

Use of the Genetic Algorithm for Variable Selection of PLS Regression in a QSAR Study on [4,5-d] pyrimidinederivatives as Antagonist of CXCR2

**Tahereh Asadollahi¹, Shayessteh Dadfarnia*¹,
Ali Mohammad Haji Shabani¹, Jahan B. Ghasemi²**

¹*Department of Chemistry, Faculty of Science, Yazd University, Yazd, I. R. Iran*

²*Department of Chemistry, Faculty of Science, K. N. Toosi University of Technology,
Tehran, I. R. Iran*

(Received July 19, 2012)

Abstract

Quantitative relationships between the structures of 52 antagonists of the CXCR2 receptors and their activities were investigated by the partial least squares (PLS) method. The genetic algorithm (GA) has been proposed for improvement of the performance of the PLS modeling by choosing the most relevant descriptors. Power prediction of the QSAR models developed with PLS and GA-PLS methods were evaluated using cross-validation, and validation through an external prediction set. A comparison between the different developed methods indicates that GA-PLS can be chosen as supreme model due to its better prediction ability than the other two methods. The developed models were found to be useful for the estimation of pIC₅₀ of CXCR2 receptors for which no experimental data is available.

*Corresponding author: sdadfarnia@yazduni.ac.ir

Introduction

The search for quantitative relations between chemical structure and biological activity is the subject of *Quantitative Structure-Activity Relationships* (QSAR), the purpose of which is to explain why a given drug produces its particular effect, and ultimately to predict the effect of newly synthesized chemical compounds [1, 2]. This theory has an ultimate role, the proposal of a model capable of estimating the activities of compounds, by relying on the assumption that these resulting effects are the consequence of the molecular structure. Since the pioneer studies of Hansch, the use of QSAR has become helpful in understanding chemical–biological interactions in drug and pesticide research, and in various toxicology studies. In QSAR, the structure is translated into the so-called molecular descriptors, describing different relevant features of the compounds, through mathematical formula obtained from the chemical graph theory, information theory, quantum mechanics and so forth. More than thousand descriptors are available and are reported in the literature. Although, the structure descriptors may be easily produced by modern software, but many of these descriptors are probably irrelevant to the targeted biological activities. Thus, one has to decide how to select the descriptors that characterize the property under consideration in the best possible manner. Removing the irrelevant descriptors or selecting the most relevant and highly correlated ones is a vital step in QSAR studies. The removal of irrelevant features often improves the performance of learning models, which provides faster and more cost-effective prediction and may provide a better understanding of what properties of target and ligand are relevant to the biological procedure.

One of the important features of mathematical model is its ability to predict the activity of molecules not yet synthesized or those with limited *in vivo* and *in vitro* experimental information, due to the economic or ethical reasons. Furthermore, the mathematical models can give information about activity of a molecule and explain receptors binding places (enzyme, ionic channel etc.) through correlations with molecular descriptors [3-10]. The more often used techniques in construction of QSAR model were PLS, principle component regression (PCR) and artificial neural network (ANN). The partial least squares (PLS) is insensitive to co-linearity among the predictor variables and allows one to handle data set where the number of variables is larger than number of observations [11]. Thus, for large data sets PLS is preferable. In addition, PLS analysis provides equation describing the relationship between one or more dependent variable and a group of explanatory variables.

However, PLS is highly sensitive to extreme values of variables, which do not contribute to a predictive model. The situation gets worse when more variables are introduced to the models. Thus, the larger the number of variables, the less the predicted ability of the developed model[12]. Therefore, a variable selection step is necessary prior to building PLS models. To solve this problem GA and PLS had been combined in variable selection of QSAR and QSPR modeling [13, 14]. Genetic algorithms are best known for their ability to efficiently search large spaces and had been widely applied in different fields. Thus, GA can be used for improving the prediction of QSAR modeling.

Thiazolo[4,5-d]pyrimidine derivatives display a plethora of biological activities. It had been reported that [4,5-d]pyrimidine analogues have anti-inflammatory activities, due to TNF inhibition [15]. These compounds have been described to have anticancer[16], anti-inflammatory and antimicrobial activity[17]. Recently, 2,7-substituted-thiazolo[4,5-d]pyrimidine have also been described as ATP-competitive inhibitors of several protein kinases (EGFR, cSrc, HER-2, Lyn and c-Abl) [18]. Thus, these structures represent a class of molecules capable of binding to multiple target proteins (receptors or enzymes) with high affinity. So, from molecular drug discovery point of view, the molecule antagonists of the CXCR2 receptor such as pyrimidine derivatives are attractive biological targets. In this study, the QSAR was used to construct a suitable model to predict the drug properties of new pyrimidine derivatives as antagonist of CXCR2 receptor compounds such as the half-maximal inhibitory concentration (IC_{50}). Furthermore, in order to predict the activity of pyrimidine derivatives, an important class of bioactive compounds, the linear regression methods of PLS and GA-PLS were used to construct the QSAR model

Materials and methods

Experimental Data

The structure and biological data of 52 molecule antagonists of the CXCR2 receptor were obtained from literature [19, 20] and are shown in Table 1-3. The theoretical molecular descriptors were derived from the chemical structure of the compounds. The 3D-structures of the molecules were drawn using HyperChem version 8.05. The resulting geometries were further refined by means of semi-empirical method AM1 and the molecular structures were optimized using the Polak-Rebiere algorithm until the root mean square gradient reaches 0.001 kJ (mol Å). Then it was transferred into the Dragon program package (Milano

Chemometrics group, version 3) [21] to obtain all of the descriptors of 0D, 1D, 2D and 3D classes of twenty different categories which involved a total of 1497 descriptors [22-24].

After the calculation of descriptors, they were analyzed to decrease the redundancy existing in the descriptor data matrix. Finally, the following steps were applied into the data matrix:

- (1) In order to detect the homogeneities in the data set and to recognize the potential outliers in all of the molecules under study the principal components analysis (PCA) was applied. According to the results of PCA, the original dataset was split into training set (including 42 compounds) which was used for QSAR modeling; and test set (10 compounds), which was used for evaluation of the predictive result of the final model.
- (2) The constant and near constant descriptors were deleted by application of the feature selection methods for production of suitable model i.e. one of the any two descriptors with an inter-correlation greater than 0.95 was removed to reduce the redundant and useless information.
- (3) In order to select the best descriptors as input for the next step, the genetic algorithm was run.
- (4) Finally the PLS model was performed by the remaining descriptors with 7 latent variables, it was examined on the test set and the model performance was evaluated.

Another key step in QSAR modeling is the evaluation of the model's stability and predictive ability. In order to evaluate the suitability of the developed models for the prediction of the activity of the compounds, different statistical parameters were used [25] including cross validation coefficient (Q^2 or R^2_{cv}), root-mean square error of prediction (RMSEP) and relative error percent of prediction sets ($REP_{pred}\%$).

Cross-validated coefficient of $R^2_{cv}(LOO-Q^2)$ was calculated according to the following formula:

$$Q_{ext}^2 = 1 - \frac{\sum_{i=1}^{Pred} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{Pred} (y_i - \bar{y}_{tr})^2} \quad (1)$$

where \hat{y}_i , y_i and \bar{y}_{tr} are the predicted value, the experimental value (over the prediction set) and the averaged value of the dependent variable for the training set, respectively.

The REP% was calculated according to the following equation:

$$REP (\%) = 100 / \bar{y} \left[\frac{1}{n} (y_i - \hat{y}_i)^2 \right]^{0.5} \quad (2)$$

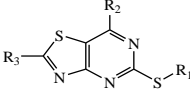
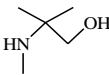
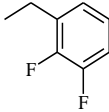
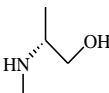
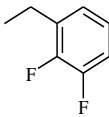
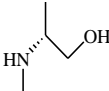
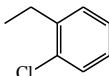
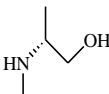
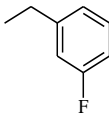
where \hat{y}_i , y_i , \bar{y} and n are the predicted value, the experimental value, the mean of the experimental value in the prediction set and the number of samples, respectively.

The root-mean square error of prediction (RMSEP) was defined as a measure of the average difference between the predicated and experimental values at the predication stage. The RMSEP is a frequently used measure of the differences between the predicted values by a model or an estimator and the actually observed values from the objects being estimated. The RMSEP was defined as follows:

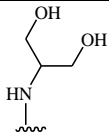
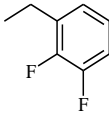
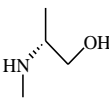
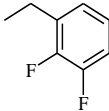
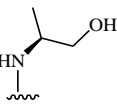
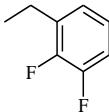
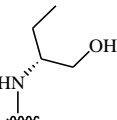
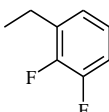
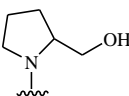
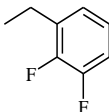

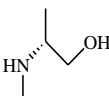
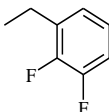

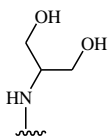
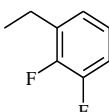

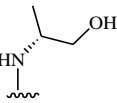
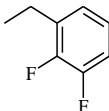

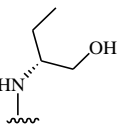
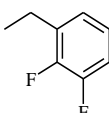
$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \quad (3)$$

where \hat{y}_i , y_i and n are the prediction value, the measured value and the number of measurements at the prediction set, respectively.

Table 1. Structures and biological activities of thiazolo [4,5-d]pyrimidines derivatives

				
No	R ₃	R ₂	R ₁	pIC ₅₀
1	NH ₂			7.6
2	NH ₂			8.4
3	NH ₂			7.9
4	NH ₂			7.8

5	NH ₂			7.5
6	NH ₂			7.9
7	NH ₂			8.4
8	NH ₂			7.8
9	NH ₂			7.0
10	NH ₂			7.5
11	NH ₂			6.0
12	NH ₂			6.9
13	NH ₂			7.8
14	NH ₂			7.1

15	NH ₂			7.5
16	NH ₂			8.4
17	NH ₂			7.2
18	NH ₂			7.6
19	NH ₂			<6.0
20				7.7
21				7.0
22				7.5
23				7.0


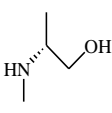
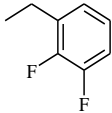
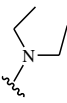
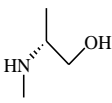
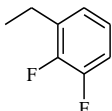
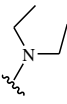
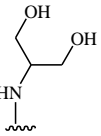
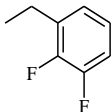
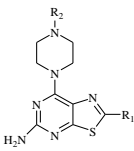

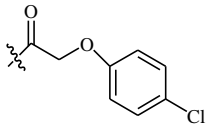

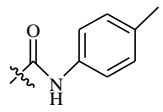

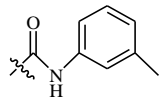

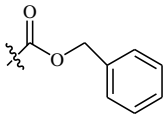

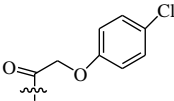
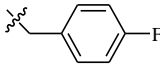
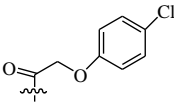
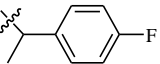
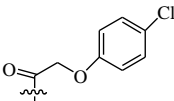
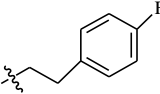
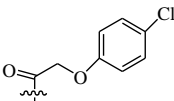
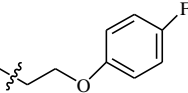
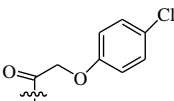
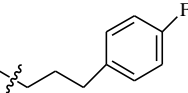
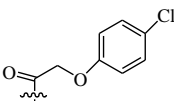
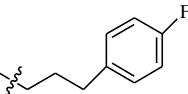
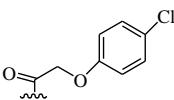
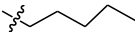
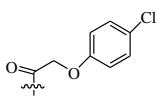
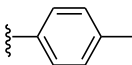
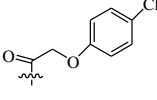
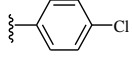
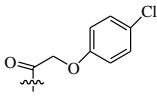

24				8.1
25				6.4
26				5.8

Table 2. Structures and biological activities of thiazolo [4,5-d]pyrimidines derivatives

				
No.	R ₂	R ₁	pIC ₅₀	
27	H		5.4	
28			6.3	
29			6.1	
30			6.5	

31			5.3
32			6.2
33			5.4
34			5.4
35			5.4
36			6.0
37			6.0
38			6.2
39			6.5
40			6.6
41			7.4

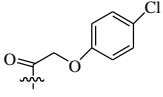

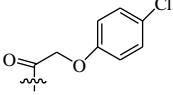

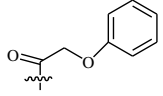
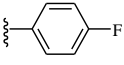
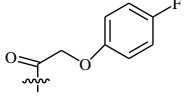
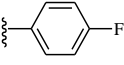
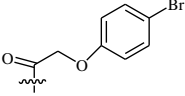
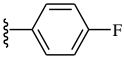
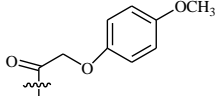
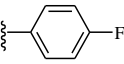
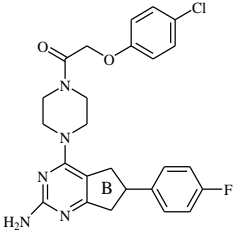





42			7.3
43			4.4
44			7.0
45			7.2
46			5.4
47			7.4

Table 3. Structures and biological activities of piperazinyl substituent of 5-amino-2-(4-fluorophenyl)-thiazolo[5,4-d]pyrimidines derivatives.

	No.	B	pIC ₅₀
	48		6.6
	49		7.0
	50		7.2
	51		6.4
	52		7.4

Variable Selection Procedure

Removing Constant and Nearly Constant Variables

First, an identical test was carried out to remove the constant variables. This procedure removes descriptors that had a constant value for a user specified percentage of the dataset. Identical tests were performed to remove those descriptors that have zero or identical information across a user defined percentage of the training set compound space (80% to 90%). Then the nearly constant variable indicated by standard deviation was removed because, in the context of QSAR, descriptors with low standard deviation are not particularly relevant for regression modeling. In other word, such descriptors do not have the required variance to be able to capture the variance of the interested property of modeling.

Genetic Algorithm

Another problem in selection of the set of molecular descriptors is the co-linearity within them. Thus, in order to find the more convenient set of descriptors, the variable selection method of the GA-PLS was performed. The QSAR model was derived by performing the GA analysis with partial least squares (PLS)-regression method on the population size of 64 and mutation rate of 0.003. The parameters and the results of R^2 , REP%, RMSEP and Q^2 for prediction set of GA-PLS study are summarized in Table 4 and 5 respectively. As the results revealed, the use of GA prior to the calibration, results in regression model with improve predictive power.

Table 4. Parameters of genetic algorithm GA.

Cross validation	Random subset
Number of subset	4
Window width	2
Initial term %	20%
Maximumgeneration	100
Convergence (%)	80
Cross-over	Double

Table 5.The statistical parameters calculated for PLS model

Parameters	GA-PLS	PLS
R^2	0.85	0.77
Q^2	0.79	0.75
RMSEP	0.552	0.601
REP %	8.313	8.939

PLS Method

The PLS is a robust, multivariate statistical extension of multiple linear regression (MLR), in handling regression tasks with too many variables [26]. PLS is a linear multivariate method for relating two data matrices, X and Y to each other. It expressed a projection to latent structures by means of the partial least squares, and improved the precision with increasing the numbers of X variables.

The general idea of the PLS method was to extract information, the Latent variables from the data, which account for most of the variance in the response and construct predictive model. Generally the following equation demonstrated the PLS method:

$$Y = b_1L_1 + b_2L_2 + b_3L_3 + \dots + b_mL_m$$

Y is the dependent variable (the biological activity); L_1 through L_m are the latent variables, and b_1 through b_m are the regression coefficients corresponding to L_1 - L_m .

The modeling by PLS method was performed in the MATLAB 7.0, using PLS Toolbox 2.

Results and Discussion

Regression Model

After the selection of suitable descriptors, all descriptor variables were preprocessed by auto scaling. Then the number of significant factors (latent variable) for the PLS algorithm was determined using the cross-validation method. With cross-validation, one sample was kept out (leave one out) of the calibration and was used for prediction. The process was repeated so that each sample was kept out once. The predicted values of left-out samples were then compared to the literature values by using prediction error sum of squares (PRESS) and are demonstrated in Fig. 1.

The PRESS was calculated according to the following equation:

$$PRESS = (y_i - y_i^{Pred})^2$$

where y_i and y_i^{Pred} are the measured and prediction value respectively.

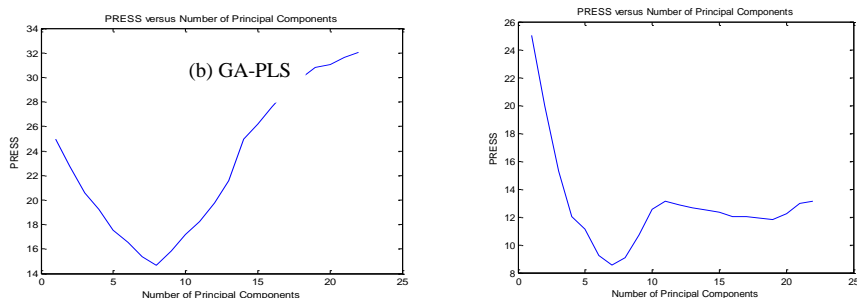


Figure 1. The PRESS versus number of principal component.

As Fig.1 revealed the most convenient PLS and GA-PLS models were produced with eight and seven latent variables, respectively. The types and definitions of 25 descriptors used in GA-PLS model are given in Table 6.

Table 6. The partial least squares regression coefficients

ID	Descriptor	Group
1	MSD, SEigZ	Topological Descriptors
2	BEHv1, BEHV2, BEHe2, BELe1, BELp1	BCUT
3	MATS1m, MATS2m,	2D autocorrelation
4	DISPe	geometrical descriptors
5	Mor19m, Mor17v	3D-MorSE descriptors
6	G1u, G3m, E3m, E1v, Dv	WHIM descriptors
7	HATS2u, HATS2m, HATS8p, R7u+, R1m+, R2v+, R8e, R2p+	GETAWAY descriptors

The calculated statistical parameters by PLS model are presented in Table 5. The low values of root-mean square error of prediction confirm the prediction ability and the accuracy of the resulted models for the external test set. The plots of predicted activity by regression

model against the experimental activity are shown in Fig. 2 and indicate that the error in the developed QSAR model is low.

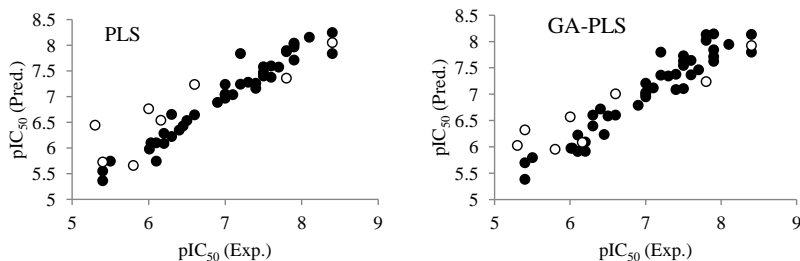


Figure 2. Predicted vs. the experimental activity (●) Training set (○) Test set

Interpretation of the Descriptors

Concerning the interpretability of the descriptors, it is important to take into account that the response of the model is the result of a series of complex biological or physico-chemical mechanisms. So, it is very difficult to attribute the mechanistic meaning to the important descriptors used in a QSAR model. Moreover, it must also be emphasized that in multivariate models such as MLR models, the interpretation of the singular molecular descriptor can be useful, however, only the combination of the selected set of descriptors is able to model the target property. On the other hand, although, the explanation of the involved descriptors in the model, by definition of chemical and physical interpretations of the selected molecular descriptors is not a simple task but, it can provide a better understanding of the relationship between the structure of the compounds and their activity. The brief explanations of descriptors used are as follow:

The first group of descriptors used was *topological descriptors* including SEigZ (Eigenvalue sum from Z weighted distance matrix) and MSD (Mean square distance index (Balaban)) index. Topological index mathematically encode information regarding the structure of molecules, which have been depicted as graphs. The molecular graph was comprised of vertices, which correspond to atoms and edges of the bonds between the atoms. Often they are sensitive to size, shape, branching, cyclicity and, to a certain extent the electronic characteristics of molecules (Table 6, 1st row).

The second group of descriptors was the *highest eigenvalue of burden matrix* that defined as eigenvalues of *Burden matrix* (B). The B matrix was defined as, the number of atoms, bond order between two atoms or the electronegativity of the atoms. The information included on the electronic environment of the atoms in the matrix should be related to the matrix eigenvalues of the electronic distribution of the whole molecule. Among the eigenvalues obtained from B matrix, the highest eigenvalues was found to reflect the relevant aspects of the molecular structure, and was therefore useful for searching the similarity among molecules. By B eigenvalue decomposition, the best structure for the molecules, e.g. the number of atoms, the number of bonds and the electronic distributions of the whole molecule were found (Table 6, 2nd row). Thus, the B eigenvalues play a good role in predictions of designation of drugs.

The third group of descriptors used was *Autocorrelation of Topological Structure*. In general the 2D-autocorrelation descriptors explain how the values of certain functions, at intervals equal to the *lag*, are correlated. The 2D-autocorrelation descriptors represent the topological structure of the compounds and in nature are more complex than the classical topological descriptors. The descriptors used to develop the proposed model, lag was the topological distance *d* and the atomic properties were the functions correlated. The 2D-autocorrelation descriptors MATS1m (Moran autocorrelation of lag 2 / weighted by atomic masses) and MATS2m (2D-(Moran) autocorrelation weighted by atomic masses) are slightly different (Table 6, 3th row) and they describe how the considered property is distributed along the topological structure.

The forth descriptors used was *WHIM descriptor*. These molecular descriptors are based on the *statistical indices* calculated on the projection of atoms along principal axes. *WHIM* descriptors were built in such a way to capture the relevant molecular 3D information regarding molecular size, shape, and symmetry and atom distribution with respect to invariant reference frames. The algorithm consists of performing a *principal components analysis* on the centered *Cartesian coordinates* of molecule by using weighted covariance obtained from different weighting schemes for the atoms. G1u (1st component symmetry directional WHIM index /unweighted), G3m (3rd component symmetry directional WHIM index /weighted by atomic mass), E3m (3rd component accessibility directional WHIM index /weighted by atomic mass), E1v (1st component accessibility directional WHIM index /weighted by van Der Waals volumes) and Dv (D total accessibility index /weighted by van Der Waals volumes) were the global WHIM descriptors that represent the total size. Size descriptors can play a significance role independently in modeling of measured directions and produce the simpler models (Table 6, 6th row).

GETAWAY (Geometry, Topology, and Atom-Weights Assembly) descriptors [27],

encode the geometrical information obtained from the molecular matrix, the topological information obtained from the molecular graph and the information obtained from atomic weights which are specially designed with the aim of matching the 3D-molecular geometry. Some of these descriptors are HATS2u (Leverage-weighted autocorrelation of lag 2 / unweighted), HATS2m (Leverage-weighted autocorrelation of lag 2/weighted by atomic masses), HATS8p (Leverage-weighted autocorrelation of lag 8/weighted by atomic polarizabilities), $R7u^+$ (R maximal autocorrelation of lag 7 / unweighted), $R1m^+$ (R maximal autocorrelation of lag 1 / weighted by atomic masses), $R2v^+$ (R maximal autocorrelation of lag 2 / weighted by atomic van der Waals volumes), $R8e$ (R autocorrelation of lag 8 / weighted by atomic Sanderson electronegativities) and $R2p^+$ (R maximal autocorrelation of lag 2 / weighted by atomic polarizabilities) (Table 6, 7th row). These were derived from the elements h_{ij} of the Molecular Influence matrix (H), obtained through the values of atomic Cartesian coordinates. The diagonal elements of $H(h_{ii})$, called leverages, and were considered to represent the influence of each atom of the molecule in determining the whole shape of the molecule. For instance, the mantle atoms always have higher h_{ii} values than the atoms near the molecule center, and each off-diagonal element h_{ij} represents the degree of accessibility of the j^{th} atom to interactions with the i^{th} atom. The influence/distance matrix (R) involves a combination of the elements of H matrix with those of the geometric matrix (G).

The *geometrical variables* (DISPe; (d COMMA2 value/weighted by atomic Sanderson electronegativities)) are among the 3D elaborated descriptors obtained from moment expansions and did not require molecular superposition or alignment for the assignment of molecular similarity. These descriptors incorporate information about the magnitude of the displacement between the molecular centroid (center of mass) and the polarizability-field (center of charge) (Table 6, 4th row) [28].

Finally, the *3D-MoRSE descriptors* (Mor19m, Mor17v) were obtained through the molecular transformation employed in electron diffraction studies [29]. The electron diffraction did not directly yield atomic coordinates, but provides diffraction patterns from which the atomic coordinates are derived by mathematical transformations. These codes were defined in order to reflect the contribution of the property under the investigation at a prescribed scattering angle to a given atomic property. The 3D MoRSE code with its fixed-length representation of 3D molecular structure allowed the comparison of datasets of molecules of different size, with different number of atoms. Thus, the 3D MoRSE code had found numerous applications in establishing relationships between molecular structure and physical, chemical, or biological properties. So the presence of a MoRSE descriptor indicates (Table 6, 5th row) that the size of the inhibitor molecule has certain effect on the extent of the

interaction between the enzyme and molecule and the larger molecules with the larger substituent's groups or cyclo-substituents have the higher activity.

Conclusion

In this study, the PLS and GA-PLS were used in construction of the QSAR model and the resulted models were compared. It was found that application of GA prior to building the regression model; improve the prediction power of the model. The accuracy and power of predictive models was evaluated through the relative percentage of error, RMSCV, RMSEP and PRESS. The results revealed that the developed model is capable of reducing the time and expenses in synthesis of new molecules, and is useful in predicting the activity of antagonist of CXCR2 receptor. Thus, the model can be used to predict the activity of compounds not yet synthesis.

References

- [1] A. Burger, *Medicinal Chemistry*, Wiley, New York, 1970.
- [2] E. J. Ariens, *Drug Design*, Academic Press, New York, 1971.
- [3] D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, S. P. De Yong, J. Levi, J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics*, Elsevier, Amsterdam, 1997.
- [4] F. D. King, *Medicinal Chemistry Principles and Practice*, Royal Soc. Chem., Cambridge, 2002.
- [5] W. Camile, *The Practice of Medicinal Chemistry*, Academic Press, Oxford, 2003.
- [6] R. Kaliszan, *Quantitative Structure-Chromatographic Retention Relationships*, Wiley, New York, 1987.
- [7] G. Ioele, M. De Luca, F. Oliverio, G. Ragno, Prediction of photosensitivity of 1,4-dihydropyridine antihypertensives by quantitative structure-property relationship, *Talanta* **79** (2009) 1418–1424.
- [8] C. Sarbu, C. Onisor, M. Poša, S. Kevrešan, K. Kuhajda, Modeling and prediction (correction) of partition coefficients of bile acids and their derivatives by multivariate regression methods, *Talanta* **75** (2008) 651–657.
- [9] M. P. Freitas, J. A. Martins, Simple and highly predictive QSAR method: application to a series of (S)-N-[(1-ethyl-2-pyrrolidinyl)methyl]-6-methoxybenzamides, *Talanta* **67** (2005) 182–186.
- [10] K. Valko, Application of high-performance liquid chromatography based measurements of lipophilicity to model biological distribution, *J. Chromatog. A* **1037** (2004) 299–310.
- [11] K. Tang, T. Li, Comparison of different partial least squares methods in QSAR, *Anal. Chem. Acta.* **476** (2003) 75–92.
- [12] H. Kubinyi, Evolutionary Variable selection in regression and PLS analyses, *J. Chemometr.* **10** (1996) 119–133.
- [13] Z. Daren, QSPR studies of PCBs by the combination of genetic algorithm and PLS analysis, *J. Comp. Chem.* **25** (2001) 197–204.

- [14] T. Asadollahi, S. Dadfarnia, A. M. Haji Shabani, J. B. Ghasemi, M. Sarkhosh, QSAR models for CXCR2 receptor antagonists based on the genetic algorithm for data preprocessing prior to application of the PLS linear regression method and design of the new compounds using in silico virtual screening, *Molecules* **16** (2011) 1928–1955.
- [15] D. Carson, WO/2000/069861.
- [16] H. T. Y. Fahmy, S. A. F. Rostom, M. N. Saudi, J. K. Zjawiony, D. Robins, Synthesis and in-vitro evaluation of the anticancer activity of novel fluorinated thiazolo[4,5-d]pyrimidines, *J. Arch. Pharm. Pharm. Med. Chem.* **336** (2003) 216–225.
- [17] A. A. Bekhit, H. T. Y. Fahmy, S. A. F. Rostom, A. M. Baraka, Design and synthesis of some substituted 1H-pyrazolyl thiazolo[4,5-d]pyrimidines as anti-inflammatory - antimicrobial agents, *Eur. J. Med. Chem.* **38** (2003) 27–36.
- [18] E. Binnun, US/2007/0185139.
- [19] M. Y. Jang, Y. Lin, S. De Jonghe, J. Gao, B. Vanderhoydonck, M. Froeyen, J. Rozenski, J. Herman, T. Louat, K. Van Belle, M. Waer, P. Herdewijn, Discovery of 7-N-Iperazinythiazolo[5,4-d]pyrimidine analogues as a novel class of immuno-suppressive agents with in vivo biological activity, *J. Med. Chem.* **54** (2011) 655–668.
- [20] F. Hunt, C. Austin, R. Austin, R. Bonnert, P. Cage, J. Christie, M. Christie, C. Dixon, S. Hill, R. Jewell, I. Martin, D. Robinson, P. Willis, SAR studies on thiazolo[4,5-d]pyrimidine based CXCR2 antagonists involving a novel tandem displacement reaction, *Bioorg. Med. Chem. Lett.* **17** (2007) 2731–2734.
- [21] R. Todeschini, Milano Chemometrics, QSPR Group, <http://WWW.disat.unimib.it/chem>.
- [22] P. R. Duchowicz, A. G. Mercader, F. M. Fernández, E. A. Castro, Prediction of Aqueous Toxicity for Heterogeneous Phenol Derivatives by QSAR, *Chemom. Intell. Lab. Syst.* **90** (2008) 97–107.
- [23] V. Consonni, R. Todeschini, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, 2000.
- [24] A. Mauri, V. Consonni, M. Pavan, R. Todeschini, DRAGON software: An easy approach to molecular descriptor calculations, *MATCH Commun. Math. Comput. Chem.* **56** (2006) 237–248.
- [25] S. Wold, L. Eriksson, Statistical validation of QSAR results, in: S. Wold, L. Eriksson, S. Clementi (Eds.), *Chemometric Methods in Molecular Design*; Wiley-VCH, Weinheim, 1995, pp. 309–338.
- [26] A. Höskuldsson, *Prediction Methods in Science and Technology Basic Theory*, Thor Publishing, Ventura, 1996.
- [27] V. Consonni, R. Todeschini, GETAWAY descriptors: New molecular descriptors combining geometrical, topological and chemical information for physico-chemical properties modelling and drug design, in: H. D. Höltje, W. Sippl (Eds.), *Rational Approaches to Drug Design*, Prous Science, Barcelona, 2001, pp. 235–240.
- [28] B. D. Silverman, Three-dimensional moments of molecular property fields, *J. Chem. Inform. Comput. Sci.* **40** (2000) 1470–1476.
- [29] J. H. Schuur, P. Selzer, J. Gasteiger, The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure - spectra correlations and studies of biological activity, *J. Chem. Inform. Comput. Sci.* **36** (1996) 334–344.