IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A PARTIAL LEAST SQUARES MODELLING APPROACH TO DEVELOP A QSAR MODEL FOR ANTICANCER ACTIVITY OF GOSSYPOL ACETIC ACID AGAINST BCL2.

[1]Varun Kumar Kashyap , [2]Dr.Rajeev Pandey

[1]Assistant professor, [2]Professor

1. *Department of Community Medicine HIMSAR New Delhi (INDIA)*

2. *Department of Statistics University of Lucknow, Lucknow, UP INDIA.*

Abstract: In recent years statistical modeling technique have grown in length, its application in various fields is inadequate if not lagging behind together. The field of drug designing is not an exception to this. In this article we develop statistical model on the basis of empirical data set has been initiated and QSAR model has been developed by using the BCL2 data set. The paper produces the study to explore a quantitative structure–activity relationship (QSAR) of an especially anticancer Protein cell data set. The method is introduced with anticancer activity of Gossypol acetic acid against BCL2 (anticancer Protein cell) target for colorectal cancer, Breast cancer, Mouth cancer. The model is developed here in this chapter using 80% of 138 virtual sample. Regression coefficient of Partial least squares (PLS) for training set using Leave-one-out (LOO) method cross validation have been computed. The remaining 20% of data set has been used as a test set for the validation of the model proposed. The detection of five influencing factors which have been identified with high degree of statistical efficiency.

Index Terms - Partial least squares (PLS), BCL2, QSAR, Gossypol acetic acid.

1. INTRODUCTION

Cancer is a group of diseases which is basically characterized by uncontrolled growth and spread of abnormal cells. If this spread is uncontrolled, it may result in death. Both external factors (tobacco, infectious organisms, chemicals, and radiation) and internal factors (inherited mutations, hormones, immune conditions, and mutations that occur from metabolism) cause cancer. In India, cancer prevalence is estimated to be around 2.5 million, with over 8, 00,000 new cases and 5, 56, 400 deaths occurring each year due to this disease in the country. Out of 1, 22, 429 study deaths, 7137 were due to cancer, corresponding to 556 400 national cancer deaths in India in 2010. 3, 95, 400 (71%) cancer deaths occurred in people aged 30-69 years (200100 men and 195 300 women). At 30—69 years, the three most common fatal cancers were oral (including lip and pharynx, 45 800 [22·9%]), stomach (25 200 [12·6%]), and lung (including trachea and larynx, 22 900 [11·4%]) in men, and cervical (33 400 [17·1%]), stomach (27 500 [14·1%]), and breast (19 900 [10·2%]) in women. Tobacco-related cancers represented 42·0% (84 000) of males and 18·3% (35 700) of female cancer deaths and there were twice as many deaths from oral cancers as lung cancers.

For this problem, Antiapoptotic BCL2 proteins played a crucial role in the treatment of tumour cell survival & thus, BCL 2 inhibitors have been developed as apoptosis inducers direct. QSARs are mathematical models and are now a days regarded as the best scientifically credible tool for basically predicting and classifying biological activities of unbiased compounds. QSAR has become inexorably embedded as an essential tool in the pharmaceutical industry, from lead discovery, optimization to lead development and computer-aided drug designing. A growing trend is to use QSAR early in the drug discovery process as a screening and enrichment tool to estimate from further development those compounds lacking drug like properties or those chemicals predicted to elicit a toxic response. The fundamental assumption of QSAR is that variations in the biological activity of a series of compounds that target a common mechanism of action are correlated and proceed in some pattern with variations in their structural, physical and chemical properties. 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.

2. MATERIALS AND METHOD:

The ordinary least squares (OLS) estimator of β, $\hat{\beta}_{OLS}$ in the model given by (2.1.) is the solution of the following optimization problem:

$$\hat{\beta}_{OLS} = \arg\max_{b} corr\{Xb, y\}^2 \qquad (2.1)$$

Multi collinearity is inevitable as a result of large number of variables collected by modern technologies of computers, networks, and sensors in many applications of multiple regressions. Despite of having desirable properties, the OLS estimator may have a very large variance and may result in imprecise prediction if the data are multi collinear. Moreover, solution of (2.1) is not unique when $n \leq p$.

One solution to deal with multi collinearity and/or dimensionality problem is regressing the response variable y on a subset of the k orthogonal (latent) vectors stored in a score matrix of size $n \times k$ by which important features of X have been retained. Score matrix is formed by taking linear combinations of columns of X. PLS regression (PLSR) constructs the columns of score matrix, $T = [t1, t2, \ldots, tk]$, by solving the following optimization problem for $h = 1, 2, \ldots\ldots k$ $(k \leq p)$:

$$r_h = \arg\max_{\|r\|=1} cov(Xr, y)^2 = \arg\max_{\|r\|=1} (r'X'yy'Xr) \qquad (2.2)$$

$$\text{Subject to } t_h' t_j = r_h' X' X_{rj} = 0 \; for \; 1 \leq j < h.$$

So, PLSR basically balances the maximal correlation criteria for OLS given in (2.1) with the requirement of explaining as much variability as in both X and y-space.

Various iterative procedures were proposed for solving nonlinear optimization problem in (2.2) such as PLS Mode A, PLS-SB, NIPALS and SIMPLS algorithms which differ by the deflation theme needed for the orthogonality of derived components.

According to **Wold. H. (1975)** who produces a PLS Model algorithm which aims at model existing relationships between variables rather than to model for prediction. PLS-SB generally computes all eigenvectors at once, and the score vectors which are obtained by this method are not necessarily orthogonal.

The most used methods are, NIPALS and SIMPLS, consist of two steps may be called calibration (deriving components) and prediction. These algorithms, are explained in the following subsections, for both univariate and multivariate responses.

In **1989 Wangen, L.E. and Kowalsky,B.R.** give the extension of two-block PLS model, where X and y (or Y for multivariate model) are block variables, to multi-block PLS model is also given in the literature but is not discussed in present research study.

2.1 SIMPLS Algorithm

In **1993 De Jong** discover SIMPLS algorithm which aims to derive PLS components directly in terms of the original data which results in faster computation with less memory requirements and with the ease of interpretation. SIMPLS deflates the cross- covariance matrix, $S_{xy} \propto X'Y$.

SIMPLS algorithm can be summarized as follows:

Step 1: Compute cross-product matrix: $S_{xy}^0 = X'Y$ (X and Y are centered),

Step 2: Repeat steps 2:1 - 2:6 for $h = 1, 2, \ldots\ldots\ldots\ldots, k$:

Step 2.1 : Compute ¯rst left singular vector of S_{xy}^{h-1} as h^{th} PLS weight vector r_h,

Step 2.2 : Compute h^{th} score, $t_h = X_{rh}$, and normalize t_h =: $t_h = t_h / \|t_h\|$,

Step 2.3 : Update h^{th} PLS weight, r_h =: $r_h = r_h / \sqrt{r_h' X' X_{rh}}$,

Step 2.4 : Compute h^{th} x-loading by regressing X on t_h: $p_h = X_{th}'$,

Step 2.5 : Store vectors r_h, t_h, and p_h into matrices $R_h = [r_1, r_2, \ldots\ldots r_h]$,

$T_h = [t, t_2, \ldots\ldots\ldots t_h]$, and $P_h = [p_1, p_2, \ldots\ldots\ldots p_h]$, respectively.

Step 2.6 : h =: $h + 1$ and $S_{xy}^{h-1} = (I_p - V_{h-1}V_{h-1}')X'y$ where columns of V_{h-1} form an orthonormal basis for P_{h-1}.

The orthogonality constraint of components is fulfilled when the PLS weight vector r_h is orthogonal to all previous x-loadings $P_{h-1} = [p_1, p_2, \ldots\ldots, p_{h-1}]$. As a result of this, the h^{th} pair of SIMPLS weight vector r_h for $h = 2, \ldots\ldots\ldots, k$ is obtained as the first left singular vector of S_{xy}^{h-2} which is projection of S_{xy}^{h-2} on a subspace orthogonal to P_{h-1}. Therefore, if the columns of $V_{h-1} = [v_1, v_2, \ldots, v_{h-1}]$ form an orthonormal basis of P_{h-1} obtained by **GramSchmidt** method, then

$$S_{xy}^{h-1} = (I_p - V_{h-1}V_{h-1}')S_{xy}^{h-2} = (I_p - V_{h-1}V_{h-1}')X'Y \qquad (2.1.1)$$

After h components are derived, the data matrix is reduced implicitly to $X(I_p - V_h V_h')$ With SIMPLS algorithm which can be seen from (2.1.1). In PLS1 algorithm, the h^{th} derived component, t_h, is equal to $E_{h-1}w_h$, where w_h is the normalized form of:

$$E_{h-1}'f_{h-1} = X'(I_n - T_{h-1}(T_{h-1}'T_{h-1})^{-1}T_{h-1}')^2 y = X'(I_n - T_{h-1}(T_{h-1}'T_{h-1})^{-1}T_{h-1}')y$$

Therefore, data matrix is reduced explicitly to $(I_n - T_h(T_h'T_h)^{-1}T_h')X$ with PLS1.

2.2. Determining the Optimal Number of Components in PLSR

The decisions related to the optimal number of components, k, are very important issue in building the PLSR model. Although, it is possible to calculate as many components as the rank of the X, it does not make sense in practice. Because data are never noise-free and some of the smaller components will only describe noise. Due to uncertain statistical behavior of PLSR, it is difficult to perform inferential tasks such as assessing the number of components. Consequently, developing as well as comparing PLS component selection rules have been and apparently continue to be subjects of active research in chemometrics. Cross validation, adjusted **Wold's** criterion and randomization test are leading methods that are proposed to seek out the optimum dimensionality of PLS models.

Among the many approaches proposed in the past, the cross-validation (CV) scheme stands out in particular. In M-fold cross-validation, the original sample is partitioned into M sub-samples. Of the M subsamples, a single sub-sample is retained as the *validation set* for testing the model, and the remaining M-1 sub-samples are used as *learning set* for estimating the model. The cross-validation process is then repeated M times (number of folds), with each of M sub-samples used exactly once as the validation set. The M results from the folds then can be combined to produce a single estimate for the optimal number of components. Particularly, the n-fold cross validation ($M = n$), where only one observation is deleted and the process is repeated as many times as samples, is called ***leave-one-out cross validation***. The resulting residual sum of squares, PRESS, is a measure of the predictive power of the components in the model. The PRESS value for h component univariate PLSR using leave-one-out cross validation is:

$$PRESS^h = \sum_{i=1}^{n} (y_i - \hat{y}_{-i(h)})^2 \qquad (2.2.1)$$

where the predicted values $\hat{y}_{-i(h)}$ are based on the parameter estimates that are obtained from the data set which does not include observation *i* using a PLSR model with *h* components. The optimal number of components is the one that yields the minimum PRESS or root mean square error, RMSE,

$$k = \underset{h}{arg\,min}\{PRESS^{(h)}\} = \underset{h}{arg\,min}\{RMSE^{(h)}\} \qquad (2.2.2)$$

Where

$$RMSE_{(h)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_{-i(h)}\right)^2} = \sqrt{\frac{1}{n}PRESS_{(h)}} \qquad (2.2.3)$$

A simple and classical method is the Wold's R criterion **([Wold S 1978])** which compares two successive values of PRESS via their proportion, that is

$$R = \frac{PRESS^{h+1}}{PRESS^{h}} \qquad (2.2.4)$$

where $PRESS^{(h)}$ is given in equation (2.2.1). When R is greater than 1, it is considered that the optimal number of components is *h*. Instead of comparing this ratio to unity, it was proposed by **([Wold S. 1978])** to fix it at 0*:*90 or 0*:*95 which is named *Adjusted Wold's Criteria*. The randomization test ([**Wiklund, S.,Nilsson, 2007 at all])** is a recent method that assesses the statistical significance of each individual component that enters the model. Theoretical approaches to achieve this goal (using a t- or F-test) have been put forth, but they are all based on some assumptions. Randomization test is a data-driven approach and therefore ideally suited for avoiding assumptions.

In 2000 Denham evaluated performances of several mean squared error (MSE) estimation approaches in terms of their accuracy and usefulness in determining the optimal number of components to include PLSR model. It is concluded that all methods perform very compatible for data sets with few variables, while the cross-validation method results in better MSE estimates for the data sets with large number of variables. One area where the method of cross-validation works poorly is design of experiments, where the randomization test should have merit **([Wiklund, S.,Nilsson, 2007 et all]).**

3. Result and discussion:

3.1 *Model development*

Gossypol acetic acid centered functional analogs containing anti-BCL2 activity was collected as an initial data set from NCBI database. Two dimensional molecular descriptors were calculated for each compound for digitization of observational data. Total 255 descriptors were calculated by PaDEL software (National University of Singapore). Which can sufficiently represent the structural characters of molecules.

Initially 255 descriptors were calculated for all compounds. Since, not all of the 255 descriptors contribute to the bioactivity, therefore following measures were taken to eliminate the less informative descriptors: (i) eliminating the descriptors with constant values, (ii) eliminating the descriptors with more than 90% zero values, (iii) eliminating the descriptors which have constant or zero variance.

After this step, highly correlated descriptors were excluded by using the correlation matrix approach. This filtering step includes selection of those descriptors which have correlation coefficient >0.3 (positively or negatively) with bioactivity vector of available datasets. As a result, only 8 descriptors came into existence for further processing.

This matrix based feature reduction was used to reduce the variable space and the chance of correlation between the descriptors. Removal of correlated descriptors reduced the noise from the data and finally we get 106 activity compound and 8 descriptors. The selected descriptors used for modeling are: MDEC.33, MDEO.11, MDEO.12, MDEO.22 and MLFER_S were obtained strongly related to activity and significant. The detail description about descriptors can be accessed from PaDEL descriptors website (http://www.ncbi.nlm.nih.gov).

The coefficient matrix is given below in which the highlighted variable represent that they are significant in model. To develop the model building process with Partial least square (PLS), we use our data set BCL2. Although there are 255 predictors and 128 Activity compounds, in which many of the predictors are highly correlated and the overall information within the predictor space may contained in a smaller number of dimensions. These predictor conditions are more favorable for applying PLS method in our data set. Cross-validation was used to determine the optimal number of components for the PLS model to retain which minimize RMSE. Here we use resampling method to generate a number of PLS regression models. Among these generated models, the model having highest coefficient of determination (R^2) with minimum RMSE value is selected as best. The best PLS model contain eight components with minimum RMSE (0.300). Using these eight components we develop a Partial lest square regression model. The PLS regression coefficients (R^2) for the training set is 0.985. The magnitudes are similar to the linear regression model that includes only those five predictors **MLFER_BH, MLFER_BO, MLFER_S, MLFER_S, nO.** 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 0.968 and RMSE is 0.242 which is very good.

The Observed and Predicted values for each compound in Training data set for PLS model is given by the following table:

S.no	Observed value for Train set	Predicted value for Train set	Residual
1	5.913503	5.940198	-0.02669
2	5.966147	5.954579	0.011568
3	8.006368	8.013408	-0.00704
4	4.49981	4.49936	0.00045
5	6.55108	6.517534	0.033546
6	6.55108	6.551881	-0.0008
7	8.948976	8.916601	0.032375
8	7.549609	7.545118	0.004491
9	5.164786	5.19371	-0.02892
10	10.63586	10.64669	-0.01084
11	9.10498	9.137346	-0.03237
12	8.81433	8.822047	-0.00772
13	10.55059	10.54126	0.009328
14	10.55059	10.54126	0.009328
15	6.975414	7.016315	-0.0409
16	8.853665	8.824177	0.029488
17	7.377759	7.383708	-0.00595
18	8.29405	8.285215	0.008835
19	5.768321	5.776074	-0.00775
20	6.49224	6.488313	0.003927
21	8.38936	8.424053	-0.03469
22	8.016318	8.034321	-0.018
23	6.016157	6.046466	-0.03031
24	4.867534	4.826292	0.041242

s.no.	Observed value for Train set	Predicted value for Train set	Residual
52	6.684612	6.673346	0.011266
53	5.075174	5.104966	-0.02979
54	5.347108	5.334954	0.012154
55	5.799093	5.800542	-0.00145
56	5.768321	5.761515	0.006806
57	7.038784	7.011479	0.027305
58	5.010635	4.978153	0.032482
59	5.828946	5.837563	-0.00862
60	5.63479	5.69657	-0.06178
61	7.727535	7.680441	0.047094
62	3.401197	3.416168	-0.01497
63	8.131531	8.16167	-0.03014
64	4.70048	4.683542	0.016938
65	9.758462	9.756189	0.002273
66	8.207947	8.230691	-0.02274
67	6.476972	6.451713	0.025259
68	5.652489	5.636384	0.016105
69	6.55108	6.506168	0.044912
70	5.347108	5.343238	0.00387
71	8.748305	8.755533	-0.00723
72	8.881836	8.91547	-0.03363
73	6.194405	6.244314	-0.04991
74	6.593045	6.612981	-0.01994
75	8.101678	8.07013	0.031548

25	2.302585	2.302727	-0.00014
26	4.60517	4.625509	-0.02034
27	6.214608	6.195057	0.019551
28	5.560682	5.551681	0.009001
29	6.956545	6.938855	0.01769
30	4.941642	4.940601	0.001041
31	7.31322	7.343754	-0.03053
32	5.799093	5.781238	0.017855
33	2.995732	3.023108	-0.02738
34	11.28978	11.2651	0.024681
35	5.075174	5.068871	0.006303
36	2.995732	3.02298	-0.02725
37	6.39693	6.419093	-0.02216
38	3.688879	3.681743	0.007136
39	5.347108	5.313217	0.033891
40	6.565265	6.608012	-0.04275
41	7.863267	7.88575	-0.02248
42	5.598422	5.590574	0.007848
43	5.669881	5.662419	0.007462
44	3.89182	3.87217	0.01965
45	7.60589	7.588642	0.017248
46	5.438079	5.455808	-0.01773
47	3.73767	3.752167	-0.0145
48	8.29405	8.28325	0.0108
49	9.249561	9.223345	0.026216
50	5.135798	5.131004	0.004794
51	5.669881	5.704243	-0.03436

76	6.507278	6.495103	0.012175
77	6.173786	6.188262	-0.01448
78	10.81978	10.80741	0.012365
79	8.632306	8.620506	0.0118
80	3.688879	3.693353	-0.00447
81	4.867534	4.852991	0.014543
82	7.21524	7.214727	0.000513
83	6.234411	6.242343	-0.00793
84	8.794825	8.819096	-0.02427
85	5.703782	5.69122	0.012562
86	8.455318	8.459963	-0.00465
87	9.769956	9.774368	-0.00441
88	1.808289	1.781124	0.027165
89	7.824046	7.828041	-0.00399
90	7.600902	7.563734	0.037168
91	5.521461	5.522867	-0.00141
92	4.70048	4.709802	-0.00932
93	6.55108	6.594836	-0.04376
94	10.66896	10.68024	-0.01128
95	5.438079	5.42188	0.016199
96	8.045588	8.043419	0.002169
97	6.39693	6.378441	0.018489
98	4.60517	4.58009	0.02508
99	3.401197	3.395543	0.005654
100	11.51293	11.52023	-0.00731
101	9.10498	9.098158	0.006822
102	6.173786	6.14116	0.032626

The Observed and Predicted values for each compound in Test data set for PLS model is given by the following table:

S.no	Observed value for Test set	Predicted value for Test set	residual
1	7.003065	6.974768	0.028297
2	8.716044	8.660108	0.055936
3	5.298317	5.319331	-0.02101
4	10.87993	10.81775	0.06218
5	5.940171	5.915273	0.024898
6	7.682482	7.620655	0.061827
7	9.277999	9.256485	0.021514
8	6.659294	6.646491	0.012803
9	5.703782	5.692473	0.011309
10	7.851661	7.794801	0.05686
11	5.010635	4.935128	0.075507
12	3.688879	3.711841	-0.02296
13	6.684612	6.651376	0.033236

S.no	Observed value for Test set	Predicted value for Test set	Residual
14	5.247024	5.26597	-0.01895
15	7.740664	7.745241	-0.00458
16	5.393628	5.255612	0.138016
17	6.697034	6.673046	0.023988
18	8.641179	8.585824	0.055355
19	8.630522	8.715967	-0.08545
20	6.063785	6.052212	0.011573
21	6.063785	6.052212	0.011573
22	8.517193	8.55671	-0.03952
23	7.244228	7.286803	-0.04258
24	2.397895	2.437982	-0.04009
25	8.948976	9.160479	-0.2115
26	3.912023	3.915889	-0.00387

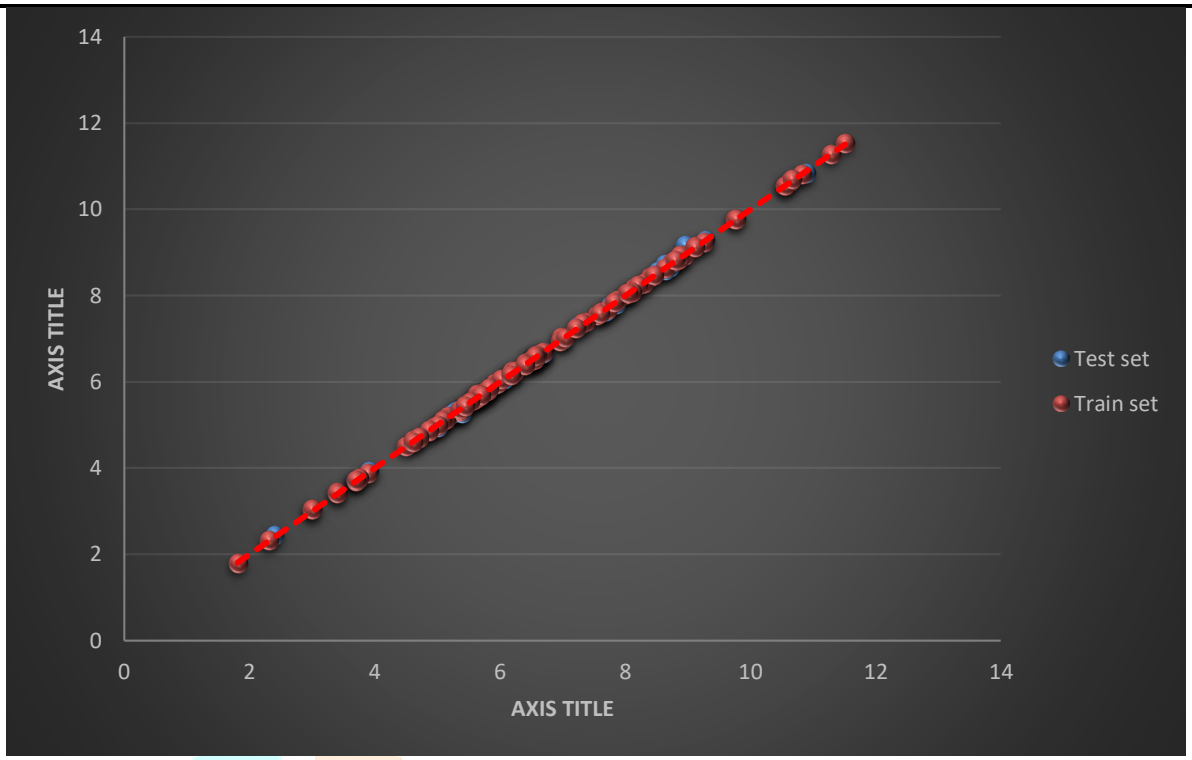

Figure 3.1. Graphical plot of multiple linear regression analysis which indicates linear relationship between experimental and predicted log IC_{50} with r^2= 0.98

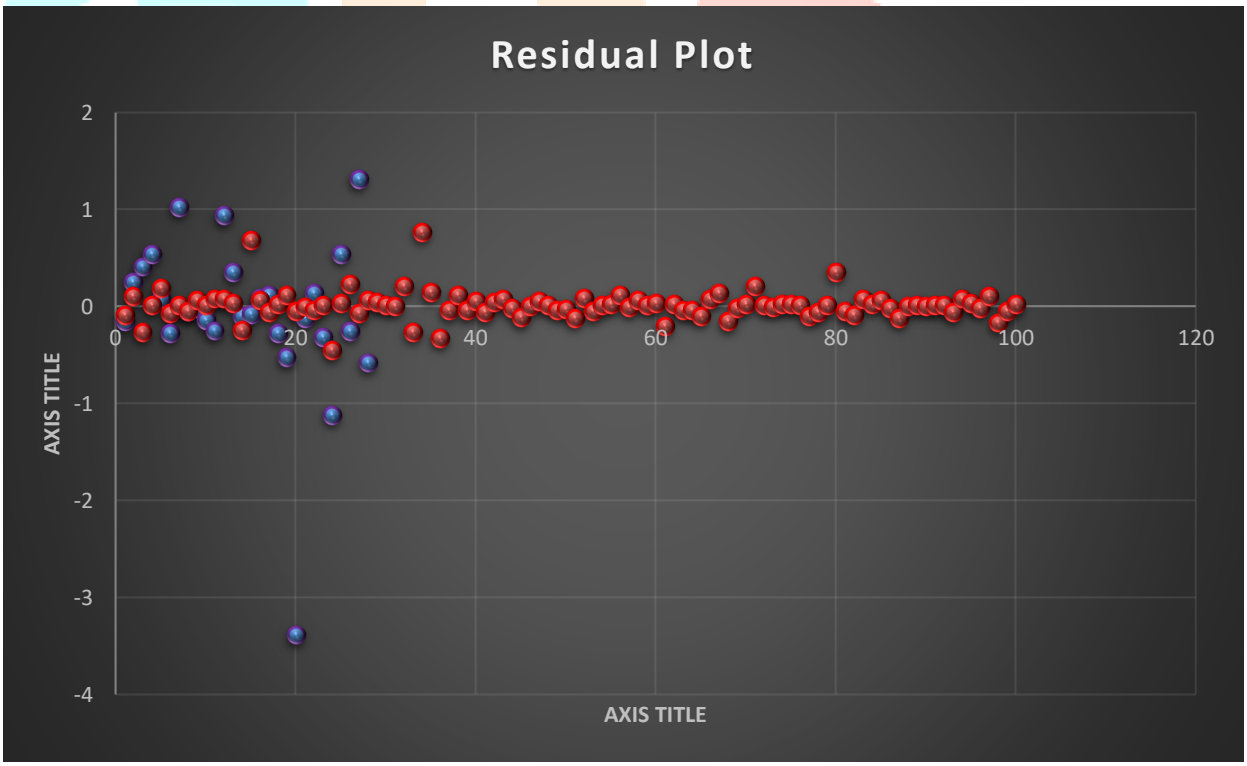

Figure 3.2. Residual plot for Train data set and Test data set.

From the above residual plot, we conclude that the compounds in test and train set are equally scattered on the marginal line and some are very far from the marginal line those are said to be outliers in our data set.

4. Conclusion:

Since the biological dataset has tremendous non-linearity. And the linear statistical methods don't behave sufficiently for modeling purposes. It was assumed that machine learning methods may provide suitable way for their modeling. Therefore, in the present study we attempted with PLS method along with BCL2 inhibitors for regression modeling. It was found that the PLS regression method is statistically sound (R2 = 0.98, R2cv = 0.96) for modeling the biological dataset. The selected descriptors used for PLS model are: "MLFER_BH ", "MLFER_BO ", " MLFER_S ", " MLFER_S " and "nO ". The developed model can be efficiently used for virtual screening of unknown Gossypol acetic acid centered functional analogs against BCL2.

References:

1. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs, [PMC free article] [PubMed], 2003 Aug; 111(10): 1361–1375.
2. Suykens, J.A.K., Brabanter, J.D., Lukas, Vandewalle, J., "Weighted least squares support vector machines: robustness and sparse approximation", *Neurocomputing*, 48 (2002), 85-105.
3. Suykens, J.A.K., Vandewalle, J., Moor, B.D., "Optimal control by least squares support vector machines", *Neural Networks*, 14(1) (2001), 23-25.
4. Damle, C. (2003). Flood Forecasting Using Time Series Data Mining. PhD thesis, Engineering Department of Industrial and Management Systems Engineering College of Engineering University of South Florida.
5. Suykens, J.A.K., Vandewalle, J., "Least squares support vector machine classifiers", *Neural Processing Letters*, 9(3) (1999), 293-300.
6. Breiman, L. and Friedman, J., "Predicting Multivariate Responses in Multiple Linear Regression", *Journal of the Royal Statistical Society, B*, 59, 3-54, 1997.
7. De Jong, S., "PLS Shrinks", *Journal of Chemometrics*, 9, 323-326, 1995.
8. De Jong, S., "PLS Fits Closer than PCR", *Journal of Chemometrics*, 7, 551-557, 1993.
9. Wangen, L.E. and Kowalsky,B.R., "A Multiblock Partial Least Squares Algorithm for Investigating Complex Chemical Systems", *Journal of Chemometrics*, 3, 3-20, 1989.
10. Geladi, P. and Kowalski, B. R., "Partial Least Squares Regression: A Tutorial", *Analytica Chimica Acta*, 185, 1-17, 1986.
11. Wold H., "Path Models with Latent Variables: The NIPALS Approach", *Quantitative Sociology: International perspectives on mathematical and statistical model building*,Academic Press, 307-357, 1975.
12. Hansch, C., A Quantitative Approach to Biochemical Structure-Activity Relationships, Acc. Chem. Res. 2, 232–239 (1969)