CHAPTER-III

Methodology of QSAR
Drugs exert their biological effect by participating in a series of events that include transport, binding with receptor and metabolism to an inactive species. Since the interaction mechanism between the molecule and the putative receptor are unknown in most cases (i.e. no bound crystal structure) it is reduced to making interference from properties which can be easily obtained i.e. molecular properties and descriptor to explain such interaction for unknown molecule. Once the relationship is defined it can be used to aid in the prediction of new or unknown molecule. There are number of approaches being applied for rational drug design which involve the different correlative methodologies. The basic objectives of such correlative methodologies are:

1. To search the relevant descriptor parameters which can account the observed biological activities of a group of compounds in a series and,

2. To determine the extent of correlation between a set of descriptors and the activity. The statistical and physical methods have been developed for this purpose. However, some factors are very important in the choice of method such as, the quality of biological data, the number of compounds to be tested, the degree of variance in results and the ratio of the time needed in synthesizing and testing the compounds for their biological activity.

In recent advances, various QSAR methodologies are developed in order to obtain more potent compound (drug) with lesser dependence on the trial and error synthesis and testing. These methods are classified as:

3.1. Classical Approach

3.1.1. Topliss Scheme

3.1.2. Fibonacci Search technique

3.1.3. Sequential simplex search
3.2. QSAR Approach

3.2.1. Extra thermodynamic approach or Hansch analysis

3.2.2. Free-Wilson approach (Additive model or De-novo approach)

3.2.3. Mixed approach of Free-Wilson and Hansch analysis.

3.3. Other statistical methods used in QSAR analysis

3.3.1 Cramer’s substructural analysis

3.3.2. Rank correlation analysis

3.3.3. Linear discriminant analysis

3.3.4. Principal component analysis (PCA)

3.3.5. Cluster analysis Partial least square (PLS) methods

3.3.6. Pattern recognition methods

3.3.7. Partial least square (PLS) methods

3.3.8. Genetic Function Approximation (GFA)

3.3.9. Genetic Partial least squares (GPLS)

3.3.10. Regression analysis

3.1. Classical Approaches

3.1.1. Topliss scheme

The scheme proposed by John Topliss (1972), is a non statistical method to automate the Hansch approach. The method is useful in planning which analogue to be synthesized if compounds are being synthesized and tested one at a time. The procedure is based on assumption that the biological activity of various analogues primarily depends on hydrophobic and electronic substituent effect i.e. on $\pi$ and $\sigma$. Any change in activity following the exchange of substituent, is therefore connected with $\pi$ and $\sigma$ values of substituent in a defined and logical manner, thus based on this logic, a decision tree can be constructed permitting the discovery of new lead. The procedure does not take steric effect into account. The method assumed that the lead compounds of interest
contained at least one phenyl ring which could serve as the template for functional group modification.

First of all unsubstituted compound is tested, regardless of which substituent is selected first, it is necessary that-

1. The substituent has no extremely hydrophobic or electronic properties
2. The substituent must be more hydrophobic than hydrogen.
3. The synthesis of corresponding compound can be achieved in a simple possible manner.
4. Increase in activity is indicated by a ‘+’ sign while decrease by a ‘–’ sign.

In a ring system the above conditions are best satisfied by the halogens as a substituent. The applicability of procedure limits in case where it is uncertain, indeed the biological activity principal depends on ‘π’ and ‘σ’ or ‘π’ and ‘Es’ combination. However methods based on ‘π’ and ‘σ’ are more successful because that combination occurs more frequently.

“Martin and dun” state that an approximately 60% economy of effort can be achieved by the method. Thus Topliss method gives fairly good results but cannot be always accepted to be successful.

3.1.2. Sequential Simplex Procedure

The biological activity of a compound can be represented in a quantitative manner, by any combination of two molecular parameters. Each compound or substituent can be represented as a point in 3D coordinate system with co-ordinates ‘BA’ (Biological activity), π and σ (Molecular parameter). These points lie in a plane that has a maximum in relation to BA at given value of π and σ and thus more active compound
can be found rapidly and without difficulty, using the simple geometric methods of ‘Darvas’\textsuperscript{7,8}.

All the compounds or substituent of potential interest are described in the $\pi$-$\sigma$ plane. Three compounds are selected for experimental analysis with their biological activity value as 3\textsuperscript{rd} co-ordinate (BA). A triangle obtained by connecting these three points is known as simplex. The point for the compound with lowest activity is connected to the midpoint of opposite side by a straight line. The compound to be chosen next is the one which is most closely to this line. ‘BA’ determined for this new compound. In this manner new simplex can be constructed until the maximum has been obtained.

The method is restricted to two molecular parameters. If more dimensions are needed, a corresponding computer variant can be used.

### 3.1.3. Fibonacci Search Technique

This technique was first applied by ‘Bustard’\textsuperscript{10}. The method can be employed whenever the biological activity within a series of compounds primarily depends on only one molecular parameter in a mono-external manner. The extremum (most) active compound is found through a definite number of steps, so called ‘Fibonacci frequency’ (i.e. No. 1, 2, 3, 4, 5, 8, 13, 12……..). The compounds to be synthesized are arranged in order of ascending values of the molecular parameter with which a mono external relation is presumed and then numbered. The first compound is assigned the number ‘0’ so that the highest number in a given sample of ‘n’ compounds is equal to ‘n-1’.

‘Deming’\textsuperscript{11} pointed the limitation of method that it fails if too much noise is present at response surface.
3.2. QSAR Approach

3.2.1. Hansch Linear Free Energy Model

The most significant approach to QSAR having high predictive ability was introduced by ‘Crowin Hansch’\textsuperscript{12-16}. This concept extends to the principle of \textit{linear free energy relationship (LEER)} to describe the effectiveness of a biological active molecule\textsuperscript{17}. It postulated that there are two distinct stages in drug action:

(A) The physicochemical aspects of drug transport i.e. a ‘random walk’ from point of administration to the site of action. It involves passage across hydrophilic and lipophilic barriers. This is related to partition coefficient i.e. $BA: f(P)$

(B) Drug receptor interaction i.e. attachment for the receptor site $K_x$ which depends on\textsuperscript{18}

(i) Shape of molecule (stereochemistry of its substituent group),

(ii) The electronic density on the attachment group.

The random walk process:

\begin{center}
\includegraphics[width=\textwidth]{random_walk_process.png}
\end{center}
\[ BA=f(\text{Transport+ Binding}) \]
\[ BA = K_1 \text{(lipo)} + K_2 \text{(lipo)} + K_3 \text{(Pol)} + K_4 \text{(elect.}) + K_5 \text{(Steric)} + K_6 \]

Hansch group suggested the linear and non-linear dependence of B.A. on different physicochemical parameters in a given group of drugs that have analogous structure and act by same mechanism, their physicochemical parameter \( (\pi, \sigma \text{ and } E_s) \) play a major role.

**(A) Hydrophobicity Constant \( (\pi) \)**

It is based on partition coefficient and analogous to Hammett’s constant.\(^{19-21}\)

\[ \pi_x = \log P_x - \log P_H \]

Where

- \( \log P_x \) = Partition coefficient of substituted compound.
- \( \log P_H \) = Partition coefficient of unsubstituted compound.

A more positive \( \pi \) value indicates higher lipophilicity of substituent and vice-versa.

**(B) Hammett’s Substitution Constant \( (\sigma) \),**

**(C) Steric effects, described By the Taft Value \( (E_s) \).**

Hansch equation is simply QSAR equation that correlates physicochemical properties to the activity in a quantitative manner. These equations are useful when correlated to B.A. in a statistical procedure known as ‘multivariate regression analyses. The linear and non-linear (parabolic) dependence of biological activity can be represented by equation (i) and (ii).\(^{22-24}\)

**eq. (i) (linear relation)**

\[ \log 1/C = K_1 (\log P) + K_2 (E_s) + K_3 (\sigma) + K_4 \]

**eq. (ii) (parabolic relation)**

\[ \log 1/C = K_1 (\log P) + K_2 (\log P)^2 + K_3 (\sigma) + K_4 (E_s) + K_5 \]
Equation (ii) is a typical Hansch equation, where $K_1$ to $K_5$ are Regression coefficients and can be determined by computer in order to get best fitting line. The variance in biological activity ($1/C$) is explained by the variance of linear free energy related to substituent constant, it describes the variance in lipophilic/Hydrophilic, electronic, steric or other molecular properties.

Since log P is present as both $(\log P)$ and $(\log P)^2$ because lipophilicity tends to follow a parabolic relation with activity. The change in electronic properties expressed in terms of Hammett constant ($\sigma$), Pka values, proton chemical shift (from NMR spectroscopy), field effect constant (f) and resonance effect constant (R). The steric influence of substituent can be expressed by Taft steric constant (Es) Molar volume (Mc) and Molar refractivity (MR).

**Hansch model is an extra thermodynamic approach?**

Though the drug transport processes and drug receptor interactions are complex in nature, yet they are essentially physicochemical and can be factored into electronic (E), Hydrophobic and steric parameter. The variance in B.A. depends on change in these physicochemical parameters and can be expressed mathematically as:

$$BA = f (\Delta H, \Delta E, \Delta S) + \text{constant} \quad \text{......................... eq.(iii)}$$

The more general form of equation (iii) can be represented by equation (iv)

$$BA = a (\text{Hydrophobic}) + b (\text{electronic}) + C (\text{steric}) + \text{constant} \quad \text{...... eq.(iv)}$$

The biological activity (B.A.) can be represented in terms of free energy change which occurs during biological response. Free-energy change can be calculated as proportional to inverse of negative logarithm of concentration of compound.
Although such linear free-energy relationship can be stated in terms of thermodynamic parameter, no thermodynamic principle states that the relationship should be true, therefore it is known as extra thermodynamic approach.

**Advantages of Hansch Approach**

1. It permitted complex biological systems to be modeled successfully using simple parameters ($\pi$, $\sigma$ and $E_S$ etc.).
2. The Model has been used successfully to predict substituent effects in a wide number of biological assays.
3. The method is quick and easier and predictions are quantitative and may be evaluated statistically.
4. Potential Extrapolation – Conclusion or results obtained may be extended to chemical substituents, not included in the original analysis.

**Limitations**

1. Large number of compounds is required to explore adequately all structural combinations (a training set for which physicochemical parameters and B.A. is available).
2. Partial protonation of drug molecules at physiological conditions is problematic (can be included in mathematical model if necessary).
3. The QSAR equations developed are as good as the data that has been entered. It is crucial to choose a valid set of compounds in order to carry out analysis. (Craig plots are useful in order to identify suitable substituents).
4. Correlation between physical descriptor *viz* Hydrophobicity will have same correlation with size and thus the Taft steric term.
5. The analysis method did not lend themselves to the consideration of

\[
\Delta G = -2.303 \, RT \, \log K = \log \frac{1}{C} = -\log \text{B.A.}
\]

OR

\[
\Delta G = -\log \text{B.A.}
\]
conformational effect.

3.2.2 Free Wilson Model

It is an additive mathematical model based on a statistical statement of typical untested assumption of drug analogue design that within a series of related compounds, modified at more than one position, the effect of a particular substituent at a specific position is independent of the effect of substituent at other positions (Free-Wilson-1964). The basic assumption is that within a homologous series of drug, individual segment of molecule makes additive and constant contribution to the biological activities. If such contribution is known, prediction/calculation of B.A. can be achieved. Generally there are two types of segments in a drug molecule –

I. ‘A constant segment’ which is identical to parent structure shared by all compounds of the series.

II. ‘A variable segment’ which are open to substituent as individual substitution sites.

Such additive concept implies that activity contributions of the parent molecule and each substituent are constant.

\[
\log \frac{1}{C} = B.A. = \mu \sum G_{ij} + \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldOTS
positions in different manner, it is possible to devise a uniform code for each variable segment on the basis of usually two digit notation $JK$, where $J$ is the substitution site and $K$ is substituent.

\[
\text{Biological Activity} = \text{Contribution of Unsubstituted Corresponding Parent Compound} + \text{Contribution of Substituents}
\]

Thus the structure of any compound of series can be shown by a vector $f^0$ and structural parameter like ‘$b_{ijk}$’ which has either a value 0 or 1. The structural parameter $b_{ijk}$ indicates whether the $i^{th}$ compound of series, the variable constant $j_k$ is present ($b_{ijk}=1$) or absent ($b_{ijk}=0$).

The contribution of parent structure and of variable segment to activity is designated by $\mu$ and $Z_{ijk}$ respectively. On the basis of Free Wilson model the B.A. of a given compound in the series can be expressed as the sum obtained from $\mu + Z_{ijk}$.

\[
\log A_i = \mu + b_{ijk} \cdot Z_{ijk}
\]

Where $A_i = \text{B.A. of } i^{th}$ compound.

When the molecule of series have ‘$p$’ substitution sites ($j=1----p$) and ‘$mj$’ substitution placed at the $i^{th}$ substitution sites ($K=1----ij$) then-

\[
\log A_i = \mu + \sum_{j=1}^{p} \sum_{k=1}^{mj} b_{ijk} \cdot Z_{ijk}
\]

or

\[
\log A_i = \mu + \sum_{j=1}^{p} \sum_{k=1}^{mj} \sum_{l=1}^{J} b_{ijk} \cdot Z_{ijk}
\]

In summary the entire Free-Wilson model can be interpreted as biological effects of individual structural elements (segments) are assessed by calculating and assigning activity contribution $Z_{ijk}$ and $\mu$, on such line that gradation of activity contribution quantifies the influence of the
segments on biological activity.

‘Purcel and Clayton’\textsuperscript{30} as well as ‘Ban and Fujita’\textsuperscript{31} supported the Free-Wilson model. In Ban-Fujita modification of Free-Wilson Model, no assumptions are made about the relevance of the model parameters to the B.A. of molecule. The effect of each substituent is considered to be independent of any other; each makes a constant contribution to the overall activity of the molecule. Therefore the method is applicable to the compound with more than one variable group.

**Advantages of Free- Wilson Analysis:**

The Free-Wilson method substantially reduces the number of analogues required explore the effect of multiple substitution pattern.

1. ‘Free-Wilson or De novo approach’ recognizes the importance of a particular substituent at specific position of the molecule. This involves the putting of an entry to the QSAR equation that contributes a specific value, when the group is present but contributes nothing when group is absent.\textsuperscript{32}

2. Method does not require use of substituent constant such as $\pi$, M.R., $\sigma$, R, f, $E_S$ etc. The equation is solved by multiple linear regressions using the presence (1) or absence (0) of different substituent as independent parameters while measured activities serve as dependent variable.

3. The contribution of each substituent at particular position can be identified with recognition of substituents that fulfil or do not fulfil the concept of additivity.

**Limitations of Free-Wilson Model:**

1. Structural variation is must in at least two different positions of substituents; otherwise meaningless group contributions would result, one for each compound.
2. It is not possible to make prediction for the substituents which are not included in the analysis.

3. Method demands that the effect of substituents on drug molecules are additive.

### 3.2.3. Combined Approach or Mixed Approach

‘Kubinyi’\(^{33-34}\) combined Hansch and Free-Wilson models as “Mixed Approach”, which offers the advantage of both individuals. It widens the applicability of QSAR.

\[
\log \frac{1}{C} = - K_1 (\log P)^2 + K_2 (\log P) + K_3 \pi + K_4 \sigma + K_5 E_s + K_6 \quad \text{...........(1)}
\]

**Hansch Model**

\[
\log \frac{1}{C} = \mu + \sum a_{ijk} \quad \text{.................................(2)}
\]

**F.W. Model**

This is based on the theoretical and numerical equivalence of Hansch linear multiple regression model and modified F.W. Model, and represents the combination of both these models (equation 3 and 4).

\[
\log \frac{1}{C} = \sum_i a_{ij} + \sum K_j \Phi_j + K \quad \text{...........................................(3)}
\]

(Representing linear form)

Where \(K_j = \text{coefficient of different Physico-chemical parameters.}\)

\[\Phi_j = \text{Physico-chemical properties (\(\pi, \sigma\) and \(E_s\) etc.)}\]

\[
\log \frac{1}{C} = K_1 \pi^2 + \sum_i a_{ij} + \sum K_j \Phi_j + K \quad \text{...........................................(4)}
\]

(Representing Parabolic form)

In this expression \(\sum a_{ij}\) is the F.W. part for substituents ‘\(X_i\)’ and ‘\(\sum K_j\)’ is the Hansch part for the parabolic dependence of \(\log \frac{1}{C}\) on lipophilic character for substituent ‘\(X_i\)’.

**Application of Mixed Approach**
It is the most significant QSAR approach, applied for present work. For the successful application of mixed approach it is highly recommended to derive Hansch equation for each subset and to compare whether they correspond to each of indicator parameter.\textsuperscript{35-36}

However F.W. Model suffers by its inability to attribute physical significance towards substituent contribution and fails to yield significant correlation in that case where Hansch’s parabolic model is applicable. Sometimes it is also difficult to correlate B.A. only in terms of known physico-chemical parameter. The use of ‘\textit{Indicator parameter (I)}’ mixed approach overcome these difficulties. The Derived QSAR equation now contains part A (with Hansch Physicochemical parameters) and part B (Indicator variables based on F.W. assumptions). It can be described as;

\[
\log \frac{1}{C} = K_1(\log P)^2 + K_2(\log P) + K_3\sigma + K_4 E_S + K_5 I_1 + K_6 I_2 + K_7 \ldots
\]

In mixed approach, indicator parameters are often used to describe pharmacophore structure, ortho-interaction, separator between two isomers (R and S) or (cis-trans) or for a substituent for steric parameters etc. It is also useful in that case where variations of similar B.A. data are large and different sets of congeners are explained by single equation.

In certain cases, the B.A. values can’t be determined accurately due to various reasons e.g. lack of sensitivity of a particular test system therefore alternative statistical techniques are used to simplify the relation. These methods involve the classification of compounds as active, partially active or inactive. The resulting data set is then searched for pattern which predicts these categories.
3.3. Other Statistical Methods used in QSAR

3.3.1. Cramer’s Substructural Analysis\(^{37-39}\)

‘Berkoff’, ‘Cramer’ and ‘Redi’ attempted to apply sub-structural analysis in field of QSAR. In this method compounds of training series are fragmented into sub-structure using library of atoms, bond and substructures. For each feature sub structural activity frequency is calculated as;

\[
SAF_j = \frac{\text{No. of active compounds containing the feature } j}{\text{Total No. of compounds containing the feature } j}
\]

Where \(SAF_j\) = Probability Contribution of \(j^{th}\) feature to overall probability.

To characterize the activity of compounds, the mean substructural activity frequency for \(i^{th}\) compound is given as;

\[
\text{MSAF}_i = \frac{1}{M_i} \sum_i b_{ij} \cdot SAF_j
\]

Where

\(M_i = \text{No. of features (fragments) occurring in the } i^{th} \text{ compound}\)

\(b_{ij} = \text{Sub structural descriptor, defined as:}\)

\(b_{ij} = 1 \text{ if } j^{th} \text{ feature is present in } i^{th} \text{ compound}\)

\(b_{ij} = 0 \text{ if } j^{th} \text{ feature is absent in } i^{th} \text{ compound}\)

Experimental studies reveal that these MSAF values are related with B.A. of compound of given series.

3.3.2. Rank Correlation Analysis
It was developed by ‘Sklenar’ and ‘Jager’. Before applying this procedure to QSAR it is assumed that in homologous series there exists a defined and monotonous relation between the level of biological action and molecular properties relevant to the action. This is less strict requirement than the linear model assured to hold in the extra thermodynamic approach which is certainly realistic especially when dealing with quantum chemical molecular parameter. Such relation can be examined by means of Rank correlation analysis in following manner;

First step comprises the transformation of BA values and molecular parameters concerned into rank numbers. These numbers indicate the position of each value when respective details are ranked in decreasing order, as:

\[ r_{sj} = 1 - 6 \frac{\sum_{i=1}^{n} d_i^2}{(n^3 - n)} \]

Where \( d_i \) denotes the difference of Rank number of parameter \( X_j \) and B.A. of \( i^{th} \) compound and \( n \) is the number of compounds in the sample. This equation serves to characterize the connection between molecular property (given) and B.A. If the correlation between A and \( X_j \) is perfect

\[ r_{sj} = 1 \]

3.3.3. Linear Discriminant Analysis

The method was first introduced by ‘Yvonne Martin’ in 1974. It is a statistical technique used to calculate the combination of predictor variables which best distinguish member of pre established groups. In drug designing strategy it is possible to examine the weightage of the statistical significance of various physical properties which might distinguish active from inactive analogues.
The discriminant functions in this analysis serve as classification of logarithms, representing weighted linear combination of features relevant for class separation. There are several types of discriminant functions, among which the non-elementary ones have optimal properties.

If \( Q \) classes occur (\( q = 1, \ldots, Q \)), \( Q-1 \) non-elementary discriminant functions \( W_k [k = 1, \ldots, (Q-1)] \) of the general form, then

\[
W_k = \sum_{j=1}^{m} a_{kj} X_j
\]

Where \( a_{kj} \) denotes weight coefficient of the \( J^{th} \) mole parameter (\( J = 1, \ldots, m \)) in the \( k^{th} \) discriminant function and \( W_k \) is being designated as non-elementary discriminant variables. There are following steps:

1. Selection of Molecular properties and formulation of approaches.
2. Check whether classes can be separated using these parameters and elimination of parameter not relevant for class separation.
3. Computation of non-elementary discriminant functions.
4. Interpretation of discriminant function and reclassification of the compounds from the training series as a means of testing of the quality of separation.
5. Classification of compounds not yet investigated so that further synthesis and tests can be planned while allowing for mechanistic conclusions from the form of discriminant functions.

### 3.3.4. Principal Component Analysis (PCA)

PCA is a statistical tool used in QSAR analysis and can be applied to a data set obtained from a series of compounds tested for their biological potency in several test systems. These quantitative activity variables i.e. Principal components may then be correlated one by one with a series of physicochemical parameters.

In fact, PCA is a data reduction method, using mathematical
techniques to identify pattern in a data matrix. PCA seeks to find simplified relationship in data by transforming the original parameters into a new set of uncorrelated variables i.e. Principal component. The overall PCA finds the set of orthogonal axes for the data which decompose variance in data. The main steps of PCA are summarized as:

1. Calculation of correlation matrix.
2. Evaluation of object and system components by principal component method.
3. Determination of the number of relevant principal components necessary to reproduce data matrix with in experimental error.
4. Uniqueness test.
5. Selection of test vectors.
6. Identification of the object components with test vectors.
7. Formulation for the system of regression equation describing the variables in terms of parameters used as test vectors.

3.3.5. Cluster Analysis\textsuperscript{44-45}

The method is useful to describe the relationships between observations which have associated with them a number of properties. Cluster analysis is used to study a large set of substituent to establish which subsets of the total were similar in physical properties \textit{(Hansch et al. 1973)}. During cluster analysis to synthesize a series of compounds with the maximum variation in minimum number of examples, one member from each subset (cluster) would be chosen.

The clustering calculations may either start from one cluster per observation or progressively coalesce clusters to form fewer and fewer or it may start with all of the observations as one large cluster and progressively form more and more clusters. The former method is agglomerative than latter one.
3.3.6. Pattern Recognition Method\textsuperscript{46-47}

It is the collection of computer-based methods used for detecting previously unknown relationships (patterns) with in large masses of multivariate data. It is done without making assumptions about the underlying statistics of the data. The method is useful in the same manner as cluster analysis, to examine distance or relatedness between observations. Because of lack of assumptions with respect to the distribution of the data, probability estimate are not made and statistical tests are not applied to the result of calculations.

3.3.7. Partial least square (PLS)\textsuperscript{49-51}

PLS is an iterative procedure that can be applied two criterions to produce its solution.

1. To extract a new compound so as to maximize the degree of common feature between the entire structural-parameters column collectively with experimental data.

2. For improvement in the ability to predict the depending variable.

Techniques used in PLS to assess the predictive ability of a QSAR is cross validation i.e. excluding temporarily the unknown compounds and then using resulting equation to predict experimental measurement of the omitted compounds. Due to large number of independent variables, cross validation procedure must be used to select the model with highest predictive ability. The chief drawback of PLS is that it can make only linear correlations, however, complex relationship between mole properties and BA are often do not show linearity.
3.3.8. Genetic Function Approximation (GFA) \(^{52}\)

GFA logarithm is a novel technique for constructing QSAR Models that are created by evolving random initial models using a genetic logarithm. GFA begins with a population of randomly constructed QSAR models; these models are rated using an error measure which estimates each model’s relative productiveness. The population is evolved by repeatedly selecting two better rated models to serve as parents, then creating a new child model using terms from each of the parent model. The worst rated model in the population is replaced by new model. As evolution proceeds, the population becomes enriched with higher and higher quality models. Experimental results against published data sets demonstrates that GFA models are comparable to and in some cases are superior to them, discovered by using standard techniques such as stepwise regression, linear regressions and partial-least square regression method.

3.3.9. Genetic Partial least squares (GPLS)

This method is the combination of PLS and GFA (Genetic Function Approximation) which retains the ease of interpretations of GFA by back transforming the PLS component to the original variable. The method results in collection of different small models that have same high predictability.

3.4. Regression Analysis \(^{53,54}\)

The verbal meanings of the term ‘Regression’ is the act of ‘returning or going back’. Regression analysis reveals the average relationship between two variables, thus helps to estimate the unknown value of one variable from the known value of the related variables. In order to explore
the causal effect of one variable upon other, the data is assembled on the underlying variables of interest and regression is employed to estimate the quantitative effect of the causal variables upon the variables that they influences.

3.4.1. Regression line

The average relationship between two variables is described by regression line. When the exact value of one variable is given the most probable value of other variable is shown in the regression line.

Properties of Regression Lines

1. If the correlation between the two variables is perfectly positive that is when the correlation coefficient ‘r’ is equal to +1, the two regression lines will coincide with each other. It means that there will only one line instead of two lines. In such situation the regression line will be as follows:

![Perfect Positive correlation](image-url)
2. On the other hand, if the correlation between two variables is perfectly negative that is when the correlation coefficient ‘r’ is equal to -1, the two regression lines will coincide and lines will be as follows:

![Perfect Negative Correlation](image)

3. When there is no correlation between two variables that is when the correlation coefficient ‘r’ is equal to 0, the regression lines will intersect each other at right angles. They will intersect each other at the point \( \bar{x}, \bar{y} \).

![No Correlation](image)

3.4.2. Role of Regression analysis in QSAR (Construction of predictive models)

Regression models help to build models, estimate their predictive ability and find underlying relationships between descriptors. For the analysis applied to a QSAR study, the main goal is to correlate a biological property, forming a column vector(y), with molecular descriptors, arranged in the columns of the so called data matrix(X). The columns are associated
with variables or descriptors, whereas the objects, in this case, the molecules are associated with the rows. The model then can be used to predict activities for new molecules, or screening a large group of molecules with unknown activities among other uses. Usually, the prediction model is elaborated using the parameters calculated for a well determined data of training set on the unknown test set. If the training set is a sufficiently representative pattern of the system, then, it can be assumed that the introduction of new elements with an unknown property will not affect their stability and that confident predictions can be attempted. A model’s ability to provide insight into the system is as important as its predictive ability. Possibly more valuable than being able to predict an activity or property is to get insight into underlying relationships between descriptors. Finally, validation methods are needed to establish the predictive capacity of a model on test data and to help determine the complexity of an equation that the amount of data justifies.

3.4.3. The significance and validity of QSAR regression equations

Once a regression equation is obtained, it is important to determine its reliability and significance. Several procedures are available to validate the regression equation. These can be used to check that the size of the model is appropriate for the quantity of data available as well as provide some estimate of how well the model can predict activity for new molecules.

In general, a regression equation can be accepted in QSAR studies,

- If the correlation coefficient $r$ is around or better than 0.9 for *in vitro* data and 0.8 for whole animal data (its value depends not only on the quality of fit but also on the overall variance of the biological data).
- If the standard deviation $s$ is not much larger than the standard deviation of the biological data.
• if its F value indicates that the overall significance level is better than 95%

The equation has to be rejected:
• If the number of variables included in the regression equation (or used for the selection of variables to be included in the equation) is unreasonably large.
• If the standard deviation s is smaller than the error in the biological data (over prediction by the model).

3.5. Terms commonly used in Regression Analysis

To express quantitatively a correlation, graph is usually a preferred method over regression analysis. In regression analysis to show how representative of the results the correlation is, additional data is to be given in terms of the correlation coefficient (r), number of compounds utilized (n), standard deviation (s) and statistical validity (F). When the number of variables exceeds 3, the results cannot be expressed in the form of either a graph or a model. A regression equation therefore remains the only method of expression which can be used in such situation.

(a) **Correlation coefficient (r)**

The correlation coefficient ‘r’ is a relative measure of the quality of the fit of the model because its value depends on the overall variance of the dependent variable. While the correlation coefficient ‘r’ of the two subsets is relatively small, but the correlation coefficient derived from the combined set is much larger, due to increase in the overall variance

\[ r = \sqrt{1 - \frac{\sum \Delta^2}{S_{yy}}} \]

where,

\[ S_{yy} \] is the overall variance i.e. \[ S_{yy} = \Sigma(y_{obs} - y_{mean}) \]
High value of regression coefficient \((r>0.9)\) indicates the statistical significance of the regression equation is high. While the low value of \(r\) indicates that the substituents constant is not important for the process under consideration.

If the ‘\(r\)’ value does not decrease significantly when a particular substituent (coefficient) constant is omitted from the equation, it means that process represented by equation is least affected by the factor symbolized by that particular substituent coefficient.

(b) **Square Of The Correlation Coefficient \((r^2)\)**

It can be envisioned as the fraction of total variance in the data which is explained by the regression model.

\[
r^2 = 1 - \frac{\Sigma \Delta^2}{Syy}
\]

Where,

\[
Syy = \Sigma (y_{obs} - y_{mean})
\]

\[
\Sigma \Delta^2 = SSQ - \Sigma (y_{obs} - y_{calc})^2
\]

Where ‘\(y_{obs}\)’ is observed biological activities, ‘\(y_{mean}\)’ is mean of biological activities value ‘\(y_{calc}\)’ is calculated biological activity used in the equation.

The squared correlation \(r^2\) is a measure of the explained variance, most often presented as a percentage value e.g. \(r = 0.8\) then \(r^2 = 0.64\) or 64\%. Data accounted by regression of that parameter, still having 36\% data yet unaccountable. Greater the value of \(r^2\) lesser is the variance that remains unaccounted by the equation.

The term explains about percent data represented by that particular equation. For example, if \(r = 0.7\) then \(r^2 = 0.49\) or 49\% data is accounted by regression of that parameters, still leaving 51\% data yet accounted. Thus the value of ‘\(r\)’ can be improved by inclusion of another parameter. Thus the term ‘\(r^2\)’ help us to understand whether other parameters should be
sought for or not. Greater the value $r^2$, lesser is the variance (data) that remains unaccounted by the equation.

**(c) Standard Deviation (SD):** This value gives us an idea about precision of that equation. This is a measure of dispersion or scatter of the observation from the mean and indicates how well the function derived by the QSAR analysis predicts the observed biological activity. Its value considers the number of object $n$ and the number of variable $k$. Therefore, SD depends not only on the quality of fit but also on the number of degrees of freedom. The smaller the value of SD the better is the QSAR.

$$\text{DF} = n-k-1.$$  
$$\text{SD} = \sqrt{\sum (y_{\text{obs}} - y_{\text{cal}})^2 / n-k-1}$$

**(d) F-value:** It is a measure of the statistical significance of the regression model, the influence of the number of variables included in the model is even larger than the standard deviation.

$$F\text{-value} = r^2 \frac{(n-k-1)/k}{(1-r^2)}$$

**(e) Number of compounds utilized (n):** For good correlation large number of compounds must be utilized.

**(f) Outliers:** An outlier may lead to a deeper insight into drug action and may even allow one to arrive at new lead compounds. The detection and follow up of outliers is an important aspect of QSAR methodology. Generally an outlier is an observation that is numerically distant from the rest of the data. In QSAR outlier is defined as the structure with a residual greater than two times the standard deviation of residual. These are the data points, which lie outside the general linear pattern of which the midline is the regression line. In order to obtain good predictive model outlier must be explored and analysis repeated with the outliers omitted.
References

15. Goodford P.J., Hudsom A.T., Sheppey G., Wooton R., Black M.,
25. Graham Patrick “Instant Notes on medicinal chemistry” BIOS Scientific Pub. Ltd. 9, Newtec Place, Magdalen Road Oxford OX4, IRE U.K.
32. Richon A.B., Young S.S.; “An Introduction to QSAR Methodology”


