO} enhances the ability of accepting electrons. Topological parameter WI, accounting for the size of the molecule, with a negative coefficient, suggests that more bulky and branchy compounds have lower hallucinogenic activity. The lipophilicity constant π_4 takes a relatively large coefficient shows that the hydrophobicity effect of the 4-position is crucial to the hallucinogenic activity. And there is evidence that substituent at the 4-position of various phenylisopropylamines might directly interact with receptors [9]. The volume of the substituents at the 2-position with a positive coefficient showed that 2-substituted compounds usually have higher activities, however, it cannot definitely concluded that larger substituent increase the activity. Atomic charge attributes to the electronic factors. A negative coefficient of the σ_N indicates that the increasing the atomic chare of the nitrogen atom will decrease the activity, while the atomic charge of the nitrogen atom was dependent on the substituents connected. Summation of the atomic charge of the carbon atoms on the phenyl ring reflects the electronic feature of the aromatic system. Indicator variable I_{Me} was introduced to explain the differences between the phenylethylamines and amphetamines. With a positive coefficient, it explains the fact that the amphetamines usually have high activities than the phenylethylamines.

IV. CONCLUSIONS

Three methods, MLR, GRNN and SVM were used to construct linear and nonlinear QSAR models to predict the hallucinogenic activities of 88 phenylzlkylamines. Seven descriptors, which represent the features of the compounds responsible for the hallucinogenic activity, were selected from stepwise regression. Inspection of the models indicated that steric, electrostatic, and the hydrophobicity of the 4-position accounted for the hallucinogenic activity. However, the whole hydrophobicity CLOGP was excluded in the stepwise regression. Among the three models, the GRNN and the SVM models developed in this study were found to outperform MLR models significantly, suggesting that they were able to make use of the nonlinear relationships in the data set. However, in prediction of the compounds in the holdout test set, GRNN did not show better result. Conversely, the RMSE value gap between the training and test compounds with SVM method was smaller than that of the GRNN method. This reflects the fact that SVM is more predictive and has generalized ability and the neural network can easily trapped into overfitting.

ACKNOWLEDGMENT

This work is supported by the Scientific Research Common Program of Beijing Municipal Commission of Education (KM200710028009) and Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality (PHR20100718)

REFERENCES

[1] C.Y. Zhao, R.S. Zhang, H.X. Zhang, C.X. Xue, H.X. Liu, Liu, M.C., Z.D. Hu, and B.T. Fan, "QSAR study of natural, synthetic and environmental endocrine disrupting compounds for binding to the androgen receptor", SAR QSAR in Environ. Res., 16(4), pp. 349-367, Aug. 2005.

[2] S. Kang,and J.P. Green, "Steric and electronic relationships among some hallucinogenic compounds", Proc.Natl. Acad. Sci. U.S.A., 67(1), pp. 62-67, Sep. 1970.

[3] S.P. Gupta, P. Singh, and M.C. Bindal, "QSAR studies on hallucinogens", Chem. Rev., 83(6), pp. 633-649, Oct. 1983.

[4] B.W. Clare, "Structure-activity correlations for psychotomimetics. 1. Phenylalkylamines: electronic, volume, and hydrophobicity parameters", J. Med. Chem., 33, pp. 687-702, Feb. 1990.

[5] B.W. Clare, "The frontier orbital phase angles: novel QSAR descriptors for benzene derivatives, applied to phenylalkylamine hallucinogens", J. Med. Chem., 41, pp.3845-3856, Sep. 1998.

[6] B.W. Clare, "QSAR of benzene derivatives: comparison of classical descriptors, quantum theoretic parameters and flip regression, exemplified by phenylalkylamine hallucinogens", J. Computer-Aided Mol. Des., Aug. 2002, 16: 611-633.

[7] A. Altun, K. Golcuk, M. Kumru, and A.F. Jalbout, Electron-conformational study for the structure–hallucinogenic activity relationships of phenylalkylamines", Bioorg. Med. Chem., 11: pp. 3861-3868, June 2003.

[8] M. Schulze-Alexandru, and K. -A. Kovar," quasiatomistic receptor surrogates for the 5-HT2A receptor: A 3D-QSAR Study on Hallucinogenic Substances", Quant. Struct.-Act. Relat., 18, pp. 548-560, Dec. 1999.

[9] Z.Y. Zhang, L.Y. An, W.X. Hu, and Y.H. Xiang, "3D-QSAR study of hallucinogenic phenylalkylamines by using CoMFA approach", J. Comput. Aided Mol. Des., 21(4), pp. 145-153, Jan. 2007.

[10] O. Kennard, C. Giacovazzo, A.S. Horn, R. Mongiorgi, L. Riva di Sanseverino, "Crystal and molecular structure of the psychotropic drug 2-(4-ethyl-2,5-dimethoxyphenyl)-1-methylethylamine (4-ethyl-2,5-dimethoxyamphetamine)", J. Chem. Soc. Perkin Trans. II., 10, pp. 1160-1163. 1974.