Med Chem Res (2014) 23:4991–5007
DOI 10.1007/s00044-014-1072-3

MEDICINAL CHEMISTRY RESEARCH

REVIEW ARTICLE

Quantitative structure–activity relationship (QSAR) studies as strategic approach in drug discovery

Harun M. Patel · Malleshappa N. Noolvi · Poonam Sharma · Varun Jaiswal · Sumit Bansal · Sandeep Lohan · Suthar Sharad Kumar · Vikrant Abbot · Saurabh Dhiman · Varun Bhardwaj

Received: 13 November 2013 / Accepted: 7 June 2014 / Published online: 26 June 2014

Abstract Drug design is a process which is driven by technological breakthroughs implying advanced experimental and computational methods. Nowadays, the techniques or the drug design methods are of paramount importance for prediction of biological profile, identification of hits, generation of leads, and moreover to accelerate the optimization of leads into drug candidates. Quantitative structure–activity relationship (QSAR) has served as a valuable predictive tool in the design of pharmaceuticals and agrochemicals. From decades to recent research, QSAR methods have been applied in the development of relationship between properties of chemical substances and their biological activities to obtain a reliable statistical model for prediction of the activities of new chemical entities. Classical QSAR studies include ligands with their binding sites, inhibition constants, rate constants, and other biological end points, in addition molecular to properties such as lipophilicity, polarizability, electronic, and steric properties or with certain structural features. 3D-QSAR has emerged as a natural extension to the classical Hansch and Free–Wilson approaches, which exploit the three-dimensional properties of the ligands to predict their biological activities using robust chemometric techniques such as PLS, G/PLS, and ANN. This paper provides an overview of 1-6 dimension-based developed QSAR methods and their approaches. In particular, we present various dimensional QSAR approaches, such as comparative molecular field analysis (CoMFA), comparative molecular similarity analysis, Topomer CoMFA, self-organizing molecular field analysis, comparative molecule/pseudo receptor interaction analysis, comparative molecular active site analysis, and FLUFF-BALL, 4D-QSAR, and G-QSAR approaches.

Keywords Drug design · QSAR · Methodology

H. M. Patel
Department of Pharmaceutical Chemistry, R.C. Patel Institute of Pharmaceutical Education and Research, Dhule, Shirpur 425405, Maharashtra, India

M. N. Noolvi
Department of Pharmaceutical Chemistry, Shree Dhanvantary Pharmacy College, Kim, Surat 394110, Gujarat, India

P. Sharma · S. Bansal · S. Lohan · S. S. Kumar · V. Abbot · S. Dhiman · V. Bhardwaj (✉)
Department of Biotechnology, Bioinformatics and Pharmacy, Jaypee University of Information Technology, Solan, Waknaghat 173234, Himachal Pradesh, India
e-mail: varunmilton@yahoo.com

V. Jaiswal
Department of Bioinformatics, Shoolini University, Solan, Himachal Pradesh, India

Introduction

Identification of promising hits and generation of high-quality leads are crucial steps in the early stages of drug discovery processes (Zhao, 2007; Lombardino and Lowe, 2004). Drug discovery is currently driven by the innovation and knowledge employing a combination of experimental and computational methods. Knowledge of the structure and function of the targets as well as the mechanism by which it interacts with potential drugs is fundamental to this approach (Guido *et al*., 2008). Drug design is an iterative process which begins with a compound that displays an interesting biological profile and ends with optimizing both the activity profile for the molecule and its chemical synthesis. The process is initiated when the chemist considers a hypothesis which relates the chemical features of the molecule or series of molecules to the biological

activity. Without a detailed understanding of the biochemical processes responsible for activity, the hypothesis generally is refined by examining structural similarities and differences for active and inactive molecules. Compounds are selected for latter based on the presence of functional groups responsible for activity. In terms of drug design and structures, refer to the properties or descriptors of the molecules, their substitution or interaction energy fields, corresponing to an experimental biological or biochemical endpoint which includes activity, binding affinity, and toxicity. Chemometric methods involve MLR, PLS, PCA, PCR, and ANN (Norinder and Bergstrom, 2006). The methods have been evolved from Hansch and free–Wilson's 1- or 2-dimensional linear-free energy relationships, Crammers's 3-dimensional QSAR to Hopfinger's 4th, and Vedani's 5 and 6 dimensions. All one- and two-dimensional and related methods are commonly referred to as "classical" QSAR methodologies. QSAR plays a vital role in modern drug design, since this represents a cost effective as well rapid alternative to the medium throughput in vitro and low throughput in vivo assays. Whenever compounds with particular biological activity are known, then the compounds can be used to design computational screening model which includes QSAR. In same context, biological activity should be obtained through reliable experiments because generated model entirely depends on biological activity profile of compounds used in model building. Nowadays, drug development is carried out by QSAR analysis, optimizing pharmacodynamic and pharmacokinetic properties (Salum and Andricopulo, 2009; Santos-Filho *et al*., 2009; Andricopulo *et al*., 2009). Applications involving genomics, proteomics, cheminformatics, bioinformatics, high-throughput X-ray crystallography, targeted combinatorial libraries, etc. largely aided to increase the output in the form of quality leads. Many efforts are still needed to reduce the time, expenditure, and attrition rate in the drug discovery cycle simultaneously addressing the huge unmet medical need across the world. Referencing the survey report, poor pharmacokinetic and preclinical toxicity were the major reasons for the failure in the drug development, in addition to the lack of efficacy and adverse effects (Stewart *et al*., 2002; Collins *et al*., 1998; Muller, 2003; Kennedy, 1997). Currently, the scenario has changed with more efforts focused on early-stage physicochemical profiling. The high-level process flow of the QSPR/QSAR modeling in the data mining (DM) environment is summarized in Fig. 1. This process flow diagram outlines the two main tasks within this environment—the QSPR/QSAR model development and the deployment of developed QSPR/QSAR models. Both of these tasks involve cooperation between different software modules within the data mining environment, such as quantum chemical (QC) calculation, molecular descriptor calculation, QSPR/QSAR

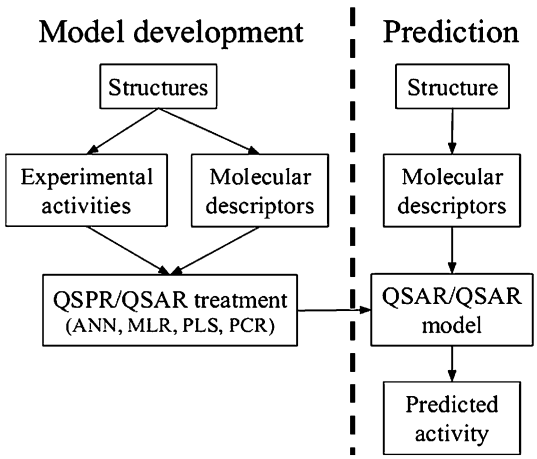

Fig. 1 Overview of QSAR/QSPR modeling

model development, and QSPR/QSAR prediction (Katritzky *et al*., 1995). Based on the way by which the descriptor values or structural representation are derived, dimensionally based methods of QSAR are categorized into the following classes:

- 1D-QSAR correlating activity with global molecular properties like *p*Ka, log P, etc.
- 2D-QSAR correlating activity with structural patterns like connectivity indices, 2D-pharmacophores, without taking into account the 3D-representation of these properties.
- 3D-QSAR correlating activity with non-covalent interaction fields surrounding the molecules.
- 4D-QSAR additionally including ensemble of ligand configurations in 3D-QSAR.
- 5D-QSAR explicitly representing different induced-fit models in 4D-QSAR.
- 6D-QSAR further incorporating different solvation models in 5D-QSAR.

Based on chemometric methods, sometimes QSAR methods are also classified depending upon the type of correlation technique employed to establish a relationship between structural properties and biological activity. This includes linear methods including linear regression (LR), multiple linear regression (MLR), partial least squares (PLS), and principal component analysis/regression (PCA/PCR). Non-linear methods consist of artificial neural networks (ANN), *k*-nearest neighbors (*k*NN), and Bayesian neural nets. As classical QSAR method is much easy to handle, they are faster and more amenable to automation. They include defined physicochemical descriptors which are suited best for the evaluation of large number of molecules and screening of molecular databases. Moreover, QSAR methods correlate macroscopic target properties with computed atom-based descriptors derived from the

spatial representation of the molecular structures. The methodology has emerged as a natural extension to the classical methods of QSAR approaches pioneered by "Hansch and Free-Wilson." Experimental assays cannot be replaced by QSAR model because of various obvious limitations in simulation of real-world situations and in vivo parameters in QSAR modeling. Although they play a decisive role in predicting and correlating the biological profile of molecules, in certain situations and conditions, they suffer from severe limitations given in the following:

- Lack of sufficient number of training molecules.
- Consideration of only two-dimensional structures.
- Insufficient parameters for relating drug–receptor interactions such as Hammett constant.
- Unavailability of specific physiochemical parameters.
- Unavailability of representation of stereochemistry.
- No unique solution with high risk of failure and chance correlations.
- Requirement of knowledge of substituent constants and chemistry utilized to design a molecule.
- Lack of suggestion to synthesize a new compound through classical QSAR equations with no graphical output.

Computational chemistry represents molecular structures as numerical models and simulates their behavior with the equations of quantum and classical physics. Available programs enable scientists to easily generate and present molecular data including geometries, energies, and associated properties (electronic, spectroscopic, and bulk). The usual paradigm for displaying and manipulating these data is a table in which compounds are defined by individual rows and molecular properties (or descriptors) by the associated columns. A QSAR attempts to find consistent relationships between the variations in the values of molecular properties and the biological activity for a series of compounds so that these "rules" can be used to evaluate new chemical entities. A QSAR generally takes the form of a linear equation

$$\text{Biological activity} = \text{Constant} + (C_1 \cdot P_1) + (C_2 \cdot P_2) + (C_3 \cdot P_3) + \ldots,$$

where the parameters P_1 through P_n are computed for each molecule in the series and the coefficients C_1through C_n are calculated by fitting variations in the parameters and the biological activity (Walpole *et al*., 1993; Kubinyi, 2004).

2D-QSAR with respect to physicochemical properties

While searching, one finds numerous hits for lead candidates, and thus lead optimization is hindered. To get more target structural information, high-throughput protein crystallization can be explored (Stewart *et al*., 2002). Lead optimization remains the most serious bottleneck. In addition, 40 % of all development candidates fails due to absorption, distribution, metabolism, excretion, and toxicity "ADMET" problems. High-throughput screening (HTS) for pharmaceutical discovery is used as a filter in order to identify the few potentially promising hits in a corporation's synthetic archive. Therefore, HTS data analysis is focused on hits, and the bulks of the non-hit data are usually ignored (Manly *et al*., 2001). Cheminformatics methods must be applied while generating data using high-throughput techniques in order to assure that good ADMET properties are achieved while making and screening compounds, this approach is called a multi-parametric optimization strategy (Baxter and Lockey, 2001). Several physicochemical properties of drug molecules such as aqueous solubility, partition coefficient (log *P*), distribution coefficient (log *D*), ionization constant (*p*Ka), and topological polar surface area (tPSA) play an important role in majority of the processes and outlined in Fig. 2. Undesirable physicochemical properties point to the potentially undesirable pharmacokinetic behavior.

Solubility

Adequate aqueous solubility is of paramount importance since dissolution of the active drug in the gastrointestinal (GI) fluids precedes its oral absorption from the GI tract. Aqueous solubility, in turn, is dependent on several factors such as size and shape of the molecule, hydrophobicity, hydrogen bonding, and crystalline/amorphous state (Wang *et al*., 2007). A detailed account on solubility prediction was reported (Jorgensen and Duffy, 2002). Two aqueous solubility prediction models, ASMS (aqueous solubility based on molecular surface) and ASMS-LOGP (aqueous solubility based on molecular surface using ClogP as a descriptor) for a diverse data set of 1,708 molecules, appeared in the literature (Wang *et al*., 2007). Both the models performed well in terms of statistics (leave-one-out $q^2 = 0.872$, RMSE = 0.748 for ASMS model and $q^2 = 0.886$, RMSE = 0.705 for ASMS-log *P* model). The authors proposed these models as possible drug-like filter which could be used in prioritizing compound libraries prior to HTS (Wang *et al*., 2007). In another study involving the development of QSPR models for predicting the aqueous solubility, random forest (RF) regression, PLS, support vector machines (SVM), and ANN methods were used (Palmer *et al*., 2007). For screening purpose, aqueous solubility is usually measured fewer than three experimental conditions PBS (pH 7.4), simulated intestinal fluid and simulated gastric fluid. These measurements can be

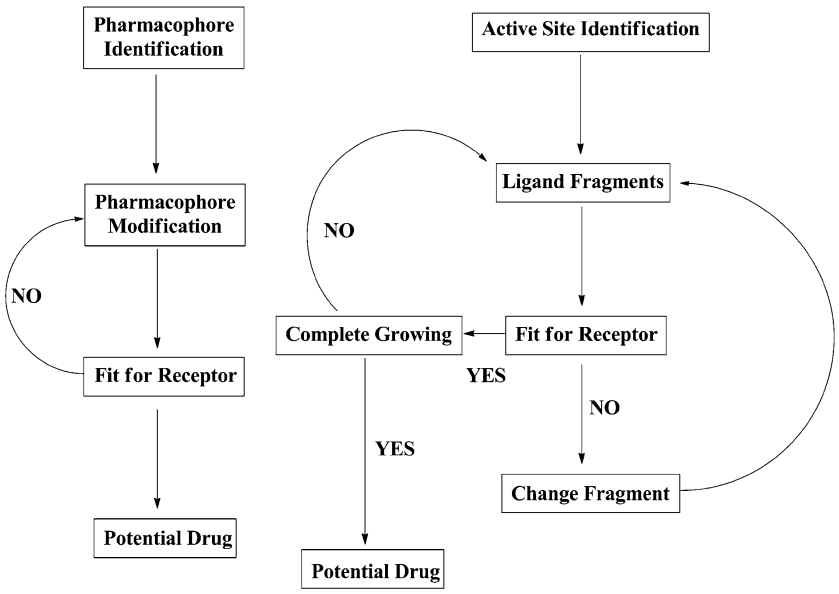

Fig. 2 General layout of drug designing

performed in high-throughput manner, and such data can be utilized for developing the predictive models.

Permeability

Two main types of permeability, namely human intestinal permeability and blood–brain barrier (BBB) permeability (important for the distribution of CNS active agents and toxicity of non-CNS drugs), are important with respect to biological profile. Out of these, permeability across human intestinal membrane represents the major step in the process of oral absorption of xenobiotics. Most of drugs cross intestinal epithelia by passive diffusion mechanism, which in turn, largely depend on the physicochemical properties of the drug. Hence, in silico models of human intestinal permeability is of great significance for ADME/T profiling. An in silico model was recently developed for the prediction of parallel artificial membrane permeability assay (PAMPA) permeability using logP, *p*Ka, and PSA descriptors (Nakao *et al*., 2009). On the other hand, BBB permeability is a crucial factor which needs careful consideration in the ADME/T profiling. CNS drugs must cross BBB to exhibit therapeutic effect, whereas non-CNS drugs are expected not to cross the BBB to avoid unwanted side effects. With reference to CNS drug study, a large data set of 1,593 molecules along with a set of 19 simple molecular descriptors was used for building BBB prediction models (Zhao *et al*., 2007). The H-bonding properties of molecules were found to modulate the BBB penetration.

*p*Ka

Prediction of *p*Ka from the molecular structure is an intense area of research as seen from the voluminous work done in this area (Lee *et al*., 2008), where the ionic state of the drug molecule at physiological pH, represented by the ionization constant, or *p*Ka, can potentially affect its pharmacokinetic behavior. The pH-dependent distribution coefficient logD (at pH 6.5, 7.4) is mainly dependent on the *p*Ka as the drugs experience varying environment (pH 1-3 in stomach, pH 5-7 in duodenum, pH 8 in jejunum and ileum) during their passage in the GIT. Thus, physicochemical properties such as aqueous solubility and lipophilicity (logD) are partly dependent on *p*Ka. The need for in silico models for *p*Ka prediction still persists. Lee *et al*. described computational methodology and applications of a computer program SPARC (SPARC Performs Automated Reasoning in Chemistry) for the prediction of ionization state of a drug (Lee *et al*., 2007). This program is based on the solid physical chemistry of reactivity models. It predicts both macroscopic and microscopic *p*Ka values for a compound simply from the molecular structure. In the reported study using 123 known drugs, SPARC predicted *p*Ka values correlated well with the experimental values ($r^2 = 0.92$ and RMSE = 0.78 log unit).

Lipophilicity

Lipophilicity, represented by logP, affects human intestinal permeability, drug absorption, distribution and clearance behavior. Several log *P* prediction programs are available. These programs are mainly divided into three categories: (a) those based on whole molecule approach, (b) fragment-based approach, and (c) atom-based approach (In *et al*., 2005). In addition to log *P*, log *D* or distribution coefficient represents more meaningful lipophilicity parameter since ionizable drugs exhibit different partition behaviors in different pH environments in the body (e.g., blood pH

7.4). Bruneau and McElroy (2006) reported predictive model for log $D_{7.4}$ using Bayesian regularized neural networks (BRNNs) employing automatic relevance determination (ARD). A data set of in-house compounds ($n = 5{,}000$) was used for developing the in silico models. BRNN with ARD has been shown to be successfully applied for $\log D_{7.4}$ prediction.

Most attempts have been made to develop a significant model as close as possible to real one and for these considerations three-dimensional paradigms have to rely on basic assumptions like molecular structure which can be measured and represented with a set of numbers usually called descriptors which encode all physical, chemical, and biological properties as there is an underlying relationship between molecular structure and biological activity. Receptor binding and biological activity are in direct proportion with differential effects on other signaling steps which usually transpire between experimentally observed response and receptor binding. Some major factors like desolvation energy, temperature, diffusion, transport, pH, and salt concentration which contribute to the overall free energy of binding are difficult to handle, and thus usually ignored. Structural properties which lead to an observable biological response are most commonly determined by non-covalent forces, mainly electrostatic and steric, and the observed biological effect is produced by the modeled ligand itself, not by its degradation product. With few exceptions, the geometry of the receptor binding site is considered rigid. Resulting QSAR model may represent one of potentially several solutions to the property–activity correlation problem (Akamatsu, 2002; Matyus and Borosy, 1998; Oprea, 2004).

Methods for building QSAR models

Statistical or chemometric techniques form the mathematic foundation for building a QSAR model. A brief history with respect to QSAR analysis (Table 1) and some of the methods are described briefly (Table 2). Most easily interpretable method was found to be linear regression analysis among various statistical methods for QSAR. These regressions represent direct correlation of independent variables (**x**) with a dependent variable (**y**). This model can be considered for prediction of ***y*** from the data of **x** variables. This can either belong to qualitative or quantitative set of system (Berk, 2003a). Inclusive variants can be SLR, MLR, and stepwise MLR. Brief explanation of these variants is as follows.

Simple linear regression (SLR)

This method performs as a standard linear regression calculation in generation of QSAR model in the form of equations

Table 1 Brief history of QSAR methodologies

Authors and year	Methodologies
Mills (1884)	Developed a QSAR model for predicting melting and boiling points in homologous series, outcomes were accurate to better than one
Hammett (1935, 1937)	Reported the effect of the substituent addition on benzoic acid with the dissociation constant, postulated electronic sigma-rho constants and established the linear free-energy relationship (LFER) principle
Albert et al. (1948)	Investigated the effects of ionization/electron distribution and steric access on the potencies of a multitude of aminoacridines
Taft (1952)	Postulated a method for separating polar, steric, and resonance effects and introduced the first steric parameter, E_S
Hansch and Fujita (1964)	Reported the combination of hydrophobic constants with Hammett's electronic constants to yield the linear Hansch equation and its many extended forms
Hansch (1969)	Developed the Hansch equation for dealing with extended hydrophobicity ranges
Free and Wilson (1964)	Formulated an additive model, where the activity is discretized as a simple sum of contributions from various substituents
Kubinyi (1976)	Determined the drugs transport via aqueous and lipoidal compartment system and further refined the parabolic equation of Hansch to develop non-linear QSAR model superior to the earlier one
Hansch and Gao (1997)	Developed comparative QSAR
Labute (1999)	Reported binary QSAR to handle binary activity measurements from high-throughput screening (e.g., active or inactive), and molecular descriptor vectors as input. Determination of probability distribution for actives and inactives was based on Bayes' Theorem

Table 2 Statistical techniques for building QSAR models

Linear regression analysis (RA)
Simple linear regression (SLR)
Multiple linear regression (MLR)
Stepwise multiple linear regression
Multivariate data analysis
Principal component analysis (PCA)
Principal components regression (PCR)
Partial least square analysis (PLS)
Genetic function approximation (GFA)
Genetic partial least squares (G/PLS)
Pattern recognition
Cluster analysis
Artificial neural networks (ANNs)
k-Nearest neighbor (*k*NN)

which include a single independent descriptors **x** and **y** as a dependent variable. This technique is found to be very promising for generating structure and activity relationships by exploring some of the most important descriptors used in governing the activity, whereas some of multiple descriptors' interaction was neglected. The simple linear regression can be expressed by the following equation (Eq. 1):

$$\mathbf{y} = \mathbf{a} + \mathbf{bx}, \quad (1)$$

where **y** is the dependent variable, **x** is the independent variable, **a** is the constant, and **b** is the regression coefficient.

Multiple linear regression (MLR)

This method is the extension of SLR to more than one dimension. In this method, standard multivariable regression calculations are performed. Identification of a drug property is carried out on all of the descriptors under investigation. Correlation possibility is checked by the value of multiple correlation coefficient (r), t-value through leave-one-out method. Correlation is checked by r^2 or q^2 values which are usually known as cross-validated correlation coefficient. This method is also known as linear free-energy relationship method (LFER). The relationship is expressed in single multiple-term linear equation (Eq. 2) as follows:

$$\mathbf{y} = b_0 + b_1\mathbf{x_1} + b_2\mathbf{x_2} + b_3\mathbf{x_3} + \cdots + b_{\mathrm{m}}\mathbf{x_m} + \mathbf{e}. \quad (2)$$

Stepwise multiple linear regression

This method commonly used variant MLR which creates a multiple-term linear equation, but not all the independent variables are used. This method has a good utilization when the number of descriptors is large and main descriptors are unknown. Moreover, orthogonal latent variables can be used in MLR (Berk, 2003b).

Partial least square (PLS)

PLS gives a statistically robust solution even when the independent variables are highly interrelated among themselves, or when the independent variables exceed the number of observations. PLS is an iterative regression method that produces its solutions based on linear transformation of a large number of original descriptors to a small number of new orthogonal terms called latent variables (Wold *et al*., 1993). Thus, this method is counted as standard statistical one.

Principal components analysis (PCA)

This method is known to create a new set of orthogonal descriptors referred to as principal components (PCs) which describe most of the information contained in the independent variables in order of decreasing variance. In this method, QSAR model is not generated but still it witness for relationship among unlike variables. PCA reduces dimensionality of data set of descriptors to actual amount of data. To generate a multiple-term linear equation, PCR applies the scores from PCA decomposition as regressors in QSAR model (Dunteman, 1989a, b).

Genetic function approximation (GFA)

This method serves as an alternative to standard regression analysis for building QSAR equations (Rogers and Hopfinger, 1994). It can build linear as well as higher order non-linear equations. Genetic partial least squares (G/PLS or GA-PLS) is an important tool which has evolved by combining the best feature of GFA and PLS. This method is extensively used by the researchers/scientists around the globe (Dunn and Rogers, 1996; Breu *et al*., 2007; Datar *et al*., 2006; Khedkar *et al*., 2007; Dhaked *et al*., 2009; Verma *et al*., 2008).

Cluster analysis

This analysis method is a pattern recognition method used to investigate the relationship between observations associated with many other properties and to partition the data set into categories into similar elements. This also implies to identify which of the subsets share similar physical properties (Aldenderfer and Blashfield, 1984).

Artificial neural networks (ANNs)

ANNs are useful tools in QSAR/QSPR studies, and particularly in case where it is difficult to specify an exact mathematical model for describing a given structure property relationship. Most of these works used neural networks based on the back-propagation learning algorithm, which has some disadvantages such as local minimum, slow convergence, time-consuming non-linear iterative optimization, and difficulty in explicit optimum network configuration (Walczak and Massart, 2000). The method of artificial neural networks originated from the real neurons that are present in an animal brain. ANNs are parallel computational systems consisting of groups of highly interconnected processing elements called neurons, which are arranged in a series of layers. The input layer as the very first one and each of its neurons receives data from user, which corresponds to one of the independent variables used as inputs in QSAR. After input layer, there are many layers of neurons, collectively known as the hidden layers. The last layer is termed to be the output layer, and its neurons handle the output from the network. Each layer

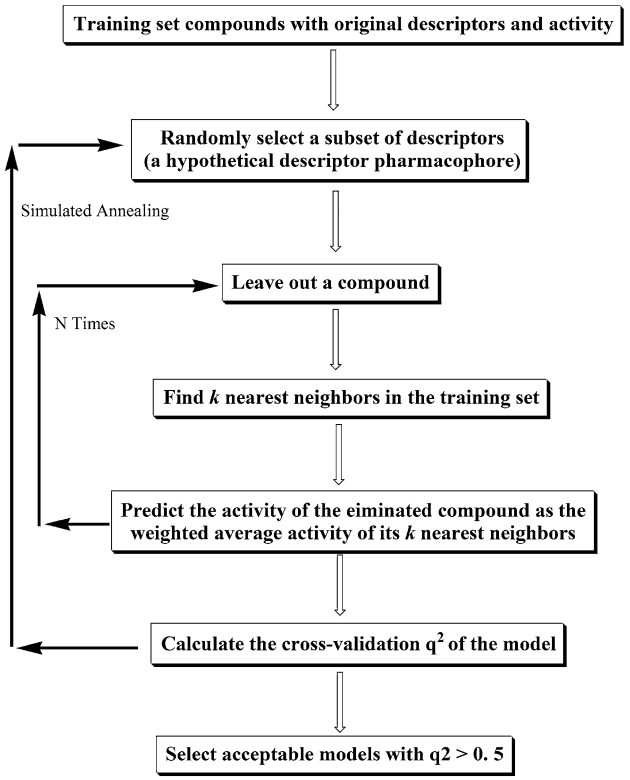

Fig. 3 Flowchart of kNN with variable selection

may make its independent computations and may pass the results yet to another layer (Baskin *et al*., 2008). In this, the result of the transfer function is communicated to the neurons in the output layer. This is the point where the results are interpreted and presented finally.

k-Nearest neighbor

The *k*NN approaches are executed by distances between an object which is unknown and all the objects in the training set. Based on the calculation of distances, the objects from training set most similar to object unknown are selected. Finally, the optimum *k* value is selected by optimization through categorizing the sample test set (Ajmani *et al*., 2006). A systematic flowchart with variable selection is represented as shown in Fig. 3.

Classification of 3D-QSAR Methods

This can be classified in various categories. Some of these are intermolecular modeling based such as ligand and receptor. These can also be classified on the basis of chemometric methods which are mainly utilized for correlation of structural properties and activities. Moreover, alignment criterion is also a base of classification. Few of them are explained briefly:

- Based on intermolecular modeling
  Ligand based e.g., CoMFA, CoMMA, CoMSIA, and GERM.
  Receptor based e.g., CoRIA, COMBINE, and AFMoC.
- Based on chemometric methods
  Linear 3D-QSAR e.g., CoMFA, CoMSIA, and GERM.
  Non-linear 3D-QSAR e.g., COMPASS and QPLS.
- Based on Alignment criterion
  Alignment dependent e.g., CoMFA, CoMMA, CoMSIA, GERM, and CoRIA.
  Alignment independent e.g., WHIM, CoSA, HQSAR, COMPASS etc.

Comparative molecular field analysis (CoMFA)

It has served as a well-deserving tool in drug designing and lead optimization for decades. Dynamic-oriented molecular modeling system which is also been known as DYLOMMS involves the utilization of PCA from the molecular interaction fields and is then correlated with biological profile (Wise *et al*., 1983). GRID and PLS were combined to DYLOMMS as modified techniques of this, in approach to make it powerful technique, which is known and termed as CoMFA. A standard procedure which is implemented in sybyl software from Tripos Inc. (Berman *et al*., 2000) has the following steps:

- Bioactive conformations of each molecule are determined.
- All the molecules are superimposed or aligned using either manual or automated methods, in a manner defined by the supposed mode of interaction with the receptor.
- The overlaid molecules are placed in the center of a lattice grid with a spacing of 2 Å.
- The algorithm compares, in three dimensions, the steric and electrostatic fields calculated around the molecules with different probe groups positioned at all intersections of the lattice.
- The interaction energy or field values are correlated with the biological activity data using PLS technique, which identifies and extracts the quantitative influence of specific chemical features of molecules on their biological activity.
- The results are articulated as correlation equations with the number of latent variable terms, each of which is a linear combination of original independent lattice descriptors.
- For visual understanding, the PLS output is presented in the form of an interactive graphics consisting of colored contour plots of coefficients of the corresponding field variables at each lattice intersection, and showing the imperative favorable and unfavorable regions in 3D-

dimensional space which are considerably associated with the biological activity.

Selection of compounds with biological profile and their optimization has a significant role in QSAR. To improve the biological activity and to reduce the side effects, QSAR has a significant value. To do so, there are many factors to be stressed on while selecting substituent for the modification of compounds. There are some important factors such as (i) the selected compounds should not be similar to the existing ones which has decisive role in minimization colinearity among the variables, (ii) maximize dissimilarity and orthogonality, and (iii) selection should be based on descriptor space and least number of compounds should be considered; moreover, synthetic selected compounds with good accessibility should be considered. 3D-QSAR can also be applied to heterogeneous set of data as some biological data accuracy should be maintained and taken under consideration. As every computational technique follows the principle of "garbage in garbage out," so QSAR techniques should be operated in such a sophisticated manner that good model is developed/generated. The model can be generated as same by having (a) compounds with specific and same mechanism of action with same binding mode; (b) compounds should correlate to their specific binding affinity; (c) molecular data for biological activity should be obtained by utilizing radioligand, cofactor, pH, activator, tissue, organism, and protein; and (d) units should be similar for the data obtained; moreover, biological data should be symmetrically distributed around the mean, and skewness can be removed or eliminated by log transforming the data and expressing it as log $1/C$ (Oprea, 2004; Hopfinger and Tokarski, 1997; Kim, 1995). This is the well-known fact that each molecule having one or more single bonds exists at every moment in many different so-called rotamers. Multiple conformations can exist at specific physiological conditions.

One of the most important problems with 3D-QSAR technique is the alignment of molecules. All the molecules in a data have common stiff core structures, where molecules can be aligned using least square fitting procedure. However, in case of structural heterogeneity in the dataset, alignment of molecules becomes very difficult; in such case, several approaches have been proposed to superimpose the molecules as accurately as possible, some of which are as follows:

- Atom overlapping-based superimposition: This method of approach is one of the most popular one and also known as "Pharmacophore approach." It basically involves atom to atom pairing between the molecules and is utilized in identifying dissimilarities between similar molecules.
- Pharmacophore-based superimposition: This method is well utilized as hypothetical pharmacophore and is useful as common target template. Molecules are directed conformationals to assume the shape obligatory for its sub-molecular features to match with either a known pharmacophore which is generated during the analysis.
- Binding sites-based superimposition: Thus, the method has application where molecular alignment is obtained. This is done by superimposing the receptor active sited or the receptor residue, which interact with ligands. This technique is found and, moreover, believed to be more conceivable.
- Multiple conformers-based superimposition: This technical method is particularly useful where the ligands may bind to a receptor in multiple and various ways. COMPASS is one of the examples, which determines and selects the best bioactive conformation and optical alignment from a set of initial poses.

Before performing the actual chemometric analysis in 3D-QSAR, the raw data are pretreated to minimize redundancy (Kim, 1995). Reduction method is based on the standard deviation cut-off. In this technique, all the energy columns with a low standard deviation are eliminated from the data, since they require longer computing time without contributing to the results. Several variable selection methods are available like in CoMFA the steric and electrostatic values are amended using cut-offs ($\pm$30 kcal/mol), depending upon the position of the lattice point. After pretreatment, the data are subjected to scaling which assigns equal weight to all the descriptors and places them on a common platform for a meaningful statistical analysis. Different scaling techniques are available and utilized effectively in 3D-QSAR approaches. Few of them are autoscaling in which the variables are scaled to zero mean and a unit standard deviation by dividing each column with its standard deviation, block-scaling such as in CoMFA standard scaling, block-adjusted scaling where the energy values are the part of analysis, etc. These techniques are found to contribute in improving the ease of interpretation and numerical stability. With respect to display of results, CoMFA generates an equation correlating the biological activity with interactive energy field's contribution at every grid point. Results are generally shown as coefficient contour plots (Hopfinger and Tokarski, 1997). Two types of contours are shown for each interaction energy field: the positive and negative contours. The contours for steric fields are shown in green (positive contours, more bulk favored) and yellow (negative contours, less bulk favored), while the electrostatic field contours are displayed in red (positive contours, electronegative substituents favored) and blue (negative contours, electropositive substituents

favored) colors. As add on, CoMFA also provides two types of plots from PLS models given as score plots and loading/weight plots.

Comparative molecular surface analysis (CoMSA)

CoMSA is a non-grid three-dimensional QSAR approach that makes use of the molecular surface for defining those regions of the compounds which are required to be compared using the mean electrostatic potentials (Polanski and Walczak, 2000; Polanski *et al*., 2002). In this method, the molecules are subjected in the data set to geometry optimization and assigning them with partial atomic charges. The Kohonen's self-organizing maps, a type of neural network, are then employed to transform the three-dimensional surface of the molecules into two-dimensional topographical maps. The partial atomic charges of the atomic molecular representations are also projected onto the two-dimensional topographical maps. The molecular electrostatic potentials are calculated at the surface points, and a mean value of the potential analogous to the respective points found in each grid cell. The calculated mean electrostatic potential values are converted into vectors, and the vectors expressing all the molecules in the series are superimposed onto a matrix, by comparing the respective topographical maps of the molecules. The parative matrix of the mean electrostatic potentials is finally used to develop a 3D-QSAR model using the PLS technique. It compares the molecular properties explaining not a discrete set of points but the average property values (MEPs) calculated for a certain area of the molecular surface. A receptor-dependent CoMSA model has been developed for sulforaphane compounds as activators of quinone reductase (Magdziarz *et al*., 2009). CoMSA application includes the modeling of *p*Ka values of benzoic acids (Gieleciak and Polanski, 2007), and hypolipidemic asarones (Magdziarz *et al*., 2006), determination of the binding mode for a series of benzoxazine oxytocin antagonists using docking and 3D-QSAR studies (Jojart *et al*., 2005).

CoMSIA

CoMSIA, Comparative Molecular Similarity Indices Analysis, was mainly developed to overcome the certain limitations of CoMFA. In CoMSIA, molecular similarity indices calculated from modified SEAL similarity fields are employed as descriptors to simultaneously consider steric, electrostatic, hydrophobic, and hydrogen bonding properties. These indices are estimated indirectly by comparing the similarity of each molecule in the dataset with a common probe atom (having a radius of 1 Å, charge of +1 and hydrophobicity of +1) positioned at the intersections of a surrounding grid/lattice. Selected examples of the applications of this methodology can be found for the following:

- Use of the Gaussian distribution of similarity indices.
- The choice of similarity probe is not only limited to either steric or electrostatic potential fields but also includes hydrophobic and hydrogen bonding including hydrogen bond acceptors and donors fields.
- Effect of the solvent entropic terms.
- The CoMSIA contours indicate those areas within the region occupied by the ligands that "favor" or "dislike" the presence of a group with a particular physicochemical property.
- The relationship between the required properties and a probable ligand shape.
- Generation of predictive 3D-QSAR models of boron containing dipeptides as proteasome inhibitors (Zhu *et al*., 2009).
- Hydroxamic acid derivatives as urease inhibitors (Ul-Haq *et al*., 2009).
- Thiazolidin-4-one derivatives as anti-HIV-1 agents (Murugesan *et al*., 2009).
- Thiazolidinediones derivatives as aldose reductase inhibitors (Liu *et al*., 2009).

In addition to these methodologies and tactical techniques, several other 3D-QSAR methodologies have been generated, out of those some are as follows:

CoRIA

- The approach which uses the descriptors that describe the thermodynamic events involved in ligand binding.
- The methodology simply consisted of calculating the non-bonded interaction energies between the ligand and the individual active site residues of the receptor, which are involved in interaction with the ligand (Datar *et al*., 2006; Dhaked *et al*., 2009).
- This approach was further extended and modified to develop two new variants of CoRIA: reverse-CoRIA and mixed-CoRIA.
- These new developed techniques, reverse-CoRIA and mixed-CoRIA, were used as independent variables that are correlated to the biological activity by G/PLS chemometric method (Verma *et al*., 2008).

Comparative molecule/pseudo receptor interaction analysis (CoMPIA)

- Based on a common template molecule, the geometry of the molecules is optimized followed by their superimposition.

- The resulting space encompassed by the set of superimposed molecules is partitioned into grids with sufficient number of lattice points to accommodate all the probe atoms.
- Nine different types of hybrid atoms/probes are distributed at each lattice point using a genetic algorithm, the steric, electrostatic, and hydrophobic interactions, and between different probes and every molecule in the set are computed and then correlated with the biological activities using PLS (Zhou *et al*., 2006).

Comparative molecular active site analysis (CoMASA)

Initially, the molecules are superimposed and their interatomic distances calculation. Afterward, the molecular representation until the distances between all the atoms/pseudo atoms is greater than the threshold value of 0.75 Å. The interaction energies such as steric, electrostatic, and hydrophobic properties are then computed for each molecule (Kotani and Higashiura, 2004).

FLUFF-BALL

- It is based on a novel field-fitting procedure called flexible ligand unified force field (FLUFF).
- A semiautomatic superimposition of the molecules is carried out.
- It is a MMFF94 force field that is customized to impart flexibility to the ligand to maximize similarity.
- The similarity between ligands and template is evaluated, and the computed steric and electrostatic descriptors are correlated with the biological activities using the PLS technique (Korhonen *et al*., 2003).

Receptor surface analysis/modeling, comparative receptor surface analysis (RSA/RSM/CoRSA)

- Molecules are optimized and superimposed in their bioactive conformation.
- A receptor-complementary surface is generated using shape fields which basically represent their aggregate molecular shape.
- Putative chemical properties of the receptor are computed.
- PLS models are developed that correlate surface properties with molecular activities (Hahn, 1995; Ivanciuc *et al*., 2000).

Self-organizing molecular field analysis (SOMFA)

In this approach, the mean activity of training set is subtracted from the activity of each molecule to obtain their mean centered activity values, and the grid values for each molecule are summed up to give the master grids. Finally, $SOMFA_{property,i}$ descriptors from the master grid values are then calculated and correlated with the log-transformed molecular activities (Robinson *et al*., 1999).

Table 3 Some important conformational analysis methods

S. no.	Conformational method	Description/utilization
1.	Grid search	Generates all possible conformations
2.	Random search	Generates a set of conformations by random change in Cartesian, bond angles and torsion angles
3.	Monte Carlo	Simulates dynamic behavior and generates conformation by structural changes and energy comparison
4.	Molecular dynamics	Follows Newton's second law of motion (Force = mass × acceleration) and thereby simulates time dependent movements and changes in conformation
5.	Simulated annealing	Method overcomes the huge energy barriers and slowly cools down the system
6.	Distance geometry algorithm	Selects random distances within each pair of upper and lower bounds to form constraints in a distance matrix
7.	Genetic algorithm	Based on biological evolution and works by forming new conformers

Hereby, with respect to the QSAR or computational study, few of the current contributing studies in support of this review are presented. Xingyan et al. reported that structure-based 3D-quantitative structure–activity relationship (QSAR) studies were performed on a series of dihydropyrazole and dihydropyrrole derivatives using comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) methods to find the correlation between Eg5 and its inhibitors. Molecular docking and 3-D QSAR studies were carried out to explore the binding mechanism of dihydropyrazole and dihydropyrrole derivatives to EG5. Good prediction of COMFA and COMSIA models was obtained with LOO cross-validation q^2 and conventional r^2 values of 0.898, 0.980, and 0.848, 0.992, respectively. The results suggested that ligands binding in the hydrophobic part of the inhibitor-binding pocket were found to be crucial for potent ligand binding and kinase selectivity (Luo *et al*., 2012). Some other recent postulates are tabulated in Table 3; in addition, Table 4 represents some work with regard to various QSAR methodologies.

4D QSAR

Hopfinger et al. (1997) proposed the 4D-QSAR formalism, which includes the conformational flexibility and the

Table 4 Recent QSAR methodologies

Author and year	Methodologies
Yadav et al. (2010)	The immuno-modulatory compound from derivatives of coumarinolignoids, a quantitative structure activity relationship (QSAR) and molecular docking studies were performed. Immunostimulatory activity was predicted through QSAR model, developed by forward feed multiple linear regression method with leave-one-out approach
	Relationship correlating measure of QSAR model was 99 % ($R^2 = 0.99$) and predictive accuracy was 96 % ($RCV^2 = 0.96$). QSAR studies indicate that dipole moment, steric energy, amide group count, lambda max (UV–visible), and molar refractivity correlate well with biological activity, while decrease in dipole moment, steric energy, and molar refractivity has negative correlation
Liu et al. (2011a)	Work was based on a dataset of 110 collected benzazepine (BAZ) DA D3 antagonists with diverse kinds of structures, a variety of in silico modeling approaches, including comparative molecular field analysis (CoMFA), comparative similarity indices analysis (CoMSIA), homology modeling, molecular docking, and molecular dynamics (MD), were carried out to reveal the requisite 3D structural features for activity
	Results showed that both the receptor-based ($Q^2 = 0.603$, $R^2_{ncv} = 0.829$, $R^2_{pre} = 0.690$, SEE = 0.316, SEP = 0.406) and ligand-based 3D-QSAR models ($Q^2 = 0.506$, $R^2_{ncv} = 0.838$, $R^2_{pre} = 0.794$, $SEE = 0.316$, $SEP = 0.296$) are reliable with proper predictive capacity
Liu et al. (2011b)	Several in silico models have been built with two classes of proteasome inhibitors (PIs) using 3D-QSAR, homology modeling, molecular docking, and molecular dynamics (MD) simulations. The study resulted in two types of satisfactory 3D-QSAR models, i.e., the CoMFA model ($Q^2 = 0.462$, $R^2_{pred} = 0.820$) for epoxyketone inhibitors (EPK) and the CoMSIA model ($Q^2 = 0.622$, $R^2_{pred} = 0.821$) for tyropeptin-boronic acid derivatives (TBA)
Tian et al. (2011)	3D-QSAR models were built using CoMFA and CoMSIA methods, and molecular docking was used to check the results. Based on the common sketch align, two good QSAR models with high predictabilities (CoMFA model: $q^2 = 0.823$, $r^2 = 0.979$; CoMSIA model: $q^2 = 0.804$, $r^2 = 0.967$) were obtained, and the contour maps obtained from both models were applied to identify the influence on the biological activity. Combined with the molecular docking results, the detail binding mode between the ligands and Tie-2 was elucidated
Fells et al. (2010)	A structurally diverse dataset of 119 compounds was used to develop and validate a 2D binary QSAR model for the LPA3 receptor. The binary QSAR model was generated using an activity threshold of greater than 15 % inhibition at 10 μM. The overall accuracy of the model on the training set was 82 %. It had accuracies of 55 % for active and 91 % for inactive compounds, respectively. The model was validated using an external test set of 10 compounds
Lowe et al. (2010)	Positive allosteric modulation of the metabotropic glutamate receptor subtype 5 was studied by conducting a comparative molecular field analysis on 118 benzoxazepine derivatives. The model with the best predictive ability retained significant cross-validated correlation coefficients of $q^2 = 0.58$ ($r^2 = 0.81$) yielding a standard error of 0.20 in pEC50 for this class of compounds. The subsequent contour maps highlight the structural features pertinent to the bioactivity values of benzoxazepines
Colosi et al. (2010)	A quantitative structure–activity relationship (QSAR) was used to streamline redesign of a model environmental catalyst, horseradish peroxidase (HRP), for enhanced reactivity toward a target pollutant, steroid hormone 17β-estradiol. This QSAR, embodying relationship between reaction rate and intermolecular binding distance, was used in silico to screen for mutations improving enzyme reactivity
Zhang et al. (2010)	Considering CoMFA and CoMSIA methodologies, homology modeling, and molecular docking, investigation of the structural determinants of Aurora B inhibitors based on three different series of derivatives of 108 molecules was performed. The resultant optimum 3D-QSAR models exhibited ($q^2 = 0.605$, $r^2_{pred} = 0.826$), ($q^2 = 0.52$, $r^2_{pred} = 0.798$) and ($q^2 = 0.582$, $r^2_{pred} = 0.971$) for MK-0457, GSK1070916 and SNS-314 classes, respectively, and the 3D contour maps generated from these models were analyzed individually
Yang et al. (2011)	Traditional Chinese Medicine Database (TCM Database@Taiwan) (http://tcm.cmu.edu.tw) to identify potential EGFR inhibitor was employed. MLR, SVM, CoMFA, and CoMSIA models were generated using a training set of EGFR ligands of known inhibitory activities. Validated MLR ($r^2 = 0.7858$) and SVM ($r^2 = 0.8754$) models predicted good bioactivity for the TCM candidates. In addition, the TCM candidates contoured well to the 3D-QSAR map derived from the CoMFA ($q^2 = 0.721$, $r^2 = 0.986$) and CoMSIA ($q^2 = 0.662$, $r^2 = 0.988$) models

Table 4 continued

Author and year	Methodologies
Natesan et al. 2012)	Reported cellular quantitative structure–activity relationship (cell-QSAR) concept that adapts ligand-based and receptor-based 3D-QSAR methods for use with cell-level activities. The unknown intracellular drug disposition is accounted for by the disposition function (DF), a model-based, non-linear function of a drug's lipophilicity, acidity, and other properties
	This was conceptually combined the DF with multispecies, multimode version of the frequently used ligand-based comparative molecular field analysis (CoMFA) method, forming a single correlation function for fitting the cell-level activities. The resulting cell-QSAR model was applied to the Selwood data on filaricidal activities of antimycin analogs. Their molecules are flexible, ionize under physiologic conditions, form different intramolecular H-bonds for neutral and ionized species, and cross several membranes to reach unknown receptors
	The calibrated cell-QSAR model is significantly more predictive than other models lacking the disposition part and provides valuable structure optimization clues by factorizing the cell-level activity of each compound into the contributions of the receptor binding and disposition.
Noolvi et al. (2010)	QSAR studies were performed on a set of 61 analogs of 4-anilino quinazoline using MDS vlife science QSAR plus module using MLR, PCR, and PLS Regression methods. A QSAR model was generated by a training set of 42 molecules with correlation coefficient (r^2) of 0.912, significant cross-validated correlation coefficient (q^2) of 0.800, F test of 60.5149, r^2 for external test set (pred_r^2) 0.6042, coefficient of correlation of predicted data set (pred_r^2se) 0.7438 and degree of freedom 38 by MLR method. Estate number, Electro-topological state indices, Bromine count, Chlorine count, and alignment-independent descriptors were found to be major contributing descriptors governing the activity
Noolvi et al. (2011)	MLR- QSAR model was found to be statistically significant with respect to training ($r^2 = 0.956$), cross-validation ($q^2 = 0.915$), and external validation (pred_$r^2 = 0.6170$). The developed MLR model suggested that Estate Contribution descriptors SaaOE-Index (30.07 %) and SsCIE-index (15.79 %) were the most important descriptors in predicting erbB-2 inhibitory activity. Electron withdrawing group at 4th position of quinazoline enhances the activity as evident by positive value of SsCIE-index (15.79)

freedom of alignment by ensemble averaging in the conventional three-dimensional descriptors found in traditional 3D-QSAR methods. Thus, the "fourth dimension" of the method is ensemble sampling the spatial features of the members of a training set. In this approach, the descriptors are the occupancy frequencies of the different atom types in the cubic grid cells during the molecular dynamics simulation (MDS) time, according to each trial alignment, corresponding to an ensemble averaging of conformational behavior (Albuquerque *et al*., 1998, 2007). The grid cell occupancy descriptors, GCODs, are generated for a number of different atom types, called interaction pharmacophore elements, IPEs. These IPEs (i.e., atom types), defined as "any type" (A or Any), "nonpolar" (NP), "polar-positive charge" (P+), "polar-negative charge" (P-), "hydrogen bond acceptor" (HA), "hydrogen bond donor" (HB), and "aromatic" (Ar), correspond to the interactions that may occur in the active site, and are related to the pharmacophore groups (Hopfinger *et al*., 1997; Albuquerque *et al*., 1998, 2007; Hopfinger, 2001). Thus, the IPEs are related to the descriptors' nature in 4D-QSAR analysis, while the GCODs are related to the coordinates of IPE mapped in a common grid. The sampling process, in turn, allows the construction of optimized dynamic spatial QSAR models in the form of 3D pharmacophores, which are dependent on conformation, alignment, and pharmacophore grouping.

One factor driving the development of 4D-QSAR analysis is the need to take into account multiple (i) conformations, (ii) alignments, and (iii) substructure groups in constructing QSAR models. These "QSAR degrees of freedom" are normally held fixed in other 3D-QSAR analysis. Insofar as 4D-QSAR analysis can meaningfully predict "active" conformations and the preferred alignment for a training set, it may actually serve as a "preprocessor" for a subsequent CoMFA and/or CoMSIA. Furthermore, the 4D-QSAR method has been proven both useful and reliable for the construction of quantitative 3D pharmacophore models for ligand–receptor data sets (Andrade *et al*., 2009; Krasowski *et al*., 2002; Thipnate *et al*., 2009). As an example, Van Daele et al. (2007) developed RI-4D-QSAR models for a set of thirty-four 5′-aryl-thiourea thymidine analogs, showing inhibitory activity against thymidine monophosphate kinase from *M. tuberculosis* (TMPKmt). This study suggested that the 4D-QSAR methodology can be used in a receptor-dependent, RD, mode when the geometry of the receptor is available as is the case here. However, RD-4D-QSAR analysis requires a relatively large and chemically diverse training set, and also definitive information on binding alignment(s), in order to achieve a non-ambiguous QSAR model. The RI-4D-QSAR analysis (Hopfinger *et al*., 1997; Andrade *et al*., 2009) was carried out, and the best 4D-QSAR model was graphically represented by plotting the

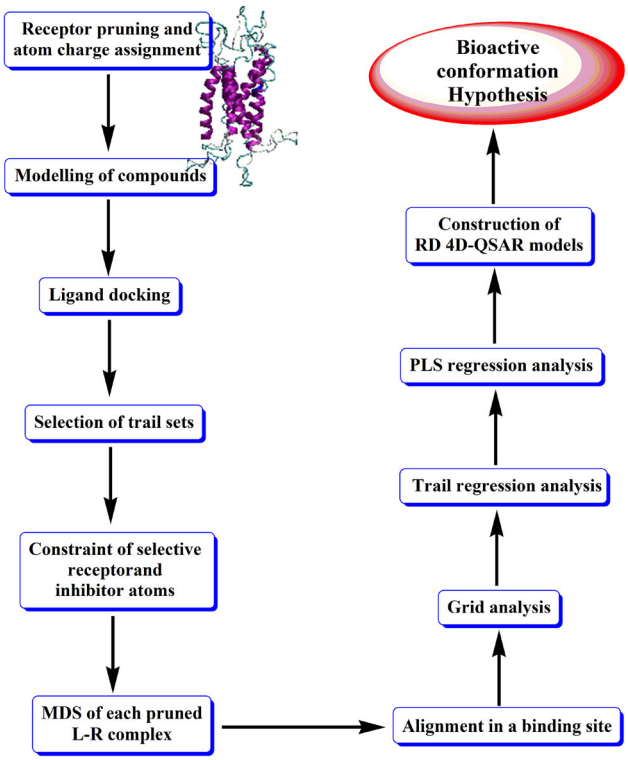

Fig. 4 Operational steps in performing a RD-4D-QSAR

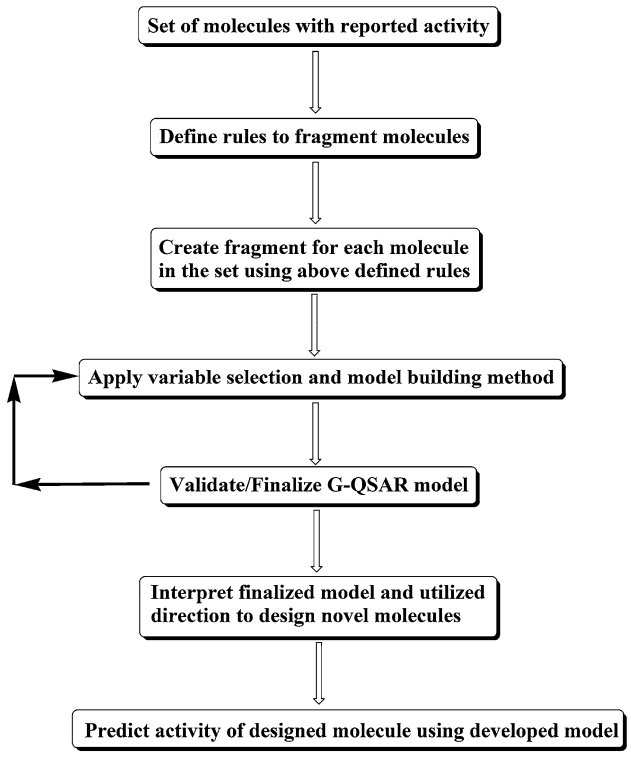

Fig. 5 Flow chart of G-QSAR methodology

significant grid cells in space along with their descriptor attributes (IPEs). The postulated "bioactive" conformation of the most potent inhibitor, according to the best 4D-QSAR model, was docked in the active site of the TMPKmt crystallographic structure. There is a solid consistency between the 3D-pharmacophore sites defined by the QSAR models and interactions with binding site residues. Operational steps in performing a RD-4D-QSAR presented in Fig. 4.

G-QSAR

The proposed new method G-QSAR (Fig. 5) differs from them in two ways: (i) In G-QSAR method, the fragmentation of each molecule in the dataset is done with a set of predefined rules, before calculating their corresponding fragment descriptors. This is unlike existing methods which search a predefined fragment (or group) in the molecule and then uses it as a descriptor either as an indicator variable, their count or corresponding index e.g., path countor molecular connectivity index. (ii) G-QSAR method considers cross/interaction terms as descriptors to account for the fragment interactions in QSAR model, whereas there is no consideration of these descriptors in existing methods.

As molecular fragmentation is a prerequisite to perform G-QSAR (VLifeMDS, 2007), software provides two methods to fragment the molecules depending on their chemical diversities.

- Congeneric series: for congeneric series or template-based series, substituent sites are used to fragment molecules. Software reads the template where substituent sites are defined as dummy atoms. It performs substructure/template search in set of molecules and then fragments the molecule where the dummy atom matches with atom of molecule.
- Noncongeneric series: for noncongeneric series or chemically diverse set molecules, fragments are derived from fragmentation along specific bonds, bonds on the ring fusion. Software provides the interface to define or select the bond for fragmentation for each molecule.

In case of VLifeMDS software, it provides the following steps to perform G-QSAR:

- Reads the set of molecules along with experimental activities.
- Fragmentation of molecules using one of the methods explained above.

- Calculation of various descriptors for each fragment of the molecule as well as fragment interaction descriptors and populating in the worksheet.
- Preprocessing of data (e.g., removing invariable).
- Selection of dependent and independent variables.
- Selection of training and test set: software provides three methods for selection (Sphere Exclusion (SE), random, and manual).

Diligence of drug design tools

The US National Cancer Institute (NCI) conducts a drug discovery program in which ~10 000 compounds are screened every year in vitro against a panel of 60 human cancer cell lines from different organs of origin. Human cancer cell lines include eight melanomas, six leukemia's, and eight cancers of breast, two of prostate, nine of lung, seven of colon, six of ovary, eight of kidney, and six of central nervous system (CNS) origin. Combinatorial libraries have also been assessed recently. Similarity in activity patterns very often indicates similarity in mechanism of action, mode of drug resistance, and molecular structure of tested compounds (Boyd and Paull, 1995; Paull *et al*., 1989). Several different algorithms have been introduced to use the activity information for discovery of anticancer drugs and for understanding of the molecular pharmacology of cancer. The COMPARE program (Boyd and Paull, 1995; Paull *et al*., 1989; Koo *et al*., 1996) has proved very useful for finding agents with activity patterns similar to that of a "seed" compound and for finding compounds with activity patterns that correlate well (positively or negatively) across the 60 cell lines with the expression levels of particular cellular targets. Back-propagation neural networks, Kohonen self-organizing maps, and principal component analysis have been used to predict mechanism of action or to organize compounds into families based on activity patterns. This "information-intensive" approach to the molecular pharmacology of cancer and anticancer drug discovery (Paull *et al*., 1989; Weinstein *et al*., 1994) has proved useful in identifying subgroups of compounds related to particular biological targets. The chemical structure (S) databases can be encoded in terms of any set of 2/3-dimensional molecular structural descriptors or experimentally measured or theoretically calculated physicochemical properties. Analysis can be carried out via database of activity patterns using the COMPARE and DISCOVERY program sets in case of compound search. As part of this process, cluster analysis leads them to identify agents belonging to cluster families.

Some QSAR-type studies were also implemented for dye–cellulose fiber interaction, as well as the qualitative SAR-type relationships, demonstrating at least an appreciable similarity of dye–fiber interactions with receptor–ligand interactions. Series of anthraquinone vat dyes, mono and bisazo, and disperse dyes were studied by several variants of classical QSAR and 3D-QSAR methods. A comparison of the results demonstrated that these methods usually agree in the prediction of structural features favorable for dyeing process. Attractive dye–cellulose interactions were generally favored along the molecular axis of the dye molecule and by the length of the molecular conjugated system. Perhaps, the most interesting result, as indicated mainly by CoMFA studies concerning the contribution of electrostatic fields, was that an increase of positive charges in the dye molecule favors dye adsorption on cellulose (Funar-Timofei *et al*., 2012).

Conclusion

Overall, the application of in silico predictive methods has shown accelerated success in recent years. It is anticipated that this will be a subject of continual development in future not only in drug design applications but also in the area of predictive toxicology. QSAR has been observed in the drug discovery area to enable the design of safe and potent drug candidates. During drug discovery and development phases, pharmacodynamic and pharmacokinetic profile of molecules can be derived using QSAR models. These in silico evaluations consist of the prediction of diverse properties (e.g., physicochemical, ADME) and activities to assist in the optimization and the prioritization of drug candidates. Numerous public, commercial, or corporate in silico tools including SAR/QSAR models, decision trees, and molecular docking have been proposed to achieve these aims. The site-specific clues along with the interpretation of descriptors provided by QSAR techniques such as G-QSAR will help medicinal chemists to design better molecules. We have provided an overview of different QSAR methods and recent development in fragment-based approaches using selected studies as an illustration. Since each QSAR method has its own advantages and disadvantages, researchers should choose appropriate methods for modeling their systems according to the information available with respect to target and ligand. However, given a wide range of choices, it is a challenging task to pick appropriate models for one's studies. This paper outlines many basic principles of new fragment-based QSAR methods as well as other three-dimensional and other dimensional QSAR models and illustrates some examples which may be helpful references to many researchers. In a nut shell, a comprehensive understanding and error-free practice of such strategies in

QSAR modeling should benefit the medicinal chemists to prioritize their experimental endeavors and considerably amplify the experimental hit rates.

References

Ajmani S, Jadhav K, Kulkarni SA (2006) Three-dimensional QSAR using the k-nearest neighbor method and its interpretation. J Chem Inf Model 46:24–31

Akamatsu M (2002) Current state and perspectives of 3D-QSAR. Curr Top Med Chem 2:1381–1394

Albert A, Goldacre R, Phillips J (1948) The strength of heterocyclic bases. J Chem Soc 2:2240–2249

Albuquerque MG, Hopfinger AJ, Barreiro EJ, De-Alencastro RB (1998) Four-dimensional quantitative structure-activity relationship analysis of a series of interphenylene 7-oxabicycloheptane oxazole thromboxane A2 receptor antagonists. J Chem Inf Comput Sci 38:925–938

Albuquerque M, Brito M, Cunha E, Alencastro R, Antunes O, Castro H, Rodrigues C (2007) Multidimensional-QSAR: beyond the third-dimension in drug design. Curr Methods Med Chem Biol Phys 1:91–100

Aldenderfer MS, Blashfield RK (1984) A review of clustering methods. In: Aldenderfer MS, Blashfield RK (eds) Cluster analysis. SAGE Publications Ltd, London, pp 33–61

Andrade CH, Pasqualoto KFM, Ferreira EI, Hopfinger AJ (2009) Rational design and 3D-pharmacophore mapping of 5′-thiourea-substituted alpha-thymidine analogues as mycobacterial TMPK inhibitors. J Chem Inf Model 49:1070–1078

Andricopulo AD, Salum LB, Abraham DJ (2009) Structure-based drug design strategies in medicinal chemistry. Curr Top Med Chem 9:771–790

Baskin II, Palyulin VA, Zefirov NS (2008) Neural networks in building QSAR models. Methods Mol Biol 458:137–158

Baxter AD, Lockey PM (2001) 'Hit' to 'lead' and 'lead' to 'candidate' optimization using multi-parametric principles. Drug Discov World 2:9–15

Berk RA (2003a) Simple linear regression. Regression analysis: a constructive critique. SAGE Publications Ltd, London, pp 21–38

Berk RA (2003b) Some popular extensions of multiple regression. Regression analysis: a constructive critique. SAGE Publications Ltd, London, pp 125–150

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242

Boyd MR, Paull KD (1995) Some practical considerations and applications of the National Cancer Institute in vitro anticancer drug discovery screen. Drug devel Res 34:91–109

Breu B, Silber K, Gohlke H (2007) Consensus adaptation of fields for molecular comparison (AFMoC) models incorporate ligand and receptor conformational variability into tailor-made scoring functions. J Chem Inf Model 47:2383–2400

Bruneau P, McElroy NR (2006) log$D_{7.4}$ modeling using Bayesian regularized neural networks. Assessment and correction of the errors of prediction. J Chem Inf Model 46:1379–1387

Collins FS, Patrinos A, Jordan E, Chakravati A, Gesteland R, Walters L (1998) New goals for the US human genome project: 1998–2003. Science 282:682–689

Colosi LM, Huang Q, Weber WJ Jr (2010) QSAR-assisted design of an environmental catalyst for enhanced estrogen remediation. Chemosphere 81:897–903

Datar PA, Khedkar SA, Malde AK, Coutinho EC (2006) Comparative residue interaction analysis (CoRIA): a 3D-QSAR approach to explore the binding contributions of active site residues with ligands. J Comput Aided Mol Des 20:343–360

Dhaked DK, Verma J, Saran A, Coutinho EC (2009) Exploring the binding of HIV-1 integrase inhibitors by comparative residue interaction analysis (CoRIA). J Mol Model 15:233–245

Dunn WJ III, Rogers D (1996) Genetic partial least squares in QSAR. In: Devillers J (ed) Genetic algorithms in molecular modeling. Academic Press, London, pp 109–130

Dunteman GH (1989a) Basic concepts of principal components analysis. In: Dunteman GH (ed) Principal components analysis. SAGE Publications Ltd, London, pp 15–22

Dunteman GH (1989b) Uses of principal components in regression analysis. In: Dunteman GH (ed) Principal components analysis. SAGE Publications Ltd., London, pp 65–74

Fells JI, Tsukahara R, Liu J, Tigyi G, Parrill AL (2010) 2D binary QSAR modeling of LPA3 receptor antagonism. J Mol Graph Model 28:828–833

Free SM Jr, Wilson JW (1964) A mathematical contribution to structure–activity studies. J Med Chem 7:395–399

Funar-Timofei S, Fabian WMF, Kurunczi L, Goodarzi M, Ali ST, Heyden YV (2012) Modelling heterocyclic azo dye affinities for cellulose fibres by computational approaches. Dyes Pigm 94:278–289

Gieleciak R, Polanski J (2007) Modeling robust QSAR. 2. Iterative variable elimination schemes for CoMSA: application for modeling benzoic acid pKa values. J Chem Inf Model 47:547–556

Guido RVC, Oliva G, Andricopulo AD (2008) Virtual screening and its integration with modern drug design technologies. Curr Med Chem 15:37–46

Hahn M (1995) Receptor surface models. 1. Definition and construction. J Med Chem 38:2080–2090

Hammett LP (1935) Some relations between reaction rates and equilibrium constants. Chem Rev 17:125–136

Hammett LP (1937) The effect of structure upon the reactions of organic compounds benzene derivatives. J Am Chem Soc 59:96–103

Hansch C (1969) Quantitative approach to biochemical structure-activity relationships. Acc Chem Res 2:232–239

Hansch C, Fujita T (1964) p-σ-π Analysis. A method for the correlation of biological activity and chemical structure. J Am Chem Soc 86:1616–1626

Hansch C, Gao H (1997) Comparative QSAR: radical reactions of benzene derivatives in chemistry and biology. Chem Rev 97:2995–3060

Hopfinger A (2001) 4D-QSAR package user's manual 3.0. The Chem21 Group Inc, Lake Forest

Hopfinger AJ, Tokarski JS (1997) Three-dimensional quantitative structure-activity relationship analysis. In: Charifson P (ed) Practical application of computer-aided drug design. Marcel Dekker, New York, pp 105–164

Hopfinger A, Wang S, Tokarski J, Jin B, Albuquerque M, Madhav P, Duraiswami C (1997) Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. J Am Chem Soc 119:10509–10524

In Y, Chai HH, No KT (2005) A partition coefficient calculation method with the SFED model. J Chem Inf Model 45:254–263

Ivanciuc O, Ivanciuc T, Cabrol-Bass D (2000) 3D quantitative structure activity relationships with CoRSA. Comparative receptor surface analysis. Application to calcium channel agonists. Analysis 28:637–642

Jojart B, Martinek TA, Marki A (2005) The 3D structure of the binding pocket of the human oxytocin receptor for benzoxazine antagonists, determined by molecular docking, scoring functions and 3D-QSAR methods. J Comput Aided Mol Des 19:341–356

Jorgensen WL, Duffy EM (2002) Prediction of drug solubility from structure. Adv Drug Deliv Rev 54:355–366

Katritzky AR, Lobanov VS, Karelson M (1995) QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. Chem Soc Rev 24:279–287

Kennedy T (1997) Managing the drug discovery/development interface. Drug Discov Today 2:436–444

Khedkar SA, Malde AK, Coutinho EC (2007) Design of inhibitors of the MurF enzyme of *Streptococcus pneumoniae* using docking, 3DQSAR, and de novo design. J Chem Inf Model 47:1839–1846

Kim KH (1995) Comparative molecular field analysis (CoMFA). In: Dean PM (ed) Molecular similarity in drug design. Blackie academic and professional, Glasgow, pp 291–331

Koo HM, Monks A, Mikheev A, Rubinstein LV, Gray-Goodrich M, McWilliams MJ, Alvord WG, Oie HK, Gazdar AF, Paull KD, Zarbl H, Vande-Woude GF (1996) Enhanced sensitivity to 1-beta-D-arabinofuranosylcytosine and topoisomerase II inhibitors in tumor cell lines harboring activated RAS oncogenes. Cancer Res 56:5211–5216

Korhonen SP, Tuppurainen K, Laatikainen R, Perakyla M (2003) FLUFF-BALL, a template-based grid-independent superposition and QSAR technique: validation using a benchmark steroid data set. J Chem Inf Comput Sci 43:1780–1793

Kotani T, Higashiura K (2004) Comparative molecular active site analysis (CoMASA). 1. An approach to rapid evaluation of 3D QSAR. J Med Chem 47:2732–2742

Krasowski MD, Hong X, Hopfinger AJ, Harrison NL (2002) 4D-QSAR analysis of a set of propofol analogues: mapping binding sites for an anesthetic phenol on the GABA (A) receptor. J Med Chem 45:3210–3221

Kubinyi H (1976) Quantitative structure-activity relationships. IV. Non-linear dependence of biological activity on hydrophobic character: a new model. Arzneimittelforschung 26:1991–1997

Kubinyi H (2004) 2D QSAR models: Hansch and Free-Wilson analyses. In: Bultinck P, Winter HD, Langenaeker W, Tollenaere JP (eds) Computational medicinal chemistry for drug discovery. Marcel Dekker, New York, pp 539–570

Labute P (1999) Binary QSAR: a new method for the determination of quantitative structure activity relationships. Pac Symp Biocomput 1999:444–455

Lee PH, Ayyampalayam SN, Carreira LA, Shalaeva M, Bhattachar S, Coselmon R, Poole S, Gifford E, Lombardo F (2007) In silico prediction of ionization constants of drugs. Mol Pharm 4:498–512

Lee AC, Yu JY, Crippen GM (2008) pKa prediction of monoprotic small molecules the SMARTS way. J Chem Inf Model 48:2042–2053

Liu HY, Liu SS, Qin LT, Mo LY (2009) CoMFA and CoMSIA analysis of 2,4-thiazolidinediones derivatives as aldose reductase inhibitors. J Mol Model 15:837–845

Liu J, Li Y, Zhang S, Xiao Z, Ai C (2011a) Studies of new fused benzazepine as selective dopamine D3 receptor antagonists using 3D-QSAR, molecular docking and molecular dynamics. Int J Mol Sci 12:1196–1221

Liu J, Zhang H, Xiao Z, Wang F, Wang X, Wang Y (2011b) Combined 3D-QSAR, molecular docking and molecular dynamics study on derivatives of peptide epoxyketone and tyropeptin-boronic acid as inhibitors against the β5 subunit of human 20S proteasome. Int J Mol Sci 12:1807–1835

Lombardino JG, Lowe JA (2004) The role of the medicinal chemist in drug discovery-then and now. Nat Rev Drug Discov 3:853–862

Lowe EW Jr, Ferrebee A, Rodriguez AL, Jeffrey-Connc P, Meiler J (2010) 3D-QSAR CoMFA study of benzoxazepine derivatives as mGluR5 positive allosteric modulators. Bioorg Med Chem Lett 20:5922–5924

Luo X, Shu S, Wang Y, Liu J, Yang W, Lin Z (2012) 3D-QSAR studies of dihydropyrazole and dihydropyrrole derivatives as inhibitors of human mitotic kinesin Eg5 based on molecular docking. Molecules 17:2015–2029

Magdziarz T, Lozowicka B, Gieleciak R, Bak A, Polanski J, Chilmonczyk Z (2006) 3D QSAR study of hypolipidemic asarones by comparative molecular surface analysis. Bioorg Med Chem 14:1630–1643

Magdziarz T, Mazur P, Polanski J (2009) Receptor independent and receptor dependent CoMSA modeling with IVE-PLS: application to CBG benchmark steroids and reductase activators. J Mol Model 15:41–51

Manly CJ, Louise-May S, Hammer JD (2001) The impact of informatics and computational chemistry on synthesis and screening. Drug Discov Today 6:1101–1110

Matyus P, Borosy AP (1998) Three dimensional structure-activity relationships. Acta Pharm Hung 68:33–38

Mills EJ (1884) On melting point and boiling point as related to composition. Philos Mag 17:173–187

Muller G (2003) Medicinal chemistry of target family-directed masterkeys. Drug Discov Today 8:681–691

Murugesan V, Prabhakar YS, Katti SB (2009) CoMFA and CoMSIA studies on thiazolidin-4-one as anti-HIV-1 agents. J Mol Graph Model 27:735–743

Nakao K, Fujikawa M, Shimizu R, Akamatsu M (2009) QSAR application for the prediction of compound permeability with in silico descriptors in practical use. J Comput Aided Mol Des 23:309–319

Natesan S, Wang T, Lukacova V, Bartus V, Khandelwal A, Subramaniam R, Balaz S (2012) Cellular quantitative structure-activity relationship (Cell-QSAR): conceptual dissection of receptor binding and intracellular disposition in antifilarial activities of selwood antimycins. J Med Chem 55:3699–3712

Noolvi MN, Patel HM, Bhardwaj V (2010) 2D QSAR studies on a series of 4-anilino quinazoline derivatives as tyrosine kinase (EGFR) inhibitor: an approach to design anticancer agents. Dig J Nanomater Bios 5:387–401

Noolvi MN, Patel HM, Bhardwaj V (2011) A comparative QSAR analysis of quinazoline analogues as tyrosine kinase (erbB-2) inhibitors. Med Chem 7:200–212

Norinder U, Bergstrom CAS (2006) Prediction of ADMET properties. ChemMedChem 1:920–937

Oprea TI (2004) 3D QSAR modeling in drug design. In: Bultinck P, Winter HD, Langenaeker W, Tollenaere JP (eds) Computational medicinal chemistry for drug discovery. Marcel Dekker, New York, pp 571–616

Palmer DS, O'Boyle NM, Glen RC, Mitchell JBO (2007) Random forest models to predict aqueous solubility. J Chem Inf Model 47:150–158

Paull KD, Shoemaker RH, Hodes L, Monks A, Scudiero DA, Rubinstein L, Plowman J, Boyd MR (1989) Display and analysis of patterns of differential activity of drugs against human tumor cell lines: development of mean graph and COMPARE algorithm. J Natl Cancer Inst 81:1088–1092

Polanski J, Walczak B (2000) The comparative molecular surface analysis (COMSA): a novel tool for molecular design. Comput Chem 24:615–625

Polanski J, Gieleciak R, Bak A (2002) The comparative molecular surface analysis (COMSA) a nongrid 3D QSAR method by a coupled neural network and PLS system: predicting *p*K(a) values of benzoic and alkanoic acids. J Chem Inf Comput Sci 42:184–191

Robinson DD, Winn PJ, Lyne PD, Richards WG (1999) Self-organizing molecular field analysis: a tool for structure-activity studies. J Med Chem 42:573–583

Rogers D, Hopfinger AJ (1994) Application of genetic function approximation to quantitative structure-activity relationships and

quantitative structure-property relationships. J Chem Inf Comput Sci 34:854–866

Salum LB, Andricopulo AD (2009) Fragment-based QSAR: perspectives in drug design. Mol Divers 13:277–285

Santos-Filho OA, Hopfinger AJ, Cherkasov A, De-Alencastro RB (2009) The receptor-dependent QSAR paradigm: an overview of the current state of the art. Med Chem 5:359–366

Stewart L, Clark R, Behnke C (2002) High-throughput crystallization and structure determination in drug discovery. Drug Discov Today 7:187–196

Taft RW (1952) Polar and steric substituent constants for aliphatic and o-benzoate groups from rates of esterification and hydrolysis of esters 1. J Am Chem Soc 74:3120–3128

Thipnate P, Liu J, Hannongbua S, Hopfinger AJ (2009) 3D pharmacophore mapping using 4D QSAR analysis for the cytotoxicity of lamellarins against human hormone-dependent T47D breast cancer cells. J Chem Inf Model 49:2312–2322

Tian Y, Xu J, Li Z, Zhu Z, Zhang J, Wu S (2011) Combined 3D-QSAR and docking modelling study on indolocarbazole series compounds as Tie-2 inhibitors. Int J Mol Sci 12:5080–5097

Ul-Haq Z, Wadood A, Uddin R (2009) CoMFA and CoMSIA 3D-QSAR analysis on hydroxamic acid derivatives as urease inhibitors. J Enzy Inhib Med Chem 24:272–278

Van Daele I, Munier-Lehmann H, Froeyen M, Balzarini J, Van Calenbergh S (2007) Rational design of 5′-thiourea-substituted alpha-thymidine analogues as thymidine monophosphate kinase inhibitors capable of inhibiting mycobacterial growth. J Med Chem 50:5281–5292

Verma J, Khedkar VM, Prabhu AS, Khedkar SA, Malde AK, Coutinho EC (2008) A comprehensive analysis of the thermodynamic events involved in ligand-receptor binding using CoRIA and its variants. J Comput Aided Mol Des 22:91–104

VLifeMDS 3.0 (2007) Molecular design suite developed by VLife Sciences Technologies Pvt Ltd. VLife Sciences Technologies Pvt Ltd., Pune

Walczak B, Massart DL (2000) Local modeling with radial basis function networks. Chemom Intell Lab Syst 50:179–198

Walpole CS, Wrigglesworth R, Bevan S, Campbell EA, Dray A, James IF, Masdin KJ, Perkins MN, Winter J (1993) Analogues of capsaicin with agonist activity as novel analgesic agents; structure-activity studies 3. The hydrophobic side-chain "C-region". J Med Chem 36:2381–2389

Wang J, Krudy G, Hou T, Zhang W, Holland G, Xu X (2007) Development of reliable aqueous solubility models and their application in drug like analysis. J Chem Inf Model 47:1395–1404

Weinstein JN, Myers T, Buolamwini J, Raghavan K, Van Osdol W, Licht J, Viswanadhan VN, Kohn KW, Rubinstein LV, Koutsoukos AD, Monks A, Scudiero DA, Anderson NL, Zaharevitz D, Chabner BA, Grever MR, Paull KD (1994) Predictive statistics and artificial intelligence in the US National Cancer Institute's drug discovery program for cancer and AIDS. Stem Cells 12:13–22

Wise M, Cramer RD, Smith D, Exman I (1983) Progress in three-dimensional drug design: the use of real time colour graphics and computer postulation of bioactive molecules in DYLOMMS. In: Dearden J (ed) Quantitative approaches to drug design. Elsevier, Amsterdam, pp 145–146

Wold S, Johansson E, Cocchi M (1993) PLS: partial least squares projections to latent structures. In: Kubinyi H (ed) 3D QSAR in drug design: theory, methods and applications. ESCOM Science Publishers, Leiden, pp 523–550

Yadav DK, Meena A, Srivastava A, Chanda D, Khan F, Chattopadhyay SK (2010) Development of QSAR model for immunomodulatory activity of natural coumarinolignoids. Drug Des Devel Ther 4:173–186

Yang S, Chang S, Chen H, Chen CY (2011) Identification of potent EGFR inhibitors from TCM Database@Taiwan. PLoS Comput Biol 7:e1002189

Zhang B, Li Y, Zhang H, Ai C (2010) 3D-QSAR and molecular docking studies on derivatives of MK-0457, GSK1070916 and SNS-314 as inhibitors against Aurora B Kinase. Int J Mol Sci 11:4326–4347

Zhao H (2007) Scaffold selection and scaffold hopping in lead generation: a medicinal chemistry perspective. Drug Discov Today 12:149–155

Zhao YH, Abraham MH, Ibrahim A, Fish PV, Cole S, Lewis ML, De Groot MJ, Reynolds DP (2007) Predicting penetration across the blood-brain barrier from simple descriptors and fragmentation schemes. J Chem Inf Model 47:170–175

Zhou P, Tong J, Tian F, Li Z (2006) A novel comparative molecule/pseudo receptor interaction analysis. Chin Sci Bull 51:1824–1829

Zhu YQ, Lei M, Lu AJ, Zhao X, Yin XJ, Gao QZ (2009) 3D-QSAR studies of boron-containing dipeptides as proteasome inhibitors with CoMFA and CoMSIA methods. Eur J Med Chem 44:1486–1499