Quantitative Structure Activity Relationship (QSAR)

VLife Sciences Technologies Pvt. Ltd.
Pride Purple Coronet, 1st floor, S No. 287, Baner Road, Pune 411 045, INDIA

The Quantitative Structure Activity Relationship (QSAR) paradigm is based on the assumption that there is an underlying relationship between the molecular structure and biological activity. On this assumption QSAR attempts to establish a correlation between various molecular properties of a set of molecules with their experimentally known biological activity.

There are two main objectives for the development of QSAR:

1) Development of predictive and robust QSAR, with a specified chemical domain, for prediction of activity of untested molecules.
2) It acts as an informative tool by extracting significant patterns in descriptors related to the measured biological activity leading to understanding of mechanisms of given biological activity. This could help in suggesting design of novel molecules with improved activity profile.

QSAR's most general mathematical form is : Activity = f (physiochemical and/or structural properties)

Data Requirement and Handling: Biological Activity

For QSAR analysis, a dataset of a series of synthesized molecules tested for its desired biological activity is required. For a QSAR to be valid and reliable, the activity of all of the chemicals covered must be elicited by a common mechanism. The quality of the model is totally dependent on the quality of the experimental data used for building the model. Biological activity can be of two types:

1) Continuous Response : MEC, IC50, ED50, % inhibition
2) Categorical Response : Active/Inactive

In order to have confidence in QSAR analysis, biological data of at least 20 molecules is recommended :

1) Preferably tested in the same lab and by the same biological assay method.
2) With wide range and uniform distribution of the activity data.
3) Activity should well-defined in terms of either real number (continuous response, and cannot be e.g. >1000 or <1000) or in a particular class (categorical response).

Molecular Descriptors

Molecular descriptors can be defined as a numerical representation of chemical information encoded within a molecular structure via mathematical procedure. Type of QSAR is based on the dimensionality of molecular descriptors used :

0D These are descriptors derived from molecular formula e.g. molecular weight, number and type of atoms etc.

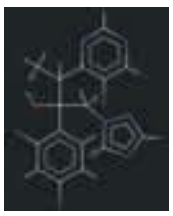
1D A substructure list representation of a molecule can be considered as a one-dimensional (1D) molecular representation and consists of a list of molecular fragments (e.g. functional groups, rings, bonds, substituents etc.).

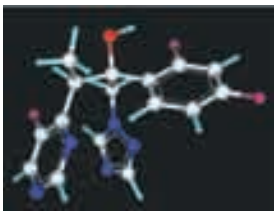
2D A molecular graph contains topological or two dimensional (2D) information. It describes how the atoms are bonded in a molecule, both the type of bonding and the interaction of particular atoms (e.g. total path count, molecular connectivity indices etc.).

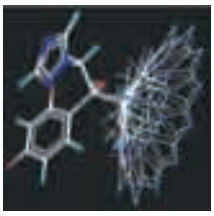
3D These are calculated starting from a geometrical or 3D representation of a molecule. These descriptors include molecular surface, molecular volume and other geometrical properties. There are different types of 3D descriptors e.g. electronic, steric, shape etc.

4D In addition to the 3D descriptors the 4th dimension is generally in terms of different conformations or any other experimental condition.


1D Representation


2D Representation


3D Representation


4D Representation

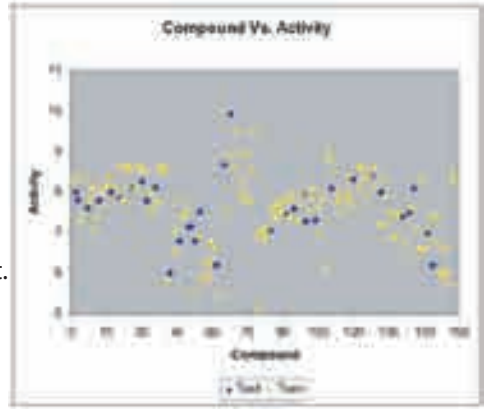
Selection of training and test set:

QSAR models are used increasingly to screen chemical databases and/or virtual chemical libraries for potentially bioactive molecules. These developments emphasize the importance of rigorous model validation to ensure that the models have both the ability to explain the variance in the biological activity (internal validation) and also the acceptable predictive power (external validation).

For model validation the dataset is required to be divided into training set (for building the QSAR model) and test set (for examining its predictive ability). For any QSAR model, it is of crucial importance that the training set selected to calibrate the model exhibits a well balanced distribution and contains representative molecules.

Following are the methods for division of the dataset into training and test set:

1) **Manual Selection** : This is done by visualizing the variation in the chemical and biological space of the given dataset.
2) **Random Selection** : This method creates training and test set by random distribution.
3) **Sphere Exclusion Method** : This is a rational method for creation of training and test set. It ensures that the points in the both the sets are uniformly distributed w.r.t. chemical and biological space.
4) **Others :**
   a) Experimental Design : full factorial, fractional factorial etc.
   b) Onion Design
   c) Cluster Analysis
   d) Principal Component Analysis
   e) Self Organizing Maps (SOM)



Distribution Plot: This plot shows the relative distribution of molecules in training and test set with respect to biological space (activity).

Variable selection methods :

There are a hundreds of molecular descriptors available for building a QSAR model. Not all of the molecular descriptors are important in determining the biological activity, and hence to find the optimal subset of the descriptors which plays an important role in determining activity, a variable selection method is required. The variable selection method could be divided mainly into two categories :

1) Systematic variable selection : These methods add and/or delete a descriptor in steps one-by-one in a model.
   a) Stepwise forward
   b) Stepwise forward-backward
   c) Stepwise backward
2) Stochastic variable selection : These methods are based on simulation of various physical or biological processes. These methods creates model starting from randomly generated model(s) and later modifying these model(s) by using different process operator(s) (e.g. perturbation, crossover etc.) to get better model(s).
   a) Simulated Annealing
   b) Genetic/Evolutionary Algorithms
   c) Modified Particle Swarm Optimization
   d) Artificial Ant Colony System

Statistical Methods :

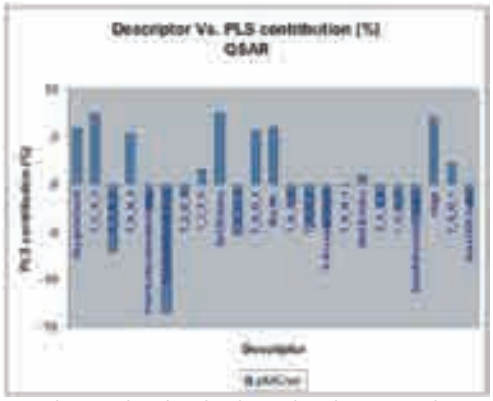
- A suitable statistical method coupled with a variable selection method allows analyses of this data in order to establish a QSAR model with the subset of descriptors that are most statistically significant in determining the biological activity.
- The statistical methods can be broadly divided into two : linear and non-linear methods. In statistics a correlation is established between dependent variable(s) (biological activity) and independent variable(s) (molecular descriptors). The linear method fits a line between the selected descriptors and activity as compared to non-linear method which fits a curve between the selected descriptors and activity.
- The statistical method to build QSAR model is decided based on the type of biological activity data. Following are few commonly used statistical methods :

Categorical Dependent Variable

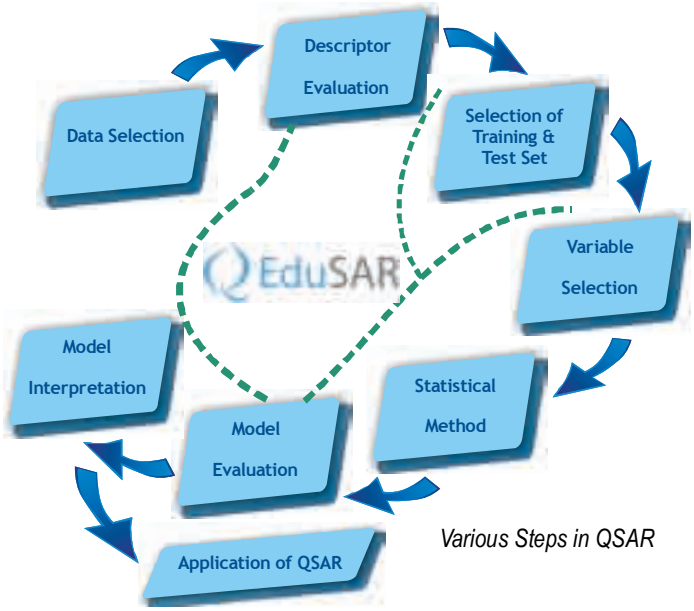
   a) Discriminant analysis
   b) Logistic regression
   c) k-Nearest Neighbor classification
   d) Decision Trees
   e) SIMCA

Continuous Dependent Variable

   a) Multiple Regression
   b) Principal Component Regression
   c) Continuum Regression
   d) Partial Least Squares Regression
   e) Canonical Correlation Analysis
   f) k-Nearest Neighbor method
   g) Neural Networks



Contribution plot: This plot shows the relative contribution of each descriptor in the model for explaining variation in the activity



Various Steps in QSAR

Evaluation of the Model :

There are various statistical measures available for evaluation of the significance of the model, following are most commonly used:

n	number of molecules (> 20 molecules)
k	number of descriptors in a model (statistically n/5 descriptors in a model)
df	degree of freedom (n-k-1) (higher is better)
r2	coefficient of determination (> 0.7)
q2	cross-validated r2 (>0.5)
pred_r2	r2 for external test set (>0.5)
SEE	standard error of estimate (smaller is better)
F-test	F-test for statistical significance of the model (higher is better, for same set of descriptors and compounds)
F_prob.	Alpha error probability (smaller is better)
Zscore	Z score calculated by the randomization test (higher is better)
best_ran_q2	highest q2 value in the randomization test (as low as compared to q2)
best_ran_r2	highest r2 value in the randomization test (as low as compared to r2)
alpha	statistical significance parameter by randomization test (<0.01)

Note: Comment in the parenthesis suggests the minimum recommended values for significant QSAR model

Fitness plot: This plot shows the significance of model predictive ability both in terms of explaining the variation in the activity as well as prediction of molecules in external test set.

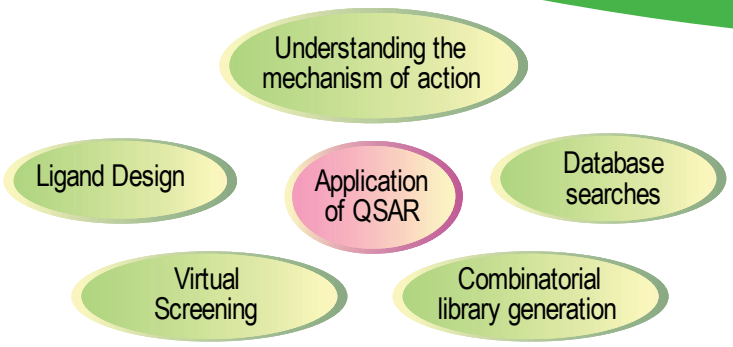
Interpretation of Model:

- Multiple regression is widely used method for building QSAR model. It is simple to interpret a regression model, in which contribution of each descriptor could be seen by the magnitude and sign of its regression coefficient.
- A descriptor coefficient magnitude shows its relative ontribution w.r.t other descriptors and sign indicates whether it is directly (+) or inversely (-) proportional to the activity.

Variety of application of QSAR :

In drug discovery and environmental toxicology, QSAR models are now regarded as a scientifically credible tool for predicting and classifying the biological activities of untested chemicals.

- Distinguishing drug-like from non drug-like molecules
- Drug resistance
- Toxicity prediction
- Physicochemical properties prediction (e.g. water solubility, lipophilicity),
- ADME properties prediction (e.g. Gastrointestinal absorption, blood brain barrier, drug metabolism),
- Activity of peptides etc.



Various application of QSAR

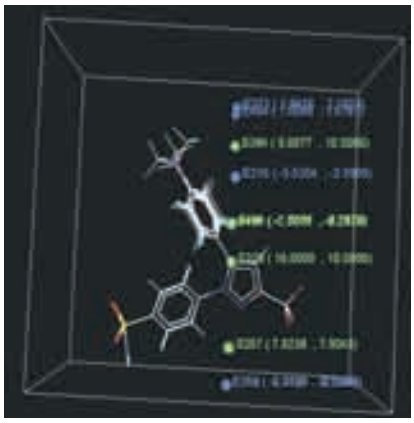
Comparison of methods in QSAR :

Here a series of substituted 1,5-diphenylpyrazoles as COX-2 inhibitors (NSAID) is selected to demonstrate the application of various QSAR methods [Molecules, 2000, 5, 945-955]. In this dataset the biological activity was expressed in terms of IC50 for COX-2 enzyme inhibition. For QSAR the dataset was divided into 25 molecules as training set and 5 molecules as test set as described in the original paper [Molecules, 2000, 5, 945-955].

A conventional QSAR model was derived from various calculated 2D molecular descriptors. At first all the descriptors were subjected to stepwise forward multiple linear regression analysis which resulted in a 5 descriptors model with good r2 (0.93) and q2 (0.87) but poor external prediction r2 (pred_r2 = -0.32). Then PLS regression method was applied on selected set of 8 descriptors which resulted in a statistically significant model with 6 PLS components (r2 = 0.93, q2 = 0.74 and pred_r2 = 0.85).

To build 3D QSAR models i.e. RSA, MFA, CoMFA and kNNMFA models, the steric fields were generated using the Tripos force field and electrostatic fields were generated using MOPAC charges obtained from PM3 optimized structures of the molecules. The alignment of the molecules was done based on the common fragment of 1,5-diphenylpyrazole.

From this study it was shown that kNN-MFA approach provides a statistically better model (q2 = 0.82 and pred_r2 = 0.90) as compared to other 2D or 3D QSAR approaches i.e. CoMFA (q2 = 0.68 and pred_r2 = 0.68).



kNN-MFA Result Plot : 3D alignment of molecules with the important steric and electrostatic points contributing to the model with ranges of values shown in parenthesis

Teaching & self-learning Software on QSAR principles and Applications

EduSAR is an excellent academic package for the new users to understand and explore the various applications of QSAR in drug discovery

- Easy to learn and use with specially designed 2D draw builder, intuitive data manager and visuzlizer
- Extensive set of physicochemical descriptors
- QSAR building with regression methods
- Useful graphs for analysis of QSAR model
- Effortless activity prediction using saved regression
- Ready to learn and use case studies and tutorials
- Handbook on QSAR as a concept and training material for new user
- Import and export facility (csv, xls, txt) for easier data handling and reports

EduSAR is developed with a view for effectively structuring a short training course around QSAR with exhaustive educational material and demonstration exercise.