International Journal of Pharmacy and Biological Sciences
ISSN: 2321-3272 (Print), ISSN: 2230-7605 (Online)
IJPBS | Volume 7 | Issue 2 | APR-JUN| 2017 | 165-172
Research Article – Biological Sciences| Open Access| UGC Approved | MCI Approved Journal

A NON-LINEAR MODELLING APPROACH TO DEVELOP A QSAR MODEL FOR ANTICANCER ACTIVITY OF GOSSYPOL ACETIC ACID AGAINST BCL2 TARGET FOR COLORECTAL CANCER

Varun Kumar Kashyap* and Rajeev Pandey[1]
Department of Statistics, University of Lucknow, Lucknow-226007 (INDIA)

***Corresponding Author Email:** varun02stat@gmail.com

ABSTRACT

Qualitative Structural Activity Relationship (QSAR) model has been widely explored for suggesting compounds as anti – BCL 2 active compounds for colorectal cancer (Breast Cancer). A method is implemented here with anticancer activity of Gossypol acetic acid with the development of QSAR model using Support Vector Regression (SVR) method, and analysis is performed by this method after collecting a data set of 255 derivatives of BCL2 from the database of NCBI. Finally, based on SVR-QSAR model, eight compounds are recommended as Anti – BCL2 active compounds with the higher value regarding coefficient of determination.

KEY WORDS

BCL2; Colorectal Cancer; QSAR (Quantitative Structural Activity Relationship); Support Vector Regression; Virtual screening and coefficient of determination.

1. Introduction:

Cancer is a leading cause of death in global in male and female. According to DeSantis et al..2011, breast cancer is the second-leading cause of woman's death in the universe. Breast carcinoma (BC) corresponds to 23 % of all cancers in women, with 1.38 million new cases and 460,000 deaths worldwide annually (Canevari RA et al., 2016). Breast cancer is a distinct disease. Estrogen is associated with the development of breast cancer by activating catalysis androgens (Yager and Davidson, 2006). Aromatase is a type of enzyme, which involves in the catalysis to estrogens. Therefore, in Breast cancer treatment, the ability of such a kind of protein plays a great therapeutic role (Narashima Murthy et al., 2004). In India, a recent global study presented by the GE Healthcare estimated that by 2030 the incidence of new cases of breast cancer would increase from today the figure of 115000 to around 200000. For this problem, Antiapoptotic BCL2 proteins played a crucial role in the treatment of breast tumor cell survival & thus, BCL 2 inhibitors have been developed as apoptosis inducers direct. ABT-199 (Veneto C lax) received breakthrough therapy designation from the FDA due to its apparent efficacy in CLL and AML. However, resistance to ABT-199 is mediated by other BCL2 proteins, including BCLXL and MCL1.

Quantitative Structure-Activity relationship models are the higher-order mathematical equation for calculation of biological activity of descriptors molecular. It may be linear or nonlinear as it depends upon the nature of biologic data set. In a real-life situation, most of the cases the data behavior is nonlinear. The present study utilizes a non –linear technique to build a QSAR model for anticancer activity of Gossypol acetic acid against BCL2. The data set to have 255 compounds, which are taken from the PubChem database of NCIB.

Support Vector Machine Technique is a relatively new in compare to existing linear and non-linear multivariate methods in the chemometric field. Vapnik was the first person who introduced this method in Statistical Machine learning theory and Structural risk minimization. This technique is deal with linear and non-linear both type of data set and reduce the over-fitting of the proposed model. We use the here kernel

function to deal with non-linearity problems, i.e. every dot product in linear SVM is replaced by a non-linear kernel which satisfies the Mercer's thermo. As SVM gives a better result and single solution. In comparison to other machine learning techniques Support Vector Regression can solve the most critical problem like ill-posed that is often singular and have real predictive power.

In 2003, this method was, firstly, introduced in QSAR (Quantitative Structural Activity Relationship) and QSPR (Quantitative structural Property Relationship). This approach was used the world wide since 2003, and there have been more than 200 articles in which 95% prefer SVR. Presently, there are numbers of data mining and modeling technique for prediction of biological activities, chemical or physical properties. In the biological data selecting a few molecular descriptors among hundreds or thousands of measured variables. Increasing the complexity in the regression models as wells in the validation of this model.

2. EXPERIMENT:

2.1. Biological data set: The study deals with Gossypol acetic acid centered functional analogs containing anti-BCL2 activity data set adapted from NCBI database. The data set contains 128 predict inhibitors and 256 descriptors. Two-dimensional molecular descriptors have been worked out for each compound for digitization of observational data. The descriptor is calculated using PaDEL software (National University of Singapore), Showing the structural properties of molecules.

2.2. Data filtering and smoothing: The anti-Breast Cancer activity was in the form pIC50 = -log IC50. Originally, IC50 was determined and randomized by inhibitors docking. Initially, 255 descriptors were calculated for each molecular compound. As all the compounds play a significant role in the bioactivity, the following measures were taken to eliminate the less important descriptor-Eliminate the descriptor and remove the descriptor with constant values.

(i) Elimination of the descriptor with more than 90% zero values.

(ii) Elimination of the descriptor which has zero variance.

(iii) Elimination of the descriptor which has consequently high correlation in the correlation matrix.

These filtration steps include the selection of those descriptors which has correlation coefficient ≥ 0.4 with the bioactive vector of available data set and hence, we get 100 Active compounds and 45 descriptors.

2.3. Training and Test set assembly: We remove errors of tremendous Non-linearity by the filtered data that is randomly partitioned into Training and Test data set with probability 80% and 20%. So, we have 80 compounds in the Training set and 20 compounds in Test data set defining the complete coverage of Chemical and Biological properties.

2.3.1. Validation of QSAR model: For the validation of QSAR model, we have following internal validation and external validation.

2.3.1(a): Internal Validation: Internal validation of developed QSAR model is carried out using leave-one-out (LOO) method. The cross-validation regression coefficient (R2) is calculated using the equation which describes the internal stability of a model, i.e.

$$R^2 = 1 - \frac{\sum\left(Y_{pred} - Y_{exp}\right)^2}{\sum\left(Y_{exp} - \hat{Y}\right)^2}$$

2.3.1(b): External Validation: For external validation, the activity of each molecule in the test set is predicted using the model developed by the training set. The following formula calculates the regression coefficient (R2) value.

$$r_{cv}^2 = 1 - \frac{\sum\left(Y_{pred(test)} - Y_{\exp(test)}\right)^2}{\sum\left(Y_{\exp(test)} - \hat{Y}_{training}\right)^2}$$

Where refers regression coefficient, Yexp (test) and Ypred (test) are respectively experimental and predictive test activity of the molecule in the training set and Ŷ is the average activity of all molecules in the training set. Both summations are over all molecules in the test set. The regression coefficient is indicative of the predictive power of the current model for the external test set. A QSAR model is considered to have a high predictive power only if the is greater than 0.6 for the test set.

3. Support Vector Regression (SVR):

The main aim of a regression problem is to determine the underlying mathematical relationship by learning between the given input observations and their output values. This relationship may be assumed to be either

linear or nonlinear. Once it is determined, using this relationship, the output for any unseen data can be predicted. By the introduction of ε − insensitive error loss function proposed by **Vapnik (2000)**, SVR methods has been successfully extended to regression problems. It is well known that the SVR formulation **(Cristianini & Shawe-Taylor, 2000; Vapnik, 2000)** leads to a QPP of minimizing the regularization term and Vapnik's ε − insensitive error loss between the observed and their predicted values. All vectors are assumed as column vectors. For any two vectors x, y in the n -dimensional real-space $\Re^n$ the inner product of the vectors will be denoted by $\mathbf{x_t y}$ where $\mathbf{x_t}$ is the ranspose of the vector x. When x is orthogonal to y i.e. x ⊥ y. The 2-norm of a vector x and a matrix Q will be denoted by || x || and || Q || respectively. For any vector x ∈ $\Re^n$, x+ is a vector in $\Re^n$ obtained by setting all the negative components of x to zero. Further we define the step function x* as:(x*) i = 1 for x_i_>_0, (x*) i = 0 if 0 < xi and (x*) i = 5.0 when xi =0 . Also, x ≥ 0 means each component of the vector x is nonnegative. diag (x) denotes the diagonal matrix of order n whose diagonal elements are the components of the vector x∈ $\Re$n. For matrices M ∈ $\Re_{mxn}$ and N ∈ $\Re_{nxl}$, the kernel matrix K of size m × l is denoted by K = K (M, N). The identity matrix of appropriate size is denoted by I and e is the column vector of ones of length m . If f is a real valued function of the variable x = $(x_1,...,x_n)t$ ∈ $\Re_n$ then the gradient of f is denoted by ∇f =(∂f/∂x1,..., ∂f/∂xn)t and the Hessian matrix of f by ∇2f =(∂2f/∂xi, ∂xj); i,j=1,….,n

4. Nonlinear SVR:

The present work uses non-linear SVR technique to develop QSAR model. In linear SVM technique, only linear hyper planes to approximate prediction function in the input space have been considered owing to which it may be unsatisfactory to predict output values. If it were allowed to transform input data from input space into a higher-dimensional feature space so that in the transformed space, the linear regression estimation is obtained.

4.1. Standard SVR formulation

Let the transformation be ϕ: $\Re^N \rightarrow \Re^N$, where N >> n. Replacing the input vector x by ϕ(x) in the feature space, the dual problem with parameters v > 0 and ε > 0 will become

$$\min_{(u_1,u_1)\in\Re^{m+m}} \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} (u_{1i} - u_{2i})\ \phi\ (x_i)\ \phi\ (x_j)(u_{1j} - u_{2j}) - \sum_{i=1}^{m}(u_{1i} - u_{2i}) + \sum_{i=1}^{m}(u_{1i} - u_{2i})\ \varepsilon$$

Subject to:

$$\sum_{i=1}^{m}(u_{1i} - u_{2i}) = 0 \quad ; \quad 0 \le u_{1i}, u_{2i} \le v, \quad \forall j\ 1,...,m \qquad (4.1.1)$$

Assume that in addition $(\phi(x_i))^t \phi(x_j) = k(x_i, x_j)$, i.e. the dot product in high dimensional feature space is equivalent to a kernel function defined on the input space. This assumption has the advantage that one need not to know explicitly the transformation function ϕ(.) as long as we need only dot product and that such kernel function exists. In fact, there are many kernel functions available that can be used. A few commonly used kernels are listed below:

(i) Polynomial kernel $k(x,y) = (1+x^ty)^d$, d >1

(ii) Radial basis kernel $k(x,y) = \exp(-\mu ||x - y||^2)$, μ > 0

(iii) Neural network kernel $k(x,y) = \tanh(\rho_1 x^t y + \rho_2)$, $\rho_1 > 0$ $\rho_2 > 0$

Where d, μ, ρ_1 and ρ_2 are kernel parameters. The Mercer's condition **(Burges, 1998)** can be used to determine if a function can be used as a kernel function and hence, for a given kernel function k (.,.), there exist a mapping ϕ (.) and an expansion $k(x_i, x_j) = (\phi(x_i))^t \phi(x_j)$ if and only if, for any g(x) such that

$\int g(x)^2 dx$ is finite and $\int k(x_i, x_j)\ g(x_i)\ g(x_j)\ dx_i\ dx_j \ge 0$ is satisfied.

On applying this kernel trick, the optimization problem (4.1.1) with parameters v > 0 and ε > 0 can be viewed in matrix form as

$$\min_{(u_1,u_1)\in\Re^{m+m}} \frac{1}{2}(u_1- u_2)^t K(A,A^t)(u_1- u_2) - (u_1-u_2)^t y + (u_1 + u_2)^t e\varepsilon,$$

subject to :

$e^t(u_1- u_2) = 0$; $0 \le u_1, u_2 \le v$, (4.1.2)

Accordingly, the approximating hyperplane in the feature space will become

$y \approx f(x) = K(x^t,A^t)(u_1 - u_2) + b$

$= \sum_{i=1}^{m}(u_{1i} - u_{2i})\ k\ k(x_i,x_j) + b$ (4.1.3)

where $K(x_t,A_t) = (k(x,x_1),....,k(x,x_m))$ is a row vector in $\Re^m$, A∈$\Re^{m\times n}$ and the $(I, j)^{th}$ element of the kernel matrix K = $K(A\ A^t)$ is given by

$K(A\ A^t)_{ij} = k(x_i,x_j) \in \Re$.

(4.1.4)

5. Results & Discussion

5.1. Architecture of developed SVM (Radial Basis kernel function) model

SVM type: eps-svr (regression)
Parameter: epsilon = 0.1 cost C = 4
Gaussian Radial Basis kernel function.
Hyper parameter: sigma = 0.136500670080402
Objective Function Value: -18.0913
Training error: 0.006221 2

5.1.1. Statistical Evaluation of Model

The developed SVR Model has been evaluated by two bases:

(a) Prediction for Training data set: R2 value 0.93 and
(b) Prediction for Test data set: R2 value 0.91

This support vector algorithm is a nonlinear form of Generalized Portrait algorithm. The implemented epsilon-SVR is designed to get a function which has a maximum of epsilon deviation from each target point of training data. In this case, it is not considerable that error is less than epsilon or not, but the deviation should not be more than the epsilon value. Here 'Radial Basis Function' (RBF) is considered as the kernel function. The output of kernel function depends on the Euclidean distance between support vector and testing data point. Support vector exists at the center of kernel function i.e. (Radial Basis Function), and the sigma represents the area of influence covered in the data space by the support vector, where the optimal values of the width hyper-parameter σ are lying in between the 0.1 and 0.9.

In the comparison of other machine learning approaches as like ANN, SVR does not have the overfitting limitations, and hence SVM based QSAR models can be considered more robust in comparison of ANN. The kernel parameter was estimated analytically to be σ = 0.0039, and the model has turned over 20 cost values between 0.25 and 131000 on the log2 scale. The final model was evaluated for cost value C= 4 and RMSE is 0.629, and the regression coefficient for training data set comes out to be 0.93.

Table 5.1: 20-tuned Model for training data set:

C	RMSE	R square	RMSE SD	R square SD
0.25	0.97	0.86	0.536	0.123
0.5	0.767	0.889	0.544	0.116
1	0.659	0.906	0.541	0.113
2	0.634	0.93	0.525	0.108
4	0.629	0.93	0.526	0.108
8	0.629	0.93	0.526	0.108
16	0.629	0.93	0.526	0.108
32	0.629	0.93	0.526	0.108
64	0.629	0.93	0.526	0.108
128	0.629	0.93	0.526	0.108
256	0.629	0.93	0.526	0.108
512	0.629	0.93	0.526	0.108
1020	0.629	0.93	0.526	0.108
2050	0.629	0.93	0.526	0.108
4100	0.629	0.93	0.526	0.108
8190	0.629	0.93	0.526	0.108
16400	0.629	0.93	0.526	0.108
32800	0.629	0.93	0.526	0.108
65500	0.629	0.93	0.526	0.108
131000	0.629	0.93	0.526	0.108

Tuning parameter 'sigma' is held constant at a value of 0.1365007
RMSE is used to select the optimal model using the smallest value.
The final values used for the model were C = 4 and sigma = 0.137.

In the next step, we test this non-linear SVR model for external data set which is not used in developing the model. Here, we find that the regression coefficient for external data set is 0.912 and RMSE is 0.478 which is significantly very good.

TABLE 5.2: The Observed and Predicted values for each compound in Training data set for SVM model

s.no.	Observed values for Training set	Predicted values for Training set	Residual	s.no.	Observed values for Training set	Predicted values for Training set	Residual
1	5.913503	6.008861	-0.09536	51	5.075174	5.197838	-0.12266
2	5.966147	5.871679	0.094468	52	5.347108	5.266451	0.080657
3	8.716044	8.987969	-0.27193	53	5.393628	5.451487	-0.05786
4	8.006368	7.999941	0.006427	54	5.799093	5.796985	0.002108
5	4.49981	4.324463	0.175347	55	5.768321	5.747584	0.020737
6	6.55108	6.629833	-0.07875	56	5.010635	4.899867	0.110768
7	6.55108	6.551173	-9.3E-05	57	5.828946	5.845527	-0.01658
8	8.948976	9.002093	-0.05312	58	5.63479	5.570844	0.063946
9	7.549609	7.487204	0.062405	59	7.727535	7.720714	0.006821
10	9.10498	9.104742	0.000238	60	6.697034	6.673101	0.023933
11	8.81433	8.742282	0.072048	61	8.641179	8.847388	-0.20621
12	10.55059	10.47883	0.071758	62	3.401197	3.385937	0.01526
13	6.975414	6.948016	0.027398	63	6.063785	6.109107	-0.04532
14	8.853665	9.102424	-0.24876	64	6.063785	6.109107	-0.04532
15	10.87993	10.20333	0.676599	65	8.131531	8.235956	-0.10442
16	7.377759	7.320531	0.057228	66	4.70048	4.635291	0.065189
17	8.29405	8.36054	-0.06649	67	9.758462	9.63013	0.128332
18	5.940171	5.921913	0.018258	68	8.207947	8.372285	-0.16434
19	5.768321	5.6596	0.108721	69	6.476972	6.499899	-0.02293
20	6.49224	6.552912	-0.06067	70	6.55108	6.53044	0.02064
21	8.016318	8.027896	-0.01158	71	5.347108	5.149052	0.198056
22	7.682482	7.731025	-0.04854	72	8.748305	8.740374	0.007931
23	6.016157	6.008254	0.007903	73	8.517193	8.529355	-0.01216
24	2.302585	2.760848	-0.45826	74	8.881836	8.870654	0.011182
25	9.277999	9.250166	0.027833	75	6.593045	6.578018	0.015027
26	4.60517	4.385416	0.219754	76	8.101678	8.096342	0.005336
27	6.214608	6.293006	-0.0784	77	6.507278	6.610657	-0.10338
28	5.560682	5.497188	0.063494	78	6.173786	6.241226	-0.06744
29	5.703782	5.666445	0.037337	79	7.244228	7.243825	0.000403
30	6.956545	6.95189	0.004655	80	10.81978	10.47273	0.347051
31	7.31322	7.3178	-0.00458	81	8.632306	8.69004	-0.05773
32	5.799093	5.595679	0.203414	82	3.688879	3.786619	-0.09774
33	2.995732	3.26386	-0.26813	83	4.867534	4.802003	0.065531
34	11.28978	10.52971	0.760073	84	7.21524	7.19422	0.02102
35	5.075174	4.937803	0.137371	85	6.234411	6.179422	0.054989
36	2.995732	3.330748	-0.33502	86	8.794825	8.823131	-0.02831
37	3.688879	3.735659	-0.04678	87	2.397895	2.524644	-0.12675
38	5.347108	5.23812	0.108988	88	8.455318	8.456554	-0.00124
39	6.565265	6.61527	-0.05	89	7.824046	7.815002	0.009044
40	7.863267	7.819319	0.043948	90	7.600902	7.601967	-0.00107
41	3.688879	3.75771	-0.06883	91	8.948976	8.949072	-9.6E-05

42	5.598422	5.565359	0.033063	92	5.521461	5.517782	0.003679
43	5.669881	5.603651	0.06623	93	4.70048	4.768063	-0.06758
44	3.89182	3.918346	-0.02653	94	10.66896	10.59309	0.075867
45	7.60589	7.722418	-0.11653	95	5.438079	5.414608	0.023471
46	8.29405	8.297024	-0.00297	96	8.045588	8.063553	-0.01797
47	5.669881	5.618599	0.051282	97	4.60517	4.509251	0.095919
48	6.684612	6.679241	0.005371	98	3.401197	3.577652	-0.17646
49	5.247024	5.288645	-0.04162	99	6.173786	6.234728	-0.06094
50	7.740664	7.770591	-0.02993	100	3.912023	3.897344	0.014679

TABLE 5.3: The Observed and Predicted values for each compound in Test data set for SVM model

S.no	Observed values for Test set	Predicted values for Test set	residual	S.no	Observed values for Test set	Predicted values for Test set	Residual
1	7.003065	7.162931	-0.15987	15	3.73767	3.825448	-0.08778
2	5.298317	5.053179	0.245138	16	9.249561	9.172455	0.077106
3	5.164786	4.764058	0.400728	17	5.135798	5.025711	0.110087
4	10.63586	10.09866	0.537195	18	7.038784	7.323686	-0.2849
5	10.55059	10.47883	0.071758	19	8.630522	9.15961	-0.52909
6	8.38936	8.669518	-0.28016	20	5.652489	9.044337	-3.39185
7	4.867534	3.850911	1.016623	21	6.194405	6.31785	-0.12345
8	6.659294	6.718832	-0.05954	22	5.703782	5.571212	0.13257
9	4.941642	4.968814	-0.02717	23	9.769956	10.09391	-0.32395
10	7.851661	8.004222	-0.15256	24	1.808289	2.933645	-1.12536
11	6.39693	6.654315	-0.25739	25	6.55108	6.017762	0.533318
12	5.010635	4.077557	0.933078	26	6.39693	6.661641	-0.26471
13	5.438079	5.09744	0.340639	27	11.51293	10.2081	1.304824
14	6.684612	6.788464	-0.10385	28	9.10498	9.692154	-0.58717

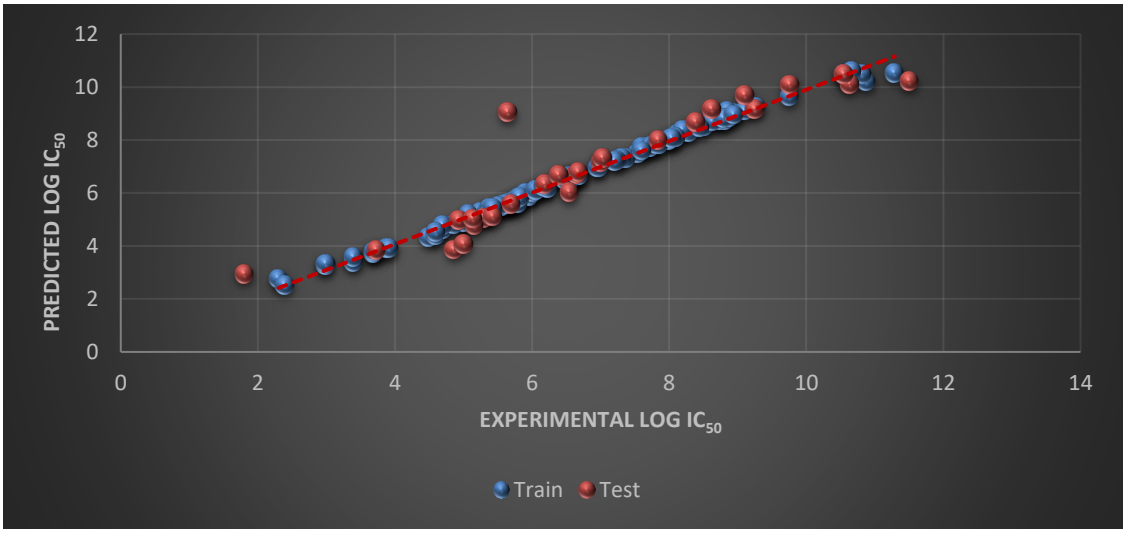

Figure 5.1: The plot of multiple linear regression analysis indicating linear relationship between Experimental and Predicted log IC_{50} with R^2= 0.93

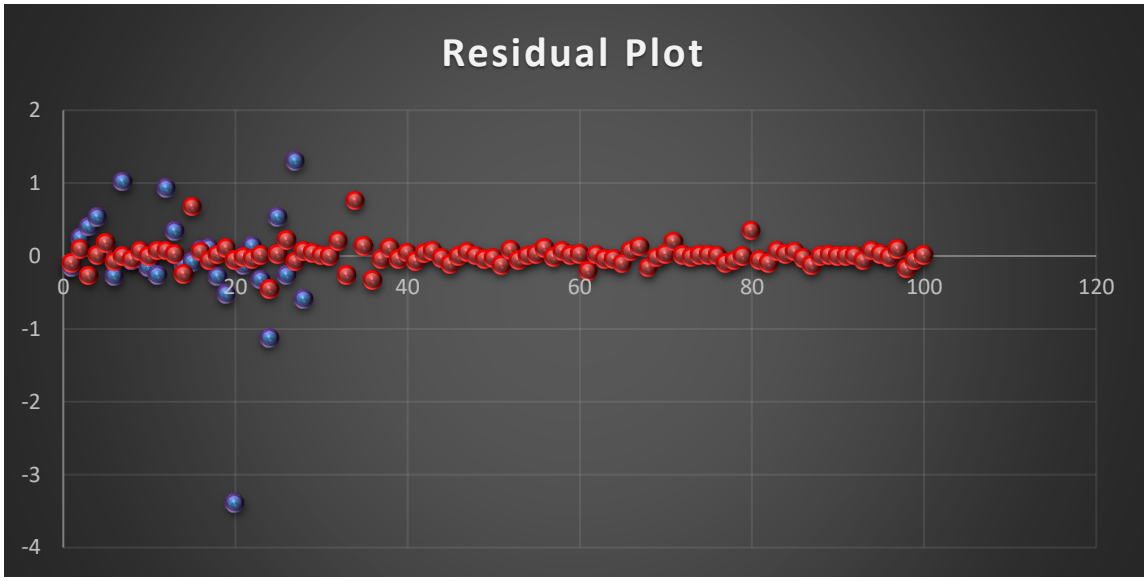

Figure 5.2: The Residual plot for Train data set and Test data set.

From the above residual plot, we conclude that the compounds in test and train set are equally scattered on the marginal line and some outliers are also observed in the data set.

6. CONCLUSION:

Verma & Hansch (2010) concluded that the inhibitory activities of BCL2 analogs against colon cancer diseases are mainly dependent on the steric and hydrophobic descriptors of their substituents, with the significant contribution coming from the molar refractivity of the substituents. As the biological dataset has tremendous non-linearity and hence the linear statistical methods do not behave sufficiently for modeling purposes. From the above study, it is inferred that machine learning techniques may give suitable way for their modeling Moreover, the present work shows that support vector machine along with BCL2 inhibitors for regression modeling provides statistically significant values of R2 = 0.91 and R2cv = 0.93 and the SVR selected descriptors used for SVM model are: "BCUTw.1h", "BCUTc.1l", "C2SP2", "SsCl", "MDEC.22", "MDEO.12", "MDEN.13" and "MDEN.22". The developed model can be efficiently used for virtual screening of unknown Gossypol acetic acid centered functional analogs against the BCL2 target for colorectal cancer (Breast Cancer).

References:

1. Vapnik, V.N., *Statistical Learning Theory*, John Wiley & Sons, New York, (1998).
2. Vapnik, V.N., *The Nature of Statistical Learning Theory*, 2nd ed., Springer, New York, (2000).
3. Cristianini, N., Shawe-Taylor, J., *An Introduction to Support Vector Machines and Kernel based Learning Methods*, Cambridge University Press, (2000).
4. Tay, F.E.H., Cao, L.J., "Application of support vector machines in financial time series with forecasting", *Omega* 29(4) (2001), 309-317.
5. Lee, Y.J., Mangasarian, O.L., "SSVM: A smooth support vector machine for classification", *Computational Optimization and Applications*, 20(1) (2001b), 5–22.
6. Le, Q.V., Smola, A.J., Gärtner, T., "Simpler knowledge-based Support Vector Machines", *ICML'06*, (2006).
7. Hansch, C., A Quantitative Approach to Biochemical Structure-Activity Relationships, Acc. Chem. Res. 2, 232–239 (1969)
8. Cristianini, N., Shawe-Taylor, J., *An Introduction to Support Vector Machines and Kernel based Learning Methods*, Cambridge University Press, (2000).
9. Demiriz, A., Bennett, K., Breneman, C., Embrechts, M., "Support vector machine regression in chemometrics", *Computing Science & Statistics* (2001).
10. Burges, C.J.C., "Geometry and invariance in kernel based methods", In *Advances in Kernel Methods- Support Vector Learning*, Bernhard Scholkopf, Cristopher J.C. Burges and Alexander J. Smola (eds.), MIT Press, Cambridge, MA (1998).
11. Balasundaram, S., Singh, R., "On finite Newton method for support vector regression", *Neural Computing and Applications*, 19(7) (2010), 967–977.

***Corresponding Author:**
Varun Kumar Kashyap*
Email: varun02stat@gmail.com