International Journal of Scientific Research in Chemistry (IJSRCH)
© 2019 IJSRCH | Volume 4 | Issue 6 | ISSN : 2456-8457

QSAR Studies as Strategic Approach in Drug Discovery

Akshay R. Yadav*, Dr. Shrinivas K. Mohite
Department of Pharmaceutical Chemistry, Rajarambapu College of Pharmacy, Kasegaon, Maharashtra, India-415404
*Corresponding author E-mail: akshayyadav24197@gmail.com

ABSTRACT

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 modeling, simulation and virtual screening of promising candidates prior to synthesis. In order to achieve a reliable statistical model for predicting the behaviors of new chemical entities, quantitative structure activity relationship (QSAR) have been used for decades to establish connections between the physicochemical properties of chemicals and their biological activities. The fundamental concept of formalism is that the biological differences in the compounds have a difference in structural properties. The atom, groups or molecular characteristics of ligands affinity to its sites, inhibition constants, frequency constants, and more biological endpoints have been linked with the classic QSAR studies such as lipophilicity, polarization, electronic and steric properties (hansch analysis) and basic structural characteristics (free wilson analysis).

Keywords: QSAR, Molecular descriptors, 2D QSAR, 3D QSAR, Genetic algorithm

I. INTRODUCTION

QSAR (Quantitative Structure Activity Relationships) have been applied for decades in the development of relationships between physicochemical properties of chemical substances and their biological activities to obtain a reliable mathematical and statistical model for prediction of the activities of new chemical entities[1]. (QSAR) have helped the scientists in the development of mathematical relationships linking chemical structures and pharmacological activity in quantitative manner of series of compound. The fundamental principle underlying the QSAR is that the difference in structural properties is responsible for the variations in biological activities of the compounds. In the classical QSAR studies, affinities of ligands to their binding sites, inhibition constants, rate constants, and other biological end points, with atomic, group or molecular properties such as lipophilicity, polarizability, electronic and steric properties (Hansch analysis) or with certain structural features (Free-Wilson analysis) have been correlated. QSAR certainly decreases the number of compounds to be synthesized by facilitating the selection of the most promising candidates. This review seeks to provide a view of the different QSAR approaches employed within the current drug discovery process to construct predictive structure–activity relationships and also discusses the limitations that are fundamental to these approaches, as well as those that might be overcome with the improved strategies[2-3].

❖ *Objectives of QSAR*

Most of the QSAR approaches concentrate on the following objectives:

- ✓ To quantitatively compare and recapitulate the relationships between trends in changes in

chemical structure and corresponding changes in biological endpoints in order to understand the chemical properties are most likely determinants of their biological activities.

- ✓ To optimizing existing leads to enhance their biological activities.
- ✓ To Predict the biological behaviors of substances that are untested and sometimes inaccessible[4].

❖ *Rationale behind QSAR modeling*

For the following reasons, QSAR becomes a useful alternative:

- ✓ Conventional methods of synthesis are costly and time-consuming.
- ✓ Biological assays are also too costly, often requiring time, sacrificing of animals or compounds in their pure forms.
- ✓ Drug failures due to poor ADMET profiles at later stages of development (or even after marketing) are highly expensive and painful. A large number of compounds are necessary for combined chemistry and HTS methods, but priority assessment is required.

❖ *Molecular Descriptors*

It is a numerical representation of the encoded chemical data by mathematical method within a molecular structure.

The structure descriptors information quality depends on two main factors:

- ➢ The compound molecular representation.
- ➢ The algorithm used for the descriptor's calculation.

  The initially suggested three main types of parameters are,

  1. Hydrophobic
  2. Electronic
  3. Steric

❖ *Classification of QSAR Methodologies*

The QSAR methods are usually classified into the following classes on the basis of their structural representation or the way in which descriptor values are derived.

- 1D-QSAR Compatibility to global molecular characteristics like pKa, log P, etc.
- 2D-QSAR Rely on structural patterns like conexion indices, 2D pharmacophores, etc.
- 3D-QSAR related behavior to non-covalent fields of molecular interaction.
- 4D-QSAR also contains a collection of 3D-QSAR ligand configurations.
- 5D-QSAR describing distinctly specific 4D-QSAR induced-fit designs.
- 6D-QSAR also integrates various solvent models in 5D-QSAR

➢ *Two-Dimensional QSAR (2D QSAR)*

2D QSAR is a powerful tool to explain the relationship between chemical and experimental observations. The system's main elements are numerical descriptors used to turn chemical structures into mathematical variables, the consistency and statistical procedures of observations and descriptors in the relationship[5].

- **2D QSAR Methods**

QSAR models are used to scan chemical repositories and/or virtual chemical libraries for molecules that are potentially bioactive. Both advances underline the importance of robust model validation to ensure that the models have both the ability to explain the biological activity variance (internal validation) and the appropriate predictive potential (external validation).

Division of the dataset into training and test set are:

1. Manual selection

This can be done by observing the variation in the given dataset's chemical and biological space.

2. Random selection

This approach provides random distribution training and testing.

3. Sphere exclusion method

This is a fair way of setting up training and testing. It ensures a consistent distribution of the points in both the sets in terms of chemical and biological volume. QSAR is the creation of a model linked with their biological activities to the molecular structures of a series of chemicals. The interpretation of the descriptors used and the degree to which they represent the structural characteristics of the molecules associated with biological activity is an essential component of any QSAR analysis. Although this method is still widely applied, hundreds of computer-generated descriptors have replaced the experimental physicochemical parameters, each encoding a particular molecular function. For example, the program CODESSA (Semichem, Shawnee, KS, USA) and the program Cerius2 (Accelrys, San Diego, CA, USA) can produce hundreds of measured descriptors to describe a molecules structure.

- **Descriptor calculation:**

Once the molecules are aligned, on a grid of points in space around the molecule, a molecular field is measured. This field describes how in the active site each molecule tends to bind. Descriptors representing the energy of steric, electrostatic and hydrophobic interaction were measured using a methyl charging probe +1 at the grid lattice points.

- **Fitness plot:**

Some descriptors can show chance correlation with activity as each variable selection method is based on descriptor and activity correlation and not on the form of data spread. Some descriptors can show chance correlation with activity as each variable selection method is based on descriptor and activity correlation and not on the form of data spread. The careful observation of the fitness plot between descriptor and operation is necessary to avoid the above described pitfall. The following are some important points that we considered when selecting the right descriptors

1. We made sure that the percentage distribution of data points on both sides of the best fit line is equivalent to 50-50%.
2. The plot slope between descriptor and operation has been carefully evaluated. Because the highly correlated descriptor showed m uch less slope at times; then such descriptors were excluded examined.
3. In the case of a topological descriptor, no particular data point occurrences were observed in the fitness plot which gave information on the frequency of occurrence of each correlation with activity also occurred in the final result of the QSAR model, but we can not take it into final consideration unless and until it shows well distributed fitness plot.
4. We found that some molecules displayed no uniform change with behavior in the descriptor value and therefore did not follow the hypothesis of QSAR. Such descriptors have been removed. In conclusion, we can say that careful observation and proper fitness plot analysis helped us to reduce the number of descriptors[6].

- **QSAR Model Generation**

If associated X variables exist, Multiple Linear Regression (MLR) is unstable. It gives a good example of why we need to look at the structure of data sets, rather than using them blindly. Examination of key components provides a way to find meaning in such data sets. This rotates the data into a new set of axes such that most of the data differences are represented by the first few axes. Through plotting the data on these axes, we can automatically detect major underlying structures. Price of each point is called the main component price when rotated to a given axis. Principal Components Analysis uses a new set of data axes. These are selected perpendicular to each other in decreasing order of variance within the results. The main components are therefore uncorrelated. So, how

about measuring main components, throwing away those that seem to only contribute noise (or constants), and using MLR on them. This process gives the modeling approach called the Regression of Principal Components. Instead of constructing a single model, as with MLR, it is possible to form a model using 1,2 components and make a decision as to how many components are optimal. Unless co-linearity was found in the initial variables, some of the components will only add noise.

- **Principal Component Regression (PCR) Method**

When there are correlated X variables, multiple linear regression (MLR) is unstable. This gives a good example why the structure must be examined in data sets rather than blindly used. Main component analysis provides a way to find structure in these data sets. It rotates the data into a new set of axes so that most of the variations are reflected in the first few axes. We can automatically detect major structures by tracing data on these axes. The value of each point is called the main component value when rotated to a given axis. Key Components Analysis selects a new set of data axes. These are selected perpendicular to each other in decreasing order of variance within the data. The main components are therefore not associated. MLR was found to lead to instability due to the associated variables. And how to measure the key components, delete those that seem to only contribute noise (or constants) and use MLR for them. The modeling method known as main components regression is provided by this process. The model can be rendered with 1,2 components and the number of components measured optimally can be estimated instead of making a single model, as with MLR. Sadly, some of the components contribute to noise even if the initial variables have co-linearity. We can guarantee a stable model as long as the models are dropped[7].

- ***Three-Dimensional QSAR (3D QSAR)***

This module generates 3D QSAR equation by using various statistical regression methods. It also provide facilities for molecular alignment and generation of steric and electrostatic interactive energies. Based on generated QSAR model this module helps to design the novel ligands. It consist of following steps:

1. Building of 3D QSAR equation using following statistical method with stepwise genetic algorithm and simulated annealing method for variable selection.
   - Multiple regression
   - Partial least square regression
   - Principal component regression

2. Using artificial intelligence method capturing non linear relationship.
   - Neural network (back propagation and pruning neural network)

3. Using K nearest neighbor principle (kNN) building 3D QSAR equation with stepwise, genetic algorithm and simulated annealing methods for different selection. This method helps generate automatic QSAR models with stepwise, genetic algorithm and simulated methods of annealing, which optimize k values.
4. Pattern recognition of descriptor via different graphs in worksheet
   - Pattern plots
   - X-Y plots
   - Fitness plot

- **Alignment rules**

In CoMFA study is problematic due to proper alignment of the molecules is critical. Definition of optimal alignment can be such that a set of molecules achieve maximum superposition of electrostatic and steric fields. Basically, collection of rules to be applied to a series of molecules to ensure that the process is

accurate. The alignment precedes one molecule at a time, which may differ from molecule to molecule based on structural similarity considerations. Alignment defines the degree to which the steric and electrostatic fields differ from one molecule to another. Therefore, alignment greatly affects model outcomes, and significant and appropriate outcomes can only be predicted for valid alignments[8].

- **3D model generation by SA-kNN-MFA**

3D QSAR were performed by generating multiple models by choosing same molecules in the respective training and test sets as in 2D QSAR by using k-Nearest Neighbor-Molecular field Analysis (kNN-MFA) methodology as most suitable method to perform 3D QSAR. The KNN methodology depends upon a simple distance learning approach. In this method, an unknown member (u) is classified according to majority of its k-Nearest Neighbors in training set. The range is determined by an acceptable metric of space. SA variable selection method is the simulation of a physical operation, 'annealing' which involves heating the device to high temperature and then slowly cooling it down to a preset temperature (e.g. room temperature). During this process, the system samples possible configuration which are distributed so that at equilibrium, low energy states are most populated. kNN-MFA Simulated Annealing: All the measured descriptors remaining after removal of invariable columns were subjected to SA algorithm combined with k-nearest neighbor (kNN) methods to construct a QSAR model based on the training set. Application validation is then performed to verify the application both internally and externally in order to test the model's efficacy and predictive ability. Internal validation is done using the let-one-out process (q^2, LOO). Through this process any molecule's biological activity is once expected, eliminating it from the system. For external validation, the behavior of each molecule in the test Set was predicted using the model developed from the training set. Pred r^2 specifies the model's ability to accurately predict the values of biological activity for an outside test range. Using following statistical measures established quantitative models is evaluated n, number of observations (molecules); k, number of variables (descriptors); number of components; number of nearest neighbours, distance between unknown object and other object is determined using Training set The object k is close to object u, those objects are selected from the training set. The object u has been labeled with the category most of the k object belongs to optimization by classifying a set of samples or by cross-validating Leave-One-Out (LOO) is used to pick an optimal k value. The variables and optimal k values were chosen by using different variable selection methods. Number of k-nearest neighbor in model; r^2, determination coefficient; q^2, cross validated r^2 (leaving one out); pred r^2, r^2 for external test set; F-test, pred r^2 se, regular test set prediction error. The r^2 and q^2 values are used as determinants for testing the model 's effectiveness[9].

- **Genetic Algorithm**

The Genetic Algorithms (GA) are efficient methods for reducing work. The models predictive error based on collection of features is optimized in the sense of descriptor selection. The genetic algorithm is the normal evolution by designing solutions for a complex population. The selected characteristics, called chromosomes, are encoded by the members of the population. The encoding typically takes the form of bit strings, with bits corresponding to the set of selected functions. Growing chromosome results in a model that is built using the encoded functions. The model's error is quantified using the training data, and serves as a fitness function. By ensuring the survival and reproduction of the fittest genes, the algorithm effectively minimizes the error function in subsequent generations. GA output depends on count of variables. In genetic feature selection the selection of the initial population is also significant. A framework based on Shannon's entropy combined with graph analysis can be used to solve this problem,

for example. Genetic Algorithms have been widely used in feature selection for QSAR with a range of mapping methods, e.g., Artificial Neural Networks, method for k-Nearest Neighbor and Random Forest.

- **Methodology**

Molecules were optimized by MMFF energy minimization method before atom dependent alignment was performed. Using a methyl charging probe +1, common rectangular grid was created and electronic interacting energies at the grid lattice point were calculated at Steric. In a similar way as in 2D QSAR, multiple training and test sets were generated. As in 2D QSAR, test sets were produced in a similar manner. The quantitative 3D QSAR models were produced using the PLS process. Statistical measures were used to evaluate the QSAR models[10].

❖ **Applications**

QSAR models are applied depending on the model's statistical significance and predictive ability. The system response prediction using QSAR is accurate only when the predicted compound falls within the applicability domain of the model. The application field is a conceptual region of a chemical space defined by model description and model response, i.e. the design of the molecules for training. Using the leverage approach, the applicability domain can be verified by a new chemical. Where p is the number of system variables +1 and n are the number of objects used to construct a template, a lever value is greater than the critical value of 3p/n a compound would be taken into account outside the applicability domain[11-12].

II. CONCLUSION

Quantitative Structure Activity Relationship (QSAR) 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 in-silico 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.

III. REFERENCES

[1]. Bajaj S, Ghode P, Singh J, Roy P, Jain S. Quantitative structure activity relationship and combinatorial design of 1,3,4-oxadiazole based thymidine phosphorylase inhibitors as potential anti-cancer agents. Cur sci. 2018; 114(10): 2063-2071.

[2]. Verma J, Khedkar V, Coutinho E. 3D-QSAR in drug design-A Review. Cur Top Med Chem. 2010; 10(2): 95-115.

[3]. Muhammad U, Uzairu A, Arthur D. Review on: quantitative structure activity relationship (QSAR) modeling. J Anal Pharm Res. 2018; 7(2): 240–242.

[4]. ain K. 3D QSAR analysis on oxadiazole derivatives as anticancer agents. Int J Pharm Sci D Res. 2011; 3(4): 230-235.

[5]. Zong G, Yan X, Bi J, Jiang R, Qin Y, Yuan H, Lu H, Dong Y, Jin S, Zhang J. Synthesis, fungicidal evaluation and 3D-QSAR studies of novel 1,3,4-thiadiazole xylofuranose derivatives. Plos One. 2017; 2(8): 1-16.

[6]. Frimayanti N. Validation of quantitative structure activity relationship (QSAR) model for photosensitizer activity prediction. Int. J. Mol. Sci. 2011; 12(2): 8626-8644.

[7]. Pathade K, Mohite S, Yadav A. 3D-QSAR And ADMET Prediction Of Triazine Derivatives For Designing Potent Anticancer Agents. Journal of University of Shanghai for Science and Technology. 2020; 22(11): 1816-1833.

[8]. Pathade K, Mohite S, Yadav A. Synthesis, Molecular Docking Studies of Novel 4-(Substituted Phenyl Amino)-6-(Substituted Aniline)-N'-Aryl-1,3,5-Triazine-2-Carbahydrazide Derivatives As Potent Antitubercular Agents. Journal of University of Shanghai for Science and Technology. 2020; 22(11): 1891-1909.

[9]. Eriksson L. Methods for reliability and uncertainty assessment and for applicability evaluations of classification and regression based QSARs. Envir Heal Pers. 2003; 10(1): 1361-1375.

[10]. Singh Y. Study of halogen substituent on docking and 3D QSAR properties of aryl substituted thiosemicarbazone as anticonvulsant. Int J Therap App. 2012; 6(4): 1-7.

[11]. Dudek A. Computational methods in developing quantitative structure activity relationships (QSAR): A Review. Comb chem high throu Scr. 2006; 9(8): 213-228.

[12]. Asirvatham S. Quantitative structure activity relationships studies of non steroidal anti-inflammatory drugs: A review. Arab J Chem. 2016; 30(5): 1-15.

Cite this article as :

Akshay R. Yadav, Dr. Shrinivas K. Mohite, "QSAR Studies as Strategic Approach in Drug Discovery", International Journal of Scientific Research in Chemistry (IJSRCH), ISSN : 2456-8457, Volume 4 Issue 6, pp. 16-22, November-December 2019. URL : http://ijsrch.com/IJSRCH19466