REVIEW ON: QUANTITATIVE STRUCTURE ACTIVITY RELATIONSHIP (QSAR) MODELING

Umma Muhammad[1], Adamu Uzairu[2] and David Ebuka Arthur[2]

[1]Department of Pre-nd Sci& Tech, School of General Studies ,Kano State Polytechnic.

[2]Department of Chemistry, Ahmadu Bello University Zaria.

Abstract

QSAR (quantitative structure activity relationship) are mathematical models that seek to predict complicated physicochemical / biological properties of chemicals from their simpler experimental or calculated properties QSAR enables the investigator to establishes a reliable quantitative relationship between structure and activity which will be used to derive an insilico model to predict the activity of novel molecules prior to their synthesis. The past few decades have witnessed much advances in the development of computational models for the prediction of a wide span of biological and chemical activities that are beneficial for screening promising compounds with robust properties. This review covers the concept, history of QSAR and also the components involved in the development of QSAR models.

Keywords: QSAR, model development, applicability domain, molecular descriptor, virtual screening

Introduction

Quantitative structure – activity relationship (QSAR) modeling pertains to the construction of predictive models of biological activities as a function of structural and molecular information of a compound library. The concept of QSAR has typically been used for drug discovery and development and has gained wide application for correlating molecular information with not only biological activities but also with other physicochemical properties, which has therefore been termed quantitative structure – property relationship (QSPR). QSAR is widely accepted predictive and diagnostic process used for finding associations between chemical structures and biological activity. QSAR has emerged and has evolved trying to fulfill the medicinal chemist' s need and desire to predict biological response (Hansch C., 1979). It first found its way into the practice of agro chemistry, pharmaceutical chemistry, and eventually most facets of chemistry.

QSAR is the final result of computational processes that start with a suitable description of molecular structure and ends with some inference, hypothesis, and predictions on the behavior of molecules in environmental, physicochemical and biological system under analysis (Eriksson et al., 2003).The final outputs of QSAR computations are set of mathematical equations relating chemical structure to biological activity (Golbraikh et al., 2003; Hansch, Sinclair, & Sinclair, 1990; Wedebye, Dybdahl, Nikolov, Jónsdóttir, & Niemelä, 2015). Multivariate QSAR analysis employs all the molecular descriptors from various representations of a molecule (1D, 2D and 3D representation) to compute a model, in a search for the best descriptors valid for the property in analysis.

This review covers the concepts, history and the steps involved in the development of QSAR models.

HISTORY OF QSAR

Cros in 1863 proposed a relationship which existed between the toxicity of primary aliphatic alcohols with their water solubility (Cros, 1863). In 1868 Crum-Brown and Fraser published an equation which is considerable to be the first generation formulation of a quantitative structure-activity relationship,in their investigations of different alkaloids (Crum-Brown A, 1868). Systematic QSAR began with the work of Cantor (2001) on the narcotic activity of various drugs (Pohorille, Wilson, New, & Chipot, 1998).Hammett in 1935 introduced a method to account for substituent effects on reaction mechanism (Hammett, 1935).Taking Hammetts model into account Taft proposed in 1956 an approach for separating polar, steric, and resonance effects of substituents in aliphatic compounds (Taft Jr, 1956). Classical approach to QSAR/QSPR was led by the pioneering works of Hanschand Fujita (1964) in the development of linear Hansch equation (Fujita, Iwasa, & Hansch, 1964).

QSAR/QSPR received a big boost with the development of newer, more complex descriptors, softwares and computers.This has been instrumental in the application of the prediction techniques that were either not feasible or were previously too time consuming.

QSAR METHODOLOGY

QSAR methodologies have the potential of decreasing substantially the time and effort required for the discovery of new medicines (Gramatica, Giani, & Papa, 2007). A major step in constructing the QSAR models is to find a set of molecular descriptors that represents variations of the structural properties of the molecule (Gramatica, 2007). The QSAR analysis employs statiscal methods to derive quantitative mathematical relationship between chemical structure and biological activity (Ghafourian & Cronin, 2005). The process of QSAR modelling can be divided into three stages: development, model validation and application.

Development

For the development of the model the compounds gathered from literature source could be divided into training and test set. The training set are used in model construction while the test set for external validation.

The structures of the complexes under study could be drawn in 2D ChemDraw. These could be converted into 3D objects using the default conversion procedure implemented in the CS Chem 3D ultra. The generated 3D structures of the complex were then subjected to energy minimization and geometry optimization using Spartan(Hehre & Huang, 1995). Molecular descriptors could be calculated using chemical software's such as Dragon (Mauri, Consonni, Pavan, & Todeschini, 2006), Gaussian (Salahub et al., 1991), paDEL (Yap, 2011) etc. Molecular descriptors can be defined as the essential information of a molecule in terms of its physicochemical properties such as constitutional, electronic, geometrical, hydrophobic, lipophilicity, solubility, steric, quantum chemical and topological descriptors (Todeschini & Consonni, 2009). Multivariate analysis such as multi linear regression, Partial least Square etc could be carried out for correlating molecular descriptors with observed activity.

Validation

Internal and external validation could be performed to validate the QSAR models.

The internal validation of derived model could be ascertained through the cross-validation index Q^2 from leave –one –out (LOO) procedure. The LOO method creates a number of modified data sets by taking away one complex from the parent data set in such a way that each observation is removed once only. Then one model is developed for each reduced data set and the response values of the deleted observations are predicted from these models. A value greater than 0.5 of Q^2 index hints towards a reasonable robust model.

For external validation,the activity of each complex in test set was computed. Goodness of fit of the models would be assessed by examining the multiple correlation coefficient (r), the standard deviation (s), the F-ratio between the variances of calculated and observed activities (F).

Internal Model Validation

The developed models were validated internally by leave- one- out (LOO) cross- validation technique. In this technique, one compound is eliminated from the data set at random in each cycle and the model is built using the rest of the compounds. The model thus formed is used for predicting the activity of the eliminated compound. The process is repeated until all the compounds are eliminated once. The Cross-validated squared correlation coefficient, R2cv (Q2) was calculated using the expression:

$$Q^2 = 1 - \frac{\sum(Y_{Obs} - Y_{Pred})^2}{\sum(Y_{Obs} - \bar{Y})^2}$$

Where Y_{OBS} represents the observed activity of the training set compounds, Y_{pred} is the predicted activity of the training set compounds and $\bar{Y}$ corresponds to the mean observed activity of the training set compounds. Also calculated was the adjusted R^2($_{adj}R^2$) which is a modification of R^2 that adjust the number of explanatory terms in a model. Unlike R^2 in which addition of descriptors to the developed QSAR model increases its value, the value of $_{adj}R^2$ increases only if the new term improves the model more than what would be expected by chance (Rudra and Kunal, 2012). Hence $_{adj}R^2$ overcomes the draw backs associated with the value of R^2 and was calculated using the expression:

$$adjR^2 = \frac{(n-1)R^2 - p}{n-p-1}$$

Where p is the number of predictor variables used in the model development. In other to judge the overall significance of the regression coefficients, the variance ratio, F value (the ratio of regression mean square to deviations mean square), was also calculated using the relation:

$$F = \frac{\left(\frac{\sum(Y_{cal} - \bar{Y})^2}{p}\right)}{\left(\frac{\sum(Y_{obs} - Y_{cal})^2}{N-P-1}\right)}$$

External Model Validation

External validation was employed in order to determine the predictive capacity of the developed model as judged by its application for the prediction of test set activity values and calculation of predictive R^2(R^2pred) value as given by the expression:

$$R^2_{pred} = 1 - \frac{\sum\left(Y_{pred\,(Test)} - Y_{(Test)}\right)^2}{\sum\left(Y_{(Test)} - \bar{Y}_{(Training)}\right)^2}$$

Where $Y_{pred\,(Test)}$ and $Y_{(Test)}$ indicate predicted and observed activity values respectively, of the test compounds. $\bar{Y}_{(Training)}$indicates mean activity value of the training set. R^2_{pred} is the predicted correlation coefficient calculated from the predicted activity of all the test set compounds. It has been observed that R2pred may not be sufficient to indicated the external predictability of a model since its value is controlled by $\sum\left(Y_{(Test\,)} - \bar{Y}_{(Training\;)}\right)^{2}$. Thus R^2_{pred} depends on the training set mean and may not truly reflect the predictive capability of the developed model with regards to a new data set (Kar & Roy, 2012). This may result in considerable numerical difference between the observed and predicted values in spite of maintaining a good overall intercorrelation.

Randomization Test

The Robustness of the developed QSAR model was checked using Y-randomization technique in which model randomization was employed. In Y-randomization, validation was performed by permutating the response values, Activity (Y) with respect to the descriptor (X) matrix which was unaltered (Roy, Kar, & Das, 2015). The deviation in the values of the squared mean correlation coefficient of the randomized model (Rr^2) from the squared correlation coefficientof the non-random model (R^2) is reflected in the value of R^2_p parameter computed from the expression (Roy and Paul, 2008):

$$R_p^2 = R^2 \times \sqrt{(R^2 - R_r^2)}$$

In an ideal case, it is observed that the average value of R^2 (R_r^2) for randomized models should be zero. This implies that the value of R_p^2 should be equal to the value of R^2 for the developed QSAR model. This led Todeschini in 2010, to suggest a correction for R_p^2 which is defined as:

$$cR_p^2 = R \times \sqrt{R^2 - R_r^2}$$

In other to penalize the developed models for the difference between the squared correlation coefficients of the randomized and the non-randomized models, the values cR_p^2 was calculated for each model. This procedure ensures that the model is not due to a chance. The Y-randomization results were generated using the program “MLR Y-Randomization Test 1.2” (Roy, Kar, & Ambure, 2015)

Application:

The application of QSAR models depends on statistical significance and predictive ability of the models. The prediction of a modeled response using QSAR is valid only if the compound being predicted is within the applicability domain of the model. The applicability domain is a theoretical region of the chemical space, defined by the model descriptors and modeled response and thus by the nature of the training set molecules(Todeschini, Consonni, & Pavan, 2007). It is possible to check whether a new chemical lies within applicability domain using the leverage approach. A compound will be considered outside the applicability domain when the leverage values is higher than the critical value of 3p/n, where p is the number of model variables plus 1 and n is the number of objects used to develop the model. Other approach includes training set interpolation by Jaworska (Jaworska, Comber, Auer, & Van Leeuwen, 2003). Cluster – based approach by Stanforth et al. (Stanforth, Kolossov, & Mirkin, 2007).

Conclusion

The QSAR models are useful for various purposes including the prediction of activities of untested chemicals. It helps in the rational design of drugs by computer aided tools via molecular modelling, simulation and virtual screening of promising candidates prior to synthesis. In this review article the concept, brief history and components involved in modelling were discussed.

References

Cantor, R. S. (2001). Breaking the Meyer-Overton rule: predicted effects of varying stiffness and interfacial activity on the intrinsic potency of anesthetics. *Biophysical journal, 80*(5), 2284-2297.

Cros, A. (1863). *Action de l'alcool amylique sur l'organisme.(Cand. AFA Cros).*

Crum-Brown A, F. T. (1868). On the connection between chemical constitution and physiological action. Pt 1. On the physiological action of the salts of the ammonium bases, derived from Strychnia, Brucia. *Thebia, Codeia, Morphia, and Nicotia. T Roy Soc Edin, 25*, 151-203.

Eriksson, L., Jaworska, J., Worth, A. P., Cronin, M. T., McDowell, R. M., & Gramatica, P. (2003). Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs. *Environmental health perspectives, 111*(10), 1361.

Fujita, T., Iwasa, J., & Hansch, C. (1964). A new substituent constant, π, derived from partition coefficients. *Journal of the American Chemical Society, 86*(23), 5175-5180.

Ghafourian, T., & Cronin, M. T. (2005). The impact of variable selection on the modelling of oestrogenicity. *SAR and QSAR in Environmental Research, 16*(1-2), 171-190.

Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y.-D., Lee, K.-H., & Tropsha, A. (2003). Rational selection of training and test sets for the development of validated QSAR models. *Journal of computer-aided molecular design, 17*(2-4), 241-253.

Gramatica, P. (2007). Principles of QSAR models validation: internal and external. *QSAR and Combinatorial Science, 26*(5), 694.

Gramatica, P., Giani, E., & Papa, E. (2007). Statistical external validation and consensus modeling: A QSPR case study for K oc prediction. *Journal of Molecular Graphics and Modelling, 25*(6), 755-766.

Hammett, L. P. (1935). Some Relations between Reaction Rates and Equilibrium Constants. *Chemical Reviews, 17*(1), 125-136.

Hansch, C., Sinclair, J. F., & Sinclair, P. R. (1990). Induction of Cytochrome P450 by Barbiturates in Chick Embryo Hepatocytes: A Quantitative Structure-Activity Analysis. *Quantitative Structure-Activity Relationships, 9*(3), 223-226.

Hansch C., L. A. (1979). Substituent constants for correlation analysis in chemistry and biology. *John Wiley and Sons, New York.*

Hehre, W. J., & Huang, W. W. (1995). *Chemistry with Computation: An introduction to SPARTAN*: Wavefunction, Inc.

Jaworska, J. S., Comber, M., Auer, C., & Van Leeuwen, C. (2003). Summary of a workshop on regulatory acceptance of (Q) SARs for human health and environmental endpoints. *Environmental health perspectives, 111*(10), 1358.

Kar, S., & Roy, K. (2012). QSAR of phytochemicals for the design of better drugs. *Expert Opinion on Drug Discovery, 7*(10), 877-902. doi:10.1517/17460441.2012.716420

Mauri, A., Consonni, V., Pavan, M., & Todeschini, R. (2006). Dragon software: An easy approach to molecular descriptor calculations. *Match, 56*(2), 237-248.

Pohorille, A., Wilson, M. A., New, M. H., & Chipot, C. (1998). Concentrations of anesthetics across the water–membrane interface; the Meyer–Overton hypothesis revisited. *Toxicology letters, 100*, 421-430.

Roy, K., Kar, S., & Ambure, P. (2015). On a simple approach for determining applicability domain of QSAR models. *Chemometrics and Intelligent Laboratory Systems, 145*, 22-29. doi:http://dx.doi.org/10.1016/j.chemolab.2015.04.013

Roy, K., Kar, S., & Das, R. N. (2015). Chapter 1 - Background of QSAR and Historical Developments. In K. R. K. N. Das (Ed.), *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment* (pp. 1-46). Boston: Academic Press.

Salahub, D. R., Fournier, R., Młynarski, P., Papai, I., St-Amant, A., & Ushio, J. (1991). Gaussian-based density functional methodology, software, and applications *Density functional methods in chemistry* (pp. 77-100): Springer.

Stanforth, R. W., Kolossov, E., & Mirkin, B. (2007). A measure of domain of applicability for QSAR modelling based on intelligent K-means clustering. *QSAR and Combinatorial Science, 26*(7), 837.

Taft Jr, R. (1956). Separation of Polar, Steric and Resonance Effects in Reactivity in Steric Effects in Organic Chemistry. *John Wiley and Sons, New York*.

Todeschini, R., & Consonni, V. (2009). *Molecular Descriptors for Chemoinformatics, Volume 41 (2 Volume Set)* (Vol. 41): John Wiley & Sons.

Todeschini, R., Consonni, V., & Pavan, M. (2007). Milano chemometrics and QSAR research group. *KOALA-Software for Kohonen Artificial Neural Networks, Version, 1*.

Wedebye, E. B., Dybdahl, M., Nikolov, N. G., Jónsdóttir, S. Ó., & Niemelä, J. R. (2015). QSAR screening of 70,983 REACH substances for genotoxic carcinogenicity, mutagenicity and developmental toxicity in the ChemScreen project. *Reproductive Toxicology, 55*, 64-72. doi:http://dx.doi.org/10.1016/j.reprotox.2015.03.002

Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry, 32*(7), 1466-1474.