Introduction to QSAR

	QSAR
	Introduction to QSAR
	Objectives of QSAR
	Historical development of QSAR
	Molecular descriptors used in QSAR
	Methods of QSAR
	2D QSAR methods
	Introduction to genetic function approximation
	Advances in QSAR
	3D-QSAR
	4D-QSAR
	5D-QSAR
	6D-QSAR

INTRODUCTION TO QSAR

QSAR

Most molecular discoveries today are the results of an iterative, three-phase cycle of design, synthesis and test. Analysis of the results from one iteration provides information and knowledge that enables the next cycle of discovery to be initiated and further improvement to be achieved. A common feature of this analysis stage is the construction of some form of model which enables the observed activity or properties to be related to the molecular structure. Such models are often referred to as Quantitative Structure Activity Relationships.[240]

Quantitative structure-activity relationships (QSARs) studies unquestionably are of great importance in modern chemistry and biochemistry. The concept of QSAR is to transform searches for compounds with desired properties using chemical intuition and experience into a mathematically quantified and computerized form. QSAR methods are characterized by two assumptions with respect to the relationship between chemical structure and the biological potency of compounds. The first is that one can derive a quantitative measure from the structural properties significant to the biological activity of a compound. The properties assumed to be physicochemical such as partition coefficient or sub structural such as presence or absence of certain chemical features. The other assumption is that one can mathematically describe the relationship between biological property one wishes to optimize and the molecular property calculated from the structure.[241] QSAR's general mathematical form is represented by the following equation.

Biological Activity = f (Physicochemical Property)

Objective of QSAR

QSAR attempts to correlate structural, chemical, statistical and physical properties with biological activity by various approaches.[242] QSAR models are scientific credible tools for predicting and classifying biological activities of untested chemicals. QSAR is an essential tool for lead development (optimization), a growing trend is to use QSAR early in drug discovery process as a screening and enrichment tool to eliminate from further development those chemicals lacking "drug like" properties or those chemicals predicted to elicit a toxic response.[243]

Historical Development of QSAR

Over the past two decades, the center of gravity (the intellectual focus) of medicinal chemistry has shifted dramatically from, how to make a molecule, to what molecule to make.[244] The challenge now is the gathering of information to make decisions regarding the use of resources in drug design. The information feeding the drug design effort is increasingly quantitative, building upon recent developments in molecular structure description, combinatorial mathematics, statistics, and computer simulations. Collectively these areas have led to a new paradigm in drug design which has been referred to as QUANTITATIVE STRUCTURE ACTIVITY RELATIONSHIP (QSAR). It has been nearly 40 years since the QSAR paradigm first found its way into the practice of pharmaceutical chemistry.[245]

Crum-Brown and Fraser[246] published equation 1.1 in 1868, which is considered to be the first formulation of a QSAR: the "physiological activity" (Φ) was expressed as a function of the chemical structure C.

$$\Phi = f(C) \qquad (1.1)$$

A few decades later Richet[247], Meyer[248] and Overton[249] independently found linear relationship between lipophilicity expressed as solubility or oil-water partition coefficient and biological effects, like toxicity and narcotic activity.[250] In 1930's, L. Hammett correlated electronic properties of organic acids and bases with their equilibrium constants and reactivity.[251,252] Taft devised a way for separating polar, steric, and resonance effects and introducing the first steric parameter, E_S.[253] The contributions of Hammett and Taft together laid the mechanistic basis for the development of the QSAR paradigm by Hansch and Fujita.[254] They combined hydrophobic constants with Hammett's electronic constants to yield the linear Hansch equation and its many extended forms.[255]

$\text{Log } 1/C = a\sigma + b\pi + ck$ …… Linear form (1.2)

$\text{Log } 1/C = a \log P - b (\log P)^2 + c\sigma + k$ …… Non linear form (1.3)

Where,

C - Concentration required to produce a standard response

Log P - partition coefficient between 1-octanol and water

σ - Hammet substituent parameter

π - Relative hydrophobicity of substituents

a, b, c, k - Model co-efficient

Besides the Hansch approach, other methodologies were also developed to tackle structure activity questions. The Free-Wilson approach addresses structure activity studies in a congeneric series as described in Equation (1.4).[256]

$$BA = \sum a_i x_i + u \quad (1.4)$$

Where BA is the biological activity, u is the average contribution of the parent molecule, and a_i is the contribution of each structural feature; x_i denotes the presence $x_i = 1$ or absence $x_i = 0$ of a particular structural fragment. Limitations in this approach led to the more sophisticated Fujita-Ban equation that used the logarithm of activity, which brought the activity parameter in line with other free energy-related terms.[257]

$$\text{Log } BA = \sum G_i X_i + u \quad (1.5)$$

u is defined as the calculated biological activity value of the unsubstituted parent compound of a particular series. G_i represents the biological activity contribution of the substituents, whereas X_i is ascribed with a value of one when the substituent is present or zero when it is absent. Variations on this activity based approach have been extended by Klopman et. al.[258] and Enslein et al.[259] Topological methods have also been used to address the relationships between molecular structure and biological activity. The Minimum Topological Difference (MTD) method of Simon and the extensive studies on molecular connectivity by Kier and Hall have contributed to the development of quantitative structure property/activity relationships.[260,261] Recently, these electro topological indices that encode significant structural information on the topological state of atoms and fragments as well as their valence electron content have been applied to biological and toxicity data.[262] Other recent developments in QSAR include approaches such as HQSAR (Hologram QSAR), Inverse QSAR, and Binary QSAR.[263-266]

Molecular Descriptors used in QSAR

Molecular descriptors can be defined as a numerical representation of chemical information encoded within a molecular structure via mathematical procedure.[267] This mathematical representation has to be invariant to the molecule's size and number of atoms to allow model building with statistical methods.

The information content of structure descriptors depends on two major factors:

(1) The molecular representation of compounds.

(2) The algorithm which is used for the calculation of the descriptor.[268]

The three major types of parameters initially suggested are,

(1) Hydrophobic

(2) Electronic

(3) Steric

Table 11: Molecular Descriptors used in QSAR

Type	Descriptors
Hydrophobic Parameters	Partition coefficient ; log P
	Hansch's substitution constant; π
	Hydrophobic fragmental constant; f, f'
	Distribution coefficient; log D
	Apparent log P
	Capacity factor in HPLC; log k' , log k'_W
	Solubility parameter; log S
Electronic Parameters	Hammett constant; σ, σ^+, σ^-
	Taft's inductive (polar) constant; σ*
	Swain and Lupton field parameter
	Ionization constant; pK_a , ΔpK_a
	Chemical shifts: IR, NMR
Steric Parameters	Taft's steric parameter; Es
	Molar volume; MV
	Van der waals radius
	Van der waals volume
	Molar refractivity; MR
	Parachor
	Sterimol
Quantum chemical descriptors	Atomic net charge; Q^σ, Q^π
	Superdelocalizability
	Energy of highest occupied molecular orbital; E_{HOMO}
	Energy of lowest unoccupied molecular orbital; E_{LUMO}
Spatial Descriptor	Jurs descriptors, Shadow indices, Radius of Gyration, Principle moment of inertia

Table 12: 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 (χ), kappa (K) shape indices
3D	Geometrical descriptors	Radius of gyration, E-state topological parameters, 3D Wiener index, 3D Balaban index

Methods of QSAR

Many different approaches to QSAR have been developed since Hansch's seminal works. QSAR methods can be analyzed from two view points:

(1) The types of structural parameters that are used to characterize molecular identities starting from different representation of molecules, from simple chemical formulas to 3D conformations.

(2) The mathematical procedure that is employed to obtain the quantitative relationship between these structural parameters and biological activity.[269]

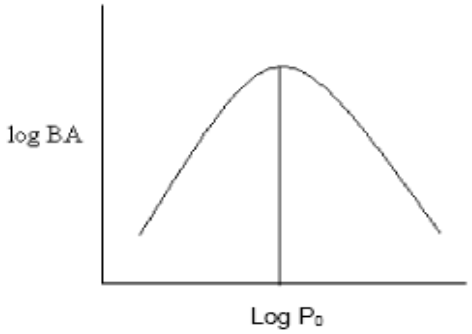
8.5.1 2D QSAR Methods

1. Free energy models
    a) Hansch analysis (Linear Free Energy Relationship, LFER)
2. Mathematical models
    a) Free Wilson analysis
    b) Fujita-Ban modification
3. Other statistical methods
    a) Discriminant Analysis (DA)
    b) Principle Component Analysis (PCA)
    c) Cluster Analysis (CA)
    d) Combine Multivariate Analysis (CMA)
    e) Factor Analysis (FA)
4. Pattern recognition
5. Topological methods
6. Quantum mechanical methods

❖ Hansch Analysis

In 1969, Corwin Hansch extends the concept of linear free energy relationships (LFER) to describe the effectiveness of a biologically active molecule. It is one of the most promising approaches to the quantification of the interaction of drug molecules with biological system. 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 biomacro molecules. 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.

1. 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 co-efficient log P (lipophilic) and can be explained by random walk theory.



2. Interaction with the receptor site which in turn depends on,
    a) Bulk of substituent groups (steric)
    b) Electron density on attachment group (electronic)

He suggested linear and non-linear dependence of biological activity on different parameters.

$$\log (1/C) = a(\log P) + b\,\sigma + cE_S + d \ldots\ldots\ldots\ldots\ldots\ldots\text{linear}$$

$$\log (1/C) = a(\log P)^2 + b(\log P) + c\,\sigma + dE_S + e \ldots\ldots\ldots\text{nonlinear}$$

Where a-e are constants determined for a particular biological activity by multiple regression analysis. Log P, σ, E_S etc, are independent variables whose values are obtained directly from experiment or from tabulations. Other parameters than those shown may also be included. If there are 'n' independent variables to be considered, then there are 2^n-1 combinations of these variables that may be used to best explain the tabulated data.[270]

❖ 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.[271-273] Indicator variables are used to denote the presence or absence of a particular structural feature. It is represented by equation 1.6.

$$BA = \Sigma a_i x_i + \mu \qquad (1.6)$$

Where BA is the biological activity, μ 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. This mathematical model incorporated symmetry equation to minimize linear dependence between variables. This approach was easy to apply; it had its drawbacks, mostly centered on the large number of parameters and subsequent loss of the statistical degree of freedom. In 1971, in an attempt to

deal with limitations of this approach, Fujita and Ban proposed a simplified approach that solely focused on the additivity of group contribution.

$$\text{Log}A/A_0 = \Sigma\, G_i X_i \qquad (1.7)$$

where A and A_0 represents the biological activity of the substituted and unsubstituted compounds respectively, while G_i is the activity of the ith sustituent, X_i had the value of 1 or 0 that corresponded to the presence or absence of that substituent.[274]

The delineation of these models led to explosive development in QSAR analysis and related approaches. The Kubinyi bilinear model is refinement of the parabolic model, and in many cases, it has proved to be superior, it is represented by equation 1.8.

$$\text{Log } 1/c = a \log P - b \log (\beta \cdot P+1) + K \qquad (1.8)$$

❖ Statistical Methods

Statistical methods are the mathematical foundation for the development of QSAR models. The application of multivariate analysis, data description, classification, and regression modeling, are combined with the ultimate goal of interpretation and prediction of non-evaluated or non-synthesized compounds.[275]

Discriminant Analysis:

The aim of discriminant analysis is to try and separate molecules into their constituent classes. Discriminant analysis finds a linear combination of factor that best discriminate between different classes. Linear discriminant analysis was used for the analysis rather than multiple linear regressions since the biological activity data were not on a continuous scale of activity but rather were classified into two groups: active and inactive.[240] It is used to obtain a qualitative association between molecular descriptor and the biological property.[240]

Cluster Analysis:

Cluster analysis is the process of dividing a collection of objects (molecules) into groups (or 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. Clustering methods can be used to select diverse subset of compounds from larger dataset. The clustering methods most widely applied to compound selection include k-means clustering, non-hierarchiral clustering and hierarchial clustering.[276]

Principle Component Analysis:

The dimensionality of a data set is the number of variables that are used to describe each object. 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.[240] PCA provides a new set of variables (the principle component) which represent most of the information contained in the independent variables.

Quantum Mechanical Methods:

Quantum mechanical techniques are usually used to obtain accurate molecular properties such as electrostatic potential or polarizabilities, which are only available with much lower resolution from classical mechanical techniques or those (ionization potential or electron affinities, etc.) that can be obtained only quantum mechanically. The methods used commonly divided into three categories: semi-empirical molecular orbital theory, density functional theory (DFT) and *ab-initio* molecular orbital theory.[277] Quantum chemical methods can be applied to quantitative structure-activity relationship by direct derivation of electronic descriptors from molecular wave function.

There is no single method that works best for all problems. Besides above mentioned methods, statistical modeling techniques aims to develop correlation models between independent variables (molecular descriptors) and dependent variable (biological property) which include simple linear regression, multiple linear regression, principle component regression, partial least squares (PLS) regression, genetic function approximation(GFA) and genetic partial least squares (G/PLS) techniques.

8.6 Introduction to Genetic Function Approximation

GFA algorithm offers a new approach to build structure-activity models. It automates the search for QSAR models by combining genetic algorithm and statistical modeling tools.[277] GFA algorithm was developed by Dr. David Rogers. GFA is genetic based method which combines Holland's genetic algorithm and Friedman's multivariate adaptive regression spines (MARS). Application of GFA algorithm may allow the construction of higher quality predictive models and make available additional information not provided by standard regression techniques.

Genetic algorithm is derived from analogy with the evolution of DNA. In this analogy, individuals are represented by a one dimensional string of bits. An initial population is created of individuals, usually with random initial bits. A fitness function is used to estimate the quality of individuals, so that the 'best' individual receives the best fitness score. Individuals with the best fitness score are more likely to be chosen for mating and to propagate their genetic material to offspring through the crossover operation in which piece of genetic material is taken from each parents and recombined to create child. After many mating steps, the average fitness of individuals in the population increases as 'good' combination of genes are discovered and spread through the population.

The GFA algorithm accomplishes the breeding of the best equations and elimination of the poorer equation by genetic algorithm. The genetic algorithm cuts and separates individual equations and recombines the fragments to form new equations. The genetic algorithm uses Friedman's Lack-of-fit (LOF) Measure to select equations for breeding and survival. Use of the LOF measure drives the population towards more parsimonious, simple model and avoid over-fitting of the data. As generations of equations are bred and mutated, the population evolves to a series of ever-increasing quality of equations.

$$LOF = LSE/ \{1 – (c + dp)/m\}^2$$

Where LSE – least square error

c - number of basis functions in the model

d - smoothing parameter

p - number of descriptors

m - number of observations in the training set

The smoothing parameter d controls the scoring bias between equations of different sizes.

The GFA method does not require the assumption that the relationship between independent and dependent variables operates over the entire variable range. GFA circumvents the need for this assumption by using spline-based terms for the construction of its regression equation.

The GFA algorithm approach has number of important advantages over the other techniques: it builds multiple models rather than a single model. It automatically selects which descriptors are to be used in basis functions and determine appropriate number of basis functions to be used by testing full-size model rather than incrementally building them; it is better at discovering combination of basis that take advantage of correlations between features; it incorporates the LOF (lack of fit) error measure that resists over-fitting. GFA can build model using not only linear polynomial, but also higher-order polynomials, spline and gaussians.[278,279]

Advances in QSAR

QSARs attempt to relate physical and chemical properties of molecules to their biological activities by simply using easily calculable descriptors and simple statistical methods like Multiple Linear Regression (MLR) to build a model which both describes the activity of the data set and can predict activities for further sets of untested compounds.

These type of descriptors often fail to take into account the three-dimensional nature of chemical structures which obviously play a part in ligand-receptor binding, and hence activity. Steric, hydrophobic and electrostatic interactions are crucial to whether a molecule will interact optimally at its active site. It is logical to model these potential interactions to find the location in space around the molecule that are both acceptable and forbidden. The preceding QSAR methods usually do not take into account the 3-D structure of the molecules or their targets such as enzymes and receptors. So, efforts have been made to explore structure-activity studies of ligands that take into account the known X-ray structures of proteins and enzymes, as well as the interaction of drugs with models of their receptors. Following are some of advanced approaches to QSAR methodology.

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.

1. 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.[280]

2. Molecular topological difference (MTD)

Simons and his coworkers developed [281] a quantitative 3D-approach, the minimal steric (topologic) difference 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.[282-285]

3. 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.[286]

4. 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 and provides predictive structure-activity relationship for a set of competitive inhibitors. Computer-assisted molecule to molecule match which makes the use of multidimensional representation of inhibitor molecules. The result of such

matching are used to construct a hypothetical active site by means of a lattice of points which is capable of modeling enzyme-inhibitor interactions.[287]

5. 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.[280]

6. 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 was introduced in 1988, is based on the assumption that since, in most cases, the drug-receptor interactions are noncovalent, the changes in biological activities or binding affinities of sample compound correlate with changes in the steric and electrostatic fields of these molecules. These field values are correlated with biological activities by partial least square (PLS) analysis.[286]

7. 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.[280]

3D 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.[288] 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.[289]

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.[290]

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.

6D-QSAR

6D-QSAR allows for the simultaneous evaluation of different solvation models. Software programme Biograf**X**, 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.[292]