CHAPTER-II

QSAR-An Overview
(with review of literature surveyed)
2.1. Introduction

To discover a new drug for a particular disease has always been a challenging process for medicinal chemists due to complexity of biological system. Traditional approaches such as ‘Trial and Error’ synthesis of compounds and random screening of them for activity have been proved to be time consuming and uneconomical. Medicinal chemists always tried to synthesize drugs with enhanced activity and fewer side effects. But such efforts involve high manufacturing cost and an unfavourable ratio between synthesized compounds and products on the market. It has been estimated that out of hundreds of compounds, synthesized in research laboratories, only one or none reach the market ultimately as a drug.

![Figure 2.1 Drug Discovery Pipeline]

During recent years, in the field of medicinal chemistry the use of organic chemistry, biology (including biophysics, spectroscopy etc.) and computer science have been proved to be useful for designing of new chemical moieties and predicting their biological activities prior to the synthesis.

In this concern, drug design is an iterative process which begins with a compound that displays an interesting profile and ends with optimizing both the activity profile for the molecule and its chemical synthesis. Such process is initiated when the chemist conceives a hypothesis which relates the chemical features of the molecule (or series of molecules) to the biological activity.\(^1\) The relationship between chemical structure and
biological activity has always been at the centre of drug research. The
development of synthetic organic chemistry and methods for structure
determination coupled with the recognition that changes in chemical
structure lead to change in biological activity had a profound effect on the
search for new medicinal compound.

It was proposed by ‘Cum-Brown’ and ‘T.Fraser’\(^2\) in 1869 that
biological activity of a compound is a function of chemical structure.
‘Crowin Hansch, Claremont and Caif’ (a group of chemist of Pomona
College) has quantified the relationship between the physicochemical
properties and its biological activity. In this context **Structure-Activity
Relationship (SAR)** is determined in which biological Activity of
compound (or series of compound) is the function of physicochemical
(physical, chemical and structural) properties.\(^3\text{--}^5\)

\[
f (B.A.) = f(\text{Physicochemical Properties})
\]
\[
= f(\text{electronic}) + f(\text{steric}) + f(\text{Hydrophobic}) + f(\text{structural}) +
\]
\[
f(\text{theoretical})
\]

From such SAR it is widely accepted that the chemical structure of
the compound contains important information about its physical and
chemical properties. Such SAR observation makes possible to design a
drug with predetermined physico-chemical properties with high therapeutic
index and low toxicity. With above idea Hansch et al. proposed a workable
methodology which involves the notation of finding, examining and
interpreting SAR in a systematic manner i.e. QSAR (Quantitative
structure Activity Relationship). In QSAR, information feeding the drug
design efforts is increasingly quantitative, building upon recent
developments in molecular structure description, combinational
mathematics, statistics and computer assistance. These areas have led to a
new paradigm in drug design i.e. QSAR\(^6\text{--}^7\)

QSAR techniques employ powerful computers, molecular graphics
and sophisticated software; they may be of enormous assistance to those trying to structure the large data base resulting from the massive efforts in drug research. Quantitative statistical methods may be useful in elucidating structural information of clinically active compound as well as processes important to drug action.\textsuperscript{8} QSAR and other related approaches have attracted broad scientific interest, particularly in pharmaceutical industry for drug discovery, in toxicology and environmental science for risk assessment.

\section*{2.2. Principle of QSAR}

QSAR is based on the well known similarity principle, which states that “\textit{similar compounds have similar activity}”. Since the drug interactions are determined by physicochemical parameters of the drug such as polarity, ionization, electrical density etc. All such properties of a drug are determined by the atomic composition, shape, size and chemical structure of drug molecule. Since these physicochemical parameters can be measured experimentally and expressed in quantitative terms, intermolecular binding forces acting as a function of structural multiple also have numerical value. Therefore if the biological action of a drug with in a series is identical, it should be possible to calculate a Quantitative-Structure activity relationship which explains how the physicochemical parameters of a series of compound affect their biological activity.\textsuperscript{9,10}

QSAR is essentially a computerized statistical method which tries to explain the observed variance in biological activity of certain classes of compounds as a function of molecular change caused by the substituent. It involves the derivation of a mathematical formula which relates the biological activity of the group of compound to their physicochemical properties. QSAR attempts to find the consistent relationship between the variance in value of molecular properties and the biological activity for a
series of compounds so that these rules can be used to evaluate new chemical entities. This approach allows important structural requirements for activity to be identified and hence narrows the search for the optimum molecule. QSAR generally takes the form of a linear equation:\textsuperscript{11,12}

\[
\text{B.A.} = \text{Constant} + (C_1 \times P_1) + (C_2 \times P_2) + (C_3 \times P_3) + (C_n \times P_n)
\]

Where parameters $P_1 - P_n$ are computed for each molecule in the series and coefficients $C_1 - C_n$ are calculated by fitting variations in the parameters and the biological activity.

For a given series of compounds, the concept of QSAR may be summarized as: The molecular parameters (XI) and biological activity data (B.A.) are related. Since the biological activity is dependent on molecular structure and the resulting properties, mathematical analysis reveals such connection in the form of so called QSAR.

\textbf{Figure : 2.2 Principle of Quantitative Structure Activity Analysis}

In order to carry out traditional QSAR, a range of analogues having common skeleton are synthesized, their biological activities are measured
and an equation is worked out relating biological activity to the physical properties such as hydrophobicity, electronegativity, steric factor, dipole moment, H-bonding interactions, resonance effects etc.

2.3. Objectives of QSAR

Since QSAR studies are partly retrospective, their objectives are largely predetermined by the type of data collected. QSAR can be constructed for different purpose and according to different methods. Since QSAR describes the connection between magnitude of a given biological effect and drug structure in set of congeners, they can be therefore employed to optimize the effect on the basis of structure variation.\textsuperscript{13}

The goal of SAR modeling is to analyze and to detect the determining factors for the measured activity for a particular system. Such SAR model should have as good predictive abilities as possible to predict the studied biological or characterizing the compounds, whose biological activity is expressed via same mechanism. Thus QSAR attempts to find what feature of molecules affect its activity and what can be modified to enhance their activities. Since QSAR paradigm is based on the assumption that there is an underlying relationship between molecular structure and biological activity which arises from this systematic variation. Thus all the physicochemical properties of chemical substance can be computed from its molecular structure and encoded in a numerical form with the help of various descriptors.\textsuperscript{14,15}

In brief the objectives of QSAR models are to allow the prediction of biological activity of untested and sometimes unavailable compound and to provide insight of relevant physicochemical properties which serve as determinant in the biological activity of compounds.

2.4. Steps involved in QSAR \textsuperscript{16}
There are following steps to be carried out in QSAR method during quantitative drug designing with desired activity profile.

**STEP–I Formulation of a series of biologically active analogues with their biological activity:** First step consists of the selection of already synthesized analogues (lead skeleton) with their particular biological activities. All the compounds of series are structurally similar and have same mode of action. However, the compound must be dissimilar enough to cause some systematic change in their biological activity. The formation of classes of similar compounds can be achieved by dividing the series of compounds of interest into categories on the basis of their chemical structure.

**STEP–II Calculation of Various physicochemical Parameters:** Second step involves the calculation of quantitative values of various physicochemical parameters for various substituent groups present in the series. Generally $\pi$ (as hydrophobic), M.R. (as steric), $f$, $R$, $\sigma_m$ and $\sigma_p$ (as electronic) are calculated during 2D QSAR analysis.\(^{17}\)

**STEP–III determination of Correlation matrix between various physicochemical parameters and biological activity:** The correlation matrix between various physicochemical parameters and biological activity is determined that shows which particular physicochemical parameter is best correlated with biological activity.

**STEP –IV Generation of QSAR equation:** After selection of the best parameter, the QSAR equations are generated. In order to calculate best mathematical expression linking together the physicochemical descriptors and biological response, it is necessary to obtain best information regarding the essential features of chemical and biological data set.

A general form of resulted QSAR equation is;
$-\log C = a \pi + b\sigma + c.Es + e...$

Where

$C$ = Concentration of drug necessary to give specific biological response,

$\pi$ = Hydrophobicity constant,

$\sigma$ = Hammet’s substituent constant,

$Es$ = Taft steric constant,

And $a, b, c … e$ are constant.

**STEP–V  Data Analysis and Interpretation of Results for the proposal of new compound:** By using QSAR equation the biological activity of newly designed compound can be predicted. Then medicinal chemist has to synthesize only those compounds which have predictions of promising biological activity.

In fact QSAR development is an iterative cycle, in which the steps are repeated a number of times, until sufficient significant results are obtained in order to design compound with their desired activity profile. Since step III and IV involve the use of statistics, therefore the validity of QSAR equation is determined in terms of correlation coefficient ($r$), standard deviation ($S$), Fisher test value ($F$) for overall significance of the model etc. From the best fit QSAR equation we will be able to predict the biological activity of future designed compounds with suitable propose by modification in drug moiety, by changing the nature of substituent or group.
2.5. Requirements of QSAR Analysis

1. All the binding data are measured with sufficient precision and bind in a comparable manner.

2. All analogues belong to congeneric series, exert same mechanism of action and act on same target in the same manner i.e. a series of compounds with a similar basic skeleton with varying substituents.

3. A set of physicochemical parameters, that can be easily calculated and which are likely to be related to receptor affinity.

4. Same quantitative activity data set for congeneric set of compounds. This data set is generated experimentally assuming that all the compounds have same mode of activity.

5. Binding affinity is correlated to interaction energies. The effect of isosteric replacement should also be predicted.
6. Biological activity data profile is correlated to the binding affinity.
7. A method for detecting the relationship between physiochemical parameters and the binding data i.e. QSAR.
8. A method for validating the QSAR equation.

### 2.6. QSAR Model

QSAR expresses a multivariate mathematical relationship between a set of physico-chemical properties or descriptors \((X_{ij})\) and an experimental function or biological activity \((Y_i)\). The QSAR relationship is expressed and used to account for the observed activity, for a compound \(i\), the linear equation that relates molecular descriptors, \(X_1, X_2, \ldots, X_n\), to desired activity \(Y\) is:

\[
Y_i = Z_{i1} b_1 + X_{i2} b_2 + \ldots + X_{in} b_n + e_i
\]

Such QSAR equation results into more compact form, for the general case of \(n\) selected descriptors \(X_{ij}\) as:

\[
Y_i = \sum X_{ij} b_j + e_i
\]

Where \(b\) is the linear slopes and expresses the correlation of the particular molecular property \(X_{ij}\), with the activity \(Y_i\) of the compound \(i\), and \(e_i\) is a constant. The slopes and constant are often calculated using regressions analysis. The independent variables (descriptors) are usually physicochemical properties that describe some aspects of the chemical structure, which may be experimentally or theoretically determined. The strength of QSAR model depends on the quality of the variables. The final QSAR equation seeks to find the smallest number of descriptors that can adequately model the activity of compounds in study. The maximum recommended ratio is a single independent variable to five compounds.
2.7. Applications of QSAR Analysis

QSAR is a scientific achievement and economic necessity to reduce an empiricism in drug design to ensure that every drug synthesized and pharmacologically tested, should be as meaningful as possible.

1. **Reduction of Drug price**: The correlation study will hopefully minimize the number of compounds the synthetic chemist will prepare. Therefore it saves time and expenses needed for synthesis and testing of drugs. This simplification leads to reduce the price of drug, which has a strong social pressure.

2. **Diagnosis of Mechanism of Drug Action**: Interpretation of correlation between biological activity and molecular descriptors (physicochemical properties) in terms of QSAR is useful to test a mechanistic hypothesis. The plot between log 1/C versus a variety of physicochemical parameters, it may be predicted that which parameters are important for active enhancements.

3. **Prediction of Activity**: QSAR may be helpful for the prediction of activity of unknown molecule. QSAR for a ‘training set’ of compounds is developed, and then used to obtain mathematical relationship to predict the biological activity of new compounds prior to synthesis.

4. **Reduced Toxicity of Drugs**: Toxicity of drugs can be reduced by means of QSAR, several efforts are made for this purpose. In case of excess toxicity established through any means, the calculated baseline toxicity value can be used to select upper exposure limit.

5. **Paper design of Series**: Since many data sets fit a QSAR equation. It has been recognized that one should plan a series to optimize the chance that the analogues will reveal any QSAR. This planning gives a good chance of finding that which combination of properties proposed to be important in the determination of potency.
6. **QSAR and Receptor Mapping:** Drug molecules are frequently considered to consist of two parts;

- Essential pharmacophore that interact directly with receptor.

- Accessory region that can be structurally modified without destroying the drug receptor interaction. However, such modification may change the strength of interaction.

If all substitutions are in accessory region of molecule (all analogue bind in a comparable manner), it may be possible to derive a QSAR of drug-receptor interaction that quantitatively describe the force involved in such interaction.

7. **QSAR and Bioisoterism:** Since certain pairs of substituents may be bioequivalent or bioisoteric. QSAR parameterization of substituents allows one to quantitate this similarity. The concept may be useful in the early stage of investigating the structural specificity of a lead and also before the synthesis in a series is terminated.

8. **Description of molecular structure:** The description of drug molecular structure, electronic orbital reactivity and role of structural and steric components can be achieved by QSAR.

2.8. **Limitations of QSAR**

- **Lack of universal Success at Predicting Potency**

  Disappointments with QSAR come when the equations do no predict the potency of a new analogue in desired manner. It may be a result of following preventable circumstances:

1. Prediction was based on poorly designed series or invalid or ambiguous regression equations.
2. It was based on an extrapolation outside the range of physical properties represented by the original substituents.

3. There is diversity in the condition of biological test performed.

❖ **Lack of Absolute ways to describe Molecules**

Another serious limitation of QSAR lies in the description of the molecule. It is difficult to transform a 3D structural diagram into a set of members that describe the potential affinity for a receptor and the ability to trigger some biochemical events. Such a difficulty is due to the lack of fundamental understanding of how to quantitatively describe substituent effects on non-covalent intermolecular interactions.

**2.9. Lead Compound and Biological Activity**

1. **Lead Compound**: A lead is the starting point when designing a drug. The lead is a prototype compound that has desired biological property that is likely to be therapeutically useful, but may have several undesirable characteristics such as high toxicity, insolubility or metabolic abnormalities etc. The first step in QSAR analysis is the detection of some biological action in a group so as to serve as lead. Lead optimization is based on the intuition of medicinal chemist combined with traditional lead modifications. This is followed by molecular manipulation to increase or modify the activity. This is based on following considerations:

   1. Molecular structure of the drug,
   2. Behaviour of the drug in biophase,
   3. Geometry of the receptor,
   4. Drug-receptor interactions,
   5. Change in structure on binding, and
   6. Observed biological response.

   Origin of QSAR extends the intuition to become more quantitative allowing for more rational approach to lead optimization.
2. Biological Activity\textsuperscript{18} : Biological activity profile for the series of analogue compound must have following characteristics:

1. Large rang in observed activity.
2. Identical mode of action.
3. Activity data as function of concentration ($IC_{50}$).
4. Concentration in molar units and activity data in percentage.
5. Possible time dependency.

The biological activity data in QSAR analysis can be represented in terms of following parameters:

- **Isolated Receptors-**  
  - Rate constant $\log K$, $\log K$ cal  
  - Inhibition constant $\log 1/Ki$  
  - Michaelis –Mentan constant $\log 1/Km.$

- **Cellular System -**  
  - Inhibition $\log 1/IC_{50}$  
  - Cross resistance $\log CR$  
  - In vitro biological data $\log 1/c$  
  - Mutagenicity state $\log TA$

- **In Vivo System -**  
  - Bio Concentration $\log BCF$  
  - In vivo-reaction rate $\log I$ (Inductions)  
  - Pharmacodynamic rate $\log I$ (Total cleaning)

Since QSAR is based on relationship of free energy to equilibrium constants, therefore free energy changes occurring during the biological dose required to produce biological effect can be calculated as proportional to the inverse or negative logarithm of the concentration of the compound.

$$\Delta G = -2.303 \, RT \log K = - \log 1/C = - \log (B.A.)$$

or

$$\Delta G \sim \log (B.A.)$$
Although such linear free energy relationship can be stated in terms of thermodynamic parameters and is called thermodynamic, thermal or Extra thermodynamic Approach.

**Expression of biological activity:**

In general, expression of the biological activity of a compound involves two quantities: a biological response and the dose required to elicit that response.

There are two types of biologically active data, dichotomous and quantitative. Dichotomous data are derived from assays with results of response/no response, active/inactive, toxic/non-toxic for given dose levels. Quantitative biological data involve a measured biological response determined within a time interval and the dose required for that response. Accordingly the biological activity of a compound may be based on the dose required for a standard biological response i.e. DFR (dose for a fixed response). Alternatively, biological activity can be expressed as the degree of biological response that a constant amount of a test material can produce RFD (response for a fixed dose). A simple plot of dose to response generates the familiar S-shaped curve.

In QSAR studies, the DFR measurements are given by the horizontal lines as dose values $C_1$ and $C_2$. The fixed dose measurements are given by the vertical line as biological responses $BR_1$ and $BR_2$. 
2.10. QSAR: Review of Literature

QSAR initially started in 19th century when Cum-Brown and T. Fraser\textsuperscript{20} suggested that the physiological action of a molecule was a function of its chemical constitution. They proposed a general equation i.e.

\[
\text{Biological Response} = f (\text{Chemical Constitution})
\]

or

\[
B = f (C)
\]

In 1893 Richet \textsuperscript{21} suggested that the toxicities in a series of simple compounds like alcohol, ether etc. were inversely proportional to the water solubility. It was Richet who has cleared up the first relationship between toxicity and lipophilicity.\textsuperscript{22}

\textit{Bertholot and Jungfleish} (1872) did the first systematic investigation on the distribution of compound between two immiscible liquids. The contribution of \textit{Nerst} to this subject acquired much more attention, however and it is \textit{Nerst’s} name that became undetectably connected with distribution and partition. At the same time \textit{Overton}\textsuperscript{23} performed his exploration on the permeability of living plants and animal cells to a large variety of organic compounds. \textit{Overton and Meyer} in 1897 related Tadpole narcosis activity to the octanol/water partitioning of chemicals.

A strong impetus to the development of QSAR originates in physical–organic chemistry. In particular Hammet contributions were of great importance. Through the year 1937-1940 \textit{Hammet} developed his system of s constant that describes the electronic effect of substituent of the benzene ring.\textsuperscript{24} Hammet’s work got a consequent sequel in the studies of \textit{Taft} who made available a set of s values suited for the description of electronic effect caused by substituent in aliphatic structure.\textsuperscript{25}
Corwin Hansch (1962)\textsuperscript{26} gave a very reliable approach. He used a model of drug action that, in part, involves a diffusive process that includes the movement of a drug across membranes in order to enter cells, since cell membranes contain hydrophobic regions, the lipophilicity of molecules was important for their in vivo biological activity. Note that in vitro measurements often do not involve the crossing of membranes. From this he gave the hypothesis that substituents on a parent molecule have a quantitative relationship with biological activity. He also gave the fragment and additive group contribution theory. He also used the electronic substituents.

Free and Wilson (1964)\textsuperscript{27} related biological activity to the presence/absence of a specific functional group at a specific location on the parent molecule.

\[
\text{Activity} = A + \sum_i \sum_j G_{ij} X_{ij}
\]

where

- $A$ = the average biological activity for the series,
- $G_{ij} =$ the contribution to activity of a functional group $i$ in the $j^{th}$ position
- $X_{ij} =$ the presence (1.0) or absence (0.0) of the functional group $i$ in the $j^{th}$ position.

Variations on this activity-based approach have been extended by Klopman et al.\textsuperscript{28} and Enslein et al.\textsuperscript{29} Topological methods have also been used to address the relationships between molecular structure and physical/biological activity. The minimum topological difference (MTD) method of Simon and the extensive studies on molecular connectivity by Kier and Hall have contributed to the development of quantitative structure property/activity relationships.\textsuperscript{30,31}
John Topliss (1972) proposed the Topliss trees; it’s a non statistical method to automate the Hansch approach. He published a paper which has detailed methodology to automate the Hansch approach. The method assumed that the lead compound of interest contained at least one phenyl ring which could serve as the template for functional group modifications. The first modification to the template was preparation of the para-chloro derivative to examine lipophilicity. Additional substitution patterns were then made sequentially in an attempt to explore and optimize the relationship between activity and the hydrophobic and electronic character of the molecule.

Marshall et al. (1979) proposed an ‘Active analogue approach’. Marshall extended the 2-D approach to QSAR by explicitly considering the conformational flexibility of a series as reflected by their 3-D shape. The first step of the ‘Active analogue approach’ was to exhaustively search the conformations of a compound which was highly active in a particular biological assay. The result of the search was a map of interatomic distances which was used to filter the conformational searches of subsequent molecules in the series. The implicit assumption of the method was that all compounds which display similar activity profiles were able to adopt similar conformations. Once the active conformation was determined, molecular volumes for each molecule were calculated and superimposed. Regression analysis of the volumes was used to establish a relationship to biological activity. Marshall and co-workers commercialized the ‘Active analogue approach’ and a suite of other drug design techniques in the SYBYL molecular modeling programme.
Recent Advances in QSAR:

Richard Cramer (1988)\textsuperscript{34}, introduced CoMFA (Comparative Molecular Field Analysis), in this he related the shape-dependent steric and electrostatic fields for molecules to their biological activity. In 1988, Richard Cramer proposed that biological activity could be analyzed by relating the shape-dependent steric and electrostatic fields for molecules to their biological activity. Additionally, rather than limiting the analysis to fitting data to a regression line, CoMFA (Comparative Molecular Field Analysis) utilized new methods of data analysis, PLS (Partial Least Squares) and cross-validation, to develop models for activity predictions. The approach used in the CoMFA procedure requires that the scientist define alignment rules for the series which overlap the putative pharmacophore for each molecule; the active conformation and alignment rule must be specified.

In 1988, Arthur M. Doweyko\textsuperscript{35} gave the HASL (Hypothetical Active Site Lattice). According to him a superimposed set of molecules into a set of regularly spaced points (lattice) defined by Cartesian coordinates (x, y, z) and atom type.

\[ \text{Biological activity} = f(\text{Lattice points}) \]

In 1990, V. E. Golender and A. B. Rozenblit\textsuperscript{36} gave the Apex-3D expert system. They related biological activity to biophoric (pharmacophoric) and secondary sites using statistical techniques and 3-D pattern matching algorithms.

In 1994, D. Rogers and A. J. Hopfinger\textsuperscript{37} and G. Klebe, U. Abraham and T. Mietzener\textsuperscript{38} in the same year gave the concept of Comparative
Molecular Similarity Indices Analysis (CoMSIA). Hopfinger and co-workers also used 3-D shape in QSAR. In molecular shape analysis of the Baker Triazines, the common space shared by all molecules of a series and the differences in their potential energy fields were computed. When these calculations were combined with a set of rules for overlapping the series, comparative indices of the shape of different molecules were obtained. Inclusion of these shape descriptors in standard Hansch analysis schemes lead to improved descriptions relating computed parameters to biological activity such that no compound in the original data set had to be eliminated from the calculations. It’s an alternate approach to the computation of molecular potential fields, in this the indices replace the distance functions of Lennard-Jones and Coulomb-type potential with Gaussian-type functions. The Genetic Function Approximation (GFA) logarithm is a novel technique for constructing QSAR models. It was specifically developed for use with data sets containing many more variables than samples, or data sets which contain nonlinear relationships between the variables and the activity.

Although much work had already been done in the field of QSAR, still more work is going on as QSAR is an emerging field of computational drug designing and attracts attention of medicinal chemists.
References


20. Ashutosh Kar; ‘Book of Medicinal Chemistry’, 1-10


