EJMA

DOI: 10.14744/ejma.2023.91300
EJMA 2023;3(4):164–169

Research Article

Development of a QSAR model for BACE-1 inhibitors using genetic algorithm-based multiple linear regression

Sumanta Kumar Sahu, Sweta Singh, Krishna Kumar Ojha

Department of Bioinformatics, Central University of south Bihar, Gaya, Bihar, India

Abstract

Objectives: BACE-1 (β-enzyme) is the main therapeutic target for the treatment of Alzheimer's disease, as it actively participates in the processing of amyloid precursor protein, resulting in the creation of amyloid-β in the brain. The current work aims to investigate and build a QSAR model of BACE-1 inhibitors.

Methods: Genetic algorithm-based multiple linear regression (GA-MLR) was used to create regression models between the descriptor and pIC_{50} value of each molecule in the training set based on selected significant molecular descriptors. The most important descriptors chosen are Burden modified eigenvalue descriptors, PaDEL-weighted path descriptors, autocorrelation descriptors, topological distance matrix descriptors, MLFER descriptors, Barysz matrix descriptors, and chi path cluster descriptors. The models were validated using both internal and external validation parameters.

Results: The study determines the chemical space that the model may predict by defining an applicability domain. The regression models developed suggest a good predictive model for BACE-1 inhibitors that can predict the IC50 value of newly designed chemical compounds.

Conclusion: The information presented here suggests a good predictive model for BACE-1 inhibitors, which can be utilized to predict the IC_{50} value of newly designed chemical compounds, thereby aiding in the treatment of Alzheimer's disease.

Keywords: Alzheimer, BACE1, QSAR.

Cite This Article: *Kumar Sahu S, Singh S, Kumar Ojha K. Development of a QSAR model for BACE-1 inhibitors using genetic algorithm-based multiple linear regression. EJMA 2023;3(4):164–169.*

Alzheimer's disease (AD) is a severe neurological disorder characterized by memory loss and cognitive decline. It is often described as a mental eraser, causing people to forget their loved ones, friends, and even aspects of their own identity.[1] This disease affects elderly people around the world and places a significant financial and emotional burden on their families, resulting in a in their quality of life and contributing to social instability. The economic impact of AD is substantial, with at least 35 million people suffering from the disease worldwide, resulting in annual costs of up to $200 billion. Unfortunately, the number of people affected by AD is increasing exponentially, indicating a growing public health concern.[2,3]

The development of drugs for the effective treatment of AD is a primary focus for researchers.[4] While several medications have been tested in clinical trials, none have been successful in reducing the impact of AD. Therefore, discovering effective treatments remains crucial. AD is characterized by two primary pathological features: insoluble neurofibrillary tangles (NFT) created in cells by the tau protein and senile plaques (SPs) caused by

Address for correspondence: Sumanta Kumar Sahu, MD. Department of Bioinformatics, Central University of south Bihar, Gaya, Bihar, India
Phone: +919937209313 **E-mail:** sumanta@cusb.ac.in

Accepted Date: July 19, 2023 **Available Online Date:** July 11, 2024

the miss-aggregation of extracellular amyloid-β (Aβ) peptides. The formation of SPs leads to cell toxicity and brain dysfunction in patients, which contributes to the severity of the disease. Researchers are diligently working to find a solution that can effectively target these features and mitigate the impact of AD.[6] Therefore, a desirable and potential approach the development of AD therapeutics has been clinical intervention to lower Aβ levels in the brain.

Amyloid precursor protein (APP) is an essential transmembrane protein found in biological tissue that is primarily expressed in the brain and is necessary for normal functioning.[2] Generation of Aβ, a hallmark of AD, begins with the initial cleavage of membrane APP by membrane-anchored aspartic protease BACE-1, also known as secretase. A second cleavage at the C-terminus called β-secretase produces the matured Aβ, highlighting the need for secretase activity in the production of Aβ. However, β-secretase also performs various physiological tasks related to cell growth, and it is unclear whether inhibiting its ability to produce Aβ will have any adverse effects on these crucial processes.[7] Therefore, current medication research for AD focuses on reducing the expression of β-secretase or limiting its secretion, which is one of the primary strategies employed. In recent years, hundreds of articles and patents have been written BACE-1 inhibitors, yet current treatments for AD only stop cognitive loss, and the underlying disease process remains unknown.[1]

In scientific research, a quantitative structure-activity relationship (QSAR) is an important concept that links a molecule's physical, chemical, and biological activity.[8] To represent the various physicochemical properties of a chemical structure, numerical values called descriptors are used as independent variables, while the IC_{50} value serves as the dependent or response variable. Numerous studies have shown the successful screening of compounds for biological activity through the use of QSAR models.[9-11] In this study, a QSAR model of BACE-1 inhibitors was developed using a genetic algorithm-based multiple linear regression (GA-MLR) approach and selected relevant descriptors.

In previous studies, several QSAR and pharmacophore models were developed to predict the activity values of Alzheimer's disease inhibitors and to analyze specific scaffold mechanisms using a limited number of molecules. However, in this study, a total of 249 compounds were used to build a QSAR model. The increased number of compounds used in the model contributes to greater accuracy and reliability of the predictions.

Materials and Methods

Data Collection

This study used a dataset of 249 BACE-1 inhibitors from the literature.[12,13] All structures were drawn using the Marvin ChemAxon tool (https://chemaxon.com/marvin), cleaned and saved in MDL (.mol) format before descriptor calculation. The compound structures were carefully examined before performing any descriptor calculations. The primary objectives of this research were to identify the structural requirements for inhibiting the BACE1 enzyme and to forecast the activity of untested chemicals against the BACE1 enzyme.

Preliminary Dataset Preparation and Data Curation

During pre-processing, missing values are removed from the data set. As IC_{50} values are in the micro molar range, we convert them to pIC_{50} (-log IC_{50}) since higher values indicate greater potency. All of the descriptors, including pIC_{50}, are used as independent variables in the analysis. To ensure consistency, the molecular descriptors are scaled and normalized, with all features falling within the range of 0 to 1.[14]

In the initial step, any parameter that could not be calculated for any compound in the data set was removed, and descriptors with zero values for all compounds were eliminated.[8] To avoid redundancy and the impact of collinearity, a correlation matrix was created with a cutoff value of 0.9. Variables that displayed exact linear dependencies between subsets of the variables and multicollinearity were excluded from the analysis to avoid high multiple correlations between subsets of the variables.[15]

To select the most significant descriptors for the biological activity value, a systematic search was conducted that followed a series of tests, including missing value and zero tests, as well as eliminating descriptors that displayed multi-collinearity or exact linear dependencies. A genetic algorithm (GA) was then applied to determine the best descriptors for the model. This approach, known as MLR-GA, has been widely used in the literature as an effective search technique for selecting descriptors for QSAR modeling based on the evolutionary principles of biological systems.[14,16-19] The genetic algorithm (GA) is an evolutionary approach for variable selection inspired by natural evolution. In this study, we utilized GA with specific parameters including 500 initial equations, 100 iterations, 7 descriptors per equation, 0.3 mutation probability, and selection of the top 30 equations based on mean absolute error-based criteria.[20]

QSAR Model Building

The MLR model correlates the IC_{50} values with the descriptors and minimizes the difference between experimental and predicted biological activities. Regression analysis

Table 1. Contains the detail descriptors selected for the QSAR model building

Name	Details	Class	Type
nF10Ring	Number of 10-membered fused rings	Ring count descriptor	2D
minsssCH	Minimum atom-type E-State: >CH-	Electrotopological state atom type descriptor	2D
minHBd	Minimum E-States for (strong) Hydrogen Bond donors	Electrotopological state atom type descriptor	2D
MDEN-33	Molecular distance edge between all tertiary nitrogens	MDE descriptor	2D
MDEO-22	Molecular distance edge between all secondary oxygens	MDE descriptor	2D
AATSC8m	Average centered Broto-Moreau autocorrelation - lag 8 / weighted by mass	Autocorrelation descriptors	2D
topoRadius	Topological radius (minimum atom eccentricity)	Topological descriptor	2D

uses descriptors to determine IC_{50} as a dependent variable, while MLR analysis expands this approach to incorporate multiple variables. The models were developed using DTC-QSAR v1.0.5 and the straightforward MLR method GA-selected variables.[21]

To evaluate the QSAR models developed in this study, several statistical parameters were used, including N, K, and R^2. Furthermore, Q^2, pred R^2, and the F-test were used to determine the statistical significance of the results, as well as the correlation coefficient between the experimental and predicted values.[22] If the regression equation explains the variation in the experimental activity of the data set, the regression coefficient R^2 quantifies it. A QSAR model is considered predictive if it satisfies the following criteria: $R^2>0.6$, $Q^2>0.6$, and pred $R^2>0.5$.[15] The F-test is a measure of how much of the variance in the data is explained by the model, and it varies with the regression error. According to the estimates of the high F-test, the model is statistically significant. Furthermore, the low standard error of Q^2, pred R^2, and projected R^2 indicates that the model is highly reliable.[22]

Validation of the QSAR Model

The cross-validation technique was employed to test the internally validated QSAR equation. This method provides more insight into the expected reliability of the QSAR equation. In this study, the leave-one-out cross-validation method was used to validate the model. Additionally, to address the potential issue of increased inaccuracy as model complexity rises, the adjusted R^2 was also defined.[22,23] The validation of a model using a test set is an important step in evaluating its internal and external performance. To achieve a better QSAR model with strong predictive power, Golbraikh and Tropsha proposed certain statistical properties for the test set that should be met.[22]

I. $R^2_{pred} > 0.6$

II. $(r^2 - r^2_0)/r^2 < 0.1$

III. $0.85 < k < 1.15$ or $0.85 < k' < 1.15$

Here, r^2 represents the squared correlation coefficient between observed and predicted activities, $r^2 0$ represents the squared correlation coefficient between predicted and observed activities, r^2 represents the squared correlation coefficient between predicted and observed activities, and k and k' represent the regression slopes passing through the origin.[19]

Applicability Domain (AD)

The accuracy of any QSAR model depends on the accuracy of predictions made by unique compounds. The chemical structure space of molecules in the training set is defined by the AD of a QSAR model. In this study, Roy and Kar's[21] standardization method was used to define the AD. The optimal scenario for the descriptors in the training set is that they exhibit a normal distribution pattern. This is because approximately 99.7% of the population is expected to fall within three standard deviations (SD) from the mean. If the standardized descriptors of a compound exceed ±3 SD, it could be an outlier in the training set or be outside the AD in the test set. Therefore, it is crucial to ensure that the descriptors in the training set are in a normal distribution pattern to increase the precision of the predictions made by any QSAR model.

Result and Discussion

QSAR model and Validation

In-silico QSAR analysis selecting descriptors based on the genetic algorithm, a multi-linear regression model was developed containing fifteen optimum descriptors. The final selected MLR-GA model is:

$PIC_{50}=7.5587 + 0.5993(nF10Ring) - 2.303(minsssCH) - 6.7375(minHBd) + 1.6899(MDEN\text{-}33) + 0.80770(MDEO\text{-}22) - 0.0478(AATSC8m) - 0.207\ (topoRadius)$

Number of Training set data points: 171

Number of features selected in the model: 7

Internal Validation metrics: $R^2=0.9016$, R^2(Adjusted)=0.8974, Standard Error of Estimation (SEE)=0.7963, Q^2(LOO)=0.8904,

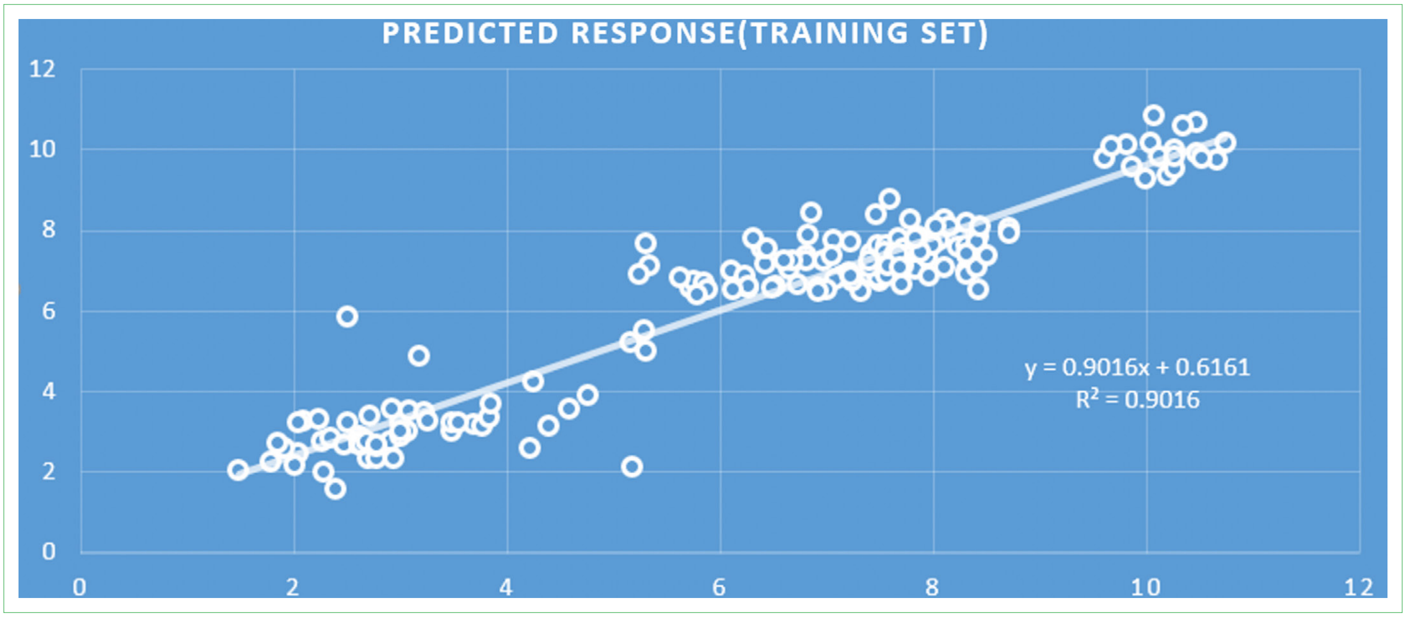

Figure 1. Defines the graph of the actual versus predicted activities of training set, with statistical parameters that support predictive ability of the model.

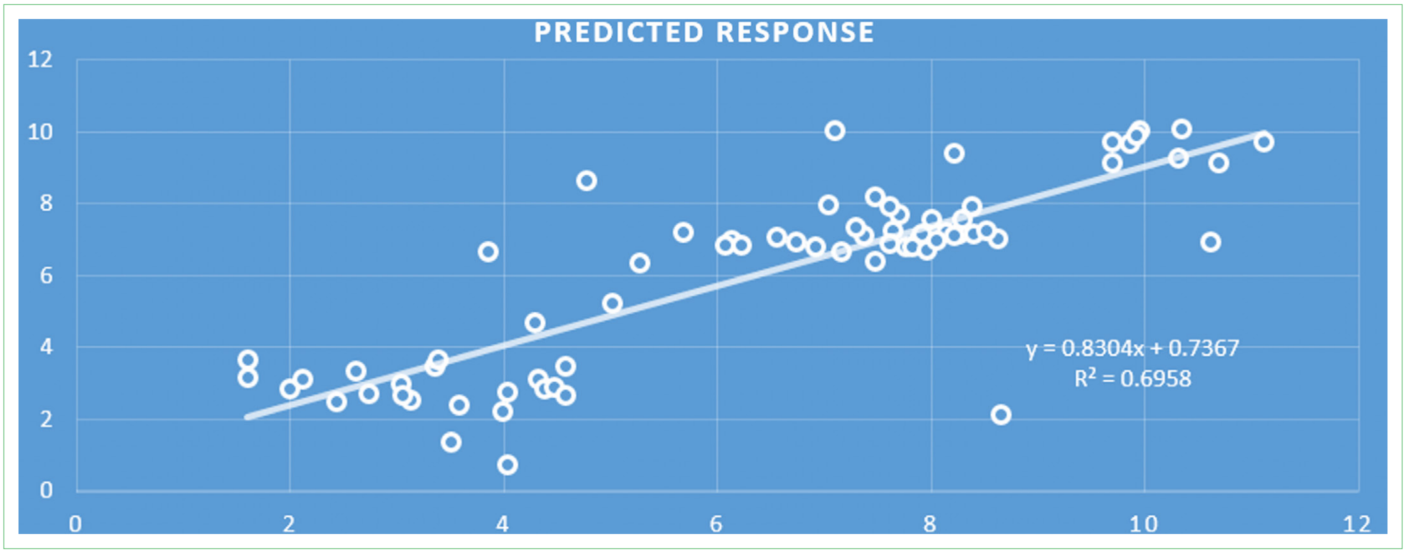

Figure 2. Defines the graph of actual verses predicted activities of test set compounds, with statistical parameters in support of predictive ability of the model.

SDEP(LOO)=0.8207, Scaled average Rm^2(LOO)= 0.8448, Scaled delta Rm^2 (LOO)=0.0787, Mean Absolute Error(-MAE) =0.6106 and External Validation metrics using a test set: Number of Test set data points: 72, Q^2(F1) Test=0.6518, Q^2(F2) Test=0.6503, Scaled average Rm^2(Test)=0.5932, Scaled delta Rm^2(Test)=0.0859, CCC (Test)=0.826, Mean Absolute Error (MAE, Test)=1.066

From the above model, it can be deduced that the 7 most significant descriptors contained the RingCountDescriptor, ElectrotopologicalStateAtomTypeDescriptor, autocorrelation descriptor and MDEDescriptor, the details are presented in Table 1.

The values of R^2 train=0.90 and R^2 test=0.69 confirm the good extrapolation between the training and test sets of data. Furthermore, the QSAR model is reliable due to the small variation between R^2 and Q^2 value. The actual and predicted activity value comparison of training and test set data is presented in Figure 1 and Figure 2.

A successful machine learning model should be able to generalize well from the training set of data. Only 7 of the best descriptors were chosen from a total of 1400 produced descriptors. The entire data set was divided into a training set (70%) and a test set (30%) at random. [24] Additionally, only compounds from the training set

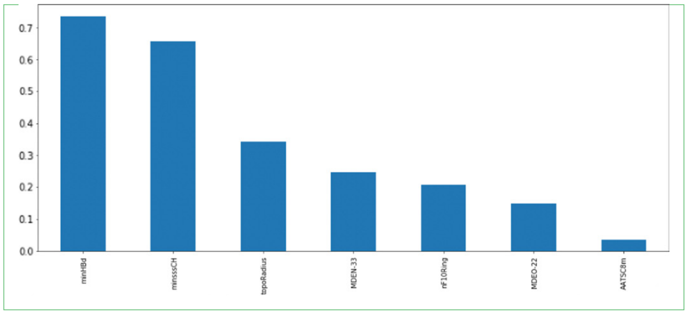

Figure 3. Feature importance plot for the built model.

are used in all calculations. Applying a GA to choose the most significant descriptors toward the biological activity value eliminated a systematic search conducted in the order of missing value test, zero tests, multilinearity, and descriptors.

Here, the model was built using freely available tools, built using 249 molecules of scaffolds taken Feature importance analysis by mutual information was used to evaluate the relative importance and contribution of each descriptor to the model (Fig. 3).[25]

Except for six compounds (63, 247, 248, 250, 257, and 268), a standardized approach to the range of AD, defined all of the compounds of the training set present within the AD. difference between observed and predicted values is small. As a result, these compounds could be regarded as influential in model performance rather than outliers being removed from the training data set. Similarly, compounds 17, 223, and 238 appear outside the AD, but the majority of the test set compounds present within the AD demonstrate confidence within the defined AD.

Conclusion

In this study, a QSAR model was built using the MLR-GA method to predict the IC_{50} of an unknown chemical compound as BACE-1 inhibitors using the data from the training The model is built using a set of 249 compounds that bind to BACE-1 and were collected from the literature. Seven optimal descriptors were chosen from the set of 1400 descriptors as having a significant impact on the value of biological activity value. The internal and external predictabilities of the model created using training and test sets are validated by cross-validation of the model (LOO), Troposha's metrics, and Rm^2 metrics. R^2 train=0.99, R^2 adjusted =0.89, and R^2 pred=0.69 for the chosen MLR-GA model. The accuracy in making predictions within the chemical domain for which it was built is further demonstrated by the evaluation of AD. The built model can help to find new inhibitors from a large database and can be used to design novel inhibitors.

Disclosures

Ethics Committee Approval: The study was approved by the Local Ethics Committee.

Peer-review: Externally peer-reviewed.

Conflict of Interest: None declared.

Authorship Contributions: Concept – S.S., S.K.S.; Design – S.K.S., K.K.O.; Supervision – K.K.O.; Materials – S.S., S.K.S.; Data Collection and/or Processing – S.S., S.K.S.; Analysis and/or Interpretation – S.S., S.K.S.; Literature Research – S.S., S.K.S.; Writing – S.S., S.K.S.; Critical Review - S.S., S.K.S.

References

1. Dzamba D, Harantova L, Butenko O, Anderova M. Glial Cells - The Key Elements of Alzheimer´s Disease. Curr Alzheimer Res. 2016;13(8):894–911.
2. Jiang C, Li G, Huang P, Liu Z, Zhao B. The Gut Microbiota and Alzheimer's Disease. J Alzheimers Dis. 2017;58(1):1–15.
3. Serrano-Pozo A, Growdon JH. Is Alzheimer's Disease Risk Modifiable? J Alzheimers Dis. 2019;67(3):795–819.
4. Calsolaro V, Edison P. Neuroinflammation in Alzheimer's disease: Current evidence and future directions. Alzheimers Dement. 2016 Jun;12(6):719–32.
5. Goyal M, Dhanjal JK, Goyal S, Tyagi C, Hamid R, Grover A. Development of dual inhibitors against Alzheimer's disease using fragment-based QSAR and molecular docking. Biomed Res Int. 2014;2014.
6. Lyketsos CG, Carrillo MC, Ryan JM, Khachaturian AS, Trzepacz P, Amatniek J, et al. Neuropsychiatric symptoms in Alzheimer's disease. Vol. 7, Alzheimer's & dementia : the journal of the Alzheimer's Association. United States; 2011. p. 532–9.
7. Ávila-Villanueva M, Gómez-Ramírez J, Ávila J, Fernández-Blázquez MA. Loneliness as Risk Factor for Alzheimer´s disease. Curr Aging Sci. 2022 Aug;15(3):293–6.
8. Ahmadi S, Habibpour E. Application of GA-MLR for QSAR Modeling of the Arylthioindole Class of Tubulin Polymerization Inhibitors as Anticancer Agents. Anticancer Agents Med Chem. 2017;17(4):552–65.
9. Kumar V, Ojha PK, Saha A, Roy K. Exploring 2D-QSAR for prediction of beta-secretase 1 (BACE1) inhibitory activity against Alzheimer's disease. SAR QSAR Environ Res [Internet]. 2020;31(2):87–133. Available from: https://doi.org/10.1080/1062936X.2019.1695226
10. Santoshi S, Naik PK, Joshi HC. Journal of Biomolecular Screening. 2011.
11. Idakwo G, Luttrell IV J, Chen M, Hong H, Gong P, Zhang C. A Review of Feature Reduction Methods for QSAR-Based Toxicity Prediction. Challenges Adv Comput Chem Phys. 2019;30:119–39.
12. Huang D, Liu Y, Shi B, Li Y, Wang G, Liang G. Comprehensive 3D-QSAR and binding mode of BACE-1 inhibitors using R-group search and molecular docking. J Mol Graph Model [Internet].

2013;45:65–83. Available from: http://dx.doi.org/10.1016/j.jmgm.2013.08.003

13. Clarke B, Demont E, Dingwall C, Dunsdon R, Faller A, Hawkins J, et al. BACE-1 inhibitors Part 1: Identification of novel hydroxy ethylamines (HEAs). Bioorg Med Chem Lett. 2008 Feb 1;18(3):1011–6.
14. Alisi IO, Uzairu A, Abechi SE, Idris SO. Quantitative structure activity relationship analysis of coumarins as free radical scavengers by genetic function algorithm. Phys Chem Res. 2018;6(1):208–22.
15. Alexander Tropsha. Best Practices for QSAR Model Development, Validation, and Exploitation. Mol Inform. 2010.
16. Sahu SK, Ojha KK, Singh VK. Development of Predictive in Silico Cytotoxic Activity Model to Predict the Cytotoxicity of a Diverse Set of Colchicine Binding Site Inhibitors. Eurasian J Med Oncol. 2022;6(2):172–81.
17. Flores MC, Márquez EA, Mora JR. Molecular modeling studies of bromopyrrole alkaloids as potential antimalarial compounds: a DFT approach. Med Chem Res [Internet]. 2018;27(3):844–56. Available from: http://dx.doi.org/10.1007/s00044-017-2107-3
18. Umar AB, Uzairu A, Shallangwa GA, Uba S. QSAR modelling and molecular docking studies for anti-cancer compounds against melanoma cell line SK-MEL-2. Heliyon [Internet]. 2020;6(3):e03640. Available from: https://doi.org/10.1016/j.heliyon.2020.e03640
19. Khan MF, Verma G, Akhtar W, Shaquiquzzaman M, Akhter M, Rizvi MA, et al. Pharmacophore modeling, 3D-QSAR, docking study and ADME prediction of acyl 1,3,4-thiadiazole amides and sulfonamides as antitubulin agents. Arab J Chem [Internet]. 2019;12(8):5000–18. Available from: http://dx.doi.org/10.1016/j.arabjc.2016.11.004
20. Leardi R, Boggia R, Terrile M. Genetic algorithms as a strategy for feature selection. J Chemom [Internet]. 1992;6(5):267–81. Available from: https://10.0.3.234/cem.1180060506
21. Roy K, Kar S, Ambure P. On a simple approach for determining applicability domain of QSAR models. Chemom Intell Lab Syst [Internet]. 2015;145:22–9. Available from: http://dx.doi.org/10.1016/j.chemolab.2015.04.013
22. Alexander DLJ, Tropsha A, Winkler DA. Beware of R2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. J Chem Inf Model. 2015;55(7):1316–22.
23. Golbraikh A, Tropsha A. Beware of q2! J Mol Graph Model. 2002 Jan 1;20(4):269–76.
24. Tyagi C, Grover S, Dhanjal JK, Goyal S, Goyal M, Grover A. Mechanistic insights into mode of action of novel natural cathepsin L inhibitors. BMC Genomics. 2013;14(SUPP 8):1–12.
25. Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. 2004;066138(June):1–16.