CHAPTER 2

QUANTITATIVE

STRUCTURE

ACTIVITY

RELATIONSHIP

CHAPTER 2

QUANTITATIVE STRUCTURE ACTIVITY RELATIONSHIP

CHAPTER 2

QUANTITATIVE STRUCTURE ACTIVITY RELATIONSHIP

2. Introduction

Traditionally, the process of drug development has revolved around a screening approach, as nobody knew which compound or approach could serve as a drug or therapy. Such almost blind screening approach is very time-consuming, costly and labourious. The alternative to this labour intensive approach to compound optimization is to develop a theory that quantitatively relates variations in biological activity to changes in molecular descriptors, which can easily be obtained for each compound. Drug discovery is a challenging process due to the complexity of biological systems. Traditional approaches such as trial and error synthesis of compounds and random screening for activity have proved to be time consuming, laborious and uneconomical. It has been estimated that out of hundreds of compounds synthesized in research laboratories only one or none reach the market ultimately as "drug" It is required that synthesis of such compounds be carefully designed and their biological activity determined on suitable test systems .The recent advances in organic and medicinal chemistry, biology (including pharmacology. toxicology, pharmacokinetics/dynamics), physics (including biophysics spectroscopy) and computer science have proved to be useful tools for designing new chemical moieties and predicting their biological activities prior to synthesis. [56] A **Quantitative Structure Activity Relationship** (QSAR) can then be utilized to help, to guide chemical synthesis. Quantitative structure-activity relationships (QSARs) attempt to correlate chemical structure with activity using statistical approaches. The QSAR models are useful for various purposes including the prediction of activities of untested chemicals. QSAR methods also provide help for better understanding of factors and processes important in drug action, thus enabling a more rational design of drug[57]. Quantitative structure-activity relationships and other related approaches have attracted broad scientific interest, particularly in the pharmaceutical industry for drug discovery and in toxicology and environmental science for risk assessment. An assortment of new QSAR methods has been developed during the past decade, most of them focused on drug discovery. This approach had its origins in the 60's and has become very important in industrial and academic drug design and basic research. Until about the late 1950's most of the structure-activity correlations were empirical and qualitative. Although attempts have been made to apply quantitative methods to biological activity since last century, a major effort in this field has been made only very recently based on the physical organic chemistry using regression analysis and computer technology.

The QSAR approach is a rational approach to lead optimization when the structure of the target is not known. The underlying premise of QSAR is that there is a relationship between the biological or pharmacological activity of a compound, and its structural, physical and chemical properties i.e. activity is a function of Structure and an equation can be developed relating activity (not yet developed) to parameters which can be determined, for example, by computer. These physicochemical descriptors, which include parameters to account for hydrophobicity, topology, electronic properties, and steric effects, are determined empirically or, more recently, by computational methods. Activities used in QSAR include chemical measurements and biological assays. An important advantage of QSAR is that it models the *in vivo* situation since it is based on activity data.

This approach allows important structural requirements for activity to be identified and hence narrows the search for the optimum molecule. Since it is a quantitative measure, how much each aspect alters activity, can also be determined. QSAR equations, once determined, also allow new compound's activity to be predicted based on structural data and hence save time, labour and money in synthesizing molecules unlikely to have good activity.

A QSAR generally takes the form of a linear equation

Biological Activity = Const + $(C_1 * P_1) + (C_2 * P_2) + (C_3 * P_3)$ +...-----------**(1)**

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

2.1 Early Developments

QSAR initially started in 19th century. In 1868, **Crum-Brown** and **Fraser**[58,59] published an equation, which is considered to be the first general equation of QSAR.The equation suggested that physiological activity depends on constitution

$$\phi = f\,(\text{const.}) \ldots\ldots\ldots \quad \textbf{(2)}$$

It was not before 1893 that **Richet** [60] followed this clue by showing that the toxicities in a series of simple compounds (alcohol, ethers and ketone) were inversely related to water solubility. The importance of these findings could not fully be understood before correct relationship between lipophilic behaviour and solubility were established. It was **Richet** who has cleared up the first relationship between toxicity and lipophilicity.[61, 62] **Bertholot** and **Jungfleish** did the first systematic investigation on the distribution of compound between two immiscible liquids in 1872. The contribution of **Nerst** to this subject acquired much more attention, however and it is **Nerst** name that became undetectably connected with distribution and partition.

Practically simultaneously with **Richet's** experiment, **Overtone** performed his exploration on the permeability of living plants and animal cell to a large variety of organic compounds. Overtone's studies largely concerned with oil-water and oil-gas partition coefficient to correlate and explain potencies of narcotic substance in tadpoles. 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 he developed his system of σ constant that describes the electronic effect of substituent of the benzene ring**.**[63] **Hammet's** work got a consequent sequel in the studies of Taft who made available a set of σ values suited for the description of electronic effect caused by substituent in aliphatic structure.[64]

The major limitations of classical QSAR's have been:

(I) Only the molecules of same congeneric series can be included in the analyses and hence it is difficult or impossible to include molecules belonging to diverse class compounds.

(II) It is difficult to account for the differences in activities observed in case of the enantiomers and different conformers of same molecules.

2.2 Steps involved in QSAR[65]

The following four steps are to be carried out in QSAR (Fig 2.1):

1. **Selection of a series of biologically active analogues with their biological activity*;*** A series of already synthesized analogues (**lead skeleton)*** is selected with their particular biological activity
2. **Calculation of various physicochemical parameters;** In the second step the calculation of quantitative values of various physicochemical parameters for various substituent groups present in the series is completed
3. **Determination of correlation matrix between various physicochemical parameters and biological activity**; The correlation matrix between various physicochemical parameter and biological activity is determined that shows which particular physicochemical parameter correlates best with biological activity
4. **Generation of QSAR equation** After selection of best parameter, the QSAR equations are generated
5. **Prediction of biological activity;** By using QSAR equation the biological activity of newly designed compound can be predicted. Then medicinal chemists synthesize only those compounds which give promising prediction of biological activity

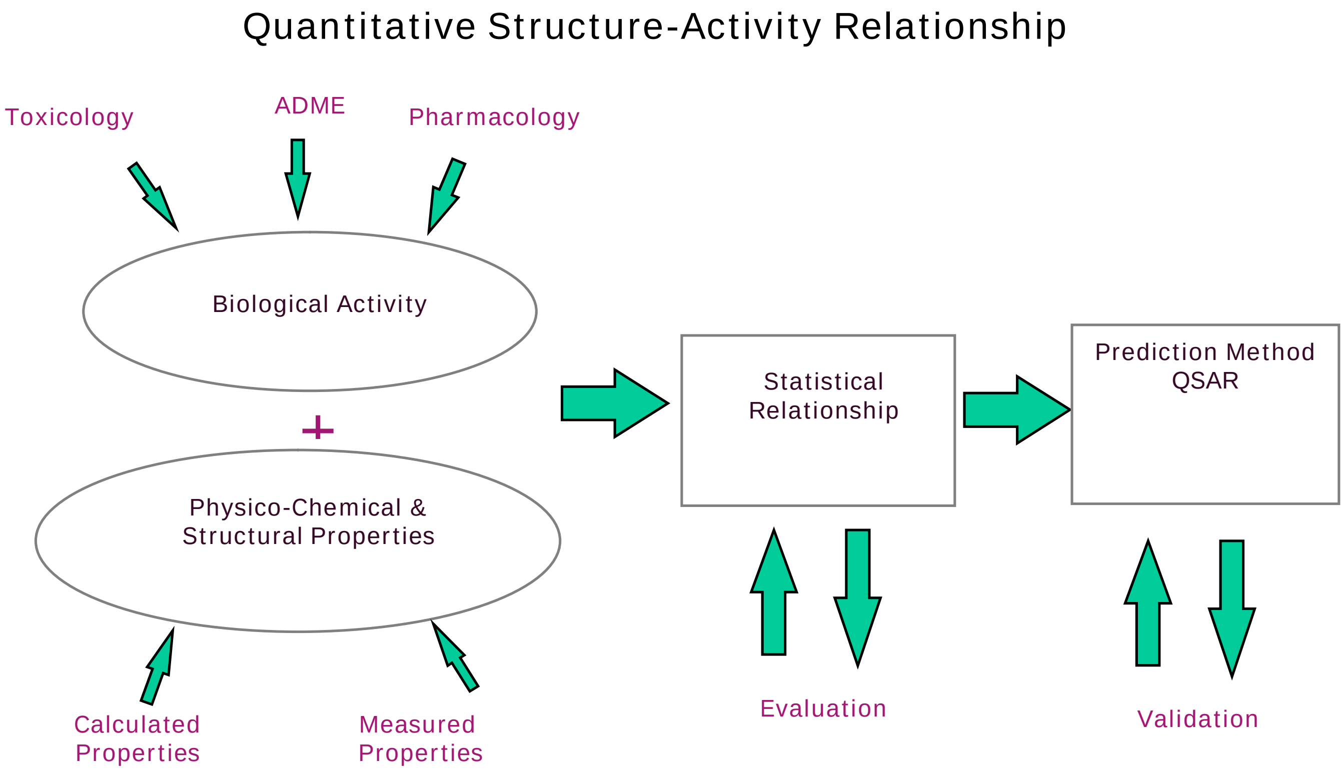


Fig 2.1 Steps involved in QSAR methodology

*The lead is a prototype compound that has desired biological or pharmacological activity, but has many undesirable characteristics eg.high toxicity, other biological activities insolubility or metabolic problems. Identification of lead nucleus depends upon the consideration of following points:

1- Molecular structure of the drug.

2- Behaviour of drug in biophase.

3- Geometry of receptor.

4- Drug receptor interaction.

5- Changes in the structure on binding.

6- The observed biological response.

After following this process, only few drugs can reach to the level of clinical applicability.

Early SAR studies, simply involved the synthesis of as many analogues as possible of the lead and their testing to determine the effect of structure on activity. Attempts were made to interpret chemical structures in terms of physical and chemical properties, transport and distribution of a drug in a biological multicompartment system, the affinity of the drug to a complementary-structurally unknown receptor and the interaction of the drug with its receptor. This adds to the cost of research for new drug. Broadly this means that if the development of new drugs is to remain economically feasible, the ratio of output to input must be increased.

2.3 Development of Physicochemical approach (QSAR)

Of the various approaches to QSAR, the most commonly used mathematical models are:

(1) The Extrathermodynamnic (Hansch) approach;

(2) The Free-Wilson (Additivity) approach, and

(3) Combined approach.

The most popular among these is the extra thermodynamic approach it is dealt here in some detail

2.3.1 The Extra thermodynamic or Hansch Approach

The Hansch approach has been the most popular due to its predictive and understanding the mechanism of action of drugs [66-70]

He proposed that the action of a drug depends upon two processes. First, the journey of drug from the point of entry in the body to the site of action and the second is its interaction with the receptor site. He suggested the linear (eq. 3) and parabolic or non-linear dependence (eq. 4) of biological activity (BA) on different parameters.

log BA = a log P + bσ + cEs + d ----------------linear **(3)**

log BA = a log P + b (log p)^2 + cσ + dEs + e -------------------parabolic **(4)**

The coefficients (a, b, c, d, e) are determined by multiple regression analysis.

Though the drug transport process and drug receptor interactions are complex in nature, yet they are essentially physicochemical and can be factored into electronic, hydrophobic and steric parameters. The variations in biological activity depend upon changes in these physicochemical parameters like hydrophobic (H), electronic (E) and steric (S) and can mathematically be expressed **[71]**by equation (5)

BA---- f (Δ H, Δ E, Δ S)

or BA=f(ΔH, ΔE, ΔS) + constant------------------- **(5)**

The more general form for the equation (5) can be represented by equation (6).

BA = a (hydrophobic) + b (electronic) + c (steric) + constant -------------**(6)**

The various free energy related parameters used for these effects are described below:

(a) Hydrophobic parameters. The most commonly used hydrophobic parameters are partition coefficients e.g. log P, π and chromatographic coefficients R_M. The chromatographic parameter R_M [72,73] [R_M = log (l/R_F -1)] is related to the logarithm of partition coefficients between the polar and non-polar phase of a chromatographic system. In addition to above parameters, distribution coefficient (D)[73] solubility parameter (δ, log δ)[74] and parachor (p)[75,76] have also been used as hydrophobic parameters.

(b) Electronic Parameters: The most commonly used electronic parameters are: Hammet constant (σ), Field effect (f) and Resonance (R) developed by Swain and Luptons respectively. Similarly σ^{+} and σ^{-} have also been used in the situations where electron withdrawing and electron donating groups interact with developing a positive or negative charges in transition state, respectively. Other σ * for aliphatic systems and $\sigma^{\cdot}$ for relative rates of phenylation have also been used.

Various modifications of δ like δ°, δ^{m}, δ^{p} which describe the electronic effect of substituent at ortho, meta and para positions, respectively, have been used.

The other commonly used electronic parameters are dipole moment (μ), ionization constant (K_a), ioniztion potential (I), hydrogenbonding parameter, NMR or CMR chemical shifts (δ ppm) values and IR frequencies, etc.

(c) Steric Parameters: The first steric parameter was suggested by Taft

Who denoted it by (E_s) and derived it from acid hydrolysis of acetate esters.This was later modified into $E_s^{o,m,p}$, for ortho, meta, and para substituted aromatic systems and by Hancock's corrected parameter for hyperconjugation (Es^c). Other parameters like molar volume (MV), vander Waal's radius (Vr), vander Waal's volume(V_w), molecular weight (MW), have also been successfully used as static parameters in the QSAR studies. In addition to these parameters, the other important parameter suggested was molar refractivity (MR), which has additive, and constitutive properties like π and can be obtained by Lorentz-Lorentz equation (7).

MR =((n^2-1)/(n^2+1))(MW/d)-------------------------- **(7)**

where MW is molecular weight, n is refractive index and d is density of compound.

In addition to the above-discussed important parameters several miscellaneous properties or combinations of other parameters are also used in LFER models.

(d) Biological data: All diverse kinds of biological data can be used in QSAR viz. affinity data, like substrate or receptor binding constants; rate constants, like association (dissociation, and Michealis Menton constants) and other in vitro and in vivo biological activity data like Ki, IC_{50}, ED_{50} and ID_{50}.

2.3.2. The Free-Wilson Approach (The Additivity Model)

The method of Free-Wilson is based upon an additive mathematical model in which particular substituent in a specific position is assumed to make an additive and constant contribution to the biological activity on a molecule in a series of chemically related molecules. This method is based upon assumption that the introduction of a particular substituent at a particular molecule position always leads to a quantitatively similar effects on biological potency of the whole molecule, as expressed by the equation (8):.

$\log BA = \mu + \Sigma\,(ai.aj)$--------------------------**(8)**

Where, ai = no. of position at which substitution occurs

aj = no. of the substituents at that position

μ = overall average

The equation is solved by multiple linear regressions using the presence (1) or absence (0) of the different substituent as independent dummy parameter, while the measured activity serves as a dependent variable.

This method differs from Hanch analysis, in that substituent constant value is based on, biological activity rather than physical properties.

2.3.3. Combined Approach

The above two approaches, Hansch and Free-Wilson, are practically interrelated and theoretically equivalent. However, Free-Wilson model suffers by its inability to attribute physical significance to the substituent contribution and fails to yield good correlations in the cases where parabolic model of Hansch is applicable. Similarly, sometimes it is difficult to correlate the biological activity only in terms of known physicochemical parameters, which has necessitated the use of dummy parameters or more commonly known as indicator variables (I) which constitutes a part of correlation equation .So the most commonly used equations now contain a part 'A' in which physicochemical (Hansch) parameters are used and part ‘B’ of indicator variables based on the assumption of Free-Wilson as described in eq. (9).

$$\text{Log I/C} = \underbrace{a_1 (\log P)^2 + a_2 (\log P) + a_3 \sigma + a_4 E_s}_{A} \;+ \underbrace{a_5 I_1 + a_6 I_2 + a_7}_{B} \text{ --------------}\mathbf{(9)}$$

The role of indicator variable can be of diverse type, sometimes it defines an important substructure (pharmacophore), ortho-interactions, a substitute for steric parameter and a separator between two isomers (R & S), etc. It may also act as ‘molecular book-keeping’ where the variation in similar type of biological activity of large and diverse sets of congeners is explained by one equation.

2.4 Major Parameters of QSAR [77,78]

In QSAR it is considered that the biological activity is connected with the physicochemical properties. So, biological activity, which is represented in either form of C, K_i, IC_{50}, ED_{50} and K_m and the physicochemical properties, which are broadly classified into electronic, lipophilic and steric parameters, are connected with a mathematical equation. The parameters selected should be orthogonal, that is, have minimal covariance.

A) Electronic Parameters:

Hammett substituent constant:

Hammett electronic parameter or substituent constant, σ_x, is the electronic effect of substituent x relative to hydrogen. σ_x is determined on the basis of influence of a substituent on the ionization of benzoic acid.

$\rho\sigma = \log K_{a\,(R)} - \log K_{a\,(H)}$ ---------------------- **(10)**

ρ is the constant for a given reaction. σ is the substituent constant, and K_a is the equilibrium constant (or rate constant, k_a) for the reaction of interest.

There are several other ways of quantifying electronic effects. For example, electronic effects can be represented as a linear combination of a field (**inductive**) effect, **F**, and a **resonance** effect, **R**:

$\sigma = aF + bR$ --------------------------------------**(11)**

Where, a and b are coefficients determined from data fitting. The use of σ has been extended to biological activity. Those values may then be applied to many types of reactions as characterized by different values of the reaction constant, $\rho = 1$, by definition, for the ionization (dissociation) of benzoic acid. Where,

$\text{Log } [K_x/K_H] = \rho\sigma_x$

Where, K_H is the equilibrium or rate constant for the parent (unsubstituted) and K_x is the equilibrium or rate constant for the derivative: measured experimentally.

Electron withdrawing groups like $-NO_2$, increases K_x and, ultimately, leads to a positive σ whereas electron donating groups like $-OCH_3$, decreases K_x and thus leads to a negative σ.

Inductive effect (electron withdrawing or donating) refers to the polarity produced in a molecule as a result of higher electro negativity of one atom compared to another. The carbon-hydrogen bond is used as a standard. Zero is assumed in this case. Atoms or groups, which donate electrons to carbon atom, are said to have a +I effect. Those atoms or groups, which withdraw electrons away from carbon atom are said to have a –I effect.

Resonance effects occur with para substituents and can lead to large magnitude of ±:**v**alues. The ± : is the reaction constant which indicates the influence of the electronic effect on the binding constant. If ± > 1 then the electronic contribution of substituents is greater than it is for the ionization of benzoic acid. If ±:< 1 then the electronic contribution of substituents is less than it is for the ionization of benzoic acid. Note that ± can be less than 0, indicating that the effect is opposite to that occurring with respect to the ionization of benzoic acid.

Highest Occupied Molecular Orbital (HOMO) and Lowest Unoccupied Molecular Orbital (LUMO):

It gives the reactivity of the molecules. Molecules with high HOMO are more readily able to donate electrons than the molecules with low (HOMO). Thus, measuring the nucleophilicity of the molecules. Molecules with low-lying LUMO can accept electrons and so, measures the electrophilicity of the molecules.

B) Lipophilic parameters:

Absorption and Distribution processes in biological systems are determined by the hydrophobic or hydrophilic properties of the molecules. As the drug has to pass in both the hydrophobic and hydrophilic media in the biological system, a parameter called partition coefficient is used to determine its hydrophobicity or lipophilicity.

Partition coefficient, based on octanol-water system is allowed for determination of hydrophobic substituent constants.The octanol water partition cooefficient is designated as P and the Hansch value ® is the effect of given substituent on log P of basic skeleton

Advantages of *n*-oatanol water system

n-oatanol water system has many important advantages as compared to other systems

- It is a suitable model of the lipid constituents of the biological membranes due to its large alkyl chain and the polar hydroxyl group
- Its hydroxyl group is a hydrogen bond donor and a hydrogen bond acceptor, interactingwith a large variety of polar groups of different solutes.
- Despite its lipophilic character it dissolves many more organic compounds than alkanes, cycloalkanes or aromatic solvent do

The reasons for choosing n-octanol as organic solvent are, it has a long saturated fatty alkyl chain, hydroxyl groups for H-bonding and dissolves water to the extent that saturated octanol contains 1.7 M water .The combination of lipophilic chains, hydrophilic hydroxyl group and water molecules appear to give *n*-octanol, properties, very similar to those of natural membrane and macromolecules. Clearly absorption and distribution processes in biological systems are determined by the **hydrophilic or hydrophobic** properties of molecules, for which the **partition coefficient, P**, **of a molecule** is used. **P** is defined by:

$$\mathbf{P} = \frac{[\mathrm{drug}]_{\mathrm{octanol}}}{[\mathrm{drug}]_{\mathrm{water}}}$$

By using a logarithmic relationship, P becomes an additive property:

$$\mathbf{logP} = \log[\mathrm{drug}]_{\mathrm{octanol}} - \log[\mathrm{drug}]_{\mathrm{water}}$$

The contribution of a substituent, X, to the logP of a molecule is defined by

$$\begin{aligned}\boldsymbol{\pi}_{\mathbf{X}} &= \log P_{\substack{\text{substituted}\\ \text{compound}}} - \log P_{\substack{\text{parent}\\ \text{compound}}}\\ &= \log \frac{P_{\mathbf{RX}}}{P_{\mathbf{RH}}}\end{aligned}$$

π is the hydrophobic **parameter for a specific substituent**. Extensive measurements and use of theoretical correlations have resulted in a large number of tabulated **logP** and π values. Where P_{Rx} is the partition coefficient of the derivative x and P_{RH} is the partition coefficient of the parent (unsubstituted) compound.

NOTE: Larger is the P, more is the hydrophobicity. Therefore, larger and positive π shows more hydrophobic compound. For a hydrophilic (polar compound) the value of P_{Rx} will be less that that of P_H, such that the ratio log P_{Rx}/P_{RH} will be a fraction leading to π being negative. When π has negative slope it indicates that as π increases (more hydrophobic) log (1/C) decreases indicating the concentration required to induce the selected biological property increases (less potent). Thus **beyond a certain range of hydrophobicity (lipophilicity), there occurs a decrease in biological activity. (Fig.2.2)** The main reason for this decrease is that due to high hydrophilic (very polar compounds) the compounds become insoluble in organic layers, so they can not cross lipid membrane and remains in aqueous layer while due to high lipophilic nature they become so lipid soluble that they can not circulate in blood stream and becomes "glued" to lipid membrane. Therefore in order to reach their targets, the compounds should posses appropriate lipophilicity so that it can cross lipophilic as well as hydrophilic barriers.

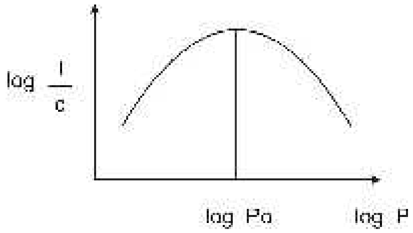


Fig.2.2 Parabolic relationship between biological activity (log1/C) and partition coefficient

Chromatographic R_m Value: When the solubility of a solute is greater in one phase than in the other,it becomes difficult to determine partition coefficient experimentally. In such conditions R_m value is used:

R_m = log [1/R_f + 1]

Log P = R_m + a constant

Where, P = Partition coefficient, R_f = Retardation factor

C) Steric parameters:

Taft Steric Substituent Constant **(Es)** The size of different substituents can clearly be important to the activity of compounds. Bulky groups may lower activity by preventing drugs from fitting properly into the binding site. On the other hand, a bulky substituent may increase activity by forcing a compound to adopt the required active conformation for binding. Measuring the steric properties of substituents is not as straightforward as the measurement of substituents hydrophobic or electronic character.

The constant, Es, is a measure of substituents size and is determined experimentally by measuring the effect; different substituents have on the rate of a chemical reaction carried out on the parent structure. Larger substituents next to the reaction center hinder the reaction more than the smaller substituents. So differences in the reaction rate lead to a measure of substituent's size. It depends on the fact that acid hydrolysis is determined almost completely by steric factors, and is defined as

$Es = \log (K/Ko)_A$

Where K =Ester hydrolysis rate constant for substituted compound.

Ko = Ester hydrolysis rate constant for methyl derivative.

A -represents acid hydrolysis.

Molar refractivity, MR: Originally proposed by **Pauling and Pressman** as a parameter for the correlation of dispersion forces involved in the binding of antigen to antibodies. It is now determined from the Lorentz-Lorentz equation.

$$MR = \frac{(n^2 - 1)\,(MW)}{(n^2 + 2\,)d}$$

Where n =refractive index, MW =the molecular weight, and d =density of a crystal,

Since refractive index doesn't change significantly for organic molecules, the term is dominated by the MW/density i.e. Volume. Larger MW, larger the steric effect and greater the density, the smaller the steric effect (the molecules tend to pack better). A smaller MR for the same MW indicates stronger interactions in the crystal (larger density indicates that the packing is better due to stronger interactions). MR can also be considered as crude steric parameter, characterizing bulk (but not shape) of the molecule or substituent.

Verloop steric parameters: Terms related to vander Waals radii of molecules that may be determined for any substituents; may be used for both symmetric and asymmetric molecules.

STERIMOL size parameters (L, B_1, B_2, B_3 and B_4) proposed by Verloop defines as dimensions of the substituent, R:

L = length along the axis of the bond joining R to the parent molecule

B_i = the four width parameters, at right angles to the axis, L, viewed in cross-section, and $B_1 < B_2 < B_3 < B_{4.}$

D) Polarizability parameters:

1) Molar Volume, MV = MW/d
2) Molar Refractivity, MR = MV $(\eta^2 - 1) / (\eta^2 + 2)$
3) Molar polarization, PM = MV. $(\varepsilon^2 - 1) / (\varepsilon^2 + 2)$
4) Parachor, PA = MV. $\gamma^{1/4}$

Where, MW = molecular weight, d=density, η=refractive index, γ=surface tension, ε =dielectric constant.

E) Indicator variable:

The role of indicator variable can be of diverse type. Sometimes it defines an important substructure (pharmacophore), ortho interactions, a substitute for steric parameter and a separator between two isomers, etc. It may also act as "molecular book-keeping" where the variation in similar type of biological activity of large and diverse sets of congeners is explained by one equation It is used to indicate the significance of any particular group or species in given series of drug molecules. An indicator variable is a descriptor that can assume only two values indicating the presence (=1) or absence (=0) of a given condition. It is often used to indicate the presence or absence of a substituent or substructure.

F) Structural parameter:

The use of structural parameter, molecular connectivity index χ, which is a measure of the molecular skeleton in terms of number and kind of atoms present and their connections with each other in a molecule, was first introduced by Randie. Kier and coworkers have studied a number of quantitative correlations of physicochemical properties and biological activities of compounds with this parameter. Hall and Kier have described the use of three types of molecular connectivity, simple connectivity (χ^s), valence connectivity (χ^v) and connectivity dependent of bond length (χ^y).

Parameters used in present thesis

The parameters considered during present studies are:

- Hydrophobic parameter (Π)
- Electronic parameter (f, R ,σ)
- Steric parameter (MR)
- Indicator parameter

2.5 Hansch Analysis

In particular Hammet contributions were of great importance. During the year 1937-1940 he developed his system of σ constant that describes the electronic effect of substituent of the benzene ring. Hammet's work got a consequent sequel in the studies of Taft who made available a set of σ values suited for the description of electronic effect caused by substituent in aliphatic structure. A few years earlier Taft has enriched the parameter arsenal with his steric effect parameter Es

QSAR based on Hammett's relationship utilize electronic properties as the descriptors of structures. Difficulties were encountered when investigators attempted to apply Hammett-type relationships to biological systems, indicating that other structural descriptors were necessary.

Against the background of numerical information on electronic and steric effect at hand the basic principle of Lipophilicity Hansch *et al* entered the scenery in 1962 with the first of a great many of pioneering contribution**[79]**. The main features from the research efforts of the Hansch group are undoubtedly the following two:

(a) The development of the hydrophobic substituent constant π. The definition of π is incorporated in equation and shows analogy with the Hammett equation:

$$\text{Log P (R-X)} = \text{log P (R-H)} + \pi\,(\text{X}) \qquad \textbf{----------------------(12)}$$

P (R-X) and P (R-H) are partition coefficients of R-X and R-H, with R-X indicating a structure derived from R-H by replacing H atom by substituent X; π (X) is the hydrophobic substituent constant, to be defined as the lipophilicity contribution of substituent X to lipophilicity when replacing H by X.**[80-82]**

(b) The multi parameter approach to QSAR **[83-85]**

Equation 13 exemplifies a correlation with the three parameters log P,σ , Es.

$$\text{Log (BA)} = a \text{ log P} + b\sigma + c \text{ Es} + d \quad \textbf{----------------------(13)}$$

The modern QSAR methodology in 1964 started by Hansch and Fujita on “q-σ-π-analysis which is method for the correlation of biological activity and chemical structure” in terms of physiochemical parameters and the other by Free and Wilson based on “the mathematical contribution of chemical (substituent) to structure activity studies **[86]**. Both contributions started the development of two new methods of quantitative structure activity relationships, later called Linear Free Energy Relationship (LFER) or Hansch analysis and Free Wilson analysis respectively.

$$\text{Log I/C=alog P +b}\sigma\text{ +c} \text{ ----------------------------- } \textbf{(14)}$$

$$\text{Log I/C=a (log P)}^{2}\text{ + blog P +c}\sigma\text{ +d} \text{ ------------- } \textbf{(15)}$$

$$\text{Log I/C=}\Sigma\text{ a}_{1}\text{ + }\mu \text{ ---------------------------------- } \textbf{(16)}$$

In these equations C is molar concentration causing a standard biological response, e.g. IC_{50}, ED_{50} or LD_{50}, P is the partition coefficient, σ is the Hammett constant, and a, b, c and d are constants determined by linear multiple regression analysis. Other physiochemical parameters can be used instead of or in addition to P and σ in equation (14) and (15). In equation (16), a_i is the value of the substituent group contributions to biological activity, and μ is regarded to be the activity contribution of the parent system (in Fujita – Ban analysis[87] μ is the theoretical biological activity value of the reference compound).

➢ Advantages of Hansch approach

A) Use of descriptors *(*Π, σ, Es etc.) from small organic molecules may be applied to biological systems.

B) Predictions are quantitative and may be evaluated statistically.

C) Quick and easy.

D) Potential extrapolation: conclusions reached *may* be extended to chemical substituents not included in the original analysis.

➢ Disadvantages of Hansch Approach

A. Large number of compounds are required (training set for which physicochemical parameters and biological activity is available).

B. Limitations associated with using small molecule descriptors, such as steric factors, on biological systems (i.e. descriptors from physical chemistry).

C. Partial protontation of drugs at physiological conditions is problematic (can be included in mathematical model if necessary).

D. Predictions limited to structural class (congeneric series).

E. Extrapolations beyond the values of descriptors used in the study are limited.

F. Correlation between physical descriptors. For example, the hydrophobicity will have some correlation with the size and, thus, the Taft steric term**.**

➢ Application of QSAR (Hansch relationships)

1) Classification

At the initial stages of a study, obtain a relationship of gross molecular features or physicochemical properties to qualitative experimental data highly active, active, inactive agonist, partial agonist and antagonist.

2) Diagnosis of Mechanism of Drug Action

Interpret correlation between biological activity and physical properties in terms of a mechanism or use QSAR to test a mechanistic hypothesis.

Meyer-Overton theory of narcosis: the accumulation of molecules in lipid biophase is the only prerequisite for activity, despite the wide variation in structure and chemical substituents.

Test of 51 alcohols, ethers and amides as narcotics on tadpoles yielded the following equivalency

$\log(1/C) = 0.94 \log P + 0.87$, $r = 0.97$, $n = 51$

Thus, the strong correlation between the biological activity and partition coefficient supports the proposed mechanism.

Overall, if a hypothesis states that the biological activity is related to one or more physicochemical/structural parameters (a, b and c) then a plot that includes the biological activity as a function of the parameters a, b and c (weighted sum of a, b and c) should yield a linear relationship. Conversely, by plotting the log(1/C) versus a variety of different physicochemical/structural parameters it may be determined which of those parameters are important for activity.

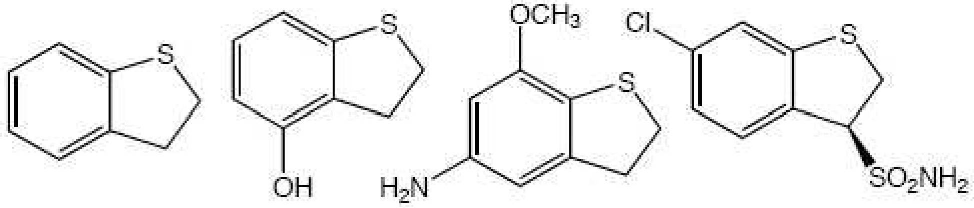
3) Prediction of Activity

Predict activity of an unknown molecule via QSAR

Develop QSAR for a "training set" of compounds, and then use the obtained mathematical relationship to predict the biological activity of new compounds prior to their synthesis.

Most accurate for **congeneric** series of compounds

Congeneric Series: Collection of structurally related compounds that vary primarily only by their substituents. For example, benzene, amino-benzene and chloro benzene would represent a congeneric series, however, indole would not be a part of the congeneric series due to the fact that it contains a different ring system. The following is an example of congeneric series of compounds



4) Lead Compound Optimization

QSAR is used to predict combination of steric, electronic and hydrophobic properties required to achieve the desired properties.

different properties must be optimized for different types of drugs.

potency: topical applications

transport properties: oral administration

type of activity: compounds with multiple activities (multiple bioassays......)

maximize desirable activities

minimize side effects

minimize toxic effects

Spectrum of selectivity (antibiotics for bacteria).

QSAR is developed with respect to a specific activity. This is done for a variety of activities for the same compounds. The analysis of the multiple QSARs is done simultaneously to find regions of substituent space, which maximize desirable properties and minimize undesirable properties.

2.6 Free Wilson analysis (the Additivity Model or De Novo Approach)[88, 89]

The Free Wilson model is a simple and efficient method for the quantitative description of structure activity relationships. It is the only numerical method which directly relates structural features with biological properties, in contrast to Hansch analysis, where physicochemical properties are correlated with biological activity values This method also assumes that the biological activity can be described by additives property of the substituents on a basic molecular structure. It is a true structure-activity relationship model based on the following assumptions:

- All the drugs tested should have the same parent structure.
- The substituents have to contribute to the biological activity, additively and in the same absence of other substituents in the molecule.
- The substituents pattern in various derivatives has to be the same.

Therefore the total activity (BA) of derivatives is the sum of constant independent of partial contributions.

$$\text{Log } 1/C = BA = \sum a_i I_j + \mu \text{----------------------} \textbf{(17)}$$

where, BA = Biological activity

a_i = Group of contribution of i^{th} substituents to the pharmacological activity

Ij = Substituents

μ = Theoretical biological activity value of reference or parent compound

The Free Wilson equations do not require the use of substituent constants such as π, σ, R, Es MR and F etc. The equation is solved by multiple linear regressions using the presence (1) or absence (0) of the different substituents as independent dummy parameters, while the measured activities serve as the dependent variable.

✓ Advantages of Free Wilson Approach

1. This is highly effective method, as the complexity of structure increases, the number of possible substituent at desired position increases
2. It is fast simple and cheap methods where no substituents constant like sigma, pi etc are considered.
3. The contribution of each substituent.at each position can clearly be identified and the substituents which fills or do not fills the principle of additivity can be recognized.
4. The method is effective especially when substituent constants are unavailable.

✓ Disdvantages of Free Wilson Approach

1. Only a small number of new analogs can be predicted from this approach
2. Predictions for substituents which are not included in the analysis are impossible
3. Structural variation is necessary in at least two different positions of substituent, otherwise, meaningless group contributions would result, one for each compound

2.7 MIXED APROACH (Combined Approach)[89]

Kubinyi has presented the combination of Hansch and Free Wilson Model as “mixed approach’ in which both models can be combined to a mixed approach in a linear and non linear form, which offers the advantages of both Hansch analysis and Free Wilson analysis, and widens their applicability in QSAR.

Log1/c = K_1®꞉+ K_2 ±꞉+ K_3Es + K-----------------------------------Hansch model(18)

Log1/c = «꞊Ｈ꞉paij---Free Wilson model (19)

Mixed approach can be written as;

Log1/c = paij +pKj¤j---(20)

Where kj represents the coefficient of different physico chemical parameters

Σaij Free Wilson part for the substituent and ϕj =π,σand Es contribution of the parent skeleton.

Mixed approach is based on the following assumptions:

- All the drugs tested have to have same parent structure.

- The substituent pattern in various derivatives has to be the same.

The substituent's contribution to the biological activity additively being independent of the presence or absence of other substituents. [90„91]

For the successful application of the mixed approach it is highly recommended to derive Hansch equation for each subset and to compare whether they correspond to each other or not, before combining them into equation with the help of indicator variables.

2.8 Demands of Biological as well as Physiological Approach

2.8.1Biological activity: [92]

1) Large range in observed activities.
2) Identical mode of action.
3) Concentration in molar units.
4) Activity data as a function of concentration (IC$_{50}$).
5) Activity data in percentage.
6) Possible time dependency.

2.8.2 Requirement of biological activity data used in QSAR analysis[93]

The biological data used in QSAR analysis can be of the following types:-

Source of Activity Biological Parameters

1. Isolated receptors

Rate constant Log K, Log K cat

Michaelis-menten constant Log1/Km

Inhibition constant Log 1/Ki

Affinity data PA2: PA1

2. Cellular systems

Inhibition constant Log 1/IC_{50}

Cross-resistance Log CR

In vitro biological data Log 1/C

Mutagenicity state Log TA_{98}

3. In vivo systems

Bio concentration factor Log BCF

In-vivo reaction rate Log I (induction)

Pharmacodynamic rates Log T (total clearance)

2.8.3Selection of physicochemical parameters[94, 95]

1. Large range in parameter space.
2. No significant inter correlation between descriptors used.
3. Homogenous distribution in the parameter space covered.

4. Ratio of number of independent variables to number of tested derivatives has to be considered (about 5 derivatives per descriptors).
5. If descriptors from complications in the literature are used; it should be checked if the values are appropriate for series in question.

2.9 Statistical Methods used in QSAR Analysis[96]

Statistical methods are an essential component of QSAR work. They help to build models, estimate a model's predictive abilities, validate an already existing model and find relationship and co-relationship among variables and activities. Data analysis methods are used to recombine data into forms and groups observations into hierarchies.

2.9.1 Regression Methods: This method is an important tool in model building. It is a mathematical procedure, which co-relates dependent (X) variable with the independent (Y) variables. There can be different forms of regression analysis:

- **Simple linear regression analysis:** An independent variable is correlated with a dependent variable and produces a linear one-term equation. It is useful for discovering some of the most important descriptors.

- **Multiple linear regression analysis (MLR):** More than one independent variables are correlated with dependent variable and a single multi term equation is formed. The number of variables should be one-fifth of the molecules in a series i.e 1 variable for every 5 compounds.

- **Stepwise linear regression analysis.** This is useful when number of independent variables is very high and is thus correlated in a stepwise manner with the dependent variable and thus producing a multi term linear equation.

2.9.2 Principal Component Analysis (PCA): Principal component analysis is a data reduction method, using mathematical techniques to identify pattern in a data matrix. The main element of this approach consist of the construction of a small set of new orthogonal, i.e. on correlated variables derived from a linear combination of the original variables.

2.9.3 Partial Least Square (PLS): PLS is an iterative procedure that applies two criterions to produce its solution.

1) To extract a new component so as to maximize the degree of commonality between all of the structure parameter columns collectively and the experimental data.

2) In the evaluation phase of a PLS iteration, the criterion for acceptance of the principal component just generated is an improvement in the ability to predict (not to reproduce) the dependent variable.

Technique used in PLS to assess the predictive ability of a QSAR is cross –validation (excluding temporarily the unknown compounds and then using the resulting equation to predict experimental measurement of the omitted compounds).

Hundreds or even thousands of independent variables (X-block) can be correlated with one or several dependent variables (Y-block). PLS is used when X data contain collinearities or N is less than 5M; where N is the number of compounds and M is the number of independent variables. Often perfect correlations are obtained in PLS analysis; due to the usually large number of X variables cross validation procedure must be used to select the model having the highest predictive values. Several PLS are performed in which one (leave one out technique, LOO) or several objects are eliminated from the data set. It is the method of choice in 3D QSAR method.

Drawback:

- ✓ no guarantee that it will uncover every relation in the training set.
- ✓ PLS analysis can only make linear correlations. However, it is conceivable that the complex relationship between molecular structure and biological activity is not linear.

2.9.4 Genetic Function Approximation (GFA): It provides multiple models that are created by evolving random initials models using a genetic algorithm. Models are improved by performing a cross over operation to recombine better sorting models; this method is used when dealing with a large numbers of descriptors.

2.9.5 Genetic Partial Least Squares (G/PLS): This method combines best of GFA and PLS. Each generation has PLS applied to instead of multiples linear regression and so each model can have more terms in it without fear of overfittng. G/PLS retains the ease of interpretations of GFA by back transforming the PLS component to the original variable. ***A major advantage of this approach*** is that a collection of diverse small models is generated that all have roughly the same high predictability. ***A disadvantage*** is that it takes too long to perform cross validation on each generation and, thus, one need to have a reasonable idea of how many terms to keep before starting

- **Regression analysis**

Regression analysis is a statistical tool for the investigation of re-lationships between variables or the causal effect of one variable upon another to explore such issues, 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 variable that they influence. This assesses the "statistical significance" of the estimated relationships, that is, the degree of confidence that the true relationship is close to the estimated relationship.. The goal of regression analysis is to determine the values of parameters for a function that cause the function to best fit a set of data observations. Regression analysis reveals average relation ship between variables and helps in estimation and prediction. It correlates independent X variables (eg physico chemical parameters, indicator variable) with dependent Y variables (biological data)

Applications of regression analysis

1. It provides estimates of values of dependent variable from values of independent variable. The device used for this estimation is regression line, which describes the average relation ship existing between variables. The equation of this line is called regression equation
2. With the help of regression coefficient we can calculate correlation Coefficient. The square of correlation coefficient (r) called coefficient of determination, measures the degree of correlation that exists between variables In general, greater the value of r^2, the better is the regression equation.

3. The third aim of regression analysis is to measure the error involved in using the regression line as basis of estimation. For this purpose the standard error of estimate is obtained

2.10 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 assist in this. 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.

A regression equation can be accepted in QSAR studies if:

1. Correlation coefficient r is around or better than 0.9
2. The standard deviation ‘s’ is not much larger than the standard deviation of the biological data
3. If its F value indicates that the overall significance is better than 95%

An equation has to be rejected if

1. The number of variables included in the regression equation is unreasonably large.
2. The standard deviation is smaller than the error in the biological data.(over prediction by the model)

2.11 Terms Commonly Used In Regression Analysis

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

High value of regression coefficient (r>.90) indicates that the statistical significance of the regression equation is high, while its low value indicates that the substituent is not important for the process under consideration

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.

$$r^2 = l - \Sigma\Delta^2/Syy$$

$$\text{Where, } Syy = \Sigma(y\ obs - y\ mean)^2$$

$$\Sigma\Delta^2 = SSQ = \Sigma(y\ obs - y\ cal)^2$$

Where ‘Syy’ is overall(total) variance,‘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.

C) Standard Deviation (SD): 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 larger the number of objects and smaller the number of variables are the smaller the standard deviation 's'The smaller the value of **SD** the better is the QSAR.

$$DF=n-k-1.$$

$$SD=\sqrt{}\Sigma(y_{obs} - y_{cal})^2 / n-k-1$$

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.

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

E) Predicted residual sum of square (PRESS):

This is the sum of the overall compounds of the square difference between the actual and the predicted value of dependent variables. Other interpretations of this acronym include predictive error sum of square. PRESS will contain one contribution from each observation.

$$P = \Sigma(y_{obs} - y_{calc})^2$$

PRESS is good estimate of the real prediction error of the model, if PRESS is smaller than the sum of squares of the response values (SSY), the model predicts better than chance and can be considerd “statistically significant”. The ratio PRESS/SSY can be used also to calculate approximate confidence intervals of prediction of new observation for reasonable QSAR model, PRESS/SSY should be smaller than 0.4, and the value of this ratio smaller than 0.1 indicates excellant model.

F) Cross validation r^2 (q^2):

This is the equation-squared correlation coefficient generated during a validation procedure.

$$Cvr^2 = SD\text{-}PRESS/SD,$$

where SD is standard deviation

A cross-validated r^2 is usually smaller than the overall r^2 for a QSAR equation. It is used as a diagnostic tool to evaluate the predictive power of an equation

G) Number of compounds utilised(n); For good corelation large number of compounds must be used.

H) Outliers: An outlier is defined as a structure with a residual greater than two times the standard deviation of residual. **Outliers** are data points, which lie outside the general linear pattern of which the midline is the regression line

I) F-ratio: Ratio of r^2 to 1.0- r^2, weighted so that the fewer the explanatory properties and the more the values of the target property, the higher the F-ratio.

J) S: Root Mean Square (RMS) or the standard error, a measure of the target property uncertainty still unexplained after the QSAR has been derived

2.12 Limitations of QSAR

The main limitation of QSAR is that we cannot use an equation for everything

There are two basic problems with using equations.

The toxicity and biodegradation of all compounds from equations cannot be predicted.

Chemicals can be divided into different families based on their structure. To develop a good QSAR, 15 to 50 (or more) effect values on chemicals from the same family are needed. That is a lot of data to generate one equation for one chemical family. With thousands of chemical families, there just is not enough data to predict the toxicity and biodegradation of them all. Additional research is ongoing and more QSARs are being developed, but it will still be a long time before a QSAR is available for all compounds.

Predictions are never as good as real data

The results may be pretty close to reality but it is never really know until the test is run. As a result, scientists only use QSARs when they need an approximate value. For example, if a chemical is going to exist in the environment at 0.001 µg/L and QSARs predict the safe concentration in the environment is 1000 µg/L, the toxicologist might decide that it is not worth the time, resources and effort to conduct a toxicity test. Even if the QSAR is close, the environment will not suffer and testing resources can be spent on compounds that pose a greater threat to the environment.

2.13. Important factors to be repeated[89]

It is the combination of so many different effects which contribute to biological activity that makes the formulation of a sound QSAR model so difficult .There is need to repeat only the most important factors:

1. Lipophilicity and ionization are responsible for the transport and the distribution of drug in the biological system.
2. The drug receptor interaction a highly specialized hydrophobic, polar,steric interaction
3. Neither the drug nor their binding sites are completely rigid systems. A flexible fit occurs during the binding of the drug.
4. Entropy effects (freezing of conformational degrees of freedom) play an important.
5. The solvation-desolvation balance may be favorable or unfavorable for binding. The insertion of water molecules between the ligand and its binding site has to be considered.