RESEARCH PAPER

Accepted 24 May 2004
Revised 10 February 2004
Received 22 September 2003
Indian J. Pharm. Sci., 2004, 66(4): 396-402

VALSTAT: Validation Program for Quantitative Structure Activity Relationship Studies

A. K. GUPTA, M. AROCKIA BABU AND S. G. KASKHEDIKAR*
Department of Pharmacy, Shri Govindram Seksaria Institute of Technology and Science, 23, Park Road, Indore-452003.

VALSTAT program has been developed using C++ in order to perform quantitative structure activity relationship analysis using stepwise multiple regression analysis. It is attempted to design and write program to select the better model for regression analysis and various quantitative structure activity relationship analysis validation methods such as cross-validation, boot strapping, randomization test, outliers, which are not available in the PC-based quantitative structure activity relationship analysis softwares. The program reproducibility and accuracy was validated using few reported series of cyclooxygenase-2 inhibitors.

Softwares such as CoMFA, CoMSIA, APEX-3D and Ceries-2 are available in the market to perform quantitative structure activity relationship (QSAR) analysis, which require silicon graphics[1-3]. This is not affordable and tangible task for all to use silicon graphics, due to high costs. To conquer thirst, number of research groups independently has developed statistical programs[4-6] which compute statistical parameters such as correlation coefficient (r), standard deviation (std) and F-test for statistical significance (F). Additional special statistical parameters such as cross- validation squared correlation coefficient (q^2) randomization test (chance) and bootstrapping squared correlation coefficient (r^2_{bs}) should be incorporated for selection and validation of better QSAR models. To fulfill this requirement, the software program that can run on PC has been developed using C++ language. Recently, Garg *et al.* have reported a comparative QSAR study of 27 series of COX-2 inhibitors using Hansch approach[7]. The program reproducibility and accuracy were validated using only two series such as terphenyl, 4,5-diaryl imidazole derivatives because the obtained models contain 4 and 2 variables in the final equations. Additionally, we have developed Fujita ban matrix for both the series in which 10 and 12 independent variables for terphenyl and 4,5-diaryl imidazole derivatives respectively appear in the final equation. It was found that the developed program gives excellent reproducibility and accuracy for Hansch analysis and it is applicable to Fujita ban matrix as well.

*For correspondence
E-mail: arunkg_73@hotmail.com

MATERIALS AND METHODS

A validation study for the two series of COX-2 inhibitors such as terphenyl, 4,5-diaryl imidazole derivatives[7-9] was carried out (Table 1 and 2 and fig. 1 and 2). In addition, we have developed Fujita ban matrix (Table 3 and 4) and performed analysis in order to validate the reproducibility and accuracy of the developed computational program, which contains large number of variables in the model.

VALSTAT was developed using c++ language[10-12]. The program has provision of stepwise multiple regression analysis with linear and parabolic relationship to generate the QSAR model, in addition to advanced statistical validation procedure to select best quantitative structure activity relationship from high populated QSAR models. Before executing the program, data must be entered in specific data files such as independent parameter data in DATA_IND.TXT file, dependent parameter data in DATA_DEP.TXT file and the name of independent parameter in DATA_NAME_IND.TXT.

The VALSTAT is interactive program, when initiated, asks for declaration in the format of (y/n). If user enters 'n', then program is terminated automatically. The main options

that appear on the screen based on the user's interest are, (i) have you entered data in specific files (y/n), (ii) do you want test and training set (y/n), (iii) do you want intercorrelation matrix (y/n), (iv) are you going for Simple Linear Regression Analysis (y/n), (v) are you going for Stepwise Linear Regression Analysis (y/n), (vi) enter the limit of r^2 for output of the data (0-1), (vii) enter correlation limit within the parameters (0-0.99), (viii) are you going for Non

TABLE 1: COX-II INHIBITORY ACTIVITY DATA AND PHYSICOCHEMICAL PROPERTIES FOR TERPHENYL DERIVATIVES

Substituent		-log IC_{50}			[a]ClogP	[b]Iy	[c]σ+X	[d]CMR
X	Y	Obs.	Cal.†	Cal.‡				
4-F	CH_3	7.85	8.05	8.05	4.67	0	-0.07	9.09
4-F	NH_2	8.40	8.41	8.41	4.50	1	-0.07	9.00
3-Cl-4-F	CH_3	8.00	8.01	8.00	5.38	0	0.30	9.59
3-Cl-4-F	NH_2	8.70	8.38	8.38	5.22	1	0.30	9.49
3-CH_3-4-F	CH_3	8.30	8.17	8.17	5.17	0	-0.14	9.56
3-CH_3-4-F	NH_2	8.70	8.54	8.54	5.00	1	-0.14	9.46
3-F-4-OCH_3	CH_3	7.68	7.53	7.53	4.52	0	-0.44	9.71
3-F-4-OCH_3	NH_2	7.89	7.91	7.91	4.37	1	-0.44	9.62
3-Cl-4-OCH_3	CH_3	7.72	7.60	7.60	5.06	0	-0.41	10.19
3-Cl-4-OCH_3	NH_2	7.89	7.98	7.99	4.91	1	-0.41	10.09
3,5-Cl_2-4-OCH_3	NH_2	7.68	7.74	7.74	5.43	1	-0.04	10.58
3-CH_3-4-OCH_3	CH_3	7.89	7.88	7.88	4.96	0	-0.85	10.16
3-CH_3-4-OCH_3	NH_2	8.30	8.24	8.25	4.79	1	-0.85	10.06
3,4-$(OCH_3)_2$	CH_3	6.47	6.68	6.68	4.17	0	-0.66	10.31
3,4-$(OCH_3)_2$	NH_2	7.19	7.07	7.07	4.02	1	-0.66	10.22
4-CH_3	CH_3	8.16	8.17	8.17	5.03	0	-0.31	9.54
4-CH_3	NH_2	8.40	8.53	8.53	4.86	1	-0.31	9.45
3-Cl-4-CH_3	CH_3	7.89	8.13	8.14	5.74	0	0.06	10.03
3-Cl-4-CH_3	NH_2	8.52	8.50	8.49	5.57	1	0.06	9.94
3,4$(CH_3)_2$	CH_3	8.30	8.60	8.60	5.31	1	-0.38	9.91
3-CH_3-4-Cl	CH_3	8.22	8.15	8.15	5.74	0	0.04	10.03
3-CH_3-4-Cl	NH_2	8.52	8.52	8.51	5.57	1	0.04	9.94
3-Cl-4-$N(CH_3)_2$	CH_3	8.10	7.91	7.91	5.31	0	-1.33	10.87
3-Cl-4-$N(CH_3)_2$	NH_2	8.22	8.29	8.29	5.15	1	-1.33	10.77

† Calculated value taken from literature, ‡ Value calculated by VALSTAT program, Obs. -Observed Value, Cal. -Calculated Value, [a]Clogp- Calculated Partition Co-efficient in octanol/water, [b]Iy- Indicative Variable for Y substituents, [c]s+X- Hemmett Electronic Parameter for X substituents and [d]CMP- Calculated Molar Refractivity.

Linear Stepwise Regression Analysis (y/n), (ix) enter number of independent variables for Multiple Linear Regression Analysis (1-24), (x) do you want correlation matrix for Multiple Linear Regression parameter (y/n) and (xi) are you going for validation of model (y/n).

Here it asks for training/test set (y/n). If answer is 'n', then program progresses to next step automatically. Now question appears on screen for correlation matrix (y/n), if reply is 'y' then program starts for finding of inter correlation within the independent parameters and the result of inter

TABLE 2: COX-II INHIBITORY ACTIVITY DATA AND PHYSICOCHEMICAL PROPERTIES FOR 4,5-DIARYL IMIDAZOLE DERIVATIVES.

Substituent X	-log IC_{50}			[a]ClogP	[b]MgVol	[c]BI_{x2}
	Obs.	Cal.†	Cal.‡			
H	6.16	6.42	6.42	3.09	2.37	1.00
3-F	6.62	6.55	6.54	3.24	2.39	1.00
4-F	6.72	6.55	6.54	3.24	2.39	1.00
2-Cl	5.89	5.89	5.90	3.56	2.49	1.80
3-Cl	7.10	6.88	6.89	3.81	2.49	1.00
2-CH_3	5.75	5.64	5.64	3.29	2.51	1.52
3-CH_3	6.22	6.48	6.48	3.59	2.51	1.00
4-CH_3	6.19	6.48	6.48	3.59	2.51	1.00
3-OCH_3	5.62	5.50	5.51	3.10	2.57	1.00
4-OCH_3	5.54	5.50	5.51	3.10	2.57	1.00
3,4-Cl_2	7.40	7.17	7.17	4.40	2.61	1.00
2,4-F_2	6.26	6.40	6.41	3.38	2.40	1.35
3,4-F_2	6.80	6.57	6.59	3.31	2.40	1.00
3-Cl-4-CH_3	6.64	6.85	6.85	4.24	2.63	1.00
2-CH_3-3-F	5.77	5.77	5.76	3.44	2.53	1.52

† Calculated value taken from literature, ‡ Value calculated by VALSTAT program, Obs.-Observed Value, Cal.-Calculated Value, [a]Clogp- Calculated Partition Co-efficient in octanol/water, [b]Mgvol- Molar Volume and cBIX2- Verloop's Sterimol Parameter for 2-X substituents.

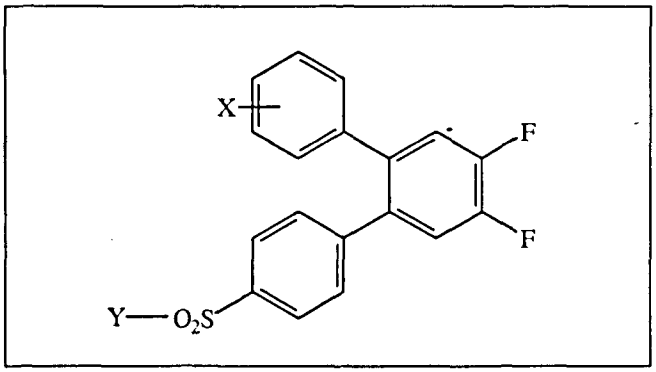


Fig. 1: Parent structure for terphenyl derivatives.

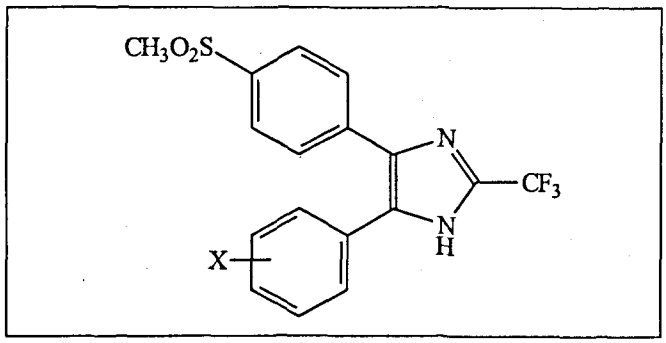
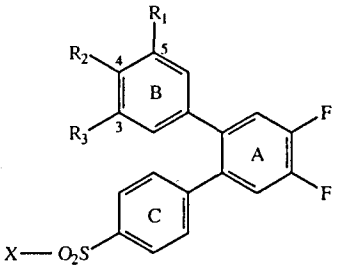


Fig. 2: Parent structure for 4,5 diaryl imidazole derivatives.

TABLE 3: FUJITA BAN MATRIX FOR TERPHENYL DERIVATIVES



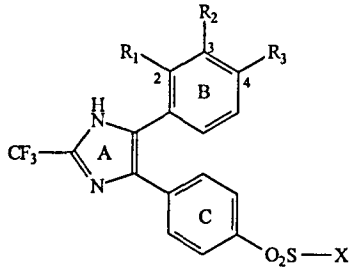
-logIC$_{50}$	[a]μ	X	R_1				R_2				R_3
		NH_2	Cl	CH_3	F	OCH_3	OCH_3	CH_3	Cl	NCH_3	Cl
7.85	1	0	0	0	0	0	0	0	0	0	0
8.40	1	1	0	0	0	0	0	0	0	0	0
8.00	1	0	1	0	0	0	0	0	0	0	0
8.70	1	1	1	0	0	0	0	0	0	0	0
8.30	1	0	0	1	0	0	0	0	0	0	0
8.70	1	1	0	1	0	0	0	0	0	0	0
7.68	1	0	0	0	1	0	1	0	0	0	0
7.89	1	1	0	0	1	0	1	0	0	0	0
7.72	1	0	1	0	0	0	1	0	0	0	0
7.89	1	1	1	0	0	0	1	0	0	0	0
7.68	1	1	1	0	0	0	1	0	0	0	1
7.89	1	0	0	1	0	0	1	0	0	0	0
8.30	1	1	0	1	0	0	1	0	0	0	0
6.47	1	0	0	0	0	1	1	0	0	0	0
7.19	1	1	0	0	0	1	1	0	0	0	0
8.16	1	0	0	0	0	0	0	1	0	0	0
8.40	1	1	0	0	0	0	0	1	0	0	0
7.89	1	0	1	0	0	0	0	1	0	0	0
8.52	1	1	1	0	0	0	0	1	0	0	0
7.64	1	0	0	1	0	0	0	1	0	0	0
8.30	1	1	0	1	0	0	0	1	0	0	0
8.22	1	0	0	1	0	0	0	0	1	0	0
8.52	1	1	0	1	0	0	0	0	1	0	0
8.10	1	0	1	0	0	0	0	0	0	1	0
8.22	1	1	1	0	0	0	0	0	0	1	0

[a]Activity for parent structure

correlation gets stored in the RESULT.TXT file in matrix form.

In next step, software inquires for stepwise regression. After entering the required information, the program is executed for stepwise regression by finding the contribution of all the independent variables towards the dependent variable with inter correlation limit within the independent parameters. The resultant equations with number of samples in regression (n), square of correlation coefficient (r^2), variance, standard deviation (std), and F-test for statistical significance (F) get transferred to RESULT.TXT.

TABLE 4: FUJITA BAN MATRIX FOR 4,5- DIARYL IMIDAZOLE DERIVATIVES



-logIC$_{50}$	$^a\mu$	R_1			R_2				R_3				X
		F	Cl	CH_3	F	Cl	CH_3	OCH_3	F	Cl	CH_3	OCH_3	NH_2
6.16	1	0	0	0	0	0	0	0	0	0	0	0	0
5.52	1	1	0	0	0	0	0	0	0	0	0	0	0
6.62	1	0	0	0	1	0	0	0	0	0	0	0	0
6.72	1	0	0	0	0	0	0	0	1	0	0	0	0
5.89	1	0	1	0	0	0	0	0	0	0	0	0	0
7.10	1	0	0	0	0	1	0	0	0	0	0	0	0
6.43	1	0	0	0	0	0	0	0	0	1	0	0	0
5.75	1	0	0	1	0	0	0	0	0	0	0	0	0
6.22	1	0	0	0	0	0	1	0	0	0	0	0	0
6.19	1	0	0	0	0	0	0	0	0	0	1	0	0
5.62	1	0	0	0	0	0	0	1	0	0	0	0	0
5.54	1	0	0	0	0	0	0	0	0	0	0	1	0
7.40	1	0	0	0	0	1	0	0	0	1	0	0	0
6.26	1	1	0	0	0	0	0	0	1	0	0	0	0
6.80	1	0	0	0	1	0	0	0	1	0	0	0	0
6.64	1	0	0	0	0	1	0	0	0	0	1	0	0
5.77	1	0	0	1	1	0	0	0	0	0	0	0	0
7.00	1	0	0	0	0	0	0	0	1	0	0	0	1
6.82	1	0	0	0	0	0	1	0	0	0	0	0	1

[a]Activity for parent structure.

It is followed by stepwise regression program for multiple regression (MLR), in MLR, software used the independent parameters that are having maximum contribution towards the equations with specified limit for inter correlation of independent parameters, Owing this step the program searches for outliers in the model. The out comes of MLR such as equation, n, r^2, variance, std, F-value, calculated biological activity of individual compounds, Z-score value and the outliers compound number are stored into RESULT.TXT file.

In next step MLR equation, goes for advanced validations like bootstrapping square of correlation coefficient (r^2_{bs}), randomization test (chance) and cross validation leave one out method (q^2). The validation data can be viewed using any editor or by the DOS command TYPE for file RESULT.TXT. Finally, software may be used for prediction of activity of test set as well as calculation of predictive square correlation coefficient (r^2_{pred}).

RESULTS AND DISSCUSSION

Many physicochemical properties of the chemical entities of COX-2 inhibitors viz., terphenyl, 4,5-diaryl imidazole have been explored. The Eqns. 1 and 2 were generated using VALSTAT program and Eqns. 1* and 2* have been reported in the literature.

The Eqns. 1 and 1* were obtained using stepwise multiple regression analysis for terphenyl derivatives employing Hansch method. It can be observed that the values of statistical equation obtained using VALSTAT program are more or less similar to that of reported values. However, the marginal difference in r^2 values is due to rounding up of the data values in the literature.

$-\log IC_{50}=1.113\ (\pm0.205)*ClogP+0.452\ (\pm0.143)*Iy-0.825\ (\pm0.274)*\sigma^{+}X-1.063\ (\pm0.242)\ *CMR+\ 12.455\ (\pm1.925)$...(Eqn. 1), n=24, r^2=0.907, variance=0.027, std=0.164, F=46.568, r^2_{bs}=0.905, chance<0.01, q^2=0.842.

$-\log IC_{50}=1.12\ (\pm0.20)*ClogP+0.45\ (\pm0.14)*Iy-0.83\ (\pm0.27)*\sigma^{+}X-1.06\ (\pm0.24)*CMR\ +12.91\ (\pm1.91)$ ... (Eqn. 1*), n=24, r^2=0.909, std=0.164, q^2=0.845

The Eqns.2 and 2* were generated for 4,5-diaryl imidazole derivatives by stepwise multiple regression analysis. It was found that the developed equation (Eqn. 2) using VALSTAT program almost coincide excellently with the reported equation.

$-\log IC_{50}=1.421\ (\pm0.394)*ClogP\ -4.615\ (\pm1.840)*MgVol\ -0.792\ (\pm0.478)*BI_{x2}+3.752\ (\pm3.969)$..(Eqn. 2), n=15, r^2=0.886, variance=0.045, std=0.212, F=28.415, r^2_{bs}=0.899, chance<0.01, q^2=0.798.

$-\log IC_{50}=1.45\ (\pm0.40)*ClogP\ -4.67\ (\pm1.84)*MgVol\ -0.78\ (\pm0.48)*BI_{x2}+3.79\ (\pm3.94)$, n=15, r^2=0.885, std =0.212,q^2=0.798...(Eqn. 2*).

In addition to Hansch approach, both the series were subjected to Fujita-Ban QSAR approach using multiple linear regression analysis.

In both the series, the variation in the substituents at each substituted position is small, a Fujita-Ban approach has been adopted to estimate the denovo contribution of substituents to the activity of the molecules. Thus following equations (Eqn. 3 and 4) were obtained for both series,

$-\log IC_{50}=-0.519\ (\pm0.827)*5Cl+0.567\ (\pm0.288)*NH_2\ -0.534\ (\pm0.477)*4OCH_3\ -0.455\ (\pm0.407)\ *4CH_3\ -0.148\ (\pm0.627)*4Cl\ -0.290\ (\pm0.627)*4N\ (CH_3)_3+0.442\ (\pm0.477)*3Cl+0.510\ (\pm0.477)\ *3CH_3+0.311\ (\pm0.748)*3F\ -0.644\ (\pm0.748)*3OCH_3+7.724\ (\pm0.432)$, n=25, r^2=0.814, std=0.326, F= 6.107...(Eqn. 3)

$-\log IC_{50}=-0.580\ (\pm0.435)*2F\ -0.361\ (\pm0.581)\ *2Cl\ -0.589\ (\pm0.446)*2CH_3\ +0.196\ (\pm0.374)\ *3F+0.779\ (\pm0.378)*3Cl\ +0.065\ (\pm0.496)*3CH_3\ -0.631(\pm0.581)*3OCH_3\ +0.438(\pm0.370)*4F+0.274\ (\pm0.452)*4Cl\ -0.226\ (\ \pm0.452)*4CH_3\ -0.711(\pm0.581)*4OCH_3\ +0.407\ (\pm0.453)*NH_2\ +\ 6.251\ (\pm0.321)$...(Eqn. 4), n=19, r^2=0.963, std=0.188, F=12.863.

The QSAR models (Eqn. 3 and 4) were obtained for terphenyl and 4,5-diaryl imidazole derivatives through Fujita-Ban approach, accounted for 81.4% and 96.3% of the variance in the biological activity with good statistical significance i.e. above 99% with $F_{(10,14)}$=6.107 ($F_{10,14\ \alpha\ 0.01}$=3.94) and $F_{(12,6)}$=12.863 ($F_{12,6\ \alpha\ 0.01}$=7.72), respectively.

In case of Fujita-Ban analysis for COX-2 inhibition of terphenyl derivatives, it can be inferred from eqn. 3 that substitution of non-polar groups such as F, Cl and OCH_3 at R_3 position of 'B' phenyl ring are more favorable. For 4,5-diaryl imidazole derivatives, it may be concluded from eqn. 4 that the groups like F, Cl in 3rd and 4th position of 'B' phenyl ring are favouable, whereas substitutions of same groups at 2nd position do not favour for the COX-2 inhibition. In both the series, the substitution of sulfonamide group is more favorable for COX-2 inhibition than methyl sulphonyl moiety at 4th position of 'C' phenyl ring substitution.

Our aim was to develop computational program for QSAR study and this has been successfully applied for generation and validation of QSAR models of two series of COX-2 inhibitors. The VALSTAT program will rationally give the way to develop new molecules by selecting better model from the population of the different QSAR models.

ACKNOWLEDGMENTS

The authors thank the Director, SGSITS, Indore, for providing facility to this research work and authors AKG and MAB are grateful to CSIR, New Delhi, for senior research fellowship.

REFERENCES

1. SYBYL 6.8, Molecular Modeling Software, Tripos Associates Inc., 1699 S. Hanley Road. St. Louis MO 63144, USA, 2002.
2. APEX-3D User Guide, Molecular Simulation Inc., San Diego, USA, 1995.
3. Cerius² Version 3.5 Biosym/Molecular Simulations Inc., 9685 Scranton Road, California, USA 92121-3752, 1995.
4. C-QSAR program, BioByte Corp., Claremont, CA, USA, 91711, 1999.
5. Systat V10.0, India soft company, Pune, 2002.
6. Chakravarti, S.K., Ajmani, S. and Chaturvedi, S.C., **Indian J. Pharm. Sci.**, 1998, 60, 371.
7. Garg, R., Kurup, A., Mekapati, S.B. and Hansch, C., **Chem. Rev.**, 2003, 103, 703.
8. Li, J.J., Norton, M.B., Reinhard, E.J., Anderson, G.D., Gregroy, S.A., Isakson, P.C., Koboldt, C.M., Masferrer, J.L., Perkins, W.E., Seibert, K., Zweifel, B.S. and Reitz, D.B., **J. Med. Chem.**, 1996, 39, 1846.
9. Barta, T.E., Steadley, M.A., Collins, P.W. and Weiner, R.M., **Bioorg. Med. Chem. Lett.**, 1998, 8, 3443.
10. Johnson, R.A., In; Probability and Statistics for Engineering, 5th Edn., Prentice Hall of India Pvt. Ltd., New Delhi, 1996, 330.
11. Freund, J.E., In; Mathematical Statistics, 5th Edn., Prentice Hall of India Pvt. Ltd., New Delhi, 1992, 494.
12. Daniel, W.W., In; Biostatistics A Foundation for Analysis in Health Sciences, 7th Edn., John Wiley and Sons, Inc., New York, 2000, 400.