QSAR study for the prediction of IC50 and Log P for 5-N-Acetyl-Beta-D-Neuraminic Acid structurally similar compounds using stepwise (multivariate) linear regression

Ponmary Pushpa Latha D.* and Jeya Sundara Sharmila D.
*Department of Computer Applications, Karunya University, Coimbatore, Tamil Nadu, India, 641114, ponmarymca@gmail.com
Department of Bioinformatics, Karunya University, Coimbatore, Tamil Nadu, India, 641114, djssharmila@gmail.com

Abstract- Multi-parametric Quantitative structure activity relationship (QSAR) study has been developed for 110 training compounds and 50 test compounds structurally similar to 5-N-ACETYL-BETA-D-NEURAMINIC ACID as inhibitors for *Clostridium tetani*. Stepwise (multi-parametric) Linear Regression QSAR models for biological activity of half maximal inhibitory concentration (IC50) and log P for octanol/water (Log P) were created with 16 different descriptors. The predictive capability of the QSAR models were evaluated by r2 , q2 LMO(TestSet) , q2LOO(TestSet) , q2BOOT(TestSet). The comparison of various external validation reveals identical q2 LMO(TestSet) , q2 LOO(TestSet) and q2 BOOT(TestSet) for IC50 (0.98), and Log P(0.7) which demonstrates the high robustness and real predictive power of IC50 and Log P model.
LMO-Leave many out, LOO-Leave one out, BOOT- bootstrapping
Keywords: neurotoxins, 5-N-ACETYL-BETA-D-NEURAMINIC ACID, Compound, Linear Regression, QSAR, *Clostridium tetani*

1. Introduction

Quantitative structure-activity relationship (QSAR) describes how a known biological activity can differ as a function of molecular descriptors derived from the chemical structure of a set of molecules. Many physiological activities of a molecule can be associated with their composition and structure. Molecular descriptors, which are numerical depictions of the molecular structures, are used for performing QSAR analysis. 5-N-ACETYL-BETA-D-NEURAMINIC ACID represents the most important class of biologically-active compounds as inhibitors of *Clostridium tetani*[1-2]. The half maximal inhibitory concentration (IC50) is the concentration of an inhibitor that is necessary for 50-percent inhibition of an enzyme in vitro. Octanol-water partition coefficient logP is used in QSAR [3-8] studies and rational drug design as a measure of molecular hydrophobicity. Hydrophobicity affects drug absorption, bioavailability, hydrophobic drug-receptor interactions, metabolism of molecules, as well as their toxicity. LogP [9] has become also a key parameter in studies of the environmental fate of chemicals. Through the review of literature, it was found that a number of QSAR [10-16] studies were developed. In the present study, QSAR studies have been carried out for 5-N-ACETYL-BETA-D-NEURAMINIC ACID and its structurally similar compounds with (>95%). We have developed the following QSAR [18-19] models for 5-N-ACETYL-BETA-D-NEURAMINIC ACID and its structurally similar compounds with (>95%) using stepwise (multivariate) linear regression method by the Poly analyst [21] software: IC50, Log P.

2. Methods

2.1 Data Set

Training set of 110 compounds and test set of 50 compounds related to 5-N-ACETYL-BETA-D-NEURAMINIC ACID "Fig. (1)" which is available in *Clostridium tetani* were collected from pubchem[20]. Table I shows the different regression summary of IC50 and Log P model Training Set.

2.2 Molecular Descriptors

Theoretical molecular descriptors are calculated using QikProp [17] program (Schrodinger, 2008a). The following descriptors are procured into consideration for developing the model: 1. Molecular Weight (MW), 2. Hydrophobic SASA (HAS), 3. Hydrophilic SASA (HLSA), 4. Molecular Volume (MV), 5. vdW Polar SA (PSA), 6. Number of Rotatable Bonds (RB), 7. Donor - Hydrogen Bonds (DHB), 8. Acceptor - Hydrogen Bonds (AHB), 9. Ionization Potential (IP), 10. Electron Affinity (EA), 11. log P for octanol/water (Log P), 12. log S for aqueous solubility (AS), 13. Human Oral Absorption (HOA), 14. Lipinski rule (LR). 15. Half maximal inhibitory concentration (IC50). 16. Number of Ring (NR).

2.3 Stepwise (Multivariate) Linear Regression methods

PolyAnalyst [21] is used to develop the model using stepwise (multi-parametric) linear regression algorithm. It automatically determines the most influencing attributes by considering subsets of the attributes selected, and then includes the most significant attributes in the final analysis. The process of creating the regression model is incremental in polyanalyst. At first, only

one-dimensional models were selected. Then the most accurate model was picked up. After that, all two-dimensional models produced by adding a new attribute to the first model were experimented. Again the best model was chosen and supplemented by a new attribute. All the considered regresion models must pass a test based on the value of Fisher statistics [19] of all their regression coefficients. If a term has the value of F-ratio less than a specified threshold, this term is removed from the model. Thus, the process of adding new terms stops either when all attributes are included in the model or when no new term can be added without violating the F-ratio criterion. The above process permits PolyAnalyst to find all influential attributes, but at the same time, it includes in the model only those attributes which do not correlate significantly with each other. When two powerful correlating attributes are tried in the same model, they both would have low values of F-ratio, and the second attribute is removed. Randomized testing is used in PolyAnalyst. While PolyAnalyst searches for the best regression model fitting the explored dataset, it solves the same problem also for several randomized datasets. Randomized datasets are prepared from the original dataset using the random permutation of the target attribute values for various records. Thus, all the values of independent and target variables remain the same, but the relation between independent and dependent attributes is broken. Only if the accuracy obtained for the real data is much higher than any randomized data then the created regression model can be considered as reliable and significant. Otherwise, the system concludes that it does not have enough data to create a reliable model. The degree of certainty is that the discovered model is not merely a result of a random statistic fluctuation in the data expressed by Significance. Randomized testing was done by the poly analyst software which is not visibile to the user. For example, in IC50 model, the influential attributes are MW, HAS, DHB and Log P. The best equation is produced by the system based on correlation coefficient, square of correlation coefficient r2. The square of correlation coefficient r2 values are given in brackets as follows: for IC50 (0.8545), Log P (0.797). The present calculated r2 values are closer to 1.0, which is a clear evidence for the best fit regression. In addition to that, PolyAnalyst performs careful significance testing on the result, comparing its significance with other models generated. q2 is calculated using the following formula. yi is the actual experimental activity, Where, the average actual experimental activity and the predicted activity of compound re computed by the predicted model. The robustness and internal predictivity of the models were evaluated by both leave-one-out cross validation (q2LOO(TestSet)) and leave-many-out cross validation (q2LMO(TestSet)). The in-house computer programs are created in Java programming to do the following cross validation techniques: Leave-some-out, Leave-one-out and Bootstrapping. In Leave-many-out, the data set was split into the sequence of six set of compounds (45,40,35,30,25,20) and the cross validation was performed. The average of q2LMO was calculated as follows: IC50(0.981896), Log P(0.734322). Leave-one-out cross validation is as follows:

1. Assign Total Compound n=50, Compound i=1
2. Leave Compound i
3. Calculate q2i
4. i=i+1
5. Repeat step 2 to 5 till i<=50
6. Find the average of q2i=1..n q2 LMO(TestSet) for IC50 and Log P Model are 0.981896 and 0.734322 respectively. Bootstrap cross validation is computed as follows:

1. Generate n random number Ri within the range of 1 to 50 where i=1..n
2. i=1
3. Remove Ri Compounds
4. Calculate q2i
5. i=i+1
6. Repeat step 3 to 5 till i<=n
7. Find the average of q2i=1..n

The average of q2BOOT was calculated as follows: IC50(0.98402102), Log P(0.78033).

3. Results and Discussion

Table IV describes the observed and predicted value of IC50 and Log P models. The studied data set comprises a total of 110 training compounds and 50 test compounds structurally highly similar (>95% similarity) to that of 5-N-ACETYL-BETA-D-NEURAMINIC ACID as inhibitors for *Clostridium tetani* neurotoxins. The objective of this study is to propose a stepwise multivariate linear QSAR models for IC50 prediction applicable to the 5-N-ACETYL-BETA-D-NEURAMINIC ACID structurally similar compounds. Several QSAR equations for training set of 110 compounds were obtained using different F-ratio value. From those equations, the best QSAR equation is generated with a selected F-ratio cut off value greater than 17, which resulted in four descriptors for IC50 model and five descriptor for Log P model. The IC50 provides good statistical measures such as correlation coefficient, standard deviation and standard error as 0.8545, 0.2932 and 0.3815 respectively (Table I). The External test set data allows the verification of the proposed models by statistical validation like q2 LMO(TestSet) , q2 LOO(TestSet) and q2 BOOT(TestSet). The statistical external validation was performed by

checking the models with 50 compounds, which did not participate to the model development. The comparison of various external validation reveals the identical q2 LMO(TestSet) , q2 LOO(TestSet) and q2 BOOT(TestSet) values for IC50 (0.98), and Log P(0.7) which demonstrates the high robustness and real predictive power of IC50 and Log P model. Thus the following equations (1) and (2) has good performance in prediction of the test set compounds. Individual model descriptions are documented below.

3.1 IC50 model

The proposed model is based on the four descriptors MW, HAS, DHB and Log P (1). The Graph of experimental verses the predicted values for the present IC50 model is displayed in "Fig. (2)". The training compound in this study shows the range of Log P between -4.777 to -1.342 , However, the range of HAS is between 193.267 and 374.856 with high F-ratio (>187). The regression coefficient of MW, DHB and Log P are calculated to be positive as 0.0153477, 0.424796 and 1.32025 respectively. This shows the increase in the above properties supporting inhibitory activity of IC50. Hence, three out of four selected descriptors influence positively the predicted IC50 values. On the other hand, HAS(-0.0227897) is correlated negatively to the affinity. Table II depicts the regression coefficient, mean and F-ratio of various parameters which were involved in building the QSAR model of IC50. Table IV describes the q2 LMO(TestSet), q2 LOO(TestSet) and q2 BOOT(TestSet) values of IC50 model. Since the values are greater than 5, The QSAR model may be considered.

Prediction rule

IC50 = +0.0153477*MW -0.0227897*HAS +0.424796*DHB +1.32025*LogP ----------(1)

3.2 Log P model

The proposed model is based on the five descriptors (2). The plot of experimental verses the predicted values for the present Log P model is shown in "Fig. (3)". The training compound in this study shows the range of IC50 between -3.511 to -0.135 , However, HLSA range is flanked by 244.665 and 407.958 with high F-ratio (>90). The regression coefficient of HAS and IC50 is calculated to be positive as 0.00709923 and 0.399870 respectively. This implies that the increase in the above properties will support in favor of inhibitory activity of Log P. Therefore, two out of five selected descriptors influence positively the predicted Log P values. HLSA(-0.00817571), IP(-0.162002) and AS(-0.473430) is associated negatively to the affinity. Table III shows the regression coefficient, mean and F-ratio of various parameters which were involved in building the QSAR model of LogP. Table IV describes the q2 LMO(TestSet), q2 LOO(TestSet) and q2 BOOT(TestSet)values of Log P model. Since the values are above 5, The QSAR model is acceptable.

Prediction rule:

LogP = +0.00709923*HAS -0.00817571*HLSA +0.399870*IC50 -0.162002*IP -0.473430*AS ----(2)

3.1 Discussion

The proposed model IC50(1) and Log P(2), which are having positive values in the regression coefficient point out that the designated descriptors supplly positively to the value of IC50 (1) and Log P (2), In other words, negative values indicate that greater the value of the descriptors, lower the value of IC50 (1) and Log P (2). In IC50 Model, raising the MW, DHB and Log P, boosts the IC50 values. Amplifying HAS will dwindle the IC50 values. In IC50 model, the presence of molecular descriptors with positive coefficient in IC50 model is manifested in the compounds 2 and 3. The occurrence of molecular descriptor with negative coefficient at HAS increases, the IC50 of drug compounds decreases. The presence of molecular descriptors with negative coefficient at HAS increases, the IC50 of drug component decreases, This is evident in the compounds 2 and 3. In Log P model, the presence of molecular descriptors with positive at HAS and IC50 increases, and the Log P of drug component decreases. This is apparent in the compounds 32 and 33. Stepwise multivariate linear regression analysis provides constructive equation (1) that can be used to predict the IC50 of compounds based upon MW, HAS, DHB and Log P parameters. The Log P (2) delivers the useful equation that can be used to predict the Log P of compounds based upon HAS, HLSA, IC50, IP and HAS. The results of both the models indicate that a strong correlation exists between Log P and IC50 for drug compounds. This procedure allows us to carry out a precise and reasonably fast method for fortitude of Log P and IC50 of 5-N-ACETYL-BETA-D-NEURAMINIC ACID and its structurally similar compounds with (>95%) and also it predicts the IC50 and Log P values with sufficient accuracy.

Conclusion

In this study it was possible to obtain a stepwise multivariate linear regression QSAR model of IC50 and Log P for a set of one hundred and ten compounds which are 95% structurally similar to 5-N-ACETYL-BETA-D-NEURAMINIC ACID as inhibitors for *Clostridium tetani* neurotoxins.The LOO, LMO and BOOT cross validation techniques show that the model is significant, robust and has good predictability. IC50 model is determined by MW, DHB, Log P and HAS. The most important parameter by considering the regression coefficients of IC50 model are as follows: Log P, HAS. The Log P for octanol/water

model is calculated by HAS, IC50, HLSA, IP and AS. The most important parameter by taking into account the regression coefficients of Log P model are as follows: IC50, HAS. IC50 and Log P models are showing minimum deviation between observed and predicted values and also having good internal and external predictive power.

References

[1] Berman H.M., Westbrook J., Feng Z., Gilliland G. Bhat T.N., Weissig ., Shindyalov I.N. and Bourne P.E. (2000) *Nucleic Acids Res.*, 28, 235–242.

[2] John Foster, Patricia Kane and Neal Speight. (2002) *Milville, NJ..*

[3] Castilho C., Guido R.V. and Andricopulo (2007) *Lett. Drug Des. Discov.,* 4, 106-113.

[4] Papa E., Dearden J.C.and Gramatica P. (2007) *Chemosphere,* 67, 351–358.

[5] Eduardo Borges de Melo and Ma´rcia Miguel Castro Ferreira (2009) *European Journal of Medicinal Chemistry*, 44(9), 3577-3583.

[6] Hazuda D. J., Felock P., Witmer M., Wolfe A., Stillmock K., Grobler J. A., Espeseth A., Gabryelski L., Schleif W., Blau C. and Miller M. D. (2000) *Science*, 287, 646-650.

[7] Jae Yoon Chung et al., (2009) *Arch Pharm Res*, 32(3), 317-32.

[8] Kamalakaran Anand Solomon, Srinivasan Sundararajan and Veluchamy Abirami (2009) *Molecules*, 14, 1448-1455

[9] Hansch C., Leo A. and Fujita T. (1964) *J. Am. Chem. Soc.*, 86, 1616-1626.

[10] Liu A., Gunag H., Zhu L. and Du G. (2007) *Sci. China Life Sci.*, 50, 726-730.

[11] Pasha F.A., Nam K.D. and Cho S. J. (2007a) *Mol. Cell. Toxicol*, 3(145-149).

[12] Pasha F. A. Srivastava H. K., Srivastava, A. and Singh, (2007b) *P. P. QSAR Comb. Sci*, 26, 69-84.

[13] Pasha F. A., Neaz M. M., Cho S. J., and Kang S. B. (2007c) *Chem Biol Drug Des*, 70, 520-529.

[14] Pasha F. A., Chung H. W., Cho S. J. and Kang S.B. (2008) *Int. J. Quant. Chem.,* 108, 391-400.

[15] Recanatini M., Cavalli A. and Hansch C.A. (1997) *Chem. Biol. Interact*. 105, 199-228.

[16] Robert et al., (2009) *J. Med. Chem*, 52(3), 737-754.

[17] Schrodinger (2008) *LLC, New York, NY*.

[18] Shen L.L., Liu G.X. and Tang Y. (2007) *Acta Pharmacol. Sin*, 28(2053-2063).

[19] Stevan Hadzivukovic and Nikolic-Doric Emilija (2005) *International Statistical Institute, 55th Session .*

[20] http://pubchem.ncbi.nlm.nih.gov/.

[21] www.Megaputer.com

[22] http://www.answers.com/topic/molecular-mass

Table I- The regression summary of different models IC50 Model

Model	r^2	$q^2_{LMO(TestSet)}$	$q^2_{LOO(TestSet)}$	$q^2_{BOOT(TestSet)}$	Significance Index	Standard deviation	Standard error	Training Set Minimum	Training Set Maximum
IC50	0.8545	0.981896	0.9828707	0.98402102	115.4	0.2932	0.3815	-3.511	-0.135
Log P	0.797	0.734322	0.7757151	0.78033	123.7	0.1937	0.4506	-4.777	-1.342

Table I- The regression summary of different models IC50 Model

name	coef.	std dev.	F-Ratio
MW	0.01535	0.001392	121.5
HAS	-0.02279	0.001099	430.1
DHB	0.4248	0.05195	66.86
LogP	1.32	0.09631	187.9

Table III- Training Set Parameter of Log P Model Observed vs Predicted Values

name	coef.	std dev.	F-Ratio
HAS	0.007099	0.001005	49.86
HLSA	-0.00818	0.000861	90.27
IC50	0.3999	0.04143	93.16
IP	-0.162	0.03783	18.34
AS	-0.4734	0.08886	28.38

Table IV- Observed vs Predicted Values of Various Model

Compound	**Obs IC50**	**Pred IC50**	**Obs LogP**	**Pred Log P**
1	-0.71	-0.71394	-2.017	-2.03
2	-0.71	-0.71389	-2.017	-2.03001
3	-0.367	-0.38894	-2.433	-2.46259
4	-0.367	-0.38933	-2.433	-2.4629
5	-0.144	-0.25931	-2.36	-2.32861
6	-0.366	-0.38723	-2.462	-2.48058
7	-0.507	-0.56883	-2.455	-2.40077
8	-0.507	-0.56897	-2.455	-2.40058
9	-0.355	-0.33108	-2.476	-2.49282
10	-0.355	-0.33106	-2.476	-2.49284
11	-2.724	-2.49898	-2.313	-2.29248

12	-2.411	-2.20872	-2.294	-2.33594
13	-2.724	-2.48023	-2.328	-2.32171
14	-2.437	-2.28093	-2.288	-2.31379
15	-0.474	-0.3588	-2.515	-2.60967
16	-0.476	-0.36571	-2.513	-2.68957
17	-0.52	-0.33565	-2.543	-2.65582
18	-0.539	-0.39537	-2.534	-2.62859
19	-0.526	-0.57306	-2.46	-2.42844
20	-0.526	-0.57309	-2.46	-2.42844
21	-2.688	-2.56066	-2.263	-2.15425
22	-2.468	-2.24777	-2.301	-2.25771
23	-2.711	-2.56175	-2.281	-2.20955
24	-2.067	-2.2558	-2.088	-1.91466
25	-0.521	-0.42852	-2.524	-2.57914
26	-0.521	-0.42852	-2.524	-2.57914
27	-0.485	-0.40821	-2.515	-2.54658
28	-0.485	-0.40825	-2.515	-2.54655
29	-2.396	-2.40706	-2.131	-2.0012
30	-2.396	-2.40706	-2.131	-2.0012
31	-0.539	-0.61231	-2.442	-2.4861
32	-2.713	-2.65524	-2.235	-2.16741
33	-0.526	-0.57304	-2.46	-2.42845
34	-0.526	-0.57306	-2.46	-2.42844
35	-2.695	-2.52809	-2.267	-2.16699
36	-2.626	-2.47035	-2.264	-2.23321
37	-0.373	-0.3624	-2.469	-2.47965
38	-0.373	-0.3624	-2.469	-2.47965
39	-0.539	-0.61272	-2.442	-2.48651
40	-0.252	-0.39993	-2.34	-2.29238
41	-0.524	-0.44793	-2.514	-2.62012
42	-0.524	-0.44791	-2.514	-2.62013
43	-0.524	-0.44791	-2.514	-2.62013
44	-0.216	-0.38829	-2.338	-2.27099
45	-0.433	-0.41422	-2.488	-2.55169
46	-0.433	-0.41422	-2.488	-2.55169
47	-2.408	-2.13646	-2.305	-2.28278
48	-2.408	-2.13646	-2.305	-2.28278
49	-2.408	-2.13646	-2.305	-2.28278
50	-0.946	-0.95278	-1.995	-1.95987

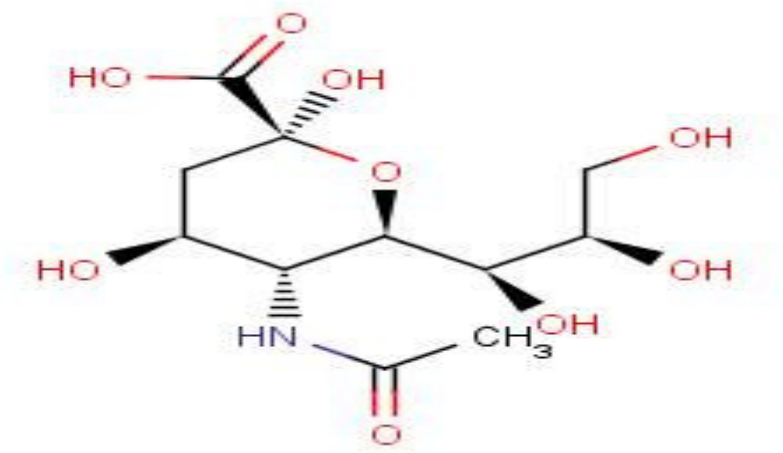

Fig. 1 5- N-Acetyl-Beta-D-Neuraminic Acid

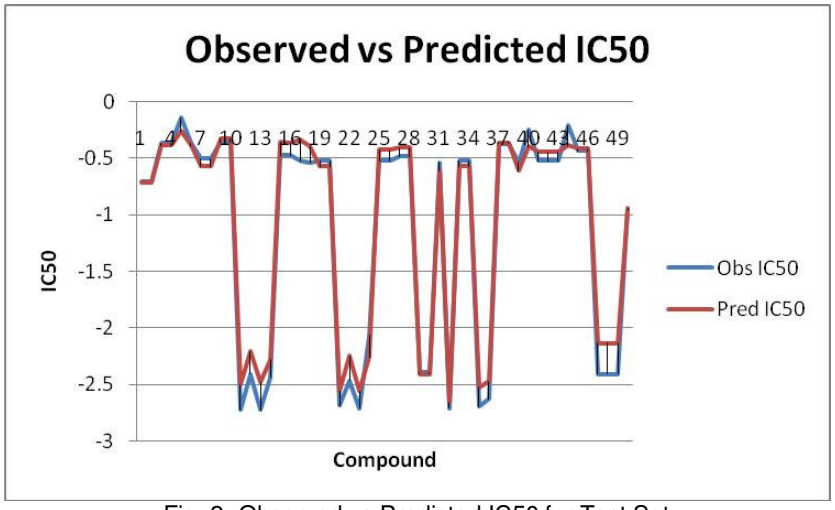

Fig. 2- Observed vs Predicted IC50 for Test Set

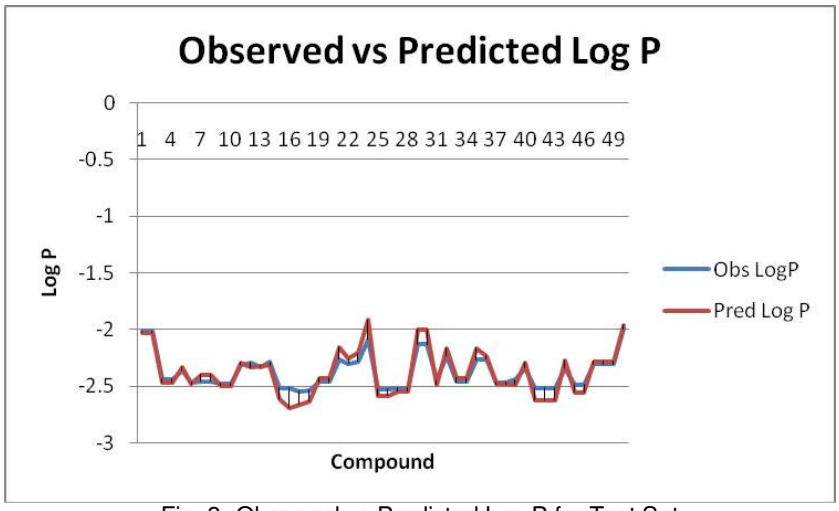

Fig. 3- Observed vs Predicted Log P for Test Set

