Journal of Advances and Scholarly Researches in Allied Education
Vol. XIV, Issue No. 2, January-2018, ISSN 2230-7540

An Analysis upon Historical Perspective and Development of QSAR Modeling: A Practical Overview

Harish Kumar Soni[1]* Dr. Praveen Kumar[2]

[1] Research Scholar of OPJS University, Churu, Rajasthan

[2] Associate Professor, OPJS University, Churu, Rajasthan

Abstract – Quantitative structure-activity relationship (QSAR) modeling pertains to the construction of predictive models of biological activities as a function of structural and molecular information of a compound library. The concept of QSAR has typically been used for drug discovery and development and has gained wide applicability for correlating molecular information with not only biological activities but also with other physicochemical properties, which has therefore been termed quantitative structure-property relationship (QSPR). Typical molecular parameters that are used to account for electronic properties, hydrophobicity, steric effects, and topology can be determined empirically through experimentation or theoretically via computational chemistry. A given compilation of data sets is then subjected to data pre-processing and data modeling through the use of statistical and/or machine learning techniques. QSAR enables the investigator to establishes a reliable quantitative relationship between structure and activity which will be used to derive an insilico model to predict the activity of novel molecules prior to their synthesis. The past few decades have witnessed much 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. This review covers the concept, history of QSAR and also the components involved in the development of QSAR models.

----------X----------

INTRODUCTION

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.

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.

Quantitative structure-activity relationship (QSAR) and quantitative structureproperty relationship (QSPR) makes it possible to predict the activities/properties of a given compound as a function of its molecular substituent. Essentially, new and untested compounds possessing similar molecular features as compounds used in the development of QSAR/QSPR models are likewise assumed to also possess similar activities/properties. Several successful QSAR/QSPR models have been published over the years which encompass a wide span of biological and physicochemical properties. QSAR/QSPR has great potential for modeling and designing novel compounds with robust properties by being able to forecast physicochemical properties as a function of structural features. The popularity of QSAR/QSPR has seen exponential growth as illustrated by a literature search in Scopus for research articles with QSAR, QSPR, structure-

activity relationship, and structure-property relationship as keywords.

It has been nearly 40 years since the quantitative structure-activity relationship (QSAR) paradigm Þrst found its way into the practice of agrochemistry, pharmaceutical chemistry, toxicology, and eventually most facets of chemistry. Its staying power may be attributed to the strength of its initial postulate that activity was a function of structure as described by electronic attributes, hydrophobicity, and steric properties as well as the rapid and extensive development in methodologies and computational techniques that have ensued to delineate and reÞne the many variables and approaches that deÞne the paradigm.

The overall goals of QSAR retain their original essence and remain focused on the predictive ability of the approach and its receptiveness to mechanistic interpretation. Rigorous analysis and Þne-tuning of independent variables has led to an expansion in development of molecular and atom-based descriptors, as well as descriptors derived from quantum chemical calculations and spectroscopy. The improvement in high-throughput screening procedures allows for rapid screening of large numbers of compounds under similar test conditions and thus minimizes the risk of combining variable test data from many sources.

The formulation of thousands of equations using QSAR methodology attests to a validation of its concepts and its utility in the elucidation of the mechanism of action of drugs at the molecular level and a more complete understanding of physicochemical phenomena such as hydrophobicity. It is now possible not only to develop a model for a system but also to compare models from a biological database and to draw analogies with models from a physical organic database. This process is dubbed *model mining* and it provides a sophisticated approach to the study of chemical-biological interactions. QSAR has clearly matured, although it still has a way to go. The previous review by Kubinyi has relevant sections covering portions of this paper as well as an extensive bibliography recommended for a more complete overview.

Quantitative structure – activity relationship (QSAR) modeling pertains to the construction of predictive models of biological activities as a function of structural and molecular information of a compound library. The concept of QSAR has typically been used for drug discovery and development and has gained wide application for correlating molecular information with not only biological activities but also with other physicochemical properties, which has therefore been termed quantitative structure – property relationship (QSPR). QSAR is widely accepted predictive and diagnostic process used for finding associations between chemical structures and biological activity. QSAR has emerged and has evolved trying to fulfill the medicinal chemist's need and desire to predict biological response.1 It found its way into the practice of agro chemistry, pharmaceutical chemistry, and eventually most facets of chemistry.2

QSAR is the final result of computational processes that start with a suitable description of molecular structure and ends with some inference, hypothesis, and predictions on the behavior of molecules in environmental, physicochemical and biological system under analysis. The final outputs of QSAR computations are set of mathematical equations relating chemical structure to biological activity. Multivariate QSAR analysis employs all the molecular descriptors from various representations of a molecule (1D, 2D and 3D representation) to compute a model, in a search for the best descriptors valid for the property in analysis. This review covers the concepts, history and the steps involved in the development of QSAR models.

BRIEF HISTORY OF QSAR

QSAR has its origins in the field of toxicology whereby Cros in 1863 proposed a relationship which existed between the toxicity of primary aliphatic alcohols with their water solubility. Likewise, Crum-Brown and Fraser postulated the linkage between chemical constitution and physiological action in their pioneering investigation in 1868 as follows:

"performing upon a substance a chemical operation which shall introduce a known change into its constitution, and then examining and comparing the physiological action of the substance before and after the change"

Shortly after, Richet (1893), Meyer (1899), and Overton (1901) separately discovered a linear correlation between lipophilicity (e. g. oil-water partition coefficients) and biological effects (e. g. narcotic effects and toxicity). By 1935, Hammett (1935, 1937) introduced a method to account for substituent effects on reaction mechanisms through the use of an equation which took two parameters into consideration namely the (i) substituent constant and the (ii) reaction constant.

Complementing the Hammett's model, Taft proposed in 1956 an approach for separating polar, steric, and resonance effects of substituents in aliphatic compounds (Taft, 1956). The contributions from Hammett and Taft set forth the mechanistic basis for QSAR/QSPR development by Hansch and Fujita (1964) in their seminal development of the linear Hansch equation which integrated hydrophobic parameters with Hammett's electronic constants. An insightful account on the development of QSAR/QSPR can be found in the excellent book by Hansch and Leo (1995).

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. 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. Crum-Brown and Fraser246 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.

$$\Phi = f(C) \quad (1)$$

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

$$\text{Log } 1/C = a\sigma + b\pi + ck \quad \text{..... Linear form} \quad (2)$$

$$\text{Log } 1/C = a \log P - b (\log P)^2 + c\sigma + k \quad \text{..... Non linear form} \quad (3)$$

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 (4).

$$BA = \sum a_i x_i + u \quad (4)$$

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; xi 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.

$$\text{Log } BA = \sum G_i X_i + u \quad (5)$$

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. and Enslein et al. 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.

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.

METHODS OF QSAR

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

(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.

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

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.

1. 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.

2. 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.

3. 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.

4. 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.

TOOLS AND TECHNIQUES OF QSAR

Biological Parameters-

In QSAR analysis, it is imperative that the biological data be both accurate and precise to develop a meaningful model. It must be realized that any resulting QSAR model that is developed is only as valid statistically as the data that led to its development. The equilibrium constants and rate constants that are used extensively in physical organic chemistry and medicinal chemistry are related to free energy values D*G*. Thus for use in QSAR, standard biological equilibrium constants such as *K*i or *K*m should be used in QSAR studies.

Likewise only standard rate constants should be deemed appropriate for a QSAR analysis. Percentage activities (e.g., % inhibition of growth at certain concentrations) are *not* appropriate

biological endpoints because of the nonlinear characteristic of dose-response relationships.

These types of endpoints may be transformed to equieffective molar doses. Only equilibrium and rate constants pass muster in terms of the free-energy relationships or inßuence on QSAR studies. Biological data are usually expressed on a logarithmic scale because of the linear relationship between response and log dose in the midregion of the log dose-response curve. Inverse logarithms for activity (log 1/*C*) are used so that higher values are obtained for more effective analogs. Various types of biological data have been used in QSAR analysis.

Biological data should pertain to an aspect of biological/biochemical function that can be measured. The events could be occurring in enzymes, isolated or bound receptors, in cellular systems, or whole animals. Because there is considerable variation in biological responses, test samples should be run in duplicate or preferably triplicate, except in whole animal studies where assay conditions (e.g., plasma concentrations of a drug) preclude such measurements.

Statistical Methods: Linear Regression Analysis-

The most widely used mathematical technique in QSAR analysis is multiple regression analysis (MRA). We will consider some of the basic tenets of this approach to gain a Þrm understanding of the statistical procedures that deÞne a QSAR. Regression analysis is a powerful means for establishing a correlation between independent variables and a dependent variable such as biological activity.

Compound Selection-

In setting up to run a QSAR analysis, compound selection is an important angle that needs to be addressed. One of the earliest manual methods was an approach devised by Craig, which involves two-dimensional plots of important physicochemical properties. Care is taken to select substituents from all four quadrants of the plot. The Topliss operational scheme allows one to start with two compounds and construct a potency tree that grows branches as the substituent set is expanded in a stepwise fashion. Topliss later proposed a batchwise scheme including certain substituents such as the 3,4-Cl2, 4-Cl, 4-CH3, 4-OCH3, and 4-H analogs (65). Other methods of manual substituent selection include the Fibonacci search method, sequential simplex strategy, and parameter focusing by Magee.

QUANTITATIVE MODELS

Linear Models-

The correlation of biological activity with physicochemical properties is often termed an *extra thermodynamic relationship*. Because it follows in the line of Hammett and Taft equations that correlate thermodynamic and related parameters, it is appropriately labeled. The Hammett equation represents relationships between the logarithms of rate or equilibrium constants and substituent constants. The linearity of many of these relationships led to their designation as linear free energy relationships. The Hansch approach represents an extension of the Hammett equation from physical organic systems to a biological milieu. It should be noted that the simplicity of the approach belies the tremendous complexity of the intermolecular interactions at play in the overall biological response.

Biological systems are a complex mix of heterogeneous phases. Drug molecules usually traverse many of these phases to get from the site of administration to the eventual site of action. Along this random-walk process, they perturb many other cellular components such as organelles, lipids, proteins, and so forth. These interactions are complex and vastly different from organic reactions in test tubes, even though the eventual interaction with a receptor may be chemical or physicochemical in nature. Thus, depending on the biological system involved isolated receptor, cell, or whole animalÑone expects the response to be multifactorial and complex. The overall process, particularly *in vitro* or *in vivo*, studies a mix of equilibrium and rate processes, a situation that deÞes easy separation and delineation.

Nonlinear Models-

Extensive studies on development of linear models led Hansch and coworkers to note that a breakdown in the linear relationship occurred when a greater range in hydrophobicity was assessed with particular emphasis placed on test molecules at extreme ends of the hydrophobicity range. Thus, Hansch et al. suggested that the compounds could be involved in a ÒrandomwalkÓ process: low hydrophobic molecules had a tendency to remain in the Þrst aqueous compartment, whereas highly hydrophobic analogs sequestered in the Þrst lipoidal phase that they encountered. This led to the formulation of a parabolic equation, relating biological activity and hydrophobicity.

CONCLUSION

The past few decades have witnessed much 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. In this review article, we have provided a brief introduction to the concepts of QSAR along with examples from our previous investigations on diverse biological and chemical systems. It should be noted that the applicability of QSAR models are only useful in the domains that they were trained and validated. As such, QSAR models spanning wider

domains of molecular diversity have the benefit of being valid for wider spans of molecules. It is also interesting to note that there are many paths for researchers in the field of QSAR/QSPR 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 has done much to enhance our understanding of fundamental processes and phenomena in medicinal chemistry and drug design. QSAR has matured over the last few decades in terms of the descriptors, models, methods of analysis, and choice of substituents and compounds. Embarking on a QSAR project may be a daunting and confusing task to a novice. However, there are many excellent reviews and tomes on this subject that can aid in the elucidation of the paradigm. Dealing with biological systems is not a simple problem and in attempting to develop a QSAR, one must always be cognizant of the biochemistry of the system analyzed and the limitations of the approach used.

REFERENCES

1. Angeli C., Bak K.L., Bakken V., et. al. (2005). DALTON, a molecular electronic structure program. Release 2.0.

2. C. Hansch and A. Leo in S. R. Heller, Ed. (2005). Exploring QSAR. Fundamentals and Applications in Chemistry and Biology, American Chemical Society, Washington, DC.

3. Furusjö E, Svenson A, Rahmberg M, Andersson M. (2006). The importance of outlier detection and training set selection for reliable environmental QSAR predictions. Chemosphere; 63: pp. 99-108.

4. Golbraikh A., Shen M., Xiao Z., et. al. (2003). Rational selection of training and test sets for the development of validated QSAR models. J Comput Aided Mol Des.; 17(2 4): pp. 241–253.

5. H. Kubinyi in M. Wolff, Ed. (2005). BurgerÕs Medicinaln Chemistry and Drug Discovery, Volume 1: Principles and Practice, John Wiley & Sons, New York, p. 497.

6. Hansch C, Leo A. (2005). Exploring QSAR. Washington, DC: American Chemical Society.

7. J. M. Blaney and C. Hansch in C. A. Ramsden, Ed. (2000). Comprehensive Medicinal Chemistry. The Rational Design, Mechanistic Study and Therapeutic Application of Chemical Compounds, Vol. 4, Quantitative Drug Design, Pergamon, Elmsford, NY, p. 459.

8. J.W. Lown in S. Neidle and M. J. Waring, Eds. (2003). Molecular Aspects of Anticancer Drug-DNA Interactions, Macmillan, Basinstoke, UK, p. 322.

9. Kar S, Roy K. (2012). QSAR of phytochemicals for the design of better drugs. Expert Opin Drug Discov.; 7(10): pp. 877–902.

10. Karelson M., Lobanov V.S., Katritzky A.R. (1996). Quantum-chemical descriptors in QSAR/ QSPR studies. Chem Rev; 96: pp. 1027- 44.

11. Kim K. (2007a). Outliers in SAR and QSAR: 2. Is a flexible binding site a possible source of outliers? J Comput Aid Mol Des; 21: pp. 421-35.

12. L.B. Kier and L. H. Hall (2009). Molecular Structure Description. The Electrotopological State, Academic Press, San Diego, CA.

13. Roy K., Kar S., Das R.N. (2015). Background of QSAR and Historical Developments. In: KRKN Das, editor. Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment. Boston: Academic Press; pp. 1–46.

14. Verma R.P. & Hansch C. (2005). An approach toward the problem of outliers in QSAR. Bioorg Med Chemistry; 13: pp. 4597-621.

Corresponding Author

Harish Kumar Soni*

Research Scholar of OPJS University, Churu, Rajasthan