(IJIRSE) International Journal of Innovative Research in Science & Engineering
ISSN (Online) 2347-3207

Quantitative structure-activity relationship investigation of pyridinone derivatives as anti-HIV prodrug

Mahmoud Saeedi Kelishami[1,*], Ghasem Ghasemi[2] , Shahab Shariati[3]

[1]Associate Professor, Department of Applied mathematics, Islamic Azad University Rasht Branch, Rasht, Iran
[2,3]Department of Chemistry, Rasht Branch, Islamic Azad University, Rasht, Iran
*Corresponding author: mskelishami@gmail.com

Abstract—**We have designed a lead HIV-1 standard transfer(ST) inhibitors strategically assembled on a pyridinone scaffold. Quantitative structure-activity relationship (QSAR) study has been done on pyridinone derivatives as anti-Hiv prodrug.Genetic algorithm (GA), Artificial neural network (ANN), Multiple linear regression (MLR), partial least squares (PLS), principal component regression (PCR), and least absolute shrinkage and selection operator (LASSO) were used to create QSAR models. The root mean square error of the calibration and R^2 using MLR method were obtained as 0.1434 and 0.95, respectively. The R^2 value using LASSO method were obtained as 0.98.The root mean square error of the calibration and R^2 using PLS method were obtained as 0.02 and 0.99, respectively. According to the obtained results, it was found that GA-PLS model is the most favorable method in comparison with other statistical methods and is suitable for use in QSAR models.**

Keywords-QSAR; pyridinone; Genetic algorithm; Artificial neural network

I. Introduction

The three viral enzymes (protease, reverse transcriptase and IN) are encoded within the HIV pol gene and translated as a polyprotein. IN (32-kDa) is released from the polyprotein by the HIV protease during maturation. The IN protein consists of three domains: N-terminal, core (or catalytic), and C-terminal domains [1]. The N-terminal domain enhances IN multimerization through zinc coordination (HHCC motif) and promotes concerted integration of the two viral cDNA ends together into a host cell chromosome. The C terminal domain is responsible for metal-independent, sequence-independent DNA binding (Fig. 1).

Each HIV-1 IN molecule contains a catalytic site within the core domain bearing three essential amino acids: Asp64, Asp116, and Glu152 (D, D-35-E motif). These acidic residues coordinate at least one and probably two divalent cations (Mg^{2+} or Mn^{2+}) that form a bridge with the DNA substrates [2]. Mutation of any of these residues abolishes IN enzymatic activities and viral replication .

In many situations we have a large number of inputs, often very correlated. The methods in this section produce a small number of linear combinations Z_m, m = 1, . . . ,M of the original inputs X_j , and the Z_m are then used in place of the X_j as inputs in the regression. The methods differ in how the linear combinations are constructed.

Principal component regression forms the derived input columns $Z_m = Xv_m$, and then regresses y on Z_1, Z_2, . . . , Z_M for some $M \leq p$. Since the z_m are orthogonal, this regression is just a sum of univariate regressions,Where p is number of steps.

As with ridge regression, principal components depend on the scaling of the inputs, so typically we first standardize them. Note that if M = p, we would just get back the usual least squares estimates, since the columns of Z = UD span the column space of X. For M < p we get a reduced regression. We see that principal components regression is very similar to ridge regression: both operate via the principal components of the input matrix. Ridge regression shrinks the coefficients of the principal components shrinking more depending on the size of the corresponding eigenvalue; principal components regression discards the p − M smallest eigenvalue components [3].

A neural network is a two-stage regression or classification model. This network applies both to regression or classification. However, these networks can handle multiple quantitative responses in a seamless fashion, so we will deal with the general case [4-6].

In some problems, the predictors belong to pre-defined groups; for example genes that belong to the same biological pathway, or collections of indicator (dummy) variables for representing the levels of a categorical predictor. In this situation it may be desirable to shrink and select the members of a group together. The grouped lasso is one way to achieve this. Suppose that the p predictors are divided into L groups, with p the number in group. For ease of notation, we use a matrix X to represent the predictors corresponding to the group, with corresponding coefficient vector. The grouped-lasso minimizes the convex criterion [7-12].

GAs perform multimodal search in complex landscapes and provide near optimal solutions for objective or fitness function of an optimization problem. GAs has also been applied to the domain of bioinformatics including that of drug design. The approach adopted is based on the use of genetic algorithms for evolving small molecules represented using a graphical structure composed of atoms as the vertices and the bonds as the edges. The task of is to determine the effectiveness of GAs in evolving a molecule that is similar to a target molecule. Thus ,knowledge about the target molecule is assumed, which may not be readily available in many situations.

Another approach for ligand design, that is based on the presence of a fixed pharmacophore and that uses the search capabilities of genetic algorithms, was studied by Goh and Foster , where the harmful protein human Rhinovirus strain14 was used as the target. This pioneering work assumed a fixed tree structure representation of the molecule on both sides of the pharmacophore [13-21].

II. Computational details

The 3D structures of the investigated Pyridinone derivatives were generated using the built optimum option of HyperChem software (version 6.0). Dragon program (version 5.5) was employed to calculate the molecular descriptors. All calculations were performed using Gaussian 03W program series. Geometry optimization of the compounds was carried out by B3LYP method employing 6–31G basis set (Table1).

In this study, the independent variables were molecular descriptors and the dependent variables were the actual half maximal inhibitory concentration (EC_{50}) values. More than 3026 theoretical descriptors were selected and calculated. Finally, Unscrambler program (version 9.7) was used for analysis of data and statistical methods.

For each compound in the training set, the correlation equation was derived with the same descriptors. Then, the obtained equation was used to predict log ($1/EC_{50}$) values for the compounds from the corresponding test sets. In the present work, the method of stepwise multiple linear regression (stepwise MLR) was used in order to select the most appropriate descriptors. Totally 3026 descriptors were generated. This value was too many to be fitted in models. So, it was necessary to reduce the number of descriptors through an objective feature selection which was performed in three steps. First, descriptors that had the same value for at least 70% of compounds within the dataset were removed. In next step, descriptors with correlation coefficients less than 0.4 with the dependent variable were regarded redundant and removed. Finally, since highly correlated descriptors provide approximately identical information, a pair wise correlation was performed. When their correlation coefficient exceeded 0.90, one of two descriptors was randomly removed.

III. Results and discussions

The structures of the pyridinone derivatives used in this study were shown in Table 1. Since, the variation in the chemical structure of the considered compounds is low, the selection of chemical descriptors, which can encode small variations between structures of molecules in data set, is very important. The four most significant descriptors which were selected are as follows. In this way, connectivity, 3-D morse index,auto correlation and WHIM descriptors are very informative , that can encode structural features of molecules. We have evaluated several layers ([3 1], [5 1], [7 1], [9 1], [11 1]) in GA and results are shown in Fig 2 and Tables 2 and 5. Multiple linear regression (MLR), partial least squares (PLS), Principal component regression (PCR), and least absolute shrinkage and selection operator (LASSO) were used to create QSAR models (Tables 2,3,4,6 and Fig. 3).

The efficiency of the QSAR model to predict log (EC50) value was also estimated using the internal cross-validation method. Considering the experimental error, the overall prediction of the log (1/EC50) values was quite satisfactory.

Two linear and non-linear variable selection methods were used to select the most significant descriptors (stepwise-MLR and GA). The selected descriptors through these methods were used to construct some linear and non-linear models using MLR, PCR, PLS and ANN methods. Based on the types of variable selection method and also the types of the feature mapping technique, these models can be shown as GA-PLS, GA-PCA, GA–MLR and GA–ANN .It revealed that the GA–PLS model was much better than other models. Statistical parameters of different constructed QSAR models are shown in Tables 2 and 5. Since the chemical variation of the considered compounds is low, the selection of chemical descriptors, which can encode small variations between structures of molecules in data set, is very important.

The eight most significant descriptors which were selected by GA-MLR and LASSO are as follows: X5AV, GATS2m, RDF030u, Mor07v, R3e+, B07 [N-Cl], E3S, B08 [O-F].

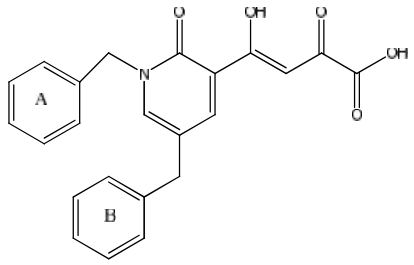
As can be seen from this table, atomic mass, electronegativity and topological distance were important descriptors in our study. In the present study, two linear and non-linear variable selection methods were used to select the most significant descriptors. The MLR, ANN and GA were used to construct a quantitative relation between activities of tricyclic pyridinone analogues and their calculated descriptors . We have evaluated several layers ([3 1], [5 1], [7 1], [9 1], [11 1]) in GA and results are shown in Fig 2 and Table 5.

IV. Conclusion

In our study, two linear and non-linear variable selection methods were used to select the most significant descriptors. The stepwise-MLR, MLR, PLS, PCR, GA and ANN were used to construct a quantitative relation between the activities of pyridinone derivatives and their calculated descriptors. MLR has been successfully used for finding a QSAR model for derivatives. It provides the best results in comparison with other studied methods. Our present attempt is to correlate the log ($1/EC_{50}$) with theoretically calculated molecular descriptors and has led to a relatively successful QSAR model that relates these derivatives. The results obtained from stepwise-MLR method, was suitable for drug design and classification. It revealed that the GA–PLS model was much better than other models.

A. Figures and Tables

Table1. The molecular structure pyridinone derivatives.



Aryl ring A Aryl ring B

Molecule	o	m	p	o	m	p
1	H	H	H	H	H	H
2	F	H	H	H	H	H
3	H	H	F	H	H	H
4	F	H	F	H	H	H
5	F	H	H	F	H	H
6	H	H	OMe	H	H	H
7	H	H	F	H	H	F

8	H	H	Me	H	H	F
9	H	Cl	F	H	H	H
10	F	H	F	F	H	F
11	H	F	H	H	H	H
12	H	Cl	H	H	Cl	H
13	H	Cl	F	H	Cl	F
14	F	H	H	H	H	F
15	H	Cl	F	H	Cl	H
16	H	H	Me	H	Cl	H
17	H	H	Me	F	H	H
18	F	Cl	F	H	H	H
19	H	H	F	H	Cl	F
20	2,6-di F	H	H	H	H	H
21	H	H	F	F	H	H
22	F	Cl	H	F	Cl	H
23	H	H	Cl	H	Cl	H
24	H	Cl	H	H	Cl	H
25	F	Cl	H	H	Me	H
26	F	H	F	F	H	H
27	H	Me	H	H	Cl	F
28	F	Cl	H	H	H	H
29	F	Cl	H	H	Cl	H
30	F	H	F	H	Cl	F

Table 2. The statistical parameters of different constructed QSAR models.

Method	RMSE c	R square
PLS	0.0221	0.99
GA – PLS[5 1]	0.0583	0.93
GA – PLS[6 1]	0.0914	0.84
PCR	0.1629	0.48
GA – PCR[5 1]	0.1920	0.28
GA – PCR[6 1]	0.1918	0.28
LASSO	-	0.98

Table 3.Experimental and predicted values of log (1/EC50) using PCR model.

PCR	GA –PCR[5 1]	GA –PCR[6 1]	observed
-2.647e-02	-2.281e-02	-2.219e-02	-0.3200
-2.231e-02	-2.792e-02	-2.754e-02	-0.0500
-4.063e-02	-4.838e-02	-4.712e-02	0
0.102	5.887e-02	6.034e-02	0.2200
6.384e-02	9.885e-02	9.906e-02	-0.1500
-9.783e-02	-0.151	-0.150	-0.2500
-3.687e-02	-5.918e-02	-5.847e-02	0.1000
-0.162	-9.634e-02	-9.562e-02	-0.2000
-6.793e-02	-4.986e-02	-4.971e-02	0.1000
0.383	7.676e-02	7.677e-02	0.5200
-5.381e-02	-2.616e-02	-2.510e-02	-0.3000
-0.290	-0.193	-0.194	-0.0400
-0.285	-0.217	-0.217	-0.0800
-4.786e-02	-4.550e-02	-4.550e-02	0.1500

-0.228	-0.221	-0.221	-0.3800
-0.229	-0.198	-0.200	-0.4000
-2.400e-02	1.685e-03	1.864e-03	0.1500
-1.659e-03	-6.762e-02	-6.659e-02	0.2200
-0.218	-0.128 8	-0.130	-0.1500
7.636e-02	-0.126	-0.126	-0.1500
7.166e-02	0.104	0.104	0.1000
-0.137	9.384e-02	9.322e-02	-0.3800
-0.257	-0.171	-0.172	-0.4000
-0.290	-0.193	-0.194	-0.1500
-3.911e-02	-0.244	-0.244	-0.1500
0.212	0.226	0.225	0.2200
-0.323	-0.278	-0.280	-0.3400
-4.611e-02	-5.770e-02	-5.710e-02	0.1500
-0.203	-0.140	-0.139	-0.1500
-4.237e-02	-0.158	-0.159	-0.1500

Table4.Experimental and predicted values of log (1/E C50) using PLS model.

PLS	GA -PLS [5 1]	GA -PLS [6 1]	Observed
-0.329	-0.286	-0.269	-0.32
-1.181e-02	4.298e-02	2.195e-02	-0.05
-2.471e-02	-3.232e-02	-4.302e-02	0.00
0.206	8.741e-02	5.710e-02	0.22
-0.157	-9.255e-02	6.081e-02	-0.15
-0.271	-0.275	-0.244	-0.25
0.122	9.245e-02	0.216	0.10
-0.187	-0.124	-0.143	-0.20
0.107	0.112	0.120	0.10
0.525	0.507	0.393	0.52
-0.278	-0.249	-0.227	-0.30
-0.101	-0.101	-0.128	-0.04
-9.549e-02	-0.100	-9.393e-02	-0.08
0.132	7.397e-02	0.104	0.15
-0.362	-0.363	-0.210	-0.38
-0.410	-0.477	-0.412	-0.40
0.158	3.248e-02	-5.678e-02	0.15
0.238	0.164	0.221	0.22
-0.142	-5.861e-02	-9.940e-02	-0.15
-0.149	-9.407e-02	-0.151	-0.15
9.832e-02	0.155	0.215	0.10
-0.371	-0.410	-0.377	-0.38
-0.401	-0.419	-0.372	-0.40
-9.162e-02	-0.112	-0.145	-0.15
-0.132	-0.141	-0.121	-0.15
0.216	0.232	0.160	0.22
-0.337	-0.359	-0.490	-0.34
0.118	0.161	7.295e-02	0.15
-0.154	-5.315e-02	-0.142	-0.15
-0.172	-0.175	-0.177	-0.15

Table 5. The statistical parameters of different constructed QSAR models with different layer..

Method	$RMSE_1$	$RMSE_2$	R Square
GA [2 1]	0.1748	0.1395	-
GA [4 1]	0.1961	0.1402	-
GA [6 1]	0.1798	0.1402	-
GA [8 1]	0.1730	0.1389	-
GA [3 1]	0.1658	0.1396	-

GA [5 1]	0.1849	0.1402	-
GA [7 1]	0.1769	0.1405	-
GA [9 1]	0.1678	0.1403	-
GA-ANN[5 1]	0.1730	0.1383	-
GA-ANN[6 1]	0.1689	0.1402	-
Jack-nife [5 1]	0.1446	0.1401	0.70
Jack-nife [6 1]	0.1351	0.1396	0.73

Table6. The results of GA-MLR and LASSO.

Descriptor	Meaning
X5AV	Average valence connectivity index chi-5
GATS2m	Geary autocorrelation –lag 2/weighted by atomic masses
RDF030u	Radial distribution function -3.0/weighted by atomic masses
Mor07v	3D-moRSE –signal 07/weighted by atomic van der walls volume
R3e+	R maximal autocorrelation of lag 3/weighted by atomic sanderson electronegativities
B07[N-Cl]	Presence absence of N-Cl at topological distance 07
E3S	3D-component accessibility directional WHIMindex/weighted by atomic electrotopological states
B08[O-F]	Presence absence of O-F at topological distance 08

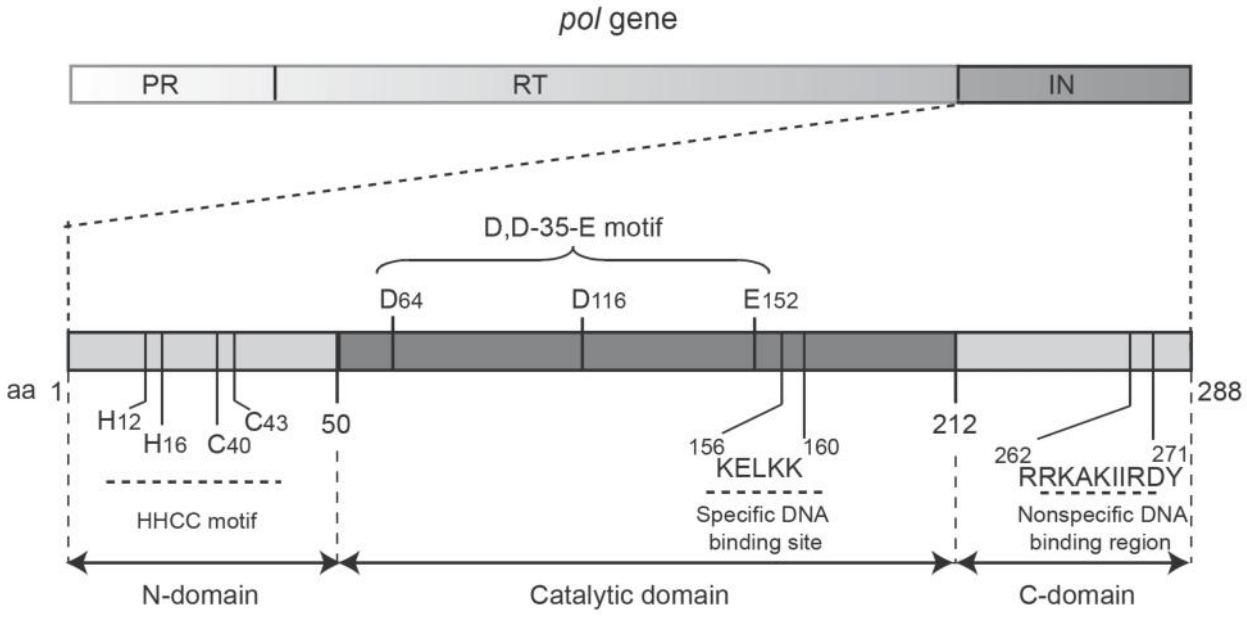


Fig. 1. Functional domains of HIV-1 integrase. IN: integrase, RT: HIV reverse transcriptase, PR: HIV protease.

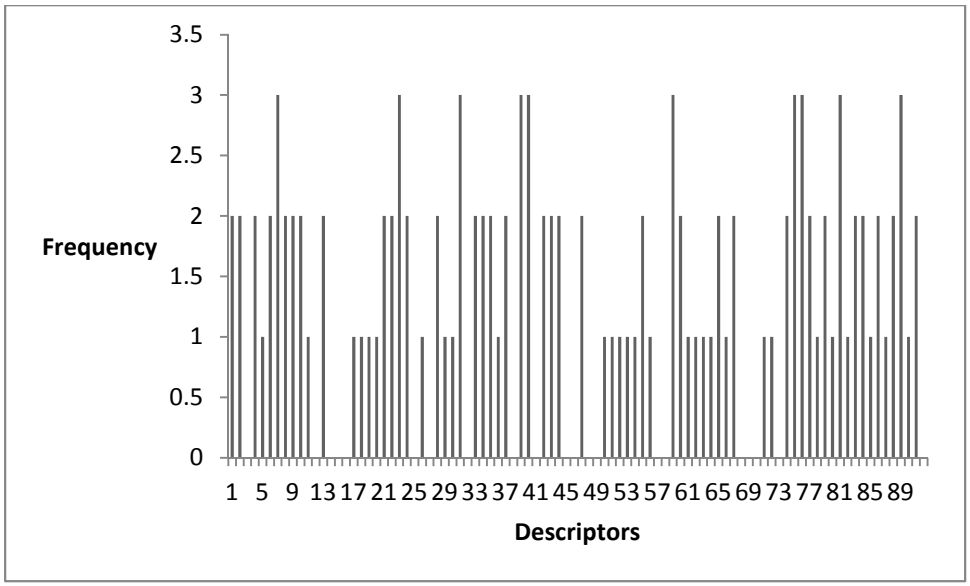


Figure 2. The results of Ga–ANN in gas phase.

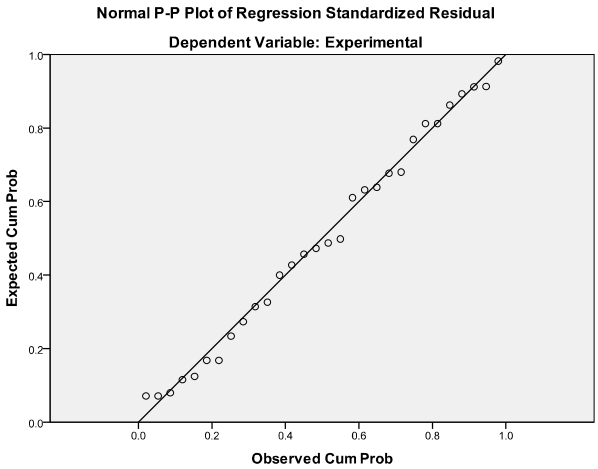


Fig. 3.Result of GA-MLR

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this article.

Acknowledgment

We thank the Research vice Presidency of Islamic Azad University, Rasht Branch for their encouragement, permission and financial support.

References

[1] T. K. Chiu, and D. R.Davies, "Structure and function of HIV-1 integrase," *Curr Top Med Chem.*, 4, pp. 965-977, 2004.
[2] C.Marchand, A. A .Johnson, E .Semenova, and Y. Pommier, "Mechanisms and inhibition of HIV, "*Drug. Discov. Today: Disease Mechanism.*,3,pp. 253-260, 2006.
[3] T.Hastie,R.Tibshirani, and J. Friedman. Data Mining,Inference,and prediction,springer ,second edition, 2008.
[4] C. Bishop. Neural Networks for Pattern Recognition, Clarendon Press, Oxford, 1995.
[5] T .Kohonen, S. Kaski, K. Lagus, A .Paatero, and A. Saarela, "Self organization of a massive document collection, " *Transactions on Neural Networks.*, 11(3), pp.574–585,2000.
[6] R .Neal. Bayesian Learning for Neural Networks. Springer, New York, 1996.
[7] J. Friedman, T .Hastie, and R .Tibshirani, "Sparse inverse covariance estimation with the graphical lasso, "*Biostatistics.*, 9, pp.432–441, 2008.
[8] K .Knight, and W .Fu "Asymptotics for Lasso-Type Estimators, " *Annals of Statistics*., 28(5), pp.1356–1378, 2000.
[9] N .Meinshausen "Relaxed lasso, " *Computational Statistics and Data Analysis*., 52(1), pp.374–393,2007.
[10] N .Meinshausen, and P. B¨uhlmann, "High-dimensional graphs and variable selection the lasso, "*Annals of Statistics.*, 34, pp.1436–1462, 2006.
[11] R .Tibshirani, and P .Wang , "Spatial smoothing and hot spot detection for cgh data using the fused lasso, "*Biostatistics* .,9, pp.18–29, 2007.
[12] H .Zou, T .Hastie, and R .Tibshirani "Degrees of freedom of the lasso, "*Annals of Statistics*., 35(5),pp. 2173–2192, 2007.
[13] D. Goldberg. Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, New York,1989.
[14] L .Davis. Handbook of Genetic Algorithms, New York:Van Nostrand Reinhold,1991.
[15] J.L.R Filho, and P.C. Treleavan, "Genetic algorithm programming environments, "*IEEE Comput*., pp.28-43,1994.
[16] S.K. Pal, and D .Bhandari, "Selection of optimum set of weights in a layered network using genetic algorithm,"*Inf. Sci*.,80, pp.213-234, 1994.
[17] S.K. Pal, D. Bhandari, and M.K .Kundu, "Genetic algorithms for optimal image enhancement, "*Pattern Recognit. Lett*.,15 , pp.261-271, 1994.
[18] S. Schulze-Kremer . Advances in Molecular Bioinformatics 258, IOS Press,1994.
[19] A.R .Leach, and V.J .Gillet. An Introduction to Chemoinformatic, Kluwer Academic Publishers, 2003.
[20] G.R .Raidl. Applications of Evolutionary Computing, Proceedings of the EvoBIO,2003.
[21] D.E .Clark . Evolutionary Algorithms in Molecular Design, Jown Wiley, 2000.