A Machine Learning Access for Selectionof Influential Variables of Several ITK Inhibitors using Regression Research

Rama Devi Chalasani, Radhika Y

ABSTRACT--- Introduction: Interleukin-2 inducible T-cell kinase (ITK) is a tyrosine kinase expressed in T-cells, NK cells and mast cells. Selective ITK inhibitors act as an immunosuppressive and anti-inflammatory agent reduces lung inflammation, eosinophil infiltration, and mucous production in response to induced allergic asthma. Methodology: A dataset of 142 ITK inhibitors as dependent variables with 32 properties of compounds as explanatory variables were studied for their multicollinearity prior multivariate regression analysis. After data normalization, an inter-correlation cutoff value of 0.75 resulted in 15 variables and regression analysis resulted in 0.641 r2 and 0.598 adjusted r2 with RMSE 0.634 respectively. As the statistical parameters are within the limits, outlying data was investigated. Results: The standardized residual analysis resulted in nine data points and a new regression model is attempted with n=133 and p=15 reported improves statistics. Further, stepwise and stepwise AIC regression followed by variance inflation factor analyzed on the dataset revealed only 7 variables as important in defining inhibitory activity of ITK. Permutation and combinations of 7 variables resulted in r2 value >0.6 for 5, 6 and 7 variables. Hence, to select the best model, FIT criterion was employed where a 5-variable model was judged as best model. Conclusion: Finally, it has been emphasized that increase in HOMO, H-Bond Donors and shape index with a concomitant decrease in number of phenyl groups and LUMO parameter favors ITK inhibition.

Keywords —regression, multicollinearity, FIT Kubinyi function, outliers, ITK

I. INTRODUCTION

Linear regression is the well-known widely used algorithm in statistics and machine learning. This machine learning regression algorithm can be trained to predict real numbered outputs. Regression analysis on a dataset can be either linear, quadratic, polynomial, non-linear, etc. The regression hypothesis is a function based on relationship between response and explanatory variables to attain statistically significant parameters. Further, selection of appropriate variables in a regression paradigm is an important part of machine learning which refers to the process of reducing the inputs for processing to find the most meaningful, promising variables [1]. The selection reduces number of explanatory variables to describe a response variable [2]. The supervised machine learning processreducesoverfitting because a model with all variables is difficult to interpret, especially when the dimensions of a dataset are large [3].

Based on the dataset, machine learning algorithms might be supervised [4], unsupervised [5] and semi-supervised [6].

In this paper, we provide analytical method to choose influential variables from a dataset of 160 ITK inhibitors with 32 variables that describe the associated properties of these inhibitors in binding to the ITK protein target. Interleukin-2 inducible T-cell kinase (ITK) is a tyrosine kinase that is expressed in T-cells, NK cells and mast cells. Studies suggest that selective ITK inhibitors should be useful as an immunosuppressiveand/or anti-inflammatory agent and reduced lung inflammation, eosinophil infiltration, and mucous production in response to induced allergic asthma [7]. The Tec (tyrosine kinase expressed in hepatocellular carcinoma) family tyrosine kinases play significant roles in mediation of signaling intracellular regions of hematopoietic cells [8]. Due to the critical role of ITK in T cell development and differentiation, disregulatedITK causes T cell related disorders.ITKknockout mice displayedcondensed Th2 cells and Th2-type cytokines. These are thought to be important in the inflammatory pathogenic diseases such as allergic asthma and atopic dermatitis [9]. Patients with allergic asthma have increased Th2 cells and Th2 cytokines which lead to lung inflammation [10]. Human immunodeficiency virus (HIV) is a retrovirus causing acquired immunodeficiency syndrome (AIDS). ITK is an important factor in regulation, infection and replication of HIV [11].*In vivo* experiments with ITK knockout mice suggest a role for ITK inhibitors in the treatment of asthma [12]. Several data from literature reported ITK inhibitors with a focus on achieving broad kinase selectivity as well as good levels of cellular activity [13].

II. MATERIALS AND METHODS

a. DATASET

A dataset of ITK inhibitors that are intended to interact and bind with specific protein target for asthma disease were extracted from literature [14, 15, 16, 17, 18 and 19]. Further, the bio activity data of 142 inhibitory compounds are treated as response (dependent) variable and nearly 32 properties of compounds comprising 2-dimensional and/or 3-dimensional features are considered as explanatory (independent) variables. These variables explain how the response variable is influenced by the change in property values. The

Revised Manuscript Received on August 19, 2019.

Rama Devi Chalasani, Reaseach Scholar, Department of CSE, GIT, Gitam Deemed to be University. Visakhapatnam.A.P, India.

DrRadhika Y, Professor, Department of CSE, GIT, Gitam Deemed to be University. Visakhapatnam.A.P, India.

influence of explanatory variables on response variable can be studied via linear regression accounting that the data is linear. Biological data are usually expressed on a logarithmic scale because of the linear relationship between response and log dose in the mid-region of the log dose-response curve. Inverse logarithms for activity (log 1/C) are used so that higher values are obtained for more effective analogs.

b. EXPLANATORY VARIABLES

Considering learning problems in data with flexible dimensions, it was reported that not all variables are applicable for predicting the outcome of interest. Some variables might represent negative effect on the accuracy of the model. Nearly 32 variables otherwise referred as independent variables were calculated for each dependent variable. They include: Molecular surface area (MSA), Molecular volume (MV), Molar Refractivity (MR), Kier indices such as Kier Chi 0 (KC0),Kier Chi 1 (KC1), Kier Chi 1 (KC2), Kier Chi 1 (KC3), Kier Chi V0 (KCV0), Kier Chi V1 (KCV1), Kappa 1 index, Kappa 2 index, Kappa 3 index, KAlpha1, KAlpha2, KAlpha3 index, Randic, Balaban, Weiner index, Shape index, Lipinski properties [20] such as Molecular Weight (MW), Hydrogen Bond Acceptors (HBA), Hydrogen Bond Donors (HBD), logP, Number of freely rotatable bonds (RB), Electrostatic properties such as Highest Occupied Molecular Orbital (HOMO), Lowest Unoccupied Molecular Orbital (LUMO), Number of 5-membered rings, 6-membered rings, methyl groups, amino groups, hydroxyl groups and phenyl moieties attached to the main molecule.

c.MULTICOLLINEARITY

It should be noted that all these physico-chemical properties of ITK inhibitors or few variables should have influential result on the biological activity. Therefore, few variable selection methods have gained prominence. However, it should be noted that few variables display collinearity with other variables. This might result to a situation where a number of independent variables in a multiple regression model are closely correlated to one another. Such mutlicollinearity leads to unreliable and unstable estimates of regression coefficients which results in unstable parameter estimates and makes it very difficult to assess the effect of independent variables on dependent variables.

d. DATA NORMALIZATION

Prior analysis, data was normalized. Because, there can be instances found in data frame where values for one feature could range between 1-100 and values for other feature might range more than 10000 or so. Such variations in values have greater impact on response variables by that particular feature which would otherwise impact prediction accuracy.

e. REGRESSION ANALYSIS

The relationship between dependent variable (log $1/IC_{50}$) and independent variables was established by linear regression analysis. Significant variables were chosen based on the statistical data of analysis. Statistical quality of the generated regression equation was judged based on the parameters like correlation coefficient (r), F-value, adjusted r2 etc.

f. OUTLIER ANALYSIS – STANDARDIZED RESIDUALS

Variance inflation factor (VIF):

Variance inflation factors (VIF) measure how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related. It is used to explain how much amount multicollinearity (correlation between predictors) exists in a regression analysis. Multicollinearity increases the variance of the regression coefficients. A VIF of four means that the variance (a measure of imprecision) of the estimated coefficients is four times higher because of correlation between the two independent variables. The general rule of thumb is that VIFs exceeding 4 warrant further investigation, while VIFs exceeding 10 are signs of serious multicollinearity requiring correction.

g. FIT KUBINYI FUNCTION

To define the statistical quality of activity prediction, the number of variables that enter in a regression model are compared by using FIT Kubinyi function [21], which is closely related to F value.

$$\text{FIT} = R^2 (n - k - 1) / (n + k^2) (1 - R^2)$$

Where*n* is the number of compounds in data set and *k* is the number of variables in the regression equation.

F value is sensitive to variations in k, it has low sensitivity when k is large. Hence, the FIT function is also sensitive towards changes in k, a significantly increase in sensitivity was possible for large k values [22].

III. RESULTS AND DISCUSSION

The ITK molecules extracted from literature are drawn as 2-dimensional structures and selected to calculate physico-chemical properties associated with them. These properties referred as independent variables explain the variance in data whereas the dependent variable is the biological activity of chemical compounds that exhibit inhibitory activity against ITK. This activity data, reported as IC_{50} micro Molar (µM) inhibition was the result of bio-analytical outcome. It was observed that to obtain a reliable and robust regression model, it is necessary to contemplate a large data set that covers reasonable chemical diversity and biological activity. Hence, a dataset of 142 ITK inhibitors and their inhibitory activities reported in terms of IC_{50} in µM were transformed into their corresponding concentration values in order to overcome overlapping data. Data given in Appendix-1.Therefore, to guarantee linear distribution of data, the ITK inhibition values were converted to negative logarithmic values and then used for subsequent analysis. Biological data are usually expressed on a logarithmic scale because of the linear relationship between response and log dose in the mid-region of the log dose-response curve. Inverse logarithms for activity (log 1/C) are used so that higher values are obtained for more effective analogs.

After data normalization, an inter-correlation cutoff value of 0.75 was induced to filter independent variables that are strongly correlated and only variables less than cut off value are selected to perform multicollinear graph. With this cross-limit cut off, out of 32 variables, only 15 variables appeared such as BALABAN , HBA, HBD, LOGP, RB, LUMO, HOMO, METHYL, AMINO, HYDROXYL, PHENYL, X5.MEM, X6.MEM, KC3 and SHAPE index. The correlation plot given in Fig. 1 suggest that the maximum correlation between variables was around 0.64 between Hydroxyl variable and KC3 variable only.

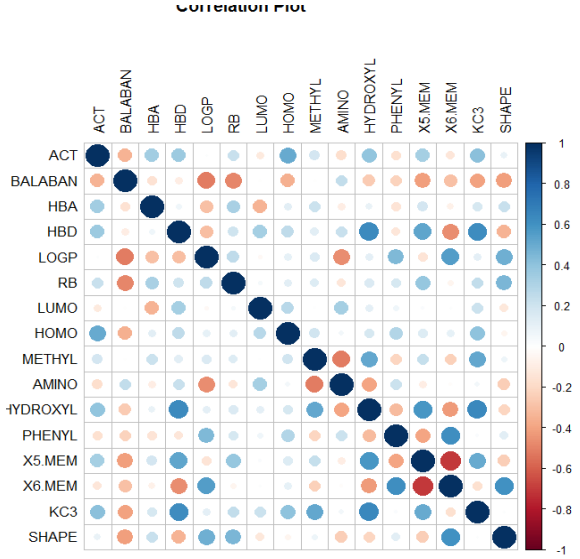

Fig. 1.Correlation plot showing 15 variables under 0.75 inter-correlation.

a. REGRESSION ANALYSIS

After rejecting few variables that are linearly correlated among one another, a multivariate linear regression analysis was calculated using 142 dependent variables with 15 independent variables. Analysis resulted in 0.641 r2 and 0.598 adjusted r2 withRMSE 0.634 respectively.

b. MODEL-1:

Log(1/C) = -0.352 X BALABAN
+0.154 X HBA
+0.274 X HBD
+0.195 X LOGP
-0.031 X RB
-0.231 X LUMO
+0.600 X HOMO
-0.439 X METHYL
-0.262 X AMINO
-0.090 X HYDROXYL
-0.331 X PHENYL
-0.800 X X5.MEM
-0.863 X X6.MEM
+0.416 X KC3
+0.239 X SHAPE
+0.00

r = 0.801; r2 = 0.641; adj r2 = 0.598; F = 15.007;
RMSE: 0.634; n=142 (1)

As the statistical parameters are within the limits, outlying data was investigated by calculating standardized residuals. Outlying data mask the influential observations and might result in low r2 values. Hence, they should be removed before finalizing the model for analytical prediction.

c. OUTLIER ANALYSIS – STANDARDIZED RESIDUALS

The standardized residual is a measure of the strength of the difference between observed and expected values. A standardized residual less than -2 and greater than +2 is considered as outlying data point. The standardized residual analysis resulted in nine data points (44, 53, 60, 66, 91, 95, 120, 134 and 135) as outlying observations (Table-I).

Table-I: Standardized residuals of dataset.

S. No.	Molecule	Activity	Predicted Activity	Residuals	Standardized Residuals
1	2_1.mol	1.455998	0.924238	0.53176	0.887671
2	2_10a.mol	0.865309	0.528101	0.337208	0.562904
3	2_10b.mol	1.596167	1.28176	0.314407	0.524842
4	2_10c.mol	1.575268	1.717554	-0.14229	-0.23752
5	2_10d.mol	1.94786	2.352376	-0.40452	-0.67526
6	2_10e.mol	1.32593	1.254861	0.071069	0.118637
7	2_10f.mol	1.666657	1.287395	0.379262	0.633105
8	2_10g.mol	1.440394	1.039435	0.400959	0.669324
9	2_10h.mol	1.93957	1.679835	0.259735	0.433578
10	2_10i.mol	1.9243	1.862195	0.062104	0.103671
11	2_10j.mol	1.617825	1.666916	-0.04909	-0.08195
12	2_9a.mol	-0.45477	0.340897	-0.79567	-1.32821
13	2_9b.mol	0.699782	1.093644	-0.39386	-0.65748
14	2_9c.mol	0.844013	1.36435	-0.52034	-0.8686
15	5_1.mol	-0.43776	-0.68382	0.24606	0.410749
16	5_17.mol	-1.05751	-1.58886	0.531353	0.886992
17	5_18.mol	-1.65394	-1.50639	-0.14755	-0.24631

18	5_19.mol	-0.95911	-0.94845	-0.01066	-0.0178
19	5_20.mol	-0.87533	-1.19931	0.323976	0.540816
20	5_21.mol	-0.28933	0.50621	-0.79554	-1.328
21	5_22.mol	-0.45121	-0.74662	0.295411	0.493131
22	5_23.mol	-0.76778	-1.02663	0.258847	0.432095
23	5_24.mol	-2.10022	-2.02904	-0.07119	-0.11883
24	5_25.mol	-0.1241	-0.90584	0.781741	1.304966
25	5_26.mol	-0.19106	-0.25132	0.060258	0.100589
26	5_30.mol	-2.76148	-2.51981	-0.24167	-0.40342
27	5_31.mol	-1.10065	-1.45719	0.356536	0.595168
28	5_32.mol	-1.43778	-0.72426	-0.71352	-1.19108
29	5_33.mol	-0.3702	-0.74553	0.375324	0.626532
30	5_34.mol	-0.92071	-0.02231	-0.8984	-1.4997
31	5_35.mol	-0.99136	-1.56055	0.569193	0.950158
32	5_36.mol	-0.9337	-0.61895	-0.31475	-0.52542
33	5_37.mol	-0.83454	-0.59207	-0.24247	-0.40475
34	5_38.mol	-0.66233	-1.2522	0.589873	0.984679
35	5_39.mol	-0.57808	-0.67513	0.097052	0.16201
36	5_40.mol	-0.15562	-0.31904	0.163416	0.272791
37	5_42.mol	-1.40214	-1.18419	-0.21795	-0.36382
38	5_43.mol	-1.74995	-1.3746	-0.37536	-0.62658
39	5_44.mol	-0.68836	-0.47958	-0.20877	-0.34851
40	5_45.mol	-0.33041	-0.48964	0.159224	0.265794
41	5_46.mol	-1.53538	-0.82806	-0.70732	-1.18073
42	5_47.mol	-1.40112	-0.74776	-0.65336	-1.09066
43	5_48.mol	-1.59813	-1.04433	-0.5538	-0.92447
44	5_49.mol	-2.71496	-1.34343	-1.37153	-2.28951
45	7_10.mol	0.819647	0.51534	0.304307	0.507982
46	7_11.mol	0.33342	0.46586	-0.13244	-0.22108
47	7_12.mol	0.565119	-0.05097	0.616089	1.028442
48	7_13.mol	0.778494	0.568191	0.210303	0.351061
49	7_14.mol	1.346048	0.618007	0.728041	1.215325
50	7_15.mol	1.242997	0.580755	0.662242	1.105486
51	7_16.mol	0.774379	0.07883	0.695549	1.161085
52	7_17.mol	0.68019	-0.0756	0.755791	1.261648
53	7_18.mol	-2.20389	-0.5813	-1.62259	-2.7086
54	7_19.mol	-0.09801	-0.20899	0.110973	0.185249
55	7_2.mol	0.084854	0.738427	-0.65357	-1.09101
56	7_20.mol	0.144795	-0.18517	0.329968	0.550819
57	7_21.mol	1.253203	0.44961	0.803593	1.341444
58	7_22.mol	1.253203	0.618526	0.634677	1.059471
59	7_23.mol	-0.22876	0.354739	-0.5835	-0.97403
60	7_24.mol	-0.91796	0.524736	-1.4427	-2.40831
61	7_28.mol	-0.85636	0.055968	-0.91232	-1.52295
62	7_29.mol	0.553659	0.477098	0.076561	0.127804
63	7_31.mol	0.165362	-0.05747	0.222832	0.371975
64	7_32.mol	1.008796	0.044239	0.964556	1.610141
65	7_33.mol	-0.41424	0.078623	-0.49286	-0.82274
66	7_34.mol	-1.47229	0.014329	-1.48662	-2.48162
67	7_36.mol	0.535905	0.282073	0.253832	0.423724
68	7_37.mol	-0.23155	0.063739	-0.29529	-0.49293
69	7_38.mol	0.819156	0.774194	0.044962	0.075056
70	7_39.mol	0.576216	0.079262	0.496954	0.829569
71	7_41.mol	-0.21135	0.397473	-0.60883	-1.01632
72	7_42.mol	1.067359	0.687754	0.379605	0.633677

73	7_43.mol	1.067408	1.012418	0.05499	0.091795
74	7_44.mol	0.392457	0.442659	-0.0502	-0.0838
75	7_47.mol	0.948987	0.474873	0.474114	0.791442
76	7_48.mol	-0.64896	-0.54862	-0.10034	-0.1675
77	7_49.mol	-0.0567	-0.27141	0.214707	0.358412
78	7_50.mol	-0.9478	-0.09476	-0.85304	-1.42399
79	7_51.mol	-0.77126	-0.23855	-0.53271	-0.88926
80	7_52.mol	0.039652	1.148741	-1.10909	-1.85141
81	7_53.mol	0.926738	1.108006	-0.18127	-0.30259
82	7_54.mol	0.428418	-0.05608	0.4845	0.808779
83	7_56.mol	1.401823	0.661563	0.74026	1.235722
84	7_8.mol	1.164873	0.713536	0.451336	0.753419
85	7_9.mol	0.81848	0.715032	0.103448	0.172686
86	9_10.mol	0.161972	-0.06232	0.224295	0.374417
87	9_11.mol	-0.24178	-0.45079	0.209008	0.348899
88	9_12.mol	-0.46119	0.289279	-0.75047	-1.25277
89	9_15.mol	-0.00467	-0.39178	0.387106	0.646199
90	9_16.mol	0.436105	-0.33236	0.768467	1.282808
91	9_17.mol	-1.73634	-0.34392	-1.39243	-2.32439
92	9_18.mol	0.034477	-0.25329	0.287763	0.480364
93	9_19.mol	0.160176	-0.42944	0.589611	0.984243
94	9_2.mol	0.005809	-0.19907	0.20488	0.342008
95	9_20.mol	1.085817	-0.16952	1.255334	2.095538
96	9_21.mol	0.066416	0.3734	-0.30698	-0.51245
97	9_22.mol	0.187642	-0.34064	0.52828	0.881861
98	9_23.mol	0.29708	0.298276	-0.0012	-0.002
99	9_24.mol	-0.00751	-0.31953	0.312018	0.520855
100	9_25.mol	0.062907	0.355823	-0.29292	-0.48897
101	9_26.mol	1.480931	0.741714	0.739217	1.233981
102	9_7.mol	-0.56774	-0.75082	0.183083	0.305621
103	9_8.mol	0.033181	0.309048	-0.27587	-0.46051
104	9_9.mol	0.006916	0.724425	-0.71751	-1.19774
105	11_1.mol	-0.67822	-0.83585	0.157622	0.26312
106	11_11.mol	-1.64626	-0.96827	-0.67799	-1.13177
107	11_12.mol	-2.09295	-1.56173	-0.53122	-0.88677
108	11_13.mol	-0.40356	-1.30667	0.903116	1.507578
109	11_14.mol	-0.10015	-0.25564	0.155488	0.259558
110	11_15.mol	-0.09117	-0.18115	0.089975	0.150196
111	11_16.mol	-1.01492	-0.12399	-0.89093	-1.48724
112	11_17.mol	-0.51824	-0.17582	-0.34242	-0.5716
113	11_18.mol	-1.14089	-0.35141	-0.78948	-1.31788
114	11_19.mol	-0.63799	-0.26466	-0.37333	-0.6232
115	11_20.mol	-0.66753	-0.2078	-0.45973	-0.76743
116	11_21.mol	-0.36388	-0.42232	0.058445	0.097563
117	11_22.mol	0.595849	0.156118	0.439732	0.734047
118	11_23.mol	0.514635	0.261046	0.253588	0.423317
119	11_24.mol	0.570415	0.13848	0.431934	0.721031
120	11_25.mol	1.608065	0.348714	1.25935	2.102244
121	11_26.mol	0.587767	-0.05956	0.647324	1.080583
122	11_27.mol	0.374709	0.062217	0.312493	0.521646
123	11_28.mol	-0.18076	-0.00097	-0.17979	-0.30013
124	11_29.mol	-0.65207	0.048318	-0.70038	-1.16916
125	11_30.mol	0.050845	0.176954	-0.12611	-0.21051
126	11_31.mol	0.023748	-0.31239	0.336141	0.561124
127	11_32.mol	0.394283	0.425002	-0.03072	-0.05128

128	11_35.mol	0.851132	0.409948	0.441184	0.736472
129	11_38.mol	0.985855	0.305881	0.679975	1.135087
130	11_39.mol	0.670739	0.477098	0.193641	0.323246
131	11_40.mol	0.705558	0.665709	0.039849	0.06652
132	11_41.mol	1.568291	0.502729	1.065561	1.77875
133	11_43.mol	-0.3469	0.35156	-0.69846	-1.16595
134	11_44.mol	-1.15927	0.37295	-1.53222	-2.55775
135	11_45.mol	-0.68003	0.672343	-1.35237	-2.25753
136	11_46.mol	-0.31596	0.172223	-0.48818	-0.81493
137	11_47.mol	0.817092	1.076509	-0.25942	-0.43305
138	11_48.mol	1.330439	0.591719	0.738719	1.23315
139	11_6.mol	0.472511	-0.3201	0.792613	1.323115
140	11_7.mol	0.512377	-0.31287	0.825243	1.377585
141	11_8.mol	-0.6827	-0.70677	0.024074	0.040187
142	11_9.mol	0.236413	-0.30659	0.543005	0.906442
			Standard Deviation	0.599051	

From the outlier study, nine outliers should be removed from analysis in order to guarantee linear distribution of correlations among parameters. By excluding these data points, regression analysis should be carried out to assess statistical significance. Therefore, a new regression model is attempted by excluding nine data points (44, 53, 60, 66, 91, 95, 120, 134 and 135) being detected as outliers with data being n=133 and p=15; and from the result given in model-2, it was evidenced that an improvement was observed in r2 value from 0.641 to 0.739 and adjusted r2 reported improved value from 0.598 to 0.706, respectively.

d. MODEL-2:

Log(1/C) = -0.257 X BALABAN
+0.210 X HBA
+0.360 X HBD
+0.268 X LOGP
-0.131 X RB
-0.148 X LUMO
+0.674 X HOMO
-0.432 X METHYL
-0.259 X AMINO
-0.129 X HYDROXYL
-0.306 X PHENYL
-0.678 X X5.MEM
-0.832 X X6.MEM
+0.342 X KC3
+0.370 X SHAPE
+0.00

r = 0.860; r2 = 0.739; adj r2 = 0.706; F = 22.095;
RMSE: 0.543; n=133 (2)

e. STEPWISE REGRESSION

Further, in order to delineate most important variables that would efficiently estimate coefficients of data are identified using step-wise regression method [23] by entering and removing variables based on p values, in a stepwise manner until there is no variable left to enter or remove any more. A variable will enter the model if p value less than 0.1 and any variable will be removed if p is >0.3 by default. Therefore, from 15 variables, only the following five variables were found to be important in defining the model with accuracy.

f. MODEL-3:

Log(1/C) = +0.390 X HBD
-0.296 X LUMO
+0.683 X HOMO
-0.327 X PHENYL
+0.321 X SHAPE
+0.00

r = 0.818; r2 = 0.669; adj r2 = 0.656; F = 51.24;
RMSE: 0.587; n=133 (3)

Based on F-value, it can be stated that model-3 reported better statistic variables and to confirm stepwise selection method, step-wise AIC regression was performed.

g. STEPWISE AIC REGRESSION

Stepwise AIC regression builds regression model from a set of 15 candidate independent variables by entering and removing predictors based on akaike information criteria [24], in a stepwise manner until there is no variable left to enter or remove any more. Analysis resulted in eleven variable model and the statistics are tabulated in Table-II.

Table-II: Stepwise Summary

Variable	*Method*	*AIC*	*RSS*	*SUM SQ*	*R-SQ*	*ADJ. R-SQ*
HOMO	Addition	327.779	87.520	44.480	0.33697	0.33191
PHENYL	Addition	302.140	71.098	6.9.2	0.46138	0.45310
SHAPE	Addition	288.230	63.082	68.918	0.52211	0.51099
HDB	Addition	268.420	53.541	78.459	0.59439	0.58171
LUMO	Addition	243.553	43.747	88.253	0.66858	0.65553
METHYL	Addition	242.702	42.820	89.180	0.67561	0.66016
AMINO	Addition	240.300	41.426	90.574	0.68617	0.66859
KC3	Addition	239.515	40.567	91.433	0.69267	0.67284
X5.MEM	Addition	238.965	39.797	92.203	0.69851	0.67645
X6.MEM	Addition	233.695	37.680	94.320	0.71455	0.69115
BALABAN	Addition	231.656	36.553	95.447	0.72308	0.69791

From the step-wise summary, it can be observed that few regression models can be obtained with combination of variables such as for a 5-variable model, the r2 value is 0.669 which is similar to the result obtained from stepwise regression analysis based on p value. Further, a 6-variable model showed 0.675 r2, a 7-variable model has 0.686 with an increasing r2 observed for 8-variable model (0.692), 9-variable model resulted in 0.698 and 10, 11-variable models showed 0.714 and 0.723 r2 values, respectively.

Variance inflation factors are evaluated for all 15 variables of model2. From Table-III, it is evidenced that the variables BALABAN, X5.MEM and X6.MEM have high VIF values and hence are excluded from regression analysis.

Table-III: Variance inflation factors of all 15 variables in the study

SNO	Variables	Tolerance	VIF
1	BALABAN	0.0794	12.6
2	HBA	0.318	3.14
3	HBD	0.213	4.71
4	LOGP	0.159	6.27
5	RB	0.162	6.18
6	LUMO	0.470	2.13
7	HOMO	0.590	1.69
8	METHYL	0.189	5.30
9	AMINO	0.169	5.90
10	HYDROXYL	0.154	6.48
11	PHENYL	0.170	5.87
12	X5.MEM	0.0493	20.3
13	X6.MEM	0.0342	29.2
14	KC3	0.118	8.45
15	SHAPE	0.135	7.43

A regression model with the remaining 12 variables resulted in better predicted statistics, however, few variables are found to be not significant statistically, where p value is greater than 0.1. Removing variables such as LOGP, RB, AMINO, HYDROXYL and KC3 and a new regression model attempted with 7 variableshas improved F-statistic parameter from 23.54 to 37.33.

h. MODEL-4:

Log(1/C) = +0.034 X HBA
+0.394 X HBD
-0.289 X LUMO
+0.700 X HOMO
-0.093 X METHYL
-0.348 X PHENYL
+0.324 X SHAPE
+0.00

r = 0.822; r2 = 0.676; adj r2 = 0.658; F = 37.33;
RMSE: 0.584; n=133 (4)

Moreover, several permutation and combination of 7 variables from model4 were carried out using olsrr package which resulted in several combinations with r2 value >0.6 for 5, 6 and 7 variables (Fig. 2).Upon careful observation of the above graph, it is evident that combination of variables for a:5-variable model numbered 99 resulted in r2: 0.669 and adjusted r2 0.656, 6-variable model 120 displayed r2: 0.676 and adjusted r2 0.660 whereas a 7-variable model numbered 127 has r2: 0.676 and adjusted r2 0.658, respectively. Regression data and statistics given below.

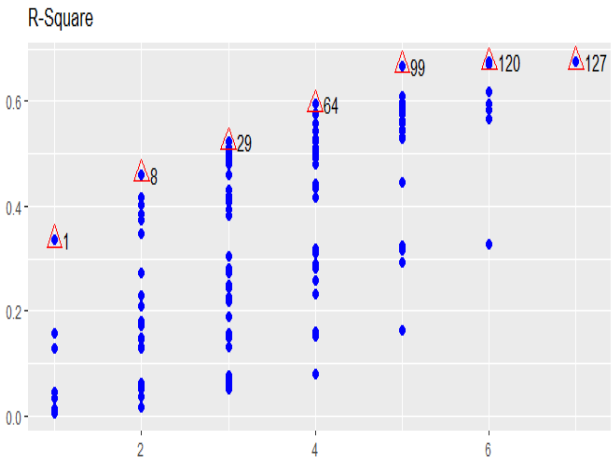

Fig. 2. Panel of fit criteria for all possible regression models

i. MODEL-5: 5-VARIABLE MODEL

Log(1/C) = +0.390 X HBD
-0.296 X LUMO
+0.683 X HOMO
-0.327 X PHENYL
+0.321 X SHAPE
+0.00

r = 0.818; r2 = 0.669; adj r2 = 0.656; F = 51.24;
RMSE: 0.587; n=133 (5)

j. MODEL-6: 6-VARIABLE MODEL

Log(1/C) = +0.401 X HBD
-0.303 X LUMO
+0.709 X HOMO
-0.091 X METHYL
-0.354 X PHENYL
+0.334 X SHAPE
+0.00

r = 0.822; r2 = 0.676; adj r2 = 0.660; F = 43.73;
RMSE: 0.583; n=133 (6)

k. MODEL-7: 7-VARIABLE MODEL

Log(1/C) = +0.034 X HBA
+0.394 X HBD
-0.290 X LUMO
+0.700 X HOMO
-0.094 X METHYL
-0.348 X PHENYL
+0.325 X SHAPE
+0.00

r = 0.822; r2 = 0.676; adj r2 = 0.658; F = 37.33;
RMSE: 0.585; n=133 (7)

The Table-IV given below presents the best possible models generated from several combinations of variables and the various statistical parameters obtained from analysis suggest that all models are within the limits of validation parameters of regression. It was observed that a low RMSE value was obtained for a 6-variable model whereas significant F-value reported by a 5-variable model. Hence, to select the best model, FIT criterion was employed.

Table-IV: Statistical data and validation parameter values of 5, 6 and 7 variable model equations.

Variable	Coefficient		
	5-variable model	6-variable model	7-variable model
HBA	-	-	+0.034
HBD	+0.390	+0.401	+0.394
LUMO	-0.296	-0.303	-0.290
HOMO	+0.683	+0.709	+0.700
METHYL	-	-0.091	-0.094
PHENYL	-0.327	-0.354	-0.348
SHAPE	+0.321	+0.334	+0.325
Intercept	-0.00	-0.00	-0.00
Statistics			
R	0.818	0.822	0.822
r^2	0.669	0.676	0.676
Adj r^2	0.656	0.660	0.658
F	51.24	43.73	37.33
N	133	133	133
RMSE	0.587	0.583	0.585
MSE	0.344	0.340	0.342
MAE	0.465	0.461	0.462

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

l. FIT KUBINYI FUNCTION

The best model was judged based on the high value obtained from FIT function and hence regression models with five, six and seven variables are applied to select best possible model (Table-V).

Table-V: FIT function of three regression models obtained for five, six and seven variables.

Variables (k)	r^2	*s*	*F*	FIT
5	0.669	0.254	51.24	1.624
6	0.676	0.255	43.73	1.555
7	0.676	0.256	37.33	1.432

The statistical FIT values of the models reported inTable-5 suggest that the model with five variables represents best model since this model has high FIT than others. Therefore, further analysis was carried out with five-variable model. The actual and predicted value of this model is given in Figure. 3.

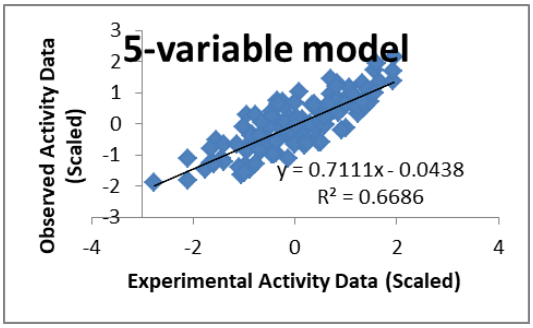

Fig. 3. Regression plot of 5-variable model data with n=133 showing r2 0.669

After rejecting nine outliers from the data set, a prominent increase in R-squared and adjusted R-squared values along with F-statistic were observed which suggests the fact that an improved regression diagnostics was achieved by five variables viz., Hydrogen Bond Donors (HBD), HOMO (Highest Occupied Molecular Orbital), LUMO (Lowest Unoccupied Molecular Orbital), No. of Phenyl groups and Shape index features in defining ITK inhibitor activity.

The highest occupied and lowest unoccupied molecular orbitals (HOMO and LUMO) are vital in envisaging the molecule reactivity. HOMO is the outmost orbital comprising the electron and LUMO is the first orbital

which is devoid of any electron. The HOMO energy measures the electron donating character and LUMO measures its electron accepting character [25]. Based on the coefficients of 5-variable model, a high value of LUMO energy contributes negatively to the activity. Electron-withdrawing substituents, such as halogens, lower the energy of LUMO. Hence, ITK inhibitors with electron-withdrawing substituents improve activity. From the coefficient values, it can be emphasized that an increase in HOMO, H-Bond Donors and shape index is required to inhibit ITK by these set of compounds whereas on the other hand, reduced number of phenyl groups and LUMO parameter favors ITK inhibition.

IV. CONCLUSION

Variable selection as a method of choice to choose most important variables which could describe the inhibitory properties of nearly 142 ligands tested against ITK protein target against 32 independent variables was studied using multivariate regression technique. With inter correlation cut off value of 0.75 resulted in 15 variables and regression resulted in 0.641 r2 and 0.598 adjusted r2 with RMSE 0.634 respectively. Eliminating nine outlying data evidenced an improvement in r2 value to 0.739 and adjusted r2 reported 0.706, whereas step-wise regression suggested only five variables as important in defining the model based on p values. Finally based on stepwise AIC regression and variance inflation factor analytics resulted in 7 variables. Several permutation and combination of these 7 variables reported r2 value >0.6 for 5, 6 and 7 variables and the statistical FIT values of all the three models suggested 5-variable model as best model since this model has high FIT than others. From the coefficient data, increase in HOMO, H-Bond Donors and shape index inhibits ITK whereas decrease in the number of phenyl groups and LUMO parameter favors ITK inhibition.

REFERENCES

1. Forman G. An extensive empirical study of feature selection metrics for text classification, J. Mach. Learn. Res., 2003, vol. 3 (pg. 1289-1305.
2. Guyon I, Elisseeff A. An introduction to variable and feature selection, J. Mach Learn Res., 2003, vol. 3 (pg. 1157-1182).
3. E. Candes and T. Tao. The dantzig selector: statistical estimation when p is much larger than n. The Annals of Statistics, pages 2313–2351, 2007.
4. J. Weston, A. Elisseff, B. Schoelkopf, and M. Tipping. Use of the zero norm with linear models and kernel methods. Journal of Machine Learning Research, 3:1439–1461, 2003.
5. J.G. Dy and C.E. Brodley. Feature selection for unsupervised learning. The Journal of Machine Learning Research, 5:845–889, 2004.
6. Z. Zhao and H. Liu. Semi-supervised feature selection via spectral analysis. In Proceedings of SIAM International Conference on Data Mining, 2007.
7. Mueller, C.; August, A. J. Immunol. 2003, 170, 5056] [Kanner, S. B.; Perez-Villar, J. J. Trends Immunol. 2003, 24, 249.
8. Tec family kinases in T lymphocyte development and function. Berg LJ, Finkelstein LD, Lucas JA, Schwartzberg PL. Annu Rev Immunol. 2005; 23():549-600.
9. Attenuation of immunological symptoms of allergic asthma in mice lacking the tyrosine kinase ITK. Mueller C, August A. J Immunol. 2003 May 15; 170(10):5056-63.
10. Reduced airway hyperresponsiveness and tracheal responses during allergic asthma in mice lacking tyrosine kinase inducible T-cell kinase. Ferrara TJ, Mueller C, Sahu N, Ben-Jebria A, August A. J Allergy ClinImmunol. 2006 Apr; 117(4):780-6.
11. Selective targeting of ITK blocks multiple steps of HIV replication. Readinger JA, Schiralli GM, Jiang JK, Thomas CJ, August A, Henderson AJ, Schwartzberg PL. ProcNatlAcadSci U S A. 2008 May 6; 105(18):6684-9.
12. August A.; Ragin M. J. Regulation of T-cell responses and diseases by Tec kinase ITK. Int. Rev. Immumol. 2012, 31, 155–165.
13. Das J.; Furch J. A.; Liu C.; Moquin R. V.; Lin J.; Spergel S. H.; McIntyre K. W.; Shuster D. J.; O'Day K. D.; Penhallow B.; Hung C. Y.; Doweyko A. M.; Kamath A.; Zhang H.; Marathe P.; Kanner S. B.; Lin T. A.; Dodd J. H.; Barrish J. C.; Wityak J. Discovery and SAR of 2-amino-5-(thioaryl)thiazoles as potent and selective ITK inhibitors. Bioorg. Med. Chem. Lett. 2006, 16, 3706–3712.
14. Doris Riether et al. 5-Aminomethylbenzimdazoles as potent ITK antagonists. Bioorganic & Medicinal Chemistry Letters 19 (2009) 1588–1591.
15. Ho Yin Lo et al. 2-Aminobenzimidazoles as potent ITK antagonists: trans-stilbene-like moieties targeting the kinase specificity pocket. Bioorganic & Medicinal Chemistry Letters 18 (2008) 6218–6221.
16. Roger J. Snow et al. Hit-to-lead studies on benzimidazole inhibitors of ITK: Discovery of a novel class of kinase inhibitors. Bioorganic & Medicinal Chemistry Letters 17 (2007) 3660–3665.
17. Kevin J. Moriarty et al. Discovery, SAR and X-ray structure of 1H-benzimidazole-5-carboxylic acid cyclohexyl-methyl-amides as inhibitors of inducible T-cell kinase (ITK). Bioorganic & Medicinal Chemistry Letters 18 (2008) 5545–5549.
18. Michael P. Winters et al. 5-Aminomethyl-1H-benzimidazoles as orally active inhibitors of inducible T-cell kinase (ITK).Bioorganic & Medicinal Chemistry Letters 18 (2008) 5541–5544.
19. Brian N. Cook et al. Discovery of potent inhibitors of interleukin-2 inducible T-cell kinase (ITK) through structure-based drug design. Bioorganic & Medicinal Chemistry Letters 19 (2009) 773–777.
20. CA Lipinski, F Lombardo, BW Dominy and P J Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Adv Drug Del Rev 46 3-26 (2001).
21. Kubinyi H. Variable selection in QSAR studies. II. A Highly Efficient Combination of Systematic Search and Evolution. Quant. Struct. Act. Relat.13, 1994, 393-401.
22. Kubinyi H. Variable selection in QSAR studies. I. An evolutionary algorithm. Quant. Struct. Act. Relat. 13, 1994, 285-294.
23. Chatterjee, Samprit and Hadi, Ali. Regression Analysis by Example. 5th ed. N.p.: John Wiley & Sons, 2012. Print.
24. Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth edition. Springer.
25. Hall LH, Mohney B, Kier LB. The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs. J. Chem. Inf. Comput. Sci. 31, 1991, 76-82.