GENERATING QUANTITATIVE STRUCTURE ACTIVITY RELATIONSHIP MODEL USING DIFFERENT APPROACHES

DR. (MRS.) Prabha Mehta, Dr. Deepika Mahajan
HOD, Department of Chemistry ,Kamla Raja Girls College , ,Jiwaji university .Gwalior(M.P.)
Email – deepikamahajan1258@gmail.com , prabhamehta6@gmail.com

Abstract: *Quantitative structure activity relationship is indirect drug design technique; QSAR is direct relation between structure and activity. 1,4 dihyropyridine are a class of calcium channel blocker and used most frequently as antihypertensive and shows antagonist activity .Data sets are taken from available literature.*
The main objective of this paper we use different QSAR approaches to find out the suitability of model using a data set with known biological activity .Through different approaches generate a comparable QSAR model and find out the best approach .we have used different statistical approaches such as Principal component analysis ,Genetic Algorithm ,Multiple linear regression analysis etc.

Key Words: *QSAR, 1, 4-dihydropyridine, MLR, PLS.*

1. INTRODUCTION:

Medicinal chemist have systematically modified lead compound with the driving force of synthetic feasibility, experience and intuition. Using traditional techniques, it may take months to synthesize a new compound for biological testing. Rational molecular design strategies like Quantitative Structure Activity Relationship (QSAR) have become an important contributor to drug discovery process by reducing experimental research. The activity of a compound is dependent on its structure and chemical composition. This leads to the concept that the structural characteristics of a molecule are responsible for its properties [(1)]. Molecular discovery process is a cyclic process of three phases of design, synthesis, and test. Analysis of the results from first cycle provides information and knowledge that enables the next cycle to be initiated and further improvements to be achieved [(2)].

Rapid developments in chemistry, screening and automation have resulted in availability of large amounts of data for a typical drug discovery process. Integration of available biological and chemical information is important for success of drug discovery and development [(3)].

Complex chemical information has always been represented in a simple way by use of names, molecular weight, graphs, etc. Computational chemistry also includes mathematical methods implemented by computer for a wide range of applications, like reproduction of chemical processes, modeling of structures, prediction of properties, activities and reaction variables, etc. [(4)] Computational chemistry has developed into an important contributor to drug design process. The mathematical approach on Structure-Activity Relationship (SAR) for biological active compounds leads to the concept of Quantitative Structure-Activity Relationship (QSAR). QSAR models exist at the intersection of chemistry, statistics and biology. Quantitative structure activity relationship (QSAR) enables the investigators to establish a reliable structure-activity and structure-property relationships in form of a mathematical equation for '*in silico'* prediction of the activity of novel molecules prior to their synthesis [(5)]. **'**In silico' research in modern drug design required for enhancing the cost-effectiveness of molecular discovery, widely relies on building extensive QSAR models [(6)]. Quantitative Structure Activity Relationship modeling results in a quantitative correlation between chemical structure and biological activity [(7)]. Elaboration of such relations requires a set of analogous molecules and their activity or property. The QSAR approach attempts to identify and quantify the physicochemical properties of a drug and to see whether any of these properties have a relationship with the drug's biological activity. QSAR model is a mathematical equation which quantifies the relationship between activity and structure.
The 1, 4 dihyropyridine are derivatives, which are well known as calcium channel blockers, are used for treatment of cardiovascular diseases such as hypertension, angina pectoris and other spastic smooth muscle disorders. These drugs act directly on the voltage dependent calcium channels and block the flux of Ca+ to the cell cytoplasm .the pharmacological effects of these compounds are characterized by tissue selectivity.[(8)].

2. MATERIALS & METHOD:

QSAR is the study of the quantitative relationship between the experimental activity of a set of compounds and their physicochemical properties using statistical methods. The experimental information associated with biological

activity, which is used as dependent variables in building a QSAR model. In this study, all computational work (2D- and 3D-QSAR) was performed using V-life MDS QSAR plus software on a HP computer with Core2 Duo processor and a window 7 operating system.

2.1 2D-QSAR modeling and dataset

Compounds were sketched using 2D draw application . 2D descriptors were calculated which encoded different aspects of molecular structure and consists of electronic, thermodynamic, spatial, and structural descriptors, e.g., retention index (chi), atomic valence connectivity index (chiV), path count, chain path count, cluster, path cluster, element count, estate number, semi-empirical, molecular weight, molecular refractivity, log*P*, and topological index.

2.2. Selection of training and test set

The dataset of 40 molecules was divided into training set (28 compounds) and test set (12 compounds) by Random method[9] for multiple linear regression (MLR), principal component regression (PCR), and partial least squares (PLS) model using IC_{50} biological activity as dependent variable and various 2D descriptors as independent variables..

2.3. Regression analysis

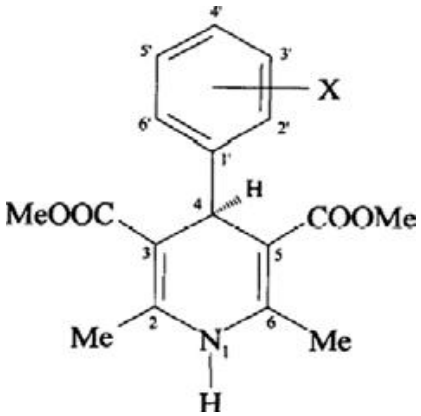
Dataset of 40 molecules was subjected to regression analysis using MLR and PLS as model building methods. QSAR models were generated using IC_{50} values as the dependent variable and various descriptors values as independent variables. The cross-correlation limit was set at 0.5, number of variables in the final equation at three in MLR and three in PLS, and term selection criteria as r^2, *F*-test 'in,' at 4 and 'out' at 3.99, r^2, and *F*-test. Variance cut-off was set at 0, scaling to auto-scaling, and number of random iterations to 10. Statistical measures were used for the evaluation of QSAR models were the number of compounds in regression *n*, regression coefficient r^2, number of descriptors in a model *k*, *F*-test (Fisher test value) for statistical significance *F*, cross-validated correlation coefficient q^2, predictive squared correlation coefficients pred_r^2, coefficient of correlation of predicted data set pred_r^2*se* and standard error (SE) of estimation r^2 *se* and q^2 *se*.

2.4. MLR analysis

MLR is a method used for modeling linear relationship between a dependent variable *Y* (IC_{50}) and independent variable *X* (2D descriptors). MLR is based on least squares: the model is fit such that sum-of-squares of differences of observed and a predicted value is minimized. MLR estimates values of regression coefficients (r^2) by applying least squares curve fitting method. The model creates a relationship in the form of a straight line (linear) that best approximates all the individual data points. In regression analysis, conditional mean of dependant variable (IC_{50}) *Y* depends on (descriptors) *X*. MLR analysis extends this idea to include more than one independent variable. Regression equation takes the form

$$Y = b1 * x1 + b2 * x2 + b3 * x3 + c$$

where *Y* is dependent variable, '*b*'s are regression coefficients for corresponding '*x*'s (independent variable), '*c*' is a regression constant or intercept[10,11] Data set of 1,4 dihyropyridine taken from literature.[12]



Structure of 1,4-dihydropyridine

Table 1: Experimental and predicted activities (-log(IC_{50}) of 1,4 dihyropyridine calcium channel antagonist

ANTAGONIST	X	EXPERIMENTAL	Log(IC50) MLR	Log(IC50) PLS
1.	3'Br	8.89	9.08	8.34
2.	2'CF3	8.82	9.61	9.63
3.	2'-Cl	8.66	9.30	8.74
4.	3'-NO2	8.40	8.62	7.96
5.	2'-C2H5	8.35	8.69	8.76
6,	2'-NO2	8.29	9.00	8.56
7.	2'-Me	8.22	8.98	8.96
8.	2'-Et	8.19	8.62	9.30
9.	2'-Br	8.12	9.28	9.14
10.	2'-CN	7.80	8.44	9.09
11	3'-Cl	7.80	9.04	8.52
12.	3'-F	7.68	9.15	8.6
13.	3'-CN	7.46	8.30	7.80
14.	3'-I	7.38	8.97	8.32
15.	2'-F	7.37	9.61	9.64
16.	2'-I	7.33	8.65	8.74
17.	2'-OMe	7.24	8.8	8.66
18.	3'-CF3	7.13	8.92	8.40
19.	3'-Me	6.96	8.74	8.50
20.	2'-OEt	6.96	8.33	8.61
21.	3'-OMe	6.72	8.14	8.23
22.	3'-NMe2	6.05	7.72	7.54
23.	3'-OH	6.00	7.29	8.25
24.	3'-NH2	5.70	7.13	8.00
25.	2'-NH2	4.40	7.39	8.29
26.	4'-F	6.89	8.89	8.19
27.	4'-Br	5.40	8.69	7.70
28.	4'-i	4.64	8.67	7.58
29.	4'-NO2	5.50	8.05	7.66
30.	4'-CN	4.00	7.99	7.99
31.	4'-Cl	5.96	8.78	8.03
32.	2'6'-Cl2	8.72	9.30	9.08
33.	2'-F 6'-Cl	8.12	8.96	9.46
34.	2'3'-Cl2	7.72	9.30	9.08
35,	2'-Cl 5'-NO2	7.52	8.29	8.60
36.	3'5'-Cl2	7.03	8.96	8.92
37.	2'-OH 5'-NO2	7.00	6.23	8.28
38.	2'5'-Me2	7.00	8.35	8.98
39.	2'4'-Cl2	6.40	9.14	8.90
40.	2'4'5'-(OMe)3	3.00	5.03	5.08

2.5 PCR method:

PCR is a data compression method based on the correlation among dependent and independent variables. PCR provides a method for finding structure in datasets. Its aim is to group correlated variables, replacing the original descriptors by new set called principal components (PCs). These PCs uncorrelated and are built as a simple linear combination of original variables. It rotates the data into a new set of axes such that first few axes reflect most of the variations within the data. First PC (PC_1) is defined in the direction of maximum variance of the whole dataset. Second PC (PC_2) is the direction that describes the maximum variance in orthogonal subspace to PC_1. Subsequent components are taken orthogonal to those previously chosen and describe maximum of remaining variance, by

plotting the data on new set of axes, it can spot major underlying structures automatically. Value of each point, when rotated to a given axis, is called the PC value. PCA selects a new set of axes for the data. These are selected in decreasing order of variance within the data. Purpose of principal component PCR is the estimation of values of a dependent variable on the basis of selected PCs of independent variables[13].

2.6 PLS regression method:

PLS analysis is a popular regression technique which can be used to relate one or more dependent variable (*Y*) to several independent (*X*) variables. PLS relates a matrix *Y* of dependent variables to a matrix *X* of molecular structure descriptors. PLS is useful in situations where the number of independent variables exceeds the number of observation, when *X* data contain colinearties or when *N* is less than 5 *M*, where *N* is number of compound and *M* is number of dependant variable. PLS creates orthogonal components using existing correlations between independent variables and corresponding outputs while also keeping most of the variance of independent variables. Main aim of PLS regression is to predict the activity (*Y*) from *X* and to describe their common structure(14).

3. RESULTS AND DISCUSSION:

Generation of 2D-QSAR models: Descriptors used in generation of 2D-QSAR models are given in Table 2. 2D-QSAR study of 1,4 dihyropyridine derivatives resulted in two QSAR models. Statistically significant QSAR models were selected for discussion.

Table 2-Molecular descriptors used in QSAR study

Constitutional Descriptor	molecular weight, H-bond donar ,H-bond acceptor, number of rot able bonds(RBN)
Topological Descriptor	Balaban J index, chi, molecular connectivity indices, distance topological.
Geometrical Descriptor	Wiener index
Quantum Chemical Descriptor	X dipole moment , y dipole moment , dipole moment , Total Dipole Moment.
Chemical Descriptor	S Log P, Molecular Volume, Polarizability
Structural Descriptor	Chi, Chiv

MLR MODEL- 1

Training test = 28 , test set =12

Log10 (IC_{50}) = Prediction [0.8661(±0.1442) + H-Donor Count[-0.1633(±0.0265) + Hosoya index[-0.0000(±0.0000) + Chiv3[0.0746(±0.0272)

Constant = 0.0997

n= 28 ,DF =23 , r^2 =0.834 , q^2 =0.590 , F test = 29.02 , r^2 Se =0.044 , q^2Se= 0.0699 , Pred_r^2 = 0.787 , Pred_r^2Se= 0.1230

MLR MODEL- 2

Training set = 28 , test set = 12

Log10(IC_{50}) = $SsNH_2$-index[-0.1478(±0.0000)] +Radius of Gyration [-0.0875(±0.0005)] + SKaverage Hydrophobicity [-12.3922(±0.5248)] + Chlorine count [0.0353(±0.0004)]

n=28, DF= 23 ,r^2 = 0.7522 ,q^2 =0.5801 , F test=17.4510 ,r^2Se= 0.0391 ,q^2Se= 0.0509 , Pred_r^2 = - 2.4075 , Pred_r^2Se=0.2719

Constant =1.4263

PLS MODEL

Training set =28 , test set = 12

$Log10(IC_{50})$ = Extrapolation(-4.86835) + Radius of Gyration (-0.07944) + Prediction (0.544345)

Constant=0 , n=28, DF= 25, r^2= 0.7061 , q^2= -0.0299 , F test = 30.028 , r^2 Se= 0.0572 , q^2Se=0.1071 ,Pred_r^2=0.2837 , Pred_r^2Se= 0.0750

In above QSAR models, r^2 is a correlation coefficient that has been multiplied by 100 gives explained variance in biological activity. . *F* value reflects ratio of variance explained by models and variance due to error in regression. High *F* value indicates that model is statistically significant. Low SE of estimation indicted by r^2 se and q^2 se suggested that models are statistically significant. . Among these two models, MLR has come out with very good results as compare to PLS model. Results of MLR analysis showed very good predictive ability as indicted by r^2, q^2, *F*-test, and pred_r^2 values.

4. CONCLUSIONS:

The QSAR models developed in this study through MLR regression method would be more stable as compare to PLS method .

REFERENCES:

1. Shamsipur, Mojtaba et al.,Highly Correlating Distance/Connectivity -based Topological Indices: QSPR studies of Alkanes, Bull.Korean Chem. Soc., 25(2), 2004, 253-259.
2. R Leach Andrew and I Gillet Valerie , An Introduction to Cheminformatics; Netherlands: Springer, 75,2007.
3. Cho, Sung Jin et al. , ADAAPT : Amgen's data access, analysis, and prediction tools , Journal of Computer Aided Molecular Design, 20, 2006, 249-261.
4. Ruiz, Irene Luque et al. , Improving the development of QSAR Prediction Models with the use of Approximate Similiarity Approach, Engineering Letters , 16(1) , 2008 .
5. S Prabhakar Yenamandra and K Manish Gupta. , Chemical Structure Indices in " In Silico" Molecular Design, Science Pharma,; 76 , 2008, 101-132.
6. S Cherkasov, 'Inductive' Descriptors: 10 Successful Years in QSAR, Current Computer-Aided Drug Design, 1, 2005, 21-42.
7. V Ravichandran et al., Prediction of HIV-1 Protease Inhibitory Activity of (4-Hydroxy-6-Phenyl-2-Oxo-2H-Pyran-3-yl) Thiomethanes: QSAR Study; Current Trends in Biotechnology and Pharmacy , 3(2), April 2009, 149-154.
8. R Fossheim, j Med Chem ,29, 1986,305.
9. 9. BD Hudson , RM Hyde ,E Ra hr ,J Wood Parameter based methods for compounds selection from chemical databases. Quant Struct Act Relat ,15, 1996,285-289
10. C Croux , K Joossens ,Influence of observations on the misclassification probability in quadratic discriminant analysis , J Multivar Anal , 96,2005,348-403.
11. J Devillers ,Neuronal network in QSAR and drug design. Academic Press, 1996 London.
12. Si Hong Zong et.al., QSAR study of 1,4 dihyropyridine calcium channel antagonists based on gene expression programming,bioorganic & medicinal chemistry ,14,2006,4834-4841.
13. JP Doucet , F Barbault , H Xia, APanaye, B Fan, Nonlinear SVM approaches to QSPR/QSAR studies and drug design, Curr Comput Aided Drug Des ,, 3,2007,263-289.
14. CJ Huberty , Applied discriminant analysis, Willey, New York, (1994).