



---

## Quantitative structure activity relationship: A tool for new drug design

Shailesh G. Jawarkar\*, Madhuri D. Game

Department of Pharmaceutical Chemistry, Vidyabharati College of Pharmacy, C.K. Naidu Road, camp, Amravati-444602

---

*Received: 19-06-2018 / Revised Accepted: 28-07-2018 / Published: 01-08-2018*

---

### ABSTRACT

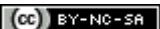
The concept of Quantitative structure activity relationship has typically been used for drug discovery and development. It has gained wide applicability for correlating molecular information with biological activities in terms of physicochemical properties. It helps to scientists in the development of reliable mathematical relationships linking chemical structures and pharmacological activity in quantitative manner of series of compound. A given compilation of data sets is then subjected to data pre-processing and data modelling through the use of statistical methods. QSAR decreases the educated guesses for synthesizing a number of compounds by facilitating the selection of the most promising candidates. The scope of this review is to highlight the QSAR studies for identification and optimization of ligands having potential to develop a new drug candidates.

**Keywords:** Drug Design, physicochemical parameters, Hansch analysis, molecular descriptors, statistical method.

---

**Address for Correspondence:** Shailesh G. Jawarkar, Department of Pharmaceutical Chemistry, Vidyabharati College of Pharmacy, C.K. Naidu Road, camp, Amravati.444602; E-mail: [sjawarkar01@rediffmail.com](mailto:sjawarkar01@rediffmail.com)

**How to Cite this Article:** Shailesh G. Jawarkar, Madhuri D. Game. Quantitative structure activity relationship: A tool for new drug design. World J Pharm Sci 2018; 6(8): 120-126.

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, which allows adapt, share and build upon the work non-commercially, as long as the author is credited and the new creations are licensed under the identical terms. 

## INTRODUCTION

Quantitative structure activity relationship (QSAR) are mathematical models that attempt to relate the structure-derived features of a compound to its biological or physicochemical activity and can be used to predict the biological activity of compounds before the actual biological testing. Thus quantitative structure activity relationship has become a tool for design of new drug. Application of computers [1-3] led the structure activity relationship from qualitative to a quantitative relationship. The most successful approach which appears involves computing molecular descriptors and collection of descriptors is then subjected to select structural building blocks to be included in creating a combinatorial library for biological evaluation.[4-5] QSAR have helped the scientists in the development of mathematical relationships linking chemical structures and pharmacological activity in quantitative manner of series of compound. The fundamental principle underlying the QSAR is that the difference in structural properties is responsible for the variations in biological activities of the compounds. QSAR certainly decreases the number of compounds to be synthesized by facilitating the selection of the most promising candidates. For design of drug requires prediction of both pharmacokinetic and pharmacodynamic properties and screening which is time consuming and expensive. These limitations would be overcome by developing a reliable model from easily measured physicochemical parameters. The mathematical and statistical analysis helps us to predict the drug activity. Thus quantitative relationship between the structure and physicochemical properties of substances and their biological activity are being used as the foundation stone in search of new medicines. This review seeks to provide a view of the different QSAR approaches employed within the current drug discovery process to construct predictive structure activity [6,7]

## PARAMETERS

**1) Lipophilic parameters:** Two parameters are commonly used to relate drug absorption and distribution with biological activity, namely the partition coefficient (P) and the lipophilic substituent Constant (p). The former parameter refers to the whole molecule whilst the latter is to pass through a number of biological membranes in order to reach its site of action. Partition coefficients were the obvious parameter to use as a measure of the movement of the drug through these membranes. The nature of the relationship obtained depends on the range of P values for the compounds used. If this range is small the results

may, be expressed as a straight line equation by the use of regression analysis having the general form:

$$\text{Log (1/C)} = k_1 \log P + k_2 \dots\dots\dots (1)$$

Where  $k_1$  and  $k_2$  are constants. This equation indicates a linear relationship between the activity of the drug and its partition coefficient.

**2) Electronic parameters:** The distribution of the electrons in a drug molecule will have an influence order to reach its target, as drug normally has to pass through a number of biological membranes. Once the drug reaches its target site, the distribution of electrons in its structure will control the type of bonds it forms with that target, which in turn affects its biological activity. The distribution of electrons within a molecule depends on the nature of the electron withdrawing and donating groups found in that structure.

### 3) Steric parameters

In order for a drug to bind effectively to its target site the dimensions of the pharmacophore of the drug must be complementary to those of the target site. The Taft steric parameter ( $E_s$ ) was the first attempt to show the relationship between a measurable parameter related to the shape and size (bulk) of a drug and the dimensions of the target site and a drug's activity. This has been followed by Charton's steric parameter, Verloop's steric parameters and the molar refractivity (MR) etc. However, in all cases the required parameter is calculated for a set of related analogues and correlated with their activity using a suitable statistical method such as regression analysis. The results of individual investigations have shown varying degrees of success in relating the biological activity to the parameter. This is probably because little is known about the finer details of the three-dimensional structures of the target sites.

## METHODS OF 2D QSAR

### 1. Free energy models

**a) Hansch Analysis:** It is also known as linear free energy (LFER) or extra thermodynamic method which assumes additive effect of various substituents in electronic, steric, hydrophobic and dispersion data in the non-covalent interaction of a drug and biomacromolecules. This method relates the biological activity within a homologous series of compounds to a set of theoretical molecular parameters which describe essential properties of the drug molecules. Hansch proposed that the action of a drug as depending on two processes.

Journey from point of entry in the body to the site of action which involves passage of series of membranes and therefore it is related to partition

coefficient log P (lipophilic) and can be explained by random walk theory.

Interaction with the receptor site which in turn depends on,

a) Bulk of substituent groups (steric)

b) Electron density on attachment group (electronic)

$$\log (1/C) = a(\log P) + b \sigma + cES + d$$

.....linear

$$\log (1/C) = a(\log P)^2 + b(\log P) + c \sigma + dES + e$$

.....nonlinear

Where a-e are constants determined for a particular biological activity by multiple regression analysis. Log P,  $\sigma$ , ES etc, are independent variables whose values are obtained directly from experiment or from tabulations<sup>[8]</sup>.

**Table 1: Molecular Descriptors used in QSAR**

Type	Descriptors	Symbol
<b>Hydrophobic Parameters</b>	i. Partition coefficient ii. Hansch's substitution constant iii. Hydrophobic fragmental constant iv. Distribution coefficient v. Apparent vi. Capacity factor in HPLC vii. Solubility parameter	log P $\pi$ $f, f'$ log D log P $\log k', \log k'W$ log S
<b>Electronic Parameters</b> <b>(A) Experimental parameter</b>	i. Hammett constant ii. Taft's inductive (polar) constant iii. Ionization constant iv. Chemical shifts: IR, NMR v. Resonance effect vi. Field effect	$\sigma, \sigma^+, \sigma^-$ $\sigma^*$ $pK_a, \Delta pK_a$ ppm R F
<b>(B) Theoretical quantum mechanical indices</b>	i. Atomic charge densities ii. Atomic net Charge iii. Super delocalizability iv. Energy of molecular orbit	$\epsilon$ $Q, QT, q^\delta \cdot Q^\delta$ $S_r^N$ $E_{LEMO}, E_{HOMO}$
<b>Steric Parameters</b>	i. Taft's steric parameter; ii. Molar volume; iii. Van der waals radius iv. Van der waals volume v. Molar refractivity; vi. Parachor vii. Sterimol	Es MV Vr Vw MR [P] L

**Table 2: Classification of descriptors based on the dimensionality of their molecular representation.**

Molecular representation	Descriptor	Example
0D	Atom count, bond counts, molecular weight, sum of atomic properties	Molecular weight, average molecular weight, number of: atoms, hydrogen atoms, carbon atoms, hetero-atoms, non-hydrogen atoms, double bonds, triple bonds, aromatic bonds, rotatable bonds, rings, 3-membered ring, 4-membered ring, 5-membered ring, 6-membered Ring.
1D	Fragments counts	Number of: primary C, secondary C, tertiary C, quaternary C, secondary carbon in ring, tertiary carbon in ring, quaternary carbon in ring, unsubstituted aromatic carbon, substituted carbon, number of H-bond donar atoms, number of H-bond acceptor atoms, unsaturation index, hydrophilic factor, molecular refractivity.

2D	Topological descriptors	Zagreb index, Wiener index, Balaban J index, connectivity indices chi ( $\chi$ ), kappa (K) shape indices
3D	Geometrical descriptors	Radius of gyration, E-state topological parameters, 3D Wiener index, 3D alaban index

## 2. Mathematical models

### a) Free Wilson Analysis

The Free-Wilson approach is truly a structure-activity based methodology because it incorporates the contribution made by various structural fragments to the overall biological activity.<sup>[9-11]</sup> Indicator variables are used to denote the presence or absence of a particular structural feature. It is represented by equation

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

Where BA is the biological activity,  $\mu$  is the overall activity,  $a_i$  is the contribution of each structural feature,  $x_i$  denotes the presence ( $x_i = 1$ ) or absence ( $x_i = 0$ ) of particular structural fragment.

### b) Fujita-Ban modification

Fujita and Ban proposed a simplified approach that solely focused on the additivity of group contribution.

$$\text{Log}A/A_0 = \sum G_i X_i$$

where A and A<sub>0</sub> represents the biological activity of the substituted and unsubstituted compounds respectively, while  $G_i$  is the activity of the substituent,  $X_i$  had the value of 1 or 0 that corresponded to the presence or absence of that substituent<sup>[12]</sup>.

**3. Statistical Methods:** A suitable statistical method coupled with a variable selection method allows analyses of this data in order to establish a QSAR model with the subset of descriptors that are most statistically significant in determining the biological activity. Following are few commonly used statistical methods:

**a) Discriminant Analysis:** The aim of discriminant analysis is to try and separate molecules into their constituent classes and finds a linear combination of factor that best discriminate between different classes. It is used to obtain a qualitative association between molecular descriptor and the biological property.

### b) Principle Component Analysis:

Principle Components Analysis (PCA) is a commonly used method for reducing the dimensionality of data set when there are significant correlations between some or all of the descriptors. PCA provides a new set of variables (the principle component) which represent most of

the information contained in the independent variables.

### c) Cluster Analysis:

Cluster analysis is the process of dividing a collection of molecules into cluster such that the objects within a cluster are highly similar whereas objects in different clusters are dissimilar. When applied to a compound dataset, the resulting clusters provide an overview of the range of structural types within the dataset and a diverse subset of compounds can be selected by choosing one or more compounds from each cluster.

### d) Combine Multivariate analysis

Combine multivariate analysis<sup>[13]</sup> is essentially an approach to quantitatively discern relationships between the independent variables and the dependent variables. The classical approach is a linear regression technique typically involving the establishment of a linear mathematical equation:

$$y = a_0 + a_1x_1 + \dots + a_nx_n$$

Where  $y$  is the dependent variable (e. g. biological/chemical property of interest).

### e) Factor Analysis (FA)

This is the combination of Factor Analysis (FA) where FA is used for initial selection of descriptors. 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.

### f) Logistic regression (LR)

LR is used to model the probability of the occurrence of some event as a linear function of a set of compounds. For example, in predicting whether an unknown compound is toxic or nontoxic.

### g) Partial Least Squares (PLS)

The basic concept of PLS regression was originally developed by Wold. In the field of QSAR, 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 analyses.<sup>[14]</sup>

**4. Quantum Mechanical Methods:** Quantum mechanical techniques are usually used to obtain accurate molecular properties such as electrostatic

potential or polarizabilities. The methods used commonly divided into three categories: semi-empirical molecular orbital theory, density functional theory (DFT) and *ab-initio* molecular orbital theory.<sup>[15]</sup>

#### a) Neural Networks (NN)

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 modeling 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.<sup>[16]</sup>

#### 3D-QSAR

Three-dimensional quantitative structure-activity relationships (3D-QSAR) involve the analysis of the quantitative relationship between the biological activity of a set of compounds and their three-dimensional properties using statistical correlation methods. 3D-QSAR uses probe-based sampling within a molecular lattice to determine three-dimensional properties of molecules (particularly steric and electrostatic values) and can then correlate these 3D descriptors with biological activity.<sup>[17]</sup>

**a) Molecular shape analysis (MSA):** Molecular shape analysis wherein matrices which include common overlap steric volume and potential energy fields between pairs of superimposed molecules were successfully correlated to the activity of series of compounds. The MSA using common volumes also provide some insight regarding the receptor-binding site shape and size.

**b) Molecular topological difference (MTD):** Simons and his coworkers developed a quantitative 3D-approach, Minimal topological difference use a hypermolecule concept for molecular alignment which correlated vertices (atoms) in the hypermolecule (a superposed set of molecules having common vertices) to activity differences in the series.

**c) Comparative molecular movement analysis (COMMA):** COMMA – a unique alignment independent approach. The 3D QSAR analysis utilizes a succinct set of descriptors that would simply characterize the three dimensional information contained in the movement descriptors of molecular mass and charge up to and inclusive of second order.

**d) Hypothetical Active Site Lattice (HASL):** Inverse grid based methodology developed in 1986-88, that allow the mathematical construction of a hypothetical active site lattice which can model

enzyme-inhibitor interaction<sup>[18]</sup> and provides predictive structure-activity relationship for a set of competitive inhibitors.

**e) Self Organizing Molecular Field Analysis (SOMFA):** SOMFA – utilizing a self-centered activity, i.e., dividing the molecule set into actives (+) and inactives (-), and a grid probe process that penetrates the overlaid molecules, the resulting steric and electrostatic potentials are mapped onto the grid points and are correlated with activity using linear regression.

**f) Comparative Molecular Field Analysis (COMFA):** The comparative molecular field analysis a grid based technique, most widely used tools for three dimensional structure-activity relationship studies. Comparative Molecular Field Analysis (CoMFA) is a mainstream and down-to earth 3D QSAR technique in the coverage of drug discovery and development. Even though CoMFA is remarkable for high predictive capacity. It's well known that the default settings in CoMFA can bring about predictive QSAR models, in the meanwhile optimized parameters was proven to provide more predictive results.<sup>[19-20]</sup>

**g) Comparative Molecular Similarity Indices (COMSIA):** COMSIA is an extension of COMFA methodology where molecular similarity indices can serve as a set of field descriptors in a novel application of 3d QSAR referred to as COMSIA.

**h) Pharmacophore modeling:** Pharmacophore modeling is powerful method to identify new potential drugs. Pharmacophore models are hypothesis on the 3D arrangement of structural properties such as hydrogen bond donor and acceptor properties, hydrophobic groups and aromatic rings of compounds that bind to the biological target.<sup>[21]</sup> The pharmacophore concept assumes that structurally diverse molecules bind to their receptor site in a similar way, with their pharmacophoric elements interacting with the same functional groups of the receptor.<sup>[22]</sup>

**4D-QSAR:** 4D-QSAR analysis incorporates conformational and alignment freedom into the development of 3D-QSAR models for training sets of structure-activity data by performing ensemble averaging, the fourth "dimension". The fourth dimension in 4-D QSAR is the possibility to represent each molecule by an ensemble of conformations, orientations and protonation states - thereby significantly reducing the bias associated with the choice of the ligand alignment. The most likely bioactive conformation/alignment is identified by the genetic algorithm.<sup>[23]</sup>

**5D-QSAR:** The fifth dimension in 5-D QSAR is the possibility to represent an ensemble of up to six different induced-fit models. The model yielding the highest predictive surrogates is selected during the simulated evolution.<sup>[24]</sup>

**6D-QSAR:** 6D-QSAR allows for the simultaneous evaluation of different solvation models. Software programme BiografX, new Unix platform combines the multi-dimensional QSAR tools Quasar, Raptor and Symposar under a single user-interface. The Macintosh version was released on March 15, 2007 and the PC/Linux version was released on September 15, 2007.

### Software for QSAR Development

With the availability of software which is relatively inexpensive, powerful and computer hardware allows for the enumeration of large virtual libraries. Application of computers led the structure activity relationship from qualitative to a quantitative relationship.<sup>[25-30]</sup>

**ChemDraw:** It is commercial software for chemical structure drawing.

**ACD/ChemSketch:** It calculation of molecular properties, 2D and 3D structure cleaning, structure naming, and prediction of logP.

**Open Babel** Open Babel is an open-source program that enables users to search, convert files, analyze or store data from molecular modeling projects.

**CORINA.** It is used for generating three-dimensional structure of small- and medium sized compounds, necessary as a pre-processing step prior to calculation of 3D molecular descriptors.

**Concord.** It is a commercial software that converts 2D inputs into 3D structures rapidly.

**ADRIANA Code.** It is one of the commercial software offered by Molecular Networks for computing molecular descriptors.

**Dragon.** version 5.5 can compute 3224 molecular descriptors which are divided into 22 blocks.

**Molconn Z.** It is commercial software for molecular descriptor calculation that works on multiple platforms.

**KNIME.** Konstanz Information Miner (KNIME) is an open-source platform with pipelining ability for data integration, processing, analysis, and exploration.

**Rapid Miner** It is an open-source system with a large collection of algorithms for data analysis and model development.

**WEKA** It has a rich compilation of modeling methods and tools for data pre-processing, classification, regression, clustering, and visualization, which are organized into different sections in the WEKA Explorer.

**Orange** It is a free program that offers tools for some simple data preparation, evaluation, visualization, classification, regression, and clustering.

**TANAGRA** It is an open-source software containing tools for data analysis, statistics, modeling, and database exploration.

**MATLAB** It is commercial software that provides an interactive system for algorithm development, data visualization, data analysis, and numeric computation with wide application in image processing, financial analysis, computational biology.

### APPLICATION OF QSAR:

- In drug discovery and environmental toxicology,
- QSAR models are now regarded as a scientifically credible tool for predicting and classifying the biological activities of untested chemicals.
- Distinguishing drug-like from non drug-like molecules
- Drug resistance
- Toxicity prediction
- Physicochemical properties prediction (e.g. water solubility, lipophilicity)
- ADME properties prediction (e.g. gastrointestinal absorption, blood brain barrier).

### CONCLUSION

The past few decades have witnessed many advances in the development of computational models for the prediction of a wide span of biological and chemical activities that are beneficial for screening promising compounds with robust properties. QSAR paradigm is based on the assumption that there is an underlying relationship between the molecular structure and biological activity. It acts as an informative tool by extracting significant patterns in descriptors related to the measured biological activity leading to understanding of mechanisms of given biological activity. This could help in suggesting design of novel molecules with a improved activity profile. It is also interesting to note that there are many paths for researchers in the field of QSAR in their quest of establishing relationships between structure and activities/properties. Such abstract nature holds the beauty of the field as there are endless possibilities in reaching the same destination of designing novel molecules with desirable properties. QSAR is thus a scientific achievement and an economic necessity to reduce empiricism in drug design to ensure that every drug synthesized and pharmacologically tested should be as meaningful.

## REFERENCES

- Hopfinger AJ. Computer-assisted drug design. *J Med Chem* 1985; 28 (9) : 1133 – 39.
- Sehgal VK et al. Computer aided drug designing. *Int J Med Dent Sci* 2017; 6(1): 1433-37.
- Kore PP et al. Computer-Aided Drug Design: An Innovative Tool for Modeling. *Open J of Med Chem* 2012; 2 : 139-48.
- Vandergraff PH et al. Multivariate quantitative structure-pharmacokinetic relationships (QSPKR) analysis of adenosine A1 receptor agonists in rat. *J Pharm Sci* 1999; 88(3): 306-12.
- Martin EJ et al. Measuring Diversity: Experimental design of combinatorial libraries for Drug Discovery. *J Med Chem* 1995; 38(9): 1431-36.
- Lowrey AH et al. Quantum chemical descriptors for linear salvation energy relationships. *Comput Chem* 1995; 19(3) : 209-15.
- Balban AT, Ivanciuc O. Historical developments of topological indices. In: Devillers J, Balban AT eds. *Topological indices and related descriptors in QSAR and QSPR*. Amsterdam: Gordon and Breach, 1999: 21-57.
- Hansch C. A quantitative approach to biochemical structure activity relationships. *Acct Chem Res* 1969; 2(8): 232-39.
- Free SM, Wilson JW. A mathematical contribution to structure-activity studies. *J Med Chem* 1964; 7: 395-99.
- Hansch C and Fujita T.  $\rho$ - $\sigma$ - $\pi$  Analysis: A method for the correlation of biological activity and chemical structure. *J Am Chem Soc* 1964; 86; 1616-26.
- Kubinyi H. Quantitative structure--activity relationships.7. The bilinear model, a new model for nonlinear dependence of biological activity on hydrophobic character. *J Med Chem*. 1977; 20(5): 625-29.
- Fujita T, Ban T. Structure activity studies of phenylethylamines as substrate of biosynthetic enzymes of sympathetic transmitters. *J Med Chem* 1971; 14(2):148-52.
- Chanin Nantasenam et al. A Practical overview of quantitative structure activity relationship. *EXCLI journal* 2009; 8: 74-88.
- Yuling An et al. Kernel-Based Partial Least Squares: Application to fingerprint-based QSAR with model visualization. *J Chem Inf Model* 2013; 53 (9): 2312–21.
- Daniel Mucs et al. The application of quantum mechanics in structure-based drug design. *J Expert Opinion on Drug Discovery* 2013; 8 : 263-76.
- Huuskonen J et al. Neural network modeling for estimation of the aqueous solubility of structurally related drugs. *J Pharm Sci* 1997; 86: 450-54.
- Jitender Verma et al. 3D-QSAR in Drug Design - A Review. *Current Topics in Med Chem* 2010; 10(1) : 95 – 115.
- Arthur M. Doweyko. The Hypothetical active lattice. An Approach to modeling active sites from data on inhibitor molecules. *J Med Chem* 1988; 31(7) : 1396-1406.
- Simon Z et al. Receptor site mapping for cardiotoxic aglicones by the minimal steric difference method. *Eur J Med Chem* 1980; 15: 521-27.
- Sandip Sen et al. CoMFA -3D QSAR approach in drug design. *Int J of Research and Development in Pharmacy and Life Sciences* 2012; 1(4) : 167-75.
- Langer T, Wolber G. Pharmacophore definition and 3D searches. *Drug Discovery Today Technol*, 2004; 1(3): 203-07.
- Jia Fei et al. Pharmacophore modeling, virtual Screening and molecular docking Studies for Discovery of Novel Akt2 Inhibitors. *Int J of Med Sci*. 2013; 10(3) : 265-75.
- Vedani A et al. Multiple conformation and protonation-state representation in 4D-QSAR: The neurokinin-1 receptor system. *J Med Chem*, 2000; 43: 4416–27.
- Vedani A, Dobler M. 5D-QSAR: The key for simulating induced fit. *J Med Chem* 2002; 45(11): 2139–49.
- Sild S et al. Open computing grid for molecular science and engineering. *J Chem Inf Model* 2006; 46: 953–59.
- Pearlman RS, Smith KM. Novel software tools for chemical diversity. *Perspect Drug Discovery Des* 1998; 9 : 339-53.
- Willett P. Chemoinformatics-similarity and diversity in chemical libraries. *Curr. Opin. Drug Discovery dev*. 2000; 11: 85-88.
- Glen RC, Rose VS. A computer programme suite for the calculation storage and manipulation of molecular property and activity descriptors. *J Mol Graph*.1987; 5:79-86.
- Hyde RM, Livingstone DJ. Perspective in QSAR computer chemistry and pattern recognition. *J Comput Aid Mol Design* 1988; 2: 145-55.
- Glen RC and Rose VS. A computer programme suite for the calculation storage and manipulation of molecular property and activity descriptors. *J Mol Graph* 1987; 5: 79-86.