A REVIEW ARTICLE
ROLE OF QSAR: SIGNIFICANCE AND USES IN MOLECULAR DESIGN

Anil Kumar Sahdev [1*], Bhawana Sethi [2], Suman Lata Rawat[3], Amrita Singh[4], Neeti Anand[5].
Innovative college of pharmacy Greater Noida Uttar Pradesh.
Corresponding author: Anil Kumar Sahdev (Assistant professor)
Innovative college of pharmacy Greater Noida Uttar Pradesh (India).

ABSTRACT: If we take a series of chemicals and attempt to form a quantitative relationship between the biological effects (i.e. the activity) and the chemistry (i.e. the structure) of each of the chemicals, then we are able to form a quantitative structure–activity relationship or QSAR. SAR is a heart of computer aided drug design (CADD), SAR is qualitative expression, when SAR express in terms of mathematical models describing important parameter like activity, structure and properties result in coining of the term QSAR. QSAR defined as quantitative correlation of biological activities with physicochemical properties. Now a days a newer technique i.e. 3 D QSAR is also used to overcome the problems of classical QSAR.

Keyword: QSAR, chemistry qualitative, 3 D structure, CoMFA, biological activities.

INTRODUCTION:

If we can understand how a molecular structure brings about a particular effect in a biological system, we have a key to unlocking the relationship and using that information to our advantage. Formal development of these relationships on this premise has proved to be the foundation for the development of predictive models. If we take a series of chemicals and attempt to form a quantitative relationship between the biological effects (i.e. the activity) and the chemistry (i.e. the structure) of each of the chemicals, then we are able to form a quantitative structure–activity relationship or QSAR. less complex, or quantitative, understanding of the role of structure to govern effects, i.e. that a fragment or sub-structure could result in a certain activity, is often simply termed a structure–activity relationship or SAR.SAR is a heart of computer aided drug design (CADD), SAR is qualitative expression, when SAR express in terms of mathematical models describing important parameter like activity, structure, properties result in coining of the term QSAR. QSAR defined as quantitative correlation of biological activities with physicochemical properties [1].

Biological activity = f (physicochemical parameter)

HISTORY OF QSAR:

1869:B.J. Richardson: narcotic effect of primary alcohol varies in proportion to their molecular weight.
1900:H.H. Meyer and C.E. Overton: lipoid theory of narcosis
1904:J. Traube: linear relation between narcosis and surface tension.
1930:L. Hammett: electronic sigma constant
1939:J. Fergusson formulated a concept linking narcotic activity, log p and thermodynamics.
1952-1956:R.W. Taft devised a procedure for separating polar, steric and resonance effect.
1964:C.Hansch and T. Fujita: QSAR
1970-1980: Development of 2 D QSAR
1980-1990: Development of 3 D QSAR
1984:P.Andrews: affinity contribution of functional group
1985:P. Goodford: GRID (hot spots at protein surface)
1988:R.Cramer: 3 D QSAR
1998: Ajay, W.P. Walters and M.A. Murcko, J.Sadowski: drug- like character

TYPE OF QSAR:

1. **Classical QSAR**-It consider only 2 D structure (Hansch and free Wilson analysis)
2. **3 D-QSAR analysis:** It has a much more scope. It start from 3 D structure and correlates biological activities with 3 D – property fields.

CLASSICAL QSAR:

- It is difficult to calculate various physicochemical parameter
- Parameter considered may not sufficient to describe drug-receptor interaction.
- It does not consider 3 D structure of molecules [1]

3 D QSAR: It consider 3 D structure of molecules.

Why 3D-QSAR IS SO ATTRACTIVE?

The era of quantitative analysis for the correlation of molecular structures with biological activities started in the 1960s from the classical equation for 2D-QSAR analysis proposed by Hansch [2]. The first applicable 3D-QSAR method was proposed by Cramer et al. in 1988. His program, CoMFA, was a major breakthrough in the field of 3D-QSAR. The primary aim of 3D-QSAR methods is to establish a correlation of biological activities of a series of structurally and biologically characterized compounds with the spatial fingerprints of numerous field properties of each molecule, such as steric demand, lipophilicity, and electrostatic interactions [3].

Typically, a 3D-QSAR analysis allows the identification of the pharmacophores arrangement of molecular features in space and provides guidelines for the design of next-generation compounds with enhanced bioactivity or selectivity. Since chemists and biologists know that 3D properties of molecules govern biological activity, it is especially informative to see a 3D picture of how structural changes influence biological activities. Approaches that do not provide such a graphical representation are often less attractive to the scientific community. An advantage of 3D-QSAR – over the traditional 2D-QSAR – method is that it takes into account the 3D structures of ligands and additionally is applicable to sets of structurally diverse compounds. The number of 3D-QSAR studies has increased exponentially over the last decade, since a variety of methods have been made commercially available in user friendly software [4]. As of the end of 2007, the number of papers dealing with 3D-QSAR is greater than 2500 when the CAS (Chemical Abstracts Service) service is searched using the keywords "3D-QSAR" or "CoMFA". However, it seems that the initial "QSAR hype" is over, as indicated by the constant number of new 3D-QSAR applications in the last few years. The major drawback of 3D-QSAR is that it is not applicable to huge data sets containing more than several thousand compounds, which are usually considered in high-throughput screening. For these kinds of studies, novel faster and simpler methods have been developed, which use the original 3D descriptors (i.e., molecular interaction fields or surface descriptors) as inputs for the generation of alignment-independent models. Examples for this kind of programs recently developed are Volsurf and Almond.

The most frequently applied methods include comparative molecular field analysis (CoMFA), comparative molecular similarity indices analysis (CoMSIA), and the GRID/GOLPE program (generating optimal linear PLS estimations). Several reviews have been published in the last few years dealing with the basic theory, the pitfalls, and the application of 3D-QSAR approaches. Apart from the commercial distribution, a major factor for the ongoing enthusiasm for CoMFA- related approaches comes from the proven ability of several of these methods to correctly estimate the biological activity of novel compounds. However very often the predictive ability of QSAR models is only tested in retrospective studies rather than taking the ability to design and develop novel bioactive molecules. Despite the known limitations of 3D-QSAR, the possibility to predict biological data is gaining respect as scientists realize that we are far away from the hoped-for fast and accurate forecast of affinity from (the structure of a) protein–ligand complexes by free-energy perturbation or empirical scoring methods[5-6].

Comparative molecule field analysis (C_OMFA):

For many years, 3D-QSAR has been used as a synonym for CoMFA, which was the first method that implemented the concept into a QSAR method, i.e., that the biological activity of a ligand can be predicted from its three-dimensional structure. Until now, CoMFA is probably the most commonly applied 3D-QSAR method. A CoMFA study normally starts with traditional pharmacophore modelling.

In order to suggest a bioactive conformation of each molecule and ways to superimpose the molecules under study. The underlying idea of CoMFA is that differences in a target property, e.g., biological activity, are often closely related to equivalent changes in shapes and strengths of non-covalent interaction fields surrounding the molecules. Or stated in a different way, the steric and electrostatic fields provide all information necessary for understanding the biological properties of a set of compounds. Hence, the molecules are placed in a cubic grid and the interaction energies between the molecule and a defined probe are calculated for each grid point. Normally, only two potentials, namely a steric potential in the form of a Lennard-Jones function and an electrostatic potential in the form of a simple Coulomb function, are used within a CoMFA study. It is obvious that the description of molecular similarity is not a trivial task nor is the description of the interaction process of ligands with corresponding biological targets. In the standard application of CoMFA, only enthalpic contributions of free energy of binding are provided by the potentials used. However, many binding effects are governed by hydrophobic and entropic contributions. Therefore, one has to characterize in advance the expected main contributions of forces and whether under these conditions CoMFA will actually be able to find realistic results. In the original CoMFA report, field values were systematically calculated for ligands at each grid point of a regularly sampled 3D grid box that extended 4 Å beyond the dimension of all molecules in the data set, using a sp3 carbon atom with +1 charge as probe. The grid resolution should be in a range to produce the field information that is necessary to describe variations in biological activity. On the other hand, introduction of too much irrelevant data to statistical analysis may result in a decrease of predictivity of the model. Typically, a resolution of 2 Å is utilized. Often, superior results are derived using a grid spacing of 2 Å as opposed to the more accurate 1 Å spacing. In addition, the CoMFA program provides a variety of other parameters (probe atoms, charges, energy scaling, energy cut-offs, etc.) which can be adjusted by the user. This flexibility in parameter settings enables the user to fit the whole procedure as closely as possible to his problem. However, it enhances the possibility of chance correlations. Interestingly, nearly all of the successful CoMFA analyses have been achieved with default parameters [7].

Molecular similarity indices in a comparative (C_OMSIA):

Due to the problems associated with the functional form of the Lennard -Jones potential used in most CoMFA methods, Klebe *et al* have developed a similarity indices-based CoMFA method named CoMSIA (comparative molecular similarity indices analysis). Instead of grid-based fields, CoMSIA is based on similarity indices that are obtained using a functional form that is adapted from the SEAL algorithm. Three different indices related to steric, electrostatic, and hydrophobic potentials were used in their study of the classical steroid benchmark data set. Models of comparable statistical quality with respect to cross-validation of the training set, as well as predictivities of a test set, were derived using CoMSIA. The advantage of this method lies in the functions used to describe the molecules studied, as well as the resulting contour maps. The contour maps obtained from CoMSIA are generally easier to interpret, compared to the ones obtained by the CoMFA approach. CoMSIA also avoids cut-off values used in CoMFA to restrict potential functions by assuming unacceptably large values [8-9].

ADVANCEMENT OF QSAR:

- **1 D QSAR:** It correlates the pKa and log P
- **2 D QSAR:** Affinity correlates to structure pattern of drug molecules.
- **3 D QSAR:** Affinity correlates 3 D structure of molecule.
- **4 D QSAR:** As with the 3 D + multiple representation of ligand conformation.
- **5 D QSAR:** As with the 4 D + multiple representation of induced fit sinerio.
- **6 D QSAR:** As with 5 D + multiple representation of salvation model [11].

PURPOSE OF QSAR:

QSAR should not be seen as an academic tool to allow for the post-rationalization of data. We wish to derive the relationships between molecular structure, chemistry and biology for good reason. From these relationships we can develop models, and with luck, good judgment and expertise these will be predictive. There are many practical purposes of a QSAR and these techniques are utilized widely in many situations. The purpose of in silico studies, therefore, includes the following:

• To predict biological activity and physico-chemical properties by rational means.
• To comprehend and rationalize the mechanisms of action within a series of chemicals.

Underlying these aims, the reasons for wishing to develop these models include

• Savings in the cost of product development (e.g. in the pharmaceutical, pesticide, etc. areas).
• Predictions could reduce the requirement for lengthy and expensive animal tests.
• Other areas of promoting green and greener chemistry to increase efficiency and eliminate waste by not following leads unlikely to be successful [1].

WHAT IS REQUIRED FOR A GOOD QSAR MODEL?

1.A defined end point.
2.An unambiguous alogrithm.
3.A defined domain of applicability.
4.Appropriate measure of goodness of fit.

Common QSAR Modelling Errors
•Uninformative descriptors •Poor descriptor selection and chance correlations •Modelling complex, nonlinear structure property relationships with linear models •Incorrectly validating QSPR models •Not understanding the domain of applicability of models

METHOD:

QSAR Modeling process consists of 5 main steps.

- Begins with the selection of molecules to be used.
- Selection of descriptor; numerical represented of molecular feature (e.g. no. of carbon)
- Original descriptor pool must be reduced in size.
- Model building
- The reliability of the model should be tested [2].

APPLICATION OF QSAR:

The ability to predict a biological activity is valuable in any number of industries. Whilst some QSARs appear to be little more than academic studies, there are a large number of applications of these models within industry, academia and governmental (regulatory) agencies. A small number of potential uses are listed below:

Chemical: One of the first historical application was to predict boiling point. It is well known for instance that within a particular family of chemical compounds, especially of organic chemistry, that these are strong correlation between structure and observed properties. A simple example is the relationship between the number of carbon in alkanes and their boiling point. There is a clear trend in the increase of boiling point with an increase in the carbon, and this serves a means for predicting the boiling point of higher alkanes.

Biological: The biological activity of a molecule is usually measured in assay to establish the level of inhibition of particular signal transduction or metabolic pathway. Drug discovery often involves the use of QSAR to identify chemical structure that could have good inhibitory effect on specific target and have low toxicity [10].

• The rational identification of new leads with pharmacological, biocidal or pesticidal activity.
• The optimization of pharmacological, biocidal or pesticidal activity.

• The rational design of numerous other products such as surface-active agents, perfumes, dyes, and fine chemicals.
• The identification of hazardous compounds at early stages of product development or the screening of inventories of existing compounds.
• The designing out of toxicity and side-effects in new compounds.
• The prediction of toxicity to humans through deliberate, occasional and occupational exposure.
• The prediction of toxicity to environmental species.
• The selection of compounds with optimal pharmacokinetic properties, whether it be stability or availability in biological systems
• The prediction of a variety of physico-chemical properties of molecules (whether they be pharmaceuticals, pesticides, personal products, fine chemicals, etc.)
• The prediction of the fate of molecules which are released into the environment.
• The rationalization and prediction of the combined effects of molecules, whether it be in mixtures or formulations [1].

PHYSICOCHEMICAL PROPERTIES: From the historical perspective of QSAR, the biological activity is a function of physicochemical properties. Physicochemical properties essentially refer to any structural, physical, chemical property of a drug. Hydrophobic, steric, electronic parameter have been studied by QSAR approach.

Hydrophobicity: The term hydrophobic means 'water fearing', from the Greek word 'hydro' means water and 'phobo' means fear. Hydrophobicity may be defined as tendency of organic molecules to partition away from water into some less polar medium. Hydrophobicity governs numerous and different biological processes such as transport, metabolism, and distribution of biological molecules. Hydrophobicity of a solute can readily determine by measuring partition coefficient.

Partition coefficient:
P= (drug) octanol / (drug) aqueous phase phosphate buffer
P is the quotient of two concentration and is normally in the form of its logarithm to base 10 (log P), because P range from 10^{-4} to 10^{-8}
Log 1/c =k_1 log P+ k_2
Log 1/c=biological property
K_1 and k_2 =constant
Generally it has been found that increasing hydrophobicity of a lead compound results in increase in biological activity. The interaction of the drug molecule with the receptor may also be governed by hydrophobic forces, because the binding site in the receptor is usually hydrophobic.
Log 1/c = - k_1 $(\log P)^2$ + k_2 log P+ k_3
NOTE: P is small the $(\log P)^2$ is very small
P is large $(\log P)^2$ becomes much more significant.

Measurement of log P : The basic procedure for obtaining a partition coefficient is to shake a weighed amount of a substance in a flask containing measured amount of water saturated octanol and octanol saturated water .The aqueous phase may be buffered with a phosphate buffer of pH 7.4 so as to mimic the physiological pH. The amount of substance dissolved in one or both the phase is determined by reverse phase HPLC, centrifugal partition, and chromatography. This is known as the "shake flask" method. Log P frequently used to estimate the membrane permeability and the bioavailability of compound, since an orally administered drug must be enough lipophilic to cross the lipid bilayer of the membrane and on the other hand it must be sufficiently water soluble to be transported in the blood and the lymph.

Hydrophilic –4.0 < logP < +8.0 lipophilic
Log p can be predicted by substituent hydrophobicity constant
$\Pi_{X=}$log px –logp_n
$\Pi_{X=}$ hydrophobicity constant with the substituent x
log $p_{x=}$logP of substituent benzene
log$p_{n=}$log P of simply benzene

Electronic effects: Electronic effect of various substituent have an effect on ionization or the polarity of a drug. This effect on ionization affect the following:
1. Passage of drug through cell membrane
2. Drug – receptor binding
Electronic properties of a molecules can be described by a wide variety of different parameter e.g. by Hammett substituent constant, field and resonance parameter, pk_a values, parameter derived from molecular spectroscopy, charge transfer constant, dipole moments, hydrogen bonding parameter.

Hammett substituent constant (σ): It is a measure of electron withdrawing or electron donating capacity of a substituent on an aromatic ring. Hammett constants (s, s1, s2) account for 7000/8500 equations in the Physical organic chemistry (PHYS) database and nearly 1600/8000 in the Biology (BIO) database, whereas quantum chemical indices such as HOMO, LUMO, BDE, and polarizability appear in 100 equations in the BIO database.
σ_x =log(k_x-k_H)=log k_x-log k_H
$\sigma_{x=}$parent molecule with electron withdrawing group at meta and para position.
Equilibrium shifted toward right.Hence σ value is positive.
σ_y =log (k_y-k_H)=log k_y -log k_H

σ_y denotes the Hammett substituent.Hence σ value is negative. Equilibrium shifted toward left σ can't be calculated for ortho position because in this case steric effect is played important role for substituent [11].

Despite the extensive and successful use in QSAR studies, there are some limitations to the Hammett equation.

1. Primary s values are obtained from the thermodynamic ionizations of the appropriate benzoic acids at 25˙ c these are reliable and easily available. Secondary values are obtained by comparison with another series of compounds and are thus subject to error because they are dependent on the accuracy of a measured series and the development of a regression line using statistical methods.
2. In some multisubstituted compounds, the lack of additivity needs to be noted. Proximal effects are operative and tend to distort electronic contributions**.**
3. Changes in mechanism or transition state cause discontinuities in Hammett plots. Nonlinear plots are often found in reactions that proceed by two concurrent pathways
4. Changes in solvent may lead to dissimilarities in reaction mechanisms. Thus extrapolation of s values from a polar solvent (e.g., CH3CN) to a nonpolar solvent such as benzene has to be approached cautiously. Solvation properties will differ considerably, particularly if the transition state is polar and/or the substituents are able to interact with the solvent.
5. A strong positional dependency of sigma makes it imperative to use appropriate values for positional, isomeric substituents. Substituents ortho to the reaction centre are difficult to describe and thus one must resort to a Fujita-Nishioka analysis [12].

Steric effect: In steric effect size and shape of molecule is involved. A group which is very small in size, may not fit at the active site properly. For a proper fitting it is therefore necessary that the size of the drug molecule and the shape of the receptor site are complementary to each other. Acid and base catalysed hydrolysis of an ester involve a tetrahedral transition state, ester hydrolysis has been used to quantify the steric parameter.

$E_s = \log k_x - \log k_0$

E_s=Taft steric factor, k_x = rate of hydrolysis of aliphatic ester with substituent

k_0= rate of hydrolysis of aliphatic ester without substituent [11].

WHAT IS MODEL?

As a common and successful research approach, quantitative structure activity/property relationship (QASR/QSPR) studies are applied extensively to chemo metrics, pharmacodynamics, pharmacokinetics, toxicology and so on. Recently, the mathematical methods used as regression tools in QSAR/QSPR analysis have been developing quickly. Multiple Linear Regression (MLR), Partial Least Squares (PLS), Neural Networks (NN), Support Vector Machine (SVM), being upgraded by improving the kernel algorithms or by combining them with other methods, but also some new methods, including Gene Expression Programming (GEP), Project Pursuit Regression (PPR) and Local Lazy Regression (LLR), are being mentioned in the current reported QSAR/QSPR studies.

1. Multiple Linear Regression (MLR):

MLR is one of the earliest methods used for constructing QSAR/QSPR models, but it is still one of the most commonly used ones to date. The advantage of MLR is its simple form and easily interpretable mathematical expression. Although utilized to great effect, MLR is vulnerable to descriptors which are correlated to one another, making it incapable of deciding which correlated sets may be more significant to the model. Some new methodologies based on MLR have been developed and reported in recent papers aimed at improving this technique. The three most important Best Multiple and commonly used of the methods are described in detail below.

- Best Multiple Linear Regression (BMLR)
- Heuristic Method (HM)
- Genetic Algorithm based Multiple Linear Regression (GA-MLR)

Best Multiple Linear Regression (BMLR):

BMLR implements the following strategy to search for the multi-parameter regression with the maximum predicting ability. All orthogonal pairs of descriptors i and j (with R2ij < R2min, default value R2ij < 0.1) are found in a given data set. The property analysed is treated by using the two parameter regression with the pairs of descriptors, obtained in the first step. The Nc (default value Nc = 400) pairs with highest regression correlation coefficients are chosen for performing the higher-order regression treatments. For each descriptor pair, obtained in the previous step, a non-collinear descriptor scale, k (with R2ik < R2nc and R2kj < R2nc, default value R2 < 0.6) is added, and the respective three-parameter regression treatment is performed. If the Fisher criterion at a given probability level, F, is smaller than that for the best two-parameter correlation, the latter is chosen as the final result. Otherwise, the Nc (default value Nc = 400) descriptor triples with highest regression correlation coefficients are chosen for the next step. For each descriptor set, chosen in the previous step, an additional non-collinear descriptor scale is added, and the respective *(n + 1)* parameter regression treatment is performed. If the Fisher criterion at the given probability level, *F*, is smaller than for the best two-parameter correlation, the latter is chosen as the final result. Otherwise, the *Nc* (default value *Nc* = 400) sets descriptor sets with highest regression correlation coefficients are chosen, and this step repeated with $n = n + 1$. Like MLR, BMLR is noted for its simple and interpretable mathematical expression. Moreover, overcoming the shortcomings of MLR, BMLR works well when the number of compounds in the training set doesn't exceed the number of molecular descriptors by at least a factor of five. However, BMLR will derive an unsatisfactory result when the structure-activity relationship is non-linear in nature. When too many descriptors are involved in a calculation, the modelling process will be time consuming. To speed up the calculations, it is advisable reject descriptors with insignificant variance within the dataset. This will significantly decrease the probability of including unrelated descriptors by chance. In addition, BMLR is unable to build a one-parameter model. [13]

Heuristic Method (HM):

HM, an advanced algorithm based on MLR, is popular for building linear QSAR/QSPR equations because of its convenience and high calculation speed. The advantage of HM is totally based on its unique strategy of selecting variables. The details of selecting descriptors are as follows:

First of all, all descriptors are checked to ensure that values of each descriptor are available for each structure. Descriptors for which values are not available for every structure in the data are discarded. Descriptors having a constant value for all structures in the data set are also discarded. Thereafter all possible one parameter regression models are tested and the insignificant descriptors are removed. As a next step, the program calculates the pair correlation matrix of descriptors and further reduces the descriptor pool by eliminating highly correlated descriptors. The details of validating inter correlation are:

(a) All quasi orthogonal pairs of structural descriptors are selected from the initial set. Two descriptors are considered orthogonal if their inter correlation coefficient *is* lower than 0.1.

(b) The pairs of orthogonal descriptors are used to compute the biparametric regression equations.

(c) To a multi-linear regression (MLR) model containing *n* descriptors, a new descriptor is added to generate a model with *n + 1* descriptors if the new descriptor is not significantly correlated with the previous *n* descriptors.

Step (c) is repeated until MLR models with a prescribed number of descriptors are obtained. The goodness of the correlation is tested by the square of coefficient regression (R^2), square of cross validate coefficient regression (*q2*), the F-test (*F*), and the standard deviation (*S*) [14].

HM is commonly used in linear QSAR and QSPR studies, and also as an excellent tool for descriptor selection before a linear or nonlinear model is built. The advantages of HM are the high speed and the absence of software restrictions on the size of the data set. HM can either quickly give a good estimation about what quality of correlation to expect from the data, or derive several best regression models. HM usually produces correlations 2 – 5 times faster than other methods with comparable quality.

Additionally, the maximum number of parameters in the resulting model can be fixed in accordance with the situation so as to save time. As a method inherited from MLR, HM is also limited in linear models [15-16].

Genetic Algorithm based Multiple Linear regression (GA-MLR):

Combining Genetic Algorithm (GA) with MLR, a new method called GA-MLR is becoming popular in currently reported QSAR and QSPR studies. In this method, GA is performed to search the feature space and select the major descriptors relevant to the activities or properties of the compounds. We give a brief summary of the main procedure of GA here in. The first step of GA is to generate a set of solutions (chromosomes) randomly, which is called an initial population. Then, a fitness function is deduced from the gene composition of a chromosome. The Friedman LOF function is commonly used as the fitness function, which was defined as follows:

$$LOF = \{SSE/(1-(c+ dp/n))\}^2$$

Where SSE is the sum of squares of errors, *c* is the number of the basis function (other than the constant term), *d* is the smoothness factor, *p* is the number of features in the model, and *n* is the number of data points from which the model is built. Unlike the *R2* error, the LOF measure cannot always be reduced by adding more terms to the regression model. By limiting the tendency to simply add more terms, the LOF measure resists over-fitting of a model. Then, crossover and mutation operations are performed to generate new individuals. In the subsequent selection stage, the fittest individuals evolve to the next generation. These steps of evolution continue until the stopping conditions are satisfied. After that, the MLR is employed to correlate the descriptors selected by GA and the values of activities or properties.

GA, a well-estimated method for parameter selection, is embedded in GA-MLR method so as to overcome the shortage of MLR in variable selection. Like the MLR method, the regression tool in GA-MLR, is a simple and classical regression method, which can provide explicit equations. The two parts have a complementation for each other to make GA-MLR a promising method in QSAR/QSPR research [18-19].

2. Partial Least Squares (PLS):

The basic concept of PLS regression was originally developed by Wold. As a popular and pragmatic methodology, PLS is used extensively in various fields. In the field of QSAR/QSPR, PLS is famous for its application to CoMFA and CoMSIA. Recently, PLS has evolved by combination with other mathematical methods to give better performance in QSAR/QSPR analyses. These evolved PLS', such as Genetic Partial Least Squares (GPLS), Factor Analysis Partial Least Squares (FA-PLS) and sections.

- Genetic Partial Least Squares (G/PLS)
- Factor Analysis Partial Orthogonal Signal Correction Partial Least Squares (OSC-PLS), are briefly introduced in the following Least Squares (FA-PLS)
- Orthogonal Signal Correction Partial Least Squares (OSC-PLS)[20].

Genetic Partial Least Squares (GPLS): GPLS is derived from two QSAR calculation methods Genetic Function Approximation (GFA) and PLS. The GPLS algorithm uses GFA to select appropriate basis functions to be used in a model of the data and PLS regression is used as the fitting technique to weigh the basis functions' relative contributions in the final model. Application of GPLS thus allows the construction of larger QSAR equations while still avoiding over-fitting and eliminating most variables. As the regression method used in Molecular Field Analysis (MFA), a well-known 3D-QSAR analysis tool, GPLS is commonly used [21].

Factor Analysis Partial Least Squares (FA-PLS):

This is the combination of Factor Analysis (FA) and PLS, where FA is used for initial selection of descriptors, after which PLS is performed. FA is a tool to find out the relationships among variables. It reduces variables into few latent factors from which important variables are selected for PLS regression. Most of the time, a leave-one-out method is used as a tool for selection of optimum number of components for PLS. [22, 23]

Orthogonal Signal Correction Partial Least Squares (OSC-PLS):

Orthogonal signal correction (OSC) was introduced by Wold et al. to remove systematic variation from the response matrix *X* that is unrelated, or orthogonal, to the property matrix *Y*. Since then, various OSC algorithms have been published in an attempt to reduce model complexity by removing orthogonal components from the signal. A pre-processing with OSC will help traditional PLS to obtain a more precise model, as proven in many studies of spectral analysis. To date, unfortunately, there are only a few reports in which OSC-PLS is applied to QSAR/QSPR studies, but more QSAR or QSPR research involving application of the OSC-PLS method are expected in the future[24].

3. Neural Networks (NN):

As an alternative to the fitting of data to an equation and reporting the coefficients derived there from, neural networks are designed to process input information and generate hidden models of the relationships. One advantage of neural networks is that they are naturally capable of modelling of nonlinear systems. Disadvantages include a tendency to over fit the data, and a significant level of difficulty in ascertaining which descriptors are most significant in the resulting model. In the recent QSAR/QSPR studies, RBFNN and GRNN are the most frequently used ones among NN.

- Radial Basis Function Neural Network (RBFNN)
- General Regression Neural Network (GRNN)[25]

4. Support Vector Machine (SVM):

As a novel type of machine learning method, is gaining popularity due to its many attractive features and promising empirical performance. Originally, SVM was developed for pattern recognition problems. After that, SVM it was applied to regression by the introduction of an alternative loss function and results appear to be very encouraging. As a developing method, new types of SVM are coming in on the stage of QSAR/QSPR, such as:

- Least Square Support Vector Machine (LS-SVM)
- Grid Search Support Vector Machine (GS-SVM)
- Potential Support Vector Machine (P-SVM) and
- Genetic Algorithms Support Vector Machine (GASVM).

LS-SVM, the most commonly used one method [26, 27].

5. Gene Expression Programming (GEP):

Gene expression programming was invented by Ferreira in 1999 and was developed from genetic algorithms and genetic programming (GP).GEP uses the same kind of diagram representation of GP, but the entities evolved by GEP (expression trees) are the expression of a genome. GEP is more simple than cellular gene progression. It mainly includes two sides: the chromosomes and the expression trees (ETs). The process of information of gene code and translation is very simple, such as a one-to-one relationship between the symbols of the chromosome and the functions or terminals they represent. The rules of GEP determine the spatial organization of the functions and terminals in the ETs and the type of interaction between sub-ETs. Therefore, the language of the genes and the ETs represents the language of GEP.

- The GEP chromosomes,
- Expression trees (ETs)
- The mapping mechanism

6. Project Pursuit Regression (PPR):

PPR, which was developed by Friedman and Stuetzle, is a powerful tool for seeking the interesting projections from high-dimensional data into lower dimensional space by means of linear projections. Therefore, it can overcome the curse of dimensionality Friedman and Stuetzle's concept of PPR avoided many difficulties experienced with other existing nonparametric regression procedures. It does not split the predictor space into two regions, thereby allowing, when necessary, more complex models. In addition, interactions of predictor variables are directly considered because linear combinations of the predictors are modelled with general smooth functions [29].

7. Local Lazy Regression (LLR):

Most QSAR/QSPR models often capture the global structure-activity/property trends which are present in a whole dataset. In many cases, there may be groups of molecules which exhibit a specific set of features which relate to their activity or property. Such a major feature can be said to represent a local structure activity/property relationship. Traditional models may not recognize such local relationships. LLR is an excellent approach which extracts a prediction by locally interpolating the neighbouring examples of the query which are considered relevant according to a distance measure, rather than considering the whole dataset. That will cause the basic core of this approach which is a simple assumption that similar compounds have similar activities or properties; that is, the activities or properties of molecules will change concurrently with the changes in the chemical structure. For one or more query points, "lazy" estimates the value of an unknown multivariate function on the basis of a set of possibly noisy samples of the function itself. Each sample is an input/output pair where the input is a vector and the output is a number. For each query point, the estimation of the input is obtained by combining different local models. Local models considered for combination by lazy are polynomials of zeroth, first, and second degree that fit a set of samples in the neighbourhood of the query point. The neighbours are selected according to either the "Manhattan" or the "Euclidean" distance [30, 31].

EVALUATION OF THE QUALITY OF QSAR MODELS:

QSAR modeling produce predictive model derived from application of statistical tools correlating biological activities or physicochemical properties. QSAR are being applied in many discipline for example: risk assessment, toxicity prediction, drug discovery and lead optimization. Obtaining a good quality QSAR model depend upon many factor such as the quality of input data , the choice of descriptor and statistical method for modeling and for validation.

For validation of QSAR models, usually various strategies are adopted:
1. Internal validation or cross – validation
2. External validation by splitting the available data set into training set for model development.
3. Blind external validation by application of model on new external data.
The success of any QSAR model depends on accuracy of the input data, selection of appropriate descriptor and statistical tools, and most importantly validation of the developed model [10].

CORRELATION OF PHYSICOCHEMICAL PARAMETER WITH BIOLOGICAL ACTIVITY:

The extra thermodynamic approach (Hansch approach):

Hansch analysis correlates biological activity values with physicochemical properties by linear, linear multiple, or nonlinear regression analysis; thus, Hansch analysis is indeed a property-property relationship model.
As practically all parameters used in Hansch analysis are linear free energy-related values ***(i.e.*** derived from rate or equilibrium constants), the terms "linear free energy-related approach" or "extra thermodynamic approach" are sometimes used as synonyms for Hansch analysis. Also the biological activity values are, if they are properly defined, linear free energy-related values **(e.g.** binding or inhibition constants, absorption and distribution rate constants, or complex data which correspond to a weighted combination of several such unit processes). Early attempts to correlate biological activity values with lipophilicity, expressed **e.g.** by solubility or partition coefficients. Only explained nonspecific structure activity relationships; the application of the concept of a general biological Hammett equation failed. The methodological breakthrough came from a suggestion by Fujita, at that time working in Hansch's group, to apply an approach used earlier by Taft. Hansch and Fujita combined different physicochemical parameters in one equation, In the equation (C = molar concentration) that produces a certain biological effect; k_1, k_2, k_3 = coefficients determined by a least squares procedure, **e.g.** linear multiple regression analysis, *to* fit the biological data).

$$\log 1/C = k_1 \log p + k_2 \sigma_+ + k_3 \quad \text{----------1}$$

For *in vivo* data eq. was extended to other eq. 2 by including a parabolic lipophilicity term. The idea behind eq. 2 was that molecules which are too hydrophilic or too lipophilic will not be able to cross lipophilic or hydrophilic barriers, respectively. Therefore, they will have a lower probability to arrive at the receptor site than molecules with intermediate lipophilicity, being readily soluble in aqueous phases as well as in lipid phases.

$$\log 1/C = -k_1(\log P)^2 + k_2 \log P + k_3 + k_4 \quad \text{----------eq. 2}$$

c is the molar concentration that evoke a standard biological activity (reciprocal of the concentration is taken because it is desirable that greater activity is achieved at a minimum dose) and the negative sign for $(\log P)^2$ signify the requirement of optimum lipophilicity [33-34].

The Additivity Model (Free Wilson Analysis):

The Free Wilson approach is a true structure-activity relationship model. An indicator variable is generated for each structural feature that deviates from an arbitrarily chosen reference compound. Values 1, indicating the presence of a certain substituent or structural feature, and 0, indicating its absence, are correlated with the biological activity values by linear multiple regression analysis. The resulting regression coefficients of the indicator variables are the biological activity contributions of the corresponding structural elements. "Mathematical model", "additivity model", or "de novo approach" are synonyms for the Free Wilson method. The free Wilson analysis assume that the parent structure and each substituent contribute an additive increment to the logarithm of the biological response, the increment from each substituent is constant and independent of interaction of substituent in other position.it means that the introduction of any particular substituent at any particular position in a molecule will always change the relative biological activity of the molecule by the same amount, independent of whatever other substituent are present in the molecule. An equation can be drawn up as such:

$$BA = \sum a_i x_i + \mu$$

Where BA=biological activity
x_i=ith substituent (if present value=1,if absent value=0)
a_i=contribution of ith substituent to the biological activity
μ=overall average activity of the parent skeleton [34].

The craig plot:it is used to visualize the properties of various substituent such as π and σ values.the y axis is taken as the value for the σ factor ,while the x axis taken as the value for π factor.The craig plot can be used for planning the selection of substituent used in QSAR study.It also shows which substituent have a positive parameter and which have a negative parameter.

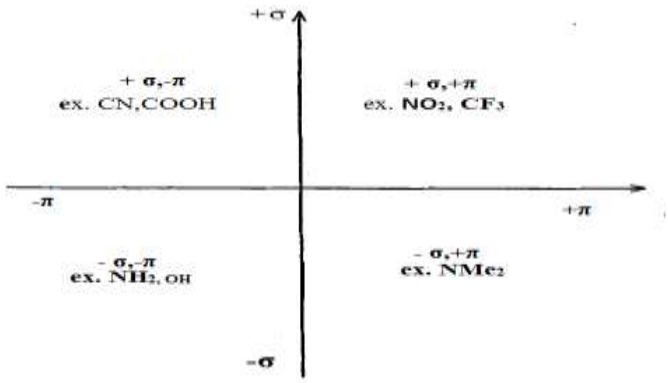

Fig. Craig plot of σ and π parameter

The topliss operational schemes: Also named as topliss decision tree**:**
Condition:when synthesis of large number of compound is difficult but biological testing is easily possible.

- It is used for synthesis of only most active compound.This is actually floe diagram of substituent.Here hydrophobic and electronic property of substituent are considered.Topliss operational scheme are two type:

1.one for aromatic substance
2.other is for aliphatic side chain substituent.

The use of this quantitative correlation analysis lies in:

1. Its ability to develop SAR
2. Predicting the properties and activities of untested molecules.
3. Optimizing the properties of a lead compound.
4. Comparing different QSAR models stastically.
5. Generating hypotheses about the characteristics of a receptor binding site.
6. Validating models of receptor binding site [11].

CONCLUSION:

The interactions of drugs with their biological counterparts are determined by intermolecular forces, *i.e.* by hydrophobic, polar, electrostatic, and steric interactions. Quantitative structure-activity relationships (QSAR) derive models which describe the structural dependence of biological activities either by physicochemical parameters (Hansch analysis), by indicator variables encoding different structural features (Free Wilson analysis), or by three-dimensional molecular property profiles of the compounds (comparative molecular field analysis, CoMFA). The classical models of quantitative structure-activity analyses do not consider the three-dimensional arrangement of functional groups, some recent Approaches (3 D QSAR) deal with this problem and describe biological activities. The methods of quantitative structure-activity relationships which have developed during the past 30 years nowadays are widely applied to describe the relationships between chemical structures of molecules and their biological activities. Many attempts have been made to understand structure-activity relationships in physicochemical terms (or in terms of structural features, using indicator variables for individual substituents and groups) and to design new drugs on a more rational basis.

REFERENCES:

[1] Borm PJ, Robbins D, Haubold S et al. (2006) The potential risks of nanomaterials: A review carried out for ECETOC. Part Fibre Toxicol 3:11
[2] Oksel ceyda, wang xue z. university of leeds.
[3] Tropsha A, Gramatica P, Gombar VK (2003) the importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci 22:69–77
[4] Chen FQ, Gerion D (2004) Fluorescent CdSe/ZnS nanocrystal-peptide conjugates for long-term, nontoxic imaging and nuclear targeting in living cells. Nano Lett 4:1827–1832
[5] Chatterjee R (2008) The challenge of regulating nanomaterials. Environ Sci Technol 42:339–343
[6] Nel A, Xia T, Madler L et al. (2006) Toxic potential of materials at the nanolevel. Science 311:622–627
[7] Yuranova T, Laub D, Kiwi J (2007) Synthesis activity and characterization of textiles showing activity under daylight irradiation. Catal Today 122:109–117
[8] Zhou K, Wang R, Xu B et al. (2006) Synthesis, characterization and catalytic properties of CuO nanocrystals with various shapes. Nanonechnology 17:3939–3943
[9] Alivisatos P (2004) the use of nanocrystals in biological detection. Nat Biotechnol 22:47–52
[10] www.wikipedia.com
[11] Y. Tsuno, T. Ibata, and Y. Yukawa, *Bull. Chem. Soc. Jpn.*, **32**, 960, 965, 971 (1959)..
[12] Katritzky, A.R.; Lobanov, V.S.; Karelson, M. Comprehensive Descriptors for Structural and Statistical Analysis (CODESSA) Ref. Man. Version 2.7.10, 2007.
[13] Yuan, Y.N.; Zhang, R.S.; Hu, R.J.; Ruan, X.F. Prediction of CCR5 receptor binding affinity of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas based on the heuristic method, support vector machine and projection pursuit regression. *Eur. J. Med. Chem.* **2009**, *44*, 25-34.
[14] Ma, W.P.; Luan, F.; Zhao, C.Y.; Zhang, X.Y.; Liu, M.C.; Hu, Z.D.; Fan, B.T. QSAR prediction of the penetration of drugs across a polydimethylsiloxane membrane. *QSAR Comb. Sci.* **2006**, *25*, 895-904.
[15] Luan, F.; Ma, W.P.; Zhang, X.Y.; Zhang, H.X.; Liu, M.C.; Hu, Z.D.; Fan, B.T. Quantitative structure-activity relationship models for prediction of sensory irritants (logRD(50)) of volatile organic chemicals. *Chemosphere* **2006**, *63*, 1142-1153.
[16] Fisz, J.J. Combined genetic algorithm and multiple linear regression (GA-MLR) optimizer: Application to multi-exponential fluorescence decay surface. *J. Phys. Chem. A* **2006**, *110*, 12977-12985.
[17] Word, H. *Research Papers in Statistics*; Wiley: New York, NY, USA, 1966.
[18] Jores-Kong, H.; Word, H. *Systems under Indirect Observation: Causality, structure, prediction*; North-Holland: Amsterdam, The Netherlands, 1982.
[19] Yin, P.Y.; Mohemaiti, P.; Chen, J.; Zhao, X.J.; Lu, X.; Yimiti, A.; Upur, H.; Xu, G.W. Serum metabolic profiling of abnormal savda by liquid chromatography/mass spectrometry. *J Chromatogr. B-Anal. Technol. Biomed. Life Sci.* **2008**, *871*, 322-327.
[20] Niazi, A.; Jafarian, B.; Ghasemi, J. Kinetic spectrophotometric determination of trace amounts of palladium by whole kinetic curve and a fixed time method using resazurine sulfide reaction. *Spectrochim. Acta A-Mol. Biomol. Spectrosc.* **2008**, *71*, 841-846.

[21] Niazi, A.; Amjadi, E.; Nori-Shargh, D.; Bozorghi, S.J. Simultaneous voltammetric determination of lead and tin by adsorptive differential pulse stripping method and orthogonal signal correction-partial least squares in water samples. *J. Chinese Chem. Soc.* **2008**, *55*, 276-285.

[22] Luan, F.; Liu, H.T.; Wen, Y.Y.; Zhang, X.Y. Prediction of quantitative calibration factors of some organic compounds in gas chromatography. *Analyst* **2008**, *133*, 881-887.

[23] Yap, C.W.; Li, Z.R.; Chen, Y.Z. Quantitative structure-pharmacokinetic relationships for drug clearance by using statistical learning methods. *J. Mol. Graph. Model.* **2006**, *24*, 383-395

[24] Chen, H.F. Quantitative predictions of gas chromatography retention indexes with support vector machines, radial basis neural networks and multiple linear regression. *Anal. Chim. Acta* **2008**, *609*, 24-36.

[25] Zhao, C.Y.; Zhang, H.X.; Zhang, X.Y.; Liu, M.C.; Hu, Z.D.; Fan, B.T. Application of support vector machine (SVM) for prediction toxic activity of different data sets. *Toxicology* **2006**, *217*, 105-119.

[26] Yuan, Y.N.; Zhang, R.S.; Hu, R.J.; Ruan, X.F. Prediction of volatile components retention time in blackstrap molasses by least-squares support vector machine. *QSAR Comb. Sci.* **2008**, *27*, 535-542.

[27] Si, H.Z.; Zhang, K.J.; Hu, Z.D.; Fan, B.T. QSAR model for prediction capacity factor of molecular imprinting polymer based on gene expression programming. *QSAR Comb. Sci.* **2007**, *26*, 41-50.

[28] Friedman, J.H.; Stuetzle, W. Projection Pursuit Regression. *J. Am. Stat. Assoc.* **1981**, *76*, 817- 823.

[29] Gunturi, S.B.; Archana, K.; Khandelwal, A.; Narayanan, R. Prediction of hERG Potassium Channel Blockade Using kNN-QSAR and Local Lazy Regression Methods. *QSAR Comb. Sci*, **2008**, *27*, 1305-1317.

[30] Guha, R.; Dutta, D.; Jurs, P.C.; Chen, T. Local lazy regression: Making use of the neighborhood to improve QSAR predictions. *J. Chem. Inf. Model.* **2006**, *46*, 1836-1847.

[31] Hansch, C., Maloney, P. P., Fujita, T., and Muir, R. M., *Nature* **194,** 178- 180 (1962)

[32] Hansch, C., and Fujita, T., *J. Am. Chem. Soc.* **86,** 1616-1626 (1964)

[33] Free Jr., S. M., and Wilson, J. W., *J. Med. Chem.* **7,** 395-399 (1964)