JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Validation of QSAR Models Using External Statistical Approach in Drug Discovery

Sameer Dixit*[1] and Arun K Sikarwar[2]

Lecturer, Associate Professor
[1]Department of Chemistry, M. J. P. Govt. Polytechnic College Khandwa,
Madhya Paradesh (INDIA)

Abstract : Quantitative structure-activity relationship (QSAR) is a statistical modelling approach mostly used in molecular modelling, property and activity prediction of new molecule, drug discovery and prediction of environment. For a best QSAR model Validation of model is an important step. Validation is a step to check the reliability and robustness of QSAR models. The use of statistical approaches in drug discovery is of more crucial due to several reasons. These approaches help in the efficient and effective identification and development of new drugs by providing robust, reliable, and quantitative methods to analyze complex biological data. In this paper some key points highlighting the significance of statistical approaches in drug discovery.

KEYWORDS : QSAR, Molecular descriptors, correlation coefficient. PRESS, SPRESS, R^2_{cv}, PSE, LSE

I. INTRODUCTION

Validation of Quantitative Structure-Activity Relationship (QSAR) models is an extremely important step in the drug discovery process, ensuring that these models are reliable and applicable for predicting the biological activity of new compounds. External validation, in particular, involves using a dataset that was not involved in the model-building process to test the model's predictive performance. This helps in assessing the generalizability of the model to unseen data.

Quantitative structure-activity relationships (QSARs) are statistically derived models based on descriptors, used to predict the physicochemical and biological (also include toxicological properties) properties of molecules from the knowledge of chemical structure. The structural features and properties are encoded within descriptors in numerical form, sometime number or matrices or graph. Descriptors support application of statistical tools generating relations which correlate activity data with descriptors (properties) in quantitative trend. The description of QSAR models has been a topic for scientific research for more than 40 years and a topic within the regulatory framework for more than 20 years[1]. In the field of QSAR, the main objective is to investigate these relationships by building mathematical models that explain the relationship in a statistical way with ultimate goal of prediction and/or mechanistic interpretation. QSARs are being applied in many disciplines like drug discovery and lead optimization, risk assessment and toxicity prediction, regulatory decisions and agrochemicals[2,3,4]. One of the major applications of QSAR models is to predict the biological activity of untested compounds from their molecular structures[5]. The estimation of accuracy of predictions is a critical problem in QSAR modeling[6].

Quantitative Structure-Activity Relationship (QSAR) is based on the hypothesis that changes in molecular structure reflect changes in the observed response or biological activity. The success of any quantitative structure–activity relationship model depends on the accuracy of the input data, selection of appropriate descriptors, statistical tools and the validation of the developed model. Validation is a crucial aspect of QSAR modelling. Validation is the process by which the reliability and significance of a procedure are established for a specific purpose. Hence in this review we focus on the importance of validation of QSAR models and different methods of validation.

II. MATERIAL AND METHODS

It is worthy to mention that the QSAR models viz. excellent statistics may not have excellent predictive power. It there are necessary to investigate predictive power of all the models discussed above. This can be done by cross validation method. As opposed to traditional regression method, Cross-validation evaluates the validity of the model by how well it predicts data rather than how well it fit data. The analysis uses a "Leave-one-out" in that model built in with N-1 compound and then N^{th} compound is predicted. Each compound is left-out of the model derivation and predicted in turn. A cross-validation study becomes very important in order to check on the application of the weights in the samples even from the same population since this procedure requires two different samples; a double cross validation study is commonly made in which each sample serves in turn of deriving weights and for testing weights.

It is interesting to mention that QSAR should be evaluated according to its ability to predict the activity and property of the molecule which were not used in the original QSAR analysis obviously, such as an evaluation can be done using cross-validation technique in

that each step-wise regression molecule are randomly or in turn excluded from the QSAR analysis. The QSAR equation is then calculated and used to predict the activity of these 'n' molecules this process is continued till the activity/property prediction of all the molecules is completed. The key statistical measures of a QSAR equations predictive ability are the followings.

1. Predictive Sum of Squares (PRESS) -

Predictive Residual Error Sum of Squares (PRESS) is calculated using the following expression

$$PRESS = \sum (Y_{est}-Y_{obs})^2$$

Where Y_{est} and Y_{obs} are the estimated and observed value of **activity or property** respectively
So The predicted values are calculated for the old-out dataset using a regression model trained on the remainder of the dataset.

$$PRESS = \sum_{i}^{N} \left(x_{pred,i} - x_{obs,i}\right)^2$$

2. Sum of Squares of the Response values (SSY) -

Sum of the squares of the response values (SSY) is calculated using the following expression

$$SSY = \sum (Y_{mean}-Y_{obs})^2$$

Ymean- mean of activity or property

Y_{obs}- obs activity or property

3. Overall Predicted ability (R^2cv or Q^2) –

The overall predictive ability which is sometimes called cross validate correlation coefficient is calculated using the following expression

$$r^2cv = R^2cv = Q^2 = SSY\text{-}PRESS/PRESS$$

4. Uncertainty of prediction (SPRESS) –

$$S_{PRESS} = (PRESS/n\text{-}k\text{-}1)^{1/2}$$

Where k is number of variables (topological indices) in the proposed model and n is the number of compounds

5. Predictive Square Error (PSE) –

Predictive Square Error – it is yet another cross-validation parameter used when uncertainty of prediction (S_{PRESS}) coincides with the standard error of estimation (Se) and is given by

$$PSE = (PRESS/n)^{1/2}$$

6. The Cross-validation correlation coefficient is the cross-validation equivalent of R^2 –

$$R^2 = 1 - \frac{PRESS}{\sum_i^N \left(x_{obs,i} - \bar{x}_{obs,i}\right)^2}$$

Where, x is the mean of variable x.

7. Adjusted R^2 or R^2adj* Adjustable R^2A –

magnitude of adjustable R i.e. R^2A values. This R^2A define by the following expression.

$$R_{adj}{}^2 = 1 - (1 - r^2)\frac{N-1}{N-P-1}$$

Where r is the correlation coefficient, N is the number of data sets (compounds) and P is the number of descriptors used in the model. In case of multiple regression analysis the multiple correlation coefficients R is used in place of r.

8. Pearson Correlation Coefficient (R) –

The Pearson correlation coefficient (R) is a measure of the correlation of two variables x and y

$$R = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

Where cov(x, y) is the covariance between variable x and y.

9. Q FACTOR –

It is worthy to mention that a model (regression equation) with excellent statistics may not necessarily have excellent predictive power. Thus, the next step of regression analysis is to examine predictive power of the proposed model this can be easily done by calculating Poglianis quality factor Q. This quality factor is defined as the ratio of correlation coefficient (r) to the standard error 'Se' (standard deviation 'sd').

$$Q = r / sd$$

Thus the higher the value of r (R) and the lower the value of sd the bigger will be the Q and the better will be the predictive power of that model.

10. Least Squares Error (LSE) –

This parameter LSE is calculated by summing the square of the residue thus we have

$$LSE = \sum (Residue)^2 = \sum (X_{obs}-X_{cal})^2$$

Where X_{obs} and X_{cal} are the observed n calculated value (property or activity) respectively.

11. Probable error of the coefficient of correlation (P.E.) –

The Probable error of the coefficient of correlation (P.E.) is an interesting parameter used in QSAR in deciding whether or not the proposed correlation is good or not. This parameter as stated in last chapter V is calculated using the following expression

$$PE = 2/3(1 - \frac{r^2}{\sqrt{n}})$$

Where r is the correlation coefficient and n is number of observation i.e. the number of compound used.

III. RESULT AND DISCUSSION

i.Predictive Sum of Squares (PRESS) -

The goal of predictive modeling is to accurately predict outcomes for new, unseen data. A lower PRESS value indicates that the model's predictions are closer to the actual observed values for these new data points. Therefore, a lower PRESS value implies better prediction accuracy. When comparing multiple models, analysts typically prefer the model with the lowest PRESS value. A lower PRESS value suggests that the model performs better in terms of prediction accuracy and generalization.

ii.Sum of Squares of the Response values (SSY)

When building regression models for prediction, rather than explanation, the absolute magnitude of SSY may be less important. Instead, what matters is how well the model is able to predict new observations. Models with high SSY may still be effective for prediction if they are able to capture the underlying patterns in the data and make accurate predictions.

PRESS and SSY:

PRESS is a good estimation of the real prediction error of the model, provided that the observations (compounds) were independent. If PRESS is smaller than the sum of squares of the response value (SSY), the model indicates better than chance and can be considered "statistically significant". The ratio PRESS/SSY can be used also to calculate approximate confidence intervals of prediction of new observations (compounds). To be reasonable QSAR model PRESS/SSY should be smaller than 0.4 and the value of this ratio smaller than 0.1indicates an excellent model.

iii.Overall Predicted ability (R^2_{cv} or Q^2) –

The overall predicted ability, often denoted as R^2_{cv} or Q^2, is a crucial metric in Quantitative Structure-Activity Relationship (QSAR) modeling. This metric is widely used to evaluate the predictive performance of QSAR models, particularly in the context of cross-validation. Here are the primary uses of R^2_{cv} or Q^2 in QSAR. It used in Model Validation, Avoiding Overfitting, Comparative Assessment of models. A high Q^2 value indicates that the model performs well on unseen data, reflecting its predictive power. It helps in identifying overfitting. If a model shows a high R^2_{cv} on the training data but a low Q^2, it suggests that the model might be overfitting to the training data and not generalizing well. Q^2 allows for the comparison of different QSAR models. Models with higher Q^2 values are considered to have better predictive capabilities, facilitating the selection of the most robust model. By analysing Q^2, researchers can optimize feature selection, ensuring that only the most relevant descriptors are included in the model.

In summary, R^2_{cv} or Q^2, is a fundamental metric in QSAR modelling, essential for validating, comparing, optimizing, and interpreting models. It ensures that the models developed are both accurate and reliable, providing confidence in their predictive abilities.

iv.Uncertainty of prediction (SPRESS) –

In Quantitative Structure-Activity Relationship (QSAR) modelling, the uncertainty of prediction, often quantified by the Standard Error of Prediction Sum of Squares (SPRESS), plays a significant role. SPRESS provides a measure of the uncertainty associated with the model's predictions and offers insights into the model's reliability and robustness. Here are the key uses of SPRESS in QSAR: it helps researchers to Model Evaluation and Model Comparison. It used in model evaluation, model comparison, Risk Assessment of model. SPRESS quantifies the uncertainty in the model's predictions. A lower SPRESS value indicates that the model's predictions are closer to the actual observed values, suggesting higher accuracy and reliability. It helps in assessing the robustness of the model. Robust models will have lower SPRESS values, indicating that the predictions are consistent and dependable across different datasets. SPRESS can be used to compare the performance of different QSAR models. Models with lower SPRESS values are preferred as they indicate more precise predictions with less uncertainty. By providing a measure of prediction uncertainty, SPRESS helps in quantifying the risk associated with the model's predictions. This is particularly important in high-stakes applications such as drug discovery and environmental risk assessment.

SPRESS is an essential metric in QSAR modeling that quantifies prediction uncertainty. It is used for model evaluation, comparison, optimization, validation, risk assessment, and regulatory compliance. By providing a measure of prediction accuracy and reliability, SPRESS helps ensure that QSAR models are robust and dependable for practical applications.

v.Predictive Square Error (PSE) –

Predictive Square Error (PSE) is a metric used to assess the predictive performance of regression models, particularly in the context of cross-validation. It is calculated based on the difference between the predicted values and the observed values for the test data during cross-validation. PSE is computed during cross-validation procedures, where the dataset is divided into training and testing subsets multiple times. During each iteration, the model is trained on the training subset and then used to predict the values of the test subset. The PSE is then calculated based on the squared differences between the predicted and observed values for each data point in the test subset. PSE provides a measure of how well the model's predictions match the observed values in the test data. Lower values of PSE indicate that the model's predictions are closer to the actual values, suggesting better predictive accuracy. Models with lower PSE values

are generally considered to have better predictive ability. PSE can also be used to guide the optimization of model parameters. By adjusting the model parameters and observing the resulting changes in PSE, researchers can identify the parameter settings that lead to the best predictive performance.

In summary, Predictive Square Error (PSE) is an essential metric in QSAR modeling for evaluating, comparing, optimizing, and validating models. It quantifies the prediction errors, helping to ensure that QSAR models are accurate, reliable, and robust for practical applications.

vi. The Cross-validation correlation coefficient is the cross-validation equivalent of R^2 –

The Cross-Validation Correlation Coefficient, often denoted as R_{cv}^2 or Q^2, is a measure of how well a regression model predicts unseen data based on cross-validation. This metric is particularly useful for evaluating the robustness and generalizability of a model. Here's how it is typically used. The Cross-Validation Correlation Coefficient (CVCC) is a critical metric in Quantitative Structure-Activity Relationship (QSAR) modelling that measures the predictive performance and robustness of a model. key uses of the Cross-Validation Correlation Coefficient in QSAR. CVCC is used in internal validation procedures, such as k-fold cross-validation or leave-one-out cross-validation, to assess the model's performance. It helps ensure that the model's predictions are consistent and reliable across different subsets of the training data. A high CVCC value indicates that the model has strong predictive power and can accurately predict the biological activity of new compounds. It helps in assessing the reliability of the model's predictions, ensuring that the model can be trusted for practical applications such as drug discovery or toxicity prediction. R_{cv}^2 or Q^2 can be used to compare the predictive performance of different models. Models with higher R_{cv}^2 or Q^2 values are preferred, as they indicate better predictive accuracy. When multiple models or different parameter settings are available, R_{cv}^2 helps in selecting the model that is likely to perform best on new data. as R_{cv}^2 or Q^2 measures how well the model generalizes to new, unseen data. High values indicate that the model predictions are close to the actual values, suggesting strong predictive performance.

vii. Adjusted R^2 or R^2_{adj}* Adjustable R^2_A –

Adjusted R-squared (R2adj) is a modified version of the R-squared (R2) metric that adjusts for the number of predictors in a regression model. Unlike R^2, which can only increase or stay the same when additional predictors are added, R^2_{adj} accounts for the model complexity and can decrease if the added predictors do not improve the model sufficiently. Here's how adjusted R-squared is typically used. Adjusted R^2 provides a measure of how well the model explains the variance in the biological activity of compounds. It adjusts for the number of predictors, offering a more accurate assessment than R^2 alone. By accounting for the number of predictors, adjusted R^2 helps in evaluating the robustness of the model. It penalizes the inclusion of non-informative descriptors, ensuring that the model is not overly complex.

In summary, adjusted R^2 is a crucial metric in QSAR modelling for evaluating, comparing, optimizing, validating, and communicating the performance of models. It accounts for the number of predictors, promoting the development of models that are both accurate and parsimonious, thereby ensuring their reliability and robustness for practical applications. High R^2_{adj} values suggest that the model is robust and captures the underlying relationship between the predictors and the response variable effectively.

viii. Pearson Correlation Coefficient (R) –

The Pearson Correlation Coefficient, often denoted as rr, measures the linear relationship between two variables. It ranges from -1 to 1, where values close to 1 indicate a strong positive linear relationship, values close to -1 indicate a strong negative linear relationship, and values around 0 indicate no linear relationship. Here are the key uses of the Pearson Correlation Coefficient. Quantifies how well the predicted biological activities correlate with the observed activities. A high R value close to 1 indicates strong predictive accuracy, while a value close to 0 indicates poor predictive performance. By assessing the strength of the linear relationship between observed and predicted values, R helps evaluate the robustness of the QSAR model. R quantifies the strength and direction of a linear relationship between two variables. A positive R indicates that as one variable increases, the other tends to increase, while a negative R indicates that as one variable increases, the other tends to decrease. It helps in identifying whether a significant linear relationship exists between two variables, which is useful in various research and analysis contexts.

ix. Q FACTOR –

In order to obtain further evidence in the favour of our results we have calculated quality factor Q for each of the QSPR models obtained as above. Note that the quality factor[7] Q is define as the ratio of correlation coefficient (r) to the standard error of estimation (sd). Thus the higher the value of r, and the lower the standard error of the estimation (sd), are the higher will be the Q values, and the better will be the proposed QSPR models.

It is worthy to mention that a model (regression equation) with excellent statistics may not necessary have excellent predictive power. Thus the next step of regression analysis is to examine predictive power of the proposed model this can be easily done by calculating Poglianis quality factor Q. This quality factor is define as the ratio of correlation coefficient (r) to the standard error of estimation (sd) i.e.

$$\mathbf{Q = r / sd}$$

Thus the higher the value of r (R) and the lower the value of sd the bigger will be the Q and the better will be the predictive power of that model.

Pogliani's quality factor Q is a metric used in Quantitative Structure-Activity Relationship (QSAR) modelling, which is a method in computational chemistry and bioinformatics that predicts the activity of chemical compounds based on their molecular structure. The quality factor Q is specifically utilized to assess the robustness and predictive power of QSAR models. Here are the key uses of Pogliani's quality factor Q in QSAR.

Pogliani's quality factor Q helps in evaluating the predictive power of QSAR models. A higher Q value indicates a model with better predictive ability. It is used to assess the robustness of the model. Robust models consistently predict the activity of compounds accurately, even when subjected to different datasets or slight variations in the input data. It is used to compare the performance of different QSAR models. This comparison helps in selecting the best model among various candidates based on their quality factor. Overall, Pogliani's quality factor Q is a crucial component in the development, evaluation, and refinement of QSAR models, ensuring that they are both accurate and reliable for predicting the biological activity of chemical compounds.

x.Least Squares Error (LSE) –

Least Squares Error (LSE), also known as the Sum of Squared Errors (SSE) or Residual Sum of Squares (RSS), is a fundamental concept in regression analysis and optimization. It quantifies the discrepancy between observed data and the values predicted by a model. Here are the primary uses of Least Squares Error.

LSE quantifies the difference between observed and predicted values by summing the squares of these differences. A lower LSE value indicates a model that better fits the data, meaning higher predictive accuracy. By measuring how well the model predictions match the actual outcomes, LSE helps in assessing the robustness of the QSAR model. Models with consistently low LSE values are considered more reliable. LSE allows for direct comparison between different QSAR models. Models with lower LSE values are preferred as they indicate better predictive performance and a closer fit to the observed data. LSE provides a measure of how well a regression model fits the data. A lower LSE indicates a better fit, meaning the model's predictions are closer to the actual data points. LSE can be used to compare the performance of different models. The model with the lower LSE is generally considered better because it indicates smaller discrepancies between observed and predicted values.

Lower LSE Indicates that the model's predictions are close to the actual data points, suggesting a good fit and accurate model. Higher LSE Indicates larger discrepancies between predicted and observed values, suggesting a poorer fit and less accurate model. In simple linear regression, the LSE is minimized to find the best-fitting line that predicts the dependent variable based on the independent variable.

xi.Probable error of the coefficient of correlation (P.E.) –

The Probable error of the coefficient of correlation (P.E.) is an interesting parameter used in QSAR in deciding whether or not the proposed correlation is good or not.

r< PE, than correlation is not significant

r> PE, than several times greater or at least three times greater than PE, the correlation is indicated

r> 6PE, than correlation is definitely good

if. In all the proposed model r was found much larger, than PE, indicating that the models are good

IV. CONCLUSION

In this paper we discussed on Special tests and methods which are used to determine the predictability of a regression model, obtained from various mathematical methods. Validation of QSAR models is a very important aspect to understand reliability of model for prediction of a new compound not present in the data set. If we consider 1000 reported QSAR models, out of which only 50 to 60 models are really predictive but it's not sure that these 60 models have been obeyed all the conditions and validation parameters discussed in this article[8].

Our opinion is, both internal and external validation strategies are important and, in fact, one should adopt all available validation strategies to check robustness of the model. Only few reported QSAR models were following all the validation characteristics mentioned in this article[9,10].

Some special tests and methods are used to determine the success of the model or which model will give better predictability in the case of more than one model. Through these methods, we can find out the significance of any type of model whether it is linear or nonlinear. The success of the regression model to a large extent depends on the descriptors and the method of selecting them. This paper refers some statistical methods factually and their application in a proper manner to understand the best model can be selected in QSAR/QSPR.

V. ACKNOWLEDGMENT

Authors are very thankful to Mr. A P Sakalle, Principal M. J. P. Govt. Polytechnic College, Khandwa for providing facilities and motivation in the work.

REFERENCES

[1] Zvinavashe E., Murk A.J., Rietjens I.M.C.M. 2008. Promises and pitfalls of quantitative structure-activity relationship approaches for predicting metabolism and toxicity. *Chem. Res. Toxicol*. 21,2229–2236.

[2] Perkins R., Fang H., Tong W., Welsh W.J. 2003. Quantitative structure-activity relationship methods: perspectives on drug discovery and toxicology. Environ. *Toxicol. Chem*,22,1666–1679.

[3] Yang G.F., Huang F. 2006. Development of Quantitative Structure-Activity Relationships and Its Application in Rational Drug Design. *Curr. Pharm. Des*. 12,601–4611.

[4] Mazzatorta P., Benfenati E., Lorenzini P., Vighi M. 2004. QSAR in ecotoxicity: an overview of modern classification techniques. *J. Chem. Inf. Comput. Sci*. 44,105–112.

[5] Konovalov D.A., Llewellyn L.E., Heyden Y.V., Coomans D.J. 2008. Robust cross-validation of linear regression QSAR models. *Chem. Inf. Model*. 48,2081–2094.

[6] Tetko I.V., Sushko I., Pandey A.K., Zhu H., Tropsha A., Papa E., Oberg T., Todeschini R., Fourches D., Varnek A. 2008. Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model*. 48,1733–1746.

[7] POGLIANI L. 1994. Structure Property Relationships of Amino Acids and Some Dipeptides, *Amino Acids*, 6,141-153.

[8] Ravichandran Veerasamy, Harish Rajak, Abhishek Jain, Shalini Sivadasan, Christapher P. Varghese and Ram Kishore Agrawal 2011. Validation of QSAR Models - Strategies and Importance *International Journal of Drug Design and Discovery*, 2(3), 511-519.

[9] Ravichandran, V.; Shalini, S.; Sundram, K.M.; Dhanaraj, S.A. 2010. Eurp. *J. Med. Chem*., 45, 2791-2797

[10] Roy, P.P.; Paul, S.; Indrani, M.; Roy, K. 2009. *Molecules*. 14, 1660-1701.