3D-QSAR in Drug Design - A Review

Jitender Verma, Vijay M. Khedkar and Evans C. Coutinho*

Department of Pharmaceutical Chemistry, Bombay College of Pharmacy, Kalina, Santacruz (E), Mumbai 400 098, India

Abstract: Quantitative structure–activity relationships (QSAR) have been applied for decades in the development of relationships between physicochemical properties of chemical substances and their biological activities to obtain a reliable statistical model for prediction of the activities of new chemical entities. The fundamental principle underlying the formalism is that the difference in structural properties is responsible for the variations in biological activities of the compounds. In the classical QSAR studies, affinities of ligands to their binding sites, inhibition constants, rate constants, and other biological end points, with atomic, group or molecular properties such as lipophilicity, polarizability, electronic and steric properties (Hansch analysis) or with certain structural features (Free-Wilson analysis) have been correlated. However such an approach has only a limited utility for designing a new molecule due to the lack of consideration of the 3D structure of the molecules. 3D-QSAR has emerged as a natural extension to the classical Hansch and Free-Wilson approaches, which exploits the three-dimensional properties of the ligands to predict their biological activities using robust chemometric techniques such as PLS, G/PLS, ANN etc. It has served as a valuable predictive tool in the design of pharmaceuticals and agrochemicals. Although the trial and error factor involved in the development of a new drug cannot be ignored completely, QSAR certainly decreases the number of compounds to be synthesized by facilitating the selection of the most promising candidates. Several success stories of QSAR have attracted the medicinal chemists to investigate the relationships of structural properties with biological activity. This review seeks to provide a bird's eye view of the different 3D-QSAR approaches employed within the current drug discovery community to construct predictive structure–activity relationships and also discusses the limitations that are fundamental to these approaches, as well as those that might be overcome with the improved strategies. The components involved in building a useful 3D-QSAR model are discussed, including the validation techniques available for this purpose.

INTRODUCTION

Quantitative structure–activity relationship (QSAR), in simplest terms, is a method for building computational or mathematical models which attempts to find a statistically significant correlation between structure and function using a chemometric technique. In terms of drug design, structure here refers to the properties or descriptors of the molecules, their substituents or interaction energy fields, function corresponds to an experimental biological/biochemical end-point like binding affinity, activity, toxicity or rate constants, while chemometric method include MLR, PLS, PCA, PCR, ANN, GA *etc*. The term 'quantitative structure–property relationship' (QSPR) is used when some property other than the biological activity is concerned. Various QSAR approaches have been developed gradually over a time span of more than a hundred years and served as a valuable predictive tool, particularly in the design of pharmaceuticals and agrochemicals. The methods have evolved from Hansch and Free-Wilson's one or two-dimensional linear free-energy relationships, via Crammer's three-dimensional QSAR to Hopfinger's fourth and Vedani's fifth and sixth-dimensions. All one and two dimensional and related methods are commonly referred to as 'classical' QSAR methodologies, and have been discussed briefly in the later sections. Irrespective of the type, all QSAR formalisms presume that every molecule included in the study binds to the same site of the same target receptor. However, the main difference between all these formalisms reside in the manner in which each one of them treats and represents structural properties of the molecules and extracts the quantitative relationships between the properties and activities. Due to the limited scope and space for this review, the author will focus only on the 3D-QSAR approaches in drug design. To present an overview of the scenario, remaining QSAR methodologies have just been outlined in brief. However, readers are recommended to go through the references provided for these methods, to comprehend them in detail.

OBJECTIVES OF QSAR

Mostly all the QSAR methods focus on the following goals:

- To quantitatively correlate and recapitulate the relationships between trends in chemical structure alterations and respective changes in biological endpoint for comprehending which chemical properties are most likely determinants for their biological activities
- To optimize the existing leads so as to improve their biological activities
- To predict the biological activities of untested and sometimes yet unavailable compounds

RATIONALE BEHIND QSAR MODELING

The extent of reliability in opting for QSAR modeling depends on the type or nature of property being predicted,

*Address correspondence to this author at the Department of Pharmaceutical Chemistry, Bombay College of Pharmacy, Kalina, Santacruz (E), Mumbai 400 098, India; Tel: +91-22-26670905; Fax: +91-22-26670816; E-mail: evans@bcpindia.org, evans@bcp.edu.in

1568-0266/10 $55.00+.00

the stage of the project and the relative ease and cost of compound synthesis and subsequent testing. More often QSAR models provide useful predictions but many times they fail, despite of good statistics generated from internal data used in training. Regardless of all such problems, QSAR becomes a useful alternative because of the following reasons:

- Conventional syntheses methods are expensive and time-consuming
- Biological assays are also too costly, often requiring time, sacrifice of animals, or compounds in their pure forms
- Drug failures due to poor ADMET profiles at later stages of development (or even after commercialization) are exceedingly expensive and painful
- Large number of compounds are available due to combinatorial chemistry and HTS approaches, but estimations are required for prioritization of synthesis and screening

EVOLUTION OF QSAR

QSAR methods originated way back in the nineteenth century. The chronological order in which these methods (1D and 2D QSAR) have evolved over a period of time is being given in Table **1**.

CLASSIFICATION OF QSAR METHODOLOGIES

Based on dimensionality - Most often the QSAR methods are categorized into following classes, based on the structural representation or the way by which the descriptor values are derived:

1D-QSAR correlating activity with global molecular properties like pK_a, log P *etc*.

2D-QSAR correlating activity with structural patterns like connectivity indices, 2D-pharmacophores *etc*., 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 the type of chemometric methods used – Sometimes QSAR methods are also classified into following two categories, depending upon the type of correlation technique employed to establish a relationship between structural properties and biological activity:

Linear methods including linear regression (LR), multiple linear regression (MLR), partial least-squares (PLS), and principal component analysis/regression (PCA/ PCR).

Non-linear methods consisting of artificial neural networks (ANN), *k*-nearest neighbors (*k*NN), and Bayesian neural nets

LIMITATIONS OF CLASSICAL QSAR METHODOLOGIES

Classical QSAR methods are much simpler, faster and more amenable to automation than 3D-QSAR approaches. They include clearly-defined physiochemical descriptors and are best suited for the analysis of large number of compounds and computational screening of molecular databases. Though they have been used for decades to correlate and predict the activity of molecules, they suffer from serious limitations in certain situations some of which are as follows [24]:

- Only 2D-structures considered
- Unavailability of appropriate physiochemical parameter (*e.g.*, numerical descriptors for new or unusual substituents), rendering the compound unfit for inclusion in QSAR analysis
- Insufficient parameters for describing drug-receptor interactions (*e.g.*, steric parameter E_s, Hammett constant σ *etc*.)
- Confined to only few substitutions in a common reference structure (simple variation of aromatic substituents) and works best with a congeneric series
- No representation of stereochemistry or 3D-structure of molecules, regardless of their availability
- Provide no unique solutions
- Higher risk of chance correlations
- High risk of failure due to 'too far outside' predictions
- No graphical output thereby making the interpretation of results in familiar chemical terms, frequently difficult if not impossible
- Requires considerable knowledge of substituent constants in physical organic chemistry to design a molecule, since classical QSAR equation do not directly suggest new compounds to synthesize

PROGRESS IN 3D-QSAR APPROACHES

3D-QSAR is a broad term encompassing all those QSAR methods which correlate macroscopic target properties with computed atom-based descriptors derived from the spatial (three-dimensional) representation of the molecular structures [25-29]. The methodology has emerged as a natural extension to the classical QSAR approaches pioneered by Hansch and Free-Wilson. The major drawback of 3D-QSAR techniques is that they all are based on various assumptions which are described in the subsequent section [29].

ASSUMPTIONS IN 3D-QSAR METHODS

No QSAR model can replace the experimental assays, though experimental techniques are also not free from errors. Because of many obvious problems in simulating the real-world situations, not every *in vivo* parameter can be included in the QSAR modeling. However, every attempt is being made to develop a model as close as possible to the real one and for this the 3D-QSAR paradigm has to rely on some basic assumptions which are given below [29]:

Table 1. A Brief History of Earlier QSAR Methodologies

Author (Year)	Contributions/Postulates
Crum-Brown and Fraser (1868)	Physiological activities of substances could be correlated with their chemical composition and constitution, but they did not show how to represent the chemical structure in a quantitative manner [1]
Richardson (1868)	Expressed the chemical structure as a function of solubility [2]
Mills (1884)	Developed a QSPR model for the prediction of melting and boiling points in homologous series, results were accurate to better than one degree [3]
Richet (1893)	Correlated toxicities of a set of alcohols, ethers and ketones with aqueous solubility and showed that their cytotoxicities are inversely related to their corresponding water solubilities [4]
Overton and Meyer (1897, 1899)	Correlated partition coefficients of a group of organic compounds with their anesthetic potencies and concluded that narcotic (depressant) activity is dependent on the lipophilicity of the molecules [5, 6]
Hammett (1935, 1937)	Correlated the effect of the addition of a substituent on benzoic acid with the dissociation constant, postulated electronic sigma-rho constants and established the linear free-energy relationship (LFER) principle [7, 8]
Ferguson (1939)	Correlated depressant action with the relative saturation of volatile compounds in their vehicle and introduced a thermodynamic generalization to the toxicity [9]
Bell and Roblin (1942)	Studied antibacterial activities of a series of sulfanilamides in terms of their ionizations [10]
Albert (1948)	Examined the effects of ionization/electron distribution and steric access on the potencies of a multitude of aminoacridines [11]
Taft (1952)	Postulated a method for separating polar, steric, and resonance effects and introduced the first steric parameter, E_S [12]
Hansch and Muir (1962)	Correlated the biological activities of plant growth regulators with Hammett constants and hydrophobicity [13]
Hansch and Fujita (1964)	Combined the hydrophobic constants with Hammett's electronic constants to yield the linear Hansch equation and its many extended forms [14]
Hansch (1969)	Developed the parabolic Hansch equation for dealing with extended hydrophobicity ranges [15]
Free and Wilson (1964)	Formulated an additive model, where the activity is discretized as a simple sum of contributions from different substituents [16]
Fujita and Ban (1971)	Simplified the Free-Wilson equation estimating the activity for the non-substituted compound of the series and postulated Fujita-Ban equation that used the logarithm of activity, which brought the activity parameter in line with other free energy-related terms [17]
Kubinyi (1976)	Investigated the transport of drugs *via* aqueous and lipoidal compartment systems and further refined the parabolic equation of Hansch to develop a superior bilinear (non-linear) QSAR model [18]
Hansch and Gao (1997)	Developed comparative QSAR (C-QSAR), incorporated in the C-QSAR program [19]
Heritage and Lowis (1997)	Developed Hologram QSAR (HQSAR), where the structures are converted into all possible fragments, which are assigned specific integers, and then hashed into a fingerprint to form the molecular hologram. The bin occupancies of these holograms are used as the QSAR descriptors, encoding the chemical and topological information of molecules [20, 21]
Cho and workers (1998)	Developed Inverse QSAR, which seeks to find values for the molecular descriptors that possess a desired activity/property value. In other words, it consists of finding the optimum sets of descriptor values best matching a target activity and then generating a focused library of candidate structures from the solution set of descriptor values [22]
Labute (1999)	Developed Binary QSAR to handle binary activity measurements from high-throughput screening (*e.g.*, pass/fail or active/inactive), and molecular descriptor vectors as input. A probability distribution for actives and inactives is then determined based on Bayes' Theorem [23]

- There is an underlying relationship between molecular structure and biological activity.
- Receptor binding is directly proportional to the biological activity. Differential effects on second messengers or other signaling steps which transpire between receptor binding and experimentally observed response, are not taken into consideration.
- Molecular structure can be measured and represented with a set of numbers usually called descriptors, which

encode all physical, chemical and biological properties of the molecule.

- Molecules with common or related structures generally have similar physicochemical properties (the similarity principle), and thus have similar binding modes and consequently comparable biological activities. The reverse also holds true. Also, molecules located in the same region of the descriptor space present similar activity (the neighborhood principle).
- Structural properties which lead to an observed biological response are most commonly determined by the non-bonding (or non-covalent) forces, mainly steric and electrostatic.
- The observed biological effect is produced by the modeled ligand itself, and not by its metabolite or degradation product.
- The lowest energy conformation of the ligand is its bioactive conformation, and it is this single conformation of the ligand which exerts the binding effects.
- With few exceptions, the geometry of the receptor binding site is considered rigid.
- The loss of translational and rotational degrees of freedom (entropy) upon binding is believed to follow a similar pattern for all the molecules.
- Total number of rotatable bonds is the only method most frequently used to estimate the entropic cost for freezing non-terminal single-bond rotors.
- For all the modeled ligands, the protein binding site is assumed to be same.
- For all the modeled compounds, the on-off rate is supposed to be similar *i.e.*, the system is considered to be in equilibrium, and kinetic aspects are usually ignored.
- Some of the major factors like desolvation energetics, temperature, diffusion, transport, pH, salt concentration *etc.* which contribute to the overall free energy of binding are difficult to handle, and thus usually ignored.
- In molecular mechanics based 3D-QSAR methods, free energy of binding is largely explained by the enthalpic component (*i.e.*, the internal energy), which is prone to the inherent force field errors.
- Resulting QSAR model may represent one of potentially several solutions to the property–activity correlation problem.

CLASSIFICATION OF 3D-QSAR APPROACHES

3D-QSAR methods can be classified on various criteria, some of which are given in Table **2**.

Table 2. Classification of 3D-QSAR approaches

Classification	Examples
On the basis of intermolecular modeling, or the information used to develop QSAR	
Ligand-based 3D-QSAR	CoMFA, CoMSIA, COMPASS, GERM, CoMMA, SoMFA
Receptor-based 3D-QSAR	COMBINE, AFMoC, HIFA, CoRIA
On the basis of alignment criterion	
Alignment-dependent 3D-QSAR	CoMFA, CoMSIA, GERM, COMBINE, AFMoC, HIFA, CoRIA
Alignment-independent 3D-QSAR	COMPASS, CoMMA, HQSAR, WHIM, EVA/CoSA, GRIND
On the basis of the chemometric technique used for correlating structural properties and activities	
Linear 3D-QSAR	CoMFA, CoMSIA, AFMoC, GERM, CoMMA, SoMFA
Non-linear 3D-QSAR	COMPASS, QPLS

CoMFA

In 1987, Cramer developed the predecessor of 3D-approaches called DYnamic Lattice-Oriented Molecular Modeling System (DYLOMMS) that involves the use of PCA to extract vectors from the molecular interaction fields, which are then correlated with biological activities [30]. Soon after he modified it by combining the two existing techniques, GRID and PLS, to develop a powerful 3D-QSAR methodology, Comparative Molecular Field Analysis (CoMFA) [31]. Today CoMFA has become a prototype of 3D-QSAR methods [32]. A standard CoMFA procedure, as implemented in the Sybyl Software [33] from Tripos Inc., follows the following sequential 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 three-dimensional space which are considerably associated with the biological activity.

Several parameters which significantly control the overall performance of the developed CoMFA model are described below; many of these are applicable to other QSAR methodologies also:

BIOLOGICAL DATA

The popular 'GIGO' (garbage in garbage out) principle applies in every computational technique. In 3D-QSAR also, one should utilize accurate activity data in order to develop a good model. Though 3D-QSAR methods can be applied to heterogeneous data sets, some considerations for maintaining the accuracy of biological data are necessary [29, 34]:

- Compounds should belong to a congeneric series (more important in case of classical QSAR).
- Compounds should have the same mechanism of action and same/comparable binding mode.
- The biological activities of compounds should correlate to their binding affinity and their enumerated biological responses should be measurable
- Biological data for molecules should be obtained using uniform protocols (radioligand, activator, cofactor, pH, buffer *etc*.) and preferably from a single source (organism/tissue/cell/protein) and single lab.
- Activity data for all the compounds should be in same units of measurement (binding/functional/IC_{50}/K_i). K_i value is preferred instead of the IC_{50} data, since it is independent of the substrate concentration.
- The ranges of biological activity covered should be as large as possible, keeping the mode of action identical. Preferably, activity range should be much larger than the standard deviations of the data; more than three logarithm units with an even spread of data is preferred.
- If possible, biological data should be symmetrically distributed around their mean, and their precision should be evenly distributed over its range of variation. If not, such skewness can be removed by log transforming the data and expressing it as log (1/C), where C refers to the molar concentration of drug producing a standard response. It is noteworthy that free energy change is proportional to the inverse log of concentration of the compound.

COMPOUND SELECTION AND SERIES OPTIMIZATION

One of the major applications of QSAR is to optimize the existing leads by structural modifications so as to improve their activity and reduce/eliminate the side-effects. However, there are many issues to be taken care of while selecting substituents for the modification of compounds; some of the important ones are given below [26, 34]:

- The compounds/substituents selected should be convincingly different from the existing ones, so as to minimize collinearity among the variables.
- The chosen compounds/substituents should have the properties which behave independent of each other, thereby maximizing dissimilarity and orthogonality.
- The selection should be done in such a manner so as to map the substituent (descriptor) space with minimum number of compounds.
- Synthetic accessibility/feasibility of the selected compounds should also be taken into consideration.

OPTIMIZATION OF 3D-STRUCTURE OF THE MOLECULES

An important issue in 3D-QSAR is how to generate and represent the starting molecular structure for analysis? The problem can be resolved both by experimental as well as computational techniques [34]. A large number of well-resolved experimentally determined crystal structures are available in databases like Cambridge Structural Database [35] and Protein Data Bank [36]. The crystal structures offer the advantage that some conformational information about the flexible molecule is included. However, molecular modeling methods are particularly useful for compounds that have not been made or cannot even exists under normal conditions. Computationally the 3D-structures can be generated by three methods: (a) manually by sketching the structures interactively in a 3D-computer graphics interface or from an existing 3D-structure included in the fragment libraries, (b) numerically by using mathematical techniques like distance geometry, quantum or molecular mechanics, and (c) by automatic methods that are often used for building 3D-structure databases.

After the generation of starting 3D-molecular structures, their geometries are refined by minimizing their conformational energies using theoretical calculation methods. Commonly used structure optimization techniques include (a) molecular mechanics methods which usually does not explicitly consider the electronic motion, and thus are fast, reasonably accurate and can be used for very large molecules like enzymes, (b) quantum mechanics or *ab initio* methods which takes into account the 3D-distribution of electrons around the nuclei, and therefore are extremely accurate but time consuming, computationally intensive and cannot handle large molecules, (c) semi-empirical methods which are basically quantum mechanical in nature but employs an extensive use of approximations as in molecular mechanics. Generally, the molecular geometry is optimized by molecular mechanics methods, and its atomic charges are calculated mostly by semi-empirical methods or less frequently by *ab initio* methods.

CONFORMATIONAL ANALYSIS OF MOLECULES

It is a well recognized fact that each compound containing one or more single bonds is existing at each moment in many different so-called rotamers or conformers. Although small molecules may have only a single lowest energy conformation but large and flexible molecules do exists in multiple conformations at physiological conditions. Therefore, it becomes necessary to include various such conformations of the molecules in a 3D-QSAR study [34]. Depending upon the type of molecules in the study, any of

the following conformational search methods can be adopted:

- *Systematic search (or grid search)* method which generates all possible conformations, by systematically varying each of the torsion angles of a molecule by some increment, keeping the bond lengths and bond angles fixed.
- *Random search* method which generates a set of conformations by repetitively and randomly changing either the Cartesian (x, y, z) or the internal (bond lengths, bond angles and torsion/dihedral angles) coordinates of a starting geometry of the molecule under consideration.
- *Monte Carlo* method which simulates dynamic behavior of a molecule and generates the conformations by making random changes in its structure, calculating and comparing its energy with that of the previous conformation and accepting it if it is unique.
- *Molecular dynamics* method which employs the Newton's second law of motion (force = mass × acceleration) to simulate the time-dependent movements and conformational changes in a molecular system, and results in a so-called trajectory showing how the positions and velocities of atoms in the molecular system vary with time.
- *Simulated annealing* which heats up the molecular system under consideration to high temperatures to overcome huge energy barriers, and after equilibrating there for sometime using molecular dynamics, cools down the system slowly and gradually to obtain the low energy conformations according to the Boltzmann distribution.
- *Distance geometry algorithm* which generates a random set of coordinates by selecting random distances within each pair of upper and lower bounds to form constraints in a distance matrix, which are the used to generate energetically feasible conformations of a set of molecules.
- *Genetic and evolutionary algorithms* which are based on the concept of biological evolution and works by first creating a population of possible solutions to the problem. The solutions with best fitness scores undergo crossovers and mutations over a time, and propagate their good characteristics down the generations to result in better solutions in the form of new conformers.

DETERMINING BIOACTIVE CONFORMATIONS OF MOLECULES

Bioactive conformation refers to that conformation of the molecule when it is bound to the receptor. Intrinsic forces between the atoms in the molecule as well as extrinsic forces between the molecule and its surrounding environment significantly influence the bioactive conformation of the molecule. Reliability of any 3D-QSAR methodology depends on the determination of bioactive conformations [25, 34]. Bioactive conformations of the molecules can be obtained both by experimental as well as theoretical techniques. Experimental methods for establishing bioactive conformations include:

X-ray crystallography: Exact 3D-structure of the macromolecules can be obtained only by this technique. Drug-receptor complexes generated by X-ray crystallography provide reasonably accurate information, but this method has several limitations like

- The protein needs to be crystallized and the constitution of crystallizing media is not usually similar to the physiological conditions
- The method produces a time-averaged structure, since data collection usually takes a long time
- Many times the structures are distorted due to crystal packing
- Because of crystal instability and active-site occlusion, it is often not possible to diffuse substrates or other biologically relevant molecules into the existing crystals
- Positions of hydrogen atoms are difficult to be resolved
- Errors in accurately determining the structure of the ligand

NMR spectroscopy: In this method the 3D-structural data is obtained in solution. It is a method of choice when the molecule cannot be crystallized through experimental ways, as in case of the membrane bound receptors or receptors which have not yet been isolated due to stability, resolution or other issues. The important features of this method are:

- Since no protein crystallization is required, the conformation of the protein is not influenced by packing forces of the crystal environment
- The solution conditions (pH, ionic strength, Substrate, temperature *etc*.) can be adjusted to match the physiological conditions. The results are also highly dependent on the solvent
- Significant information regarding dynamic aspects of molecular motion can be obtained
- Takes less time but is suitable for small molecules only
- Positions of hydrogen atoms can be resolved
- Apolar solvents may lead to an overestimation of hydrogen bonding phenomena
- Structure obtained from NMR may not be similar to the one obtained from experimental methods, and many times it may not represent the receptor bound conformation

Theoretically 3D-structural information can be obtained by a knowledge-based approach called Protein/homology modeling: In this method, the primary sequence of new protein is compared with all sequences of structurally known proteins stored in a database like PDB. Proteins in the database which are found to be homologous to the unknown are retrieved and used as templates for the structural prediction of the unknown protein. However this approach is limited only to the target proteins that are amenable to structure determination. Also the quality and applicability of this method primarily depends on the sequence similarity

between the protein of known structure (template) and the protein to be modeled (target).

ALIGNMENT OF MOLECULES

One of the most crucial problems in most of the alignment-based 3D-QSAR methods is that their results are highly sensitive to the manner in which the bioactive conformations of all the molecules are superimposed over each other [25, 34]. In cases, where all the molecules in a data set have a common rigid core structure, molecules can be aligned easily using least-square fitting procedure. However in case of structural heterogeneity in the dataset, alignment of highly flexible molecules becomes quite difficult and time consuming. 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 involves corresponding atom to atom pairing between the molecules. It is also called as the pharmacophore approach and is the most popular method, since it gives the best matching of the preselected atom positions. It is beneficial in identifying dissimilarity between similar molecules, but cannot be applied to molecules with different structural types where corresponding atoms are difficult to select.

Binding sites based superimposition: In this method, molecular alignment is obtained by superimposing the receptor active sites or the receptor residues that interact with the ligands. This approach is believed to be more conceivable, despite problems in conformational analysis due to enhanced degrees of freedom.

Fields/pseudofields based superimposition: This method perform superimposition by comparing the similarities in the calculated interaction energy fields between the molecules. Electrostatic similarity and molecular surface similarity indices have also been used by the researchers for molecular alignment.

Pharmacophore based superimposition: This method uses a hypothetical pharmacophore as a useful common target template. Each molecule is conformationally directed to assume the shape obligatory for its sub-molecular features to match with either a known pharmacophore or the one which is generated during the conformational analysis.

Multiple conformers based superimposition: This method is particularly useful in cases where the ligands may bind to a receptor in multiple ways, or when the correct binding mode is unknown and the ligands have a fair degree of conformational flexibility. For example, the 3D-QSAR method COMPASS (described in later section) iteratively determines and selects the best bioactive conformation and optimal alignment from a set of initial poses.

CALCULATION OF MOLECULAR INTERACTION ENERGY FIELDS

After superimposition, the overlaid set of molecules is positioned in the center of a lattice or grid box, to calculate interaction energies between the ligands and different probe atoms placed at each intersection of the lattice [34, 37]. Various aspects that are required to be taken care of while calculating the interaction energies in CoMFA methodology are as follows:

- The standard size of the grid spacing is 2 Å. The grid spacing is inversely proportional to the rigorousness of calculations. As the grid spacing decreases to 1 Å or less, the calculations becomes more intensive requiring much more computing time and disc storage space. The reduced grid spacing (0.5 Å) is usually employed while extracting interaction energy fields for a reference (most active) compound during molecular superimposition based on fields, as described earlier.
- The typical size of the grid box is 3 - 4 Å larger than the union surface of the overlaid molecules. Since the electrostatic/Coulombic interactions are long-rang in nature, a larger grid box may be needed. Due to inherent correlation between electrostatic energies among lattice points in close proximity, a similar size grid box can be used for steric/van der Waals interactions.
- Many times the position of the grid box considerably influences the statistics particularly the number of components in the final CoMFA model. Generally, the initial models are developed at various locations to spot the best grid position. Two approaches have been proposed to reduce the instability. The first one suggests rotating the set of overlaid molecules in a manner that they are not parallel to any of the grid edges. The second strategy recommends substituting the field value at a lattice point by average of the field values at the vertices of a cube centered on the grid point, whose side length is two-thirds of the grid spacing.
- In CoMFA, the interaction energies are calculated using probes. The probe may be a small molecule like water, or a chemical fragment such as a methyl group. The electrostatic energies are calculated with H^+ probe, whereas a sp^3 hybridized carbon atom with an effective radius of 1.53 Å and a +1.0 charge is used as probe for including the steric energies. Each probe is positioned in turn at every intersection point of the lattice, and the interaction energies between the probe and each of the compounds are calculated using different molecular force fields.
- A force field is a mathematical equation, which using a combination of bond lengths, bond angles, dihedral angles, interatomic distances along with coordinates and other parameters, empirically fit the potential energy surface. Major forces encountered in the drug-receptor intermolecular interactions include electrostatic/ Coulombic, hydrogen bonding, steric/van der Waals and hydrophobic. The electrostatic and hydrogen bonding interactions are responsible for ligand-receptor specificity, whereas hydrophobic interactions generally provide the strength for binding. The most commonly employed fields in CoMFA are steric and electrostatic, which are mainly enthalpic in nature. However, many times the entropic effects, in the form of hydrophobic interactions, are also included in the CoMFA analysis. Creativity of the research and the validity of the underlying theory are the major parameters deciding the type of field to be generated and included in a CoMFA model.

- In CoMFA, the standard (6-12) Lennard-Jones function is used to model the van der Waals interactions whereas electrostatic interactions are determined by the Coulomb's law. The slope of the Lennard-Jones potentials is very steep close to the van der Waals surface, as a result of which the potential energy at lattice points in the proximity of the surface changes significantly. This implies that a trivial difference in the mutual shift or conformational changes of the compounds may result in very large differences in energy values. Moreover, the Lennard-Jones and Coulombic potentials show singularities (unacceptably large values) at the atomic positions. Therefore to avoid all these problems in CoMFA, the cut-off values (± 30 kcal/mol) for steric and electrostatic energy are defined.

DATA PRETREATMENT AND SCALING

Before performing the actual chemometric analysis in 3D-QSAR, the raw data is usually pretreated to minimize redundancy [34]. One of the common reduction methods is based on the standard deviation cut-off, in which all the energy columns with a low standard deviation are eliminated from the data, since they require longer computing time without contributing significantly to the results. Similarly several variable selection methods are available, which can be used to reduce collinearity among the descriptors thereby preventing data over-fitting and improving the prediction performance of the model. Also, in CoMFA the steric and electrostatic values are amended by using cut-offs (± 30 kcal/mol, as mentioned earlier), depending upon the position of the lattice point.

Many times after pretreatment, the data is subjected to scaling which assigns equal weight to all the descriptors and places them on a common platform for a meaningful statistical analysis. Scaling significantly improves the signal to noise ratio and also allows ranking the relative importance of individual variables. Different scaling techniques are available and can be used effectively in 3D-QSAR approaches. For example: autoscaling scales the variables to zero mean and a unit standard deviation by dividing each column with its standard deviation, block-scaling provides each category of variables with the same weight by dividing the initial autoscaling weights of descriptors in one class by the square root of the number of descriptors in that class (CoMFA standard scaling), and block-adjusted scaling which is particularly useful when other variables are included along with the energy values in the analysis. This scaling gives other variables a comparable weight to the total variables.

Sometimes the pretreated data is subjected to centering by subtracting the column means from all the data. This does not change any coefficient values or comparative weights of the descriptors, but the number of significant components from PLS may be one less than from the data without centering. The method is supposed to improve the ease of interpretation and numerical stability.

MODEL GENERATION AND VALIDATION

After pretreatment and scaling of the descriptors (interaction energies and other variables, if necessary), they are correlated to the biological activities of the molecules, assuming a linear relationship between them [34, 37, 38]. Since the number of independent (x) variables in CoMFA is much larger than the number of compounds in the data set, the traditional linear regression analysis cannot be used to perform the fitting process. Therefore to extract a stable and best QSAR model from a range of possible solutions, the partial least-squares (PLS) technique is used. Other methods to model linear relationships include MLR, PCA, PCR *etc*. However many times the relationship between the dependent (y) and independent (x) variables is not linear or it can't be predicted, in such cases non-linear chemometric methods like neural networks are employed; these methods make no assumption about the relationship between the variables during training and model development. Most of these chemometric techniques for QSAR modeling are discussed in the later sections.

The most important criterion for judging the quality of a QSAR model is its ability to predict accurately not only the activities of molecules that form part of training set (internal prediction), but also of molecules not included in the development of the model (external prediction) [38]. The internal predictive capability of the model can be judged from cross-validated by techniques like leave-one-out and leave-group-out, whereas its external predictivity can be evaluated by using a separate set of molecules (the test set) not included in the model development. To further assess the robustness and statistical confidence of the derived models Fischer statistics, randomization (y-scrambling) and bootstrapping analysis are also performed. All these cross-validation methods have been explained in the later sections.

DISPLAY OF RESULTS

CoMFA generates an equation correlating the biological activity with the contribution of interaction energy fields at every grid point. To allow simple and easy visual interpretation, results are generally shown as coefficient (or scalar product of coefficients and standard deviation) contour plots, depicting important regions in space around the molecules where specific structural modifications significantly alters the activity [29, 34]. Generally 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.

In addition of contour plots, CoMFA also provides two types of plots from PLS models: score plots and loading/ weight plots. Score plots between biological activity (Y-scores) and latent variables (X-scores) show relationship between the activity and the structures, whereas plots of latent variables (X-scores) display the similarity/dissimilarity between the molecules, and their clustering propensities.

DRAWBACKS AND LIMITATIONS OF CoMFA

Despite of offering many advantages over classical QSAR and good performance in various practical appli-

cations, CoMFA has several pitfalls and imperfections as given below [26, 29, 34]:

- Since the time of its origin in 1988, numerous applications of the CoMFA method in different fields have been published [39]. Several data sets have been investigated; the first being the binding affinity of the steroid data set [40] for human corticosteroid-binding globulins (CBG) and testosterone-binding globulins (TBG). Many successful endeavors of CoMFA approach in the areas of enzyme Highly sensitive to bioactive conformation, different binding modes of ligands, alignment rules and number of components
- Too many adjustable parameters like overall orientation, lattice placement, step size, probe atom type *etc.*
- Uncertainty in selection of compounds and variables
- Fragmented contour maps with variable selection procedures
- Hydrophobicity not well-quantified
- Cut-off limits used
- Low signal to noise ratio due to many useless field variables
- Imperfections in potential energy functions
- Various practical problems with PLS
- Applicable only to *in vitro* data

Since the time of its origin in 1988, numerous applications of the CoMFA method in different fields have been published [39]. Several data sets have been investigated; the first being the binding affinity of the steroid data set [40] for human corticosteroid-binding globulins (CBG) and testosterone-binding globulins (TBG). Many successful endeavors of CoMFA approach in the areas of enzyme inhibition, agrochemistry (pesticides, insecticides or herbicides), physicochemistry (partition coefficients, capacity factors, enantio-separation factors and ^{13}C chemical shifts), ADME and toxicity, thermodynamics and kinetics have also been exhaustively appraised in several reviews [25, 27, 32, 41, 42].

MSA

Molecular Shape Analysis (MSA) is a ligand-based 3D-QSAR formalism which attempts to merge conformational analysis with the classical Hansch approach. It deals with the quantitative characterization, representation and manipulation of molecular shape in the construction of a QSAR model [43]. The methodology begins by subjecting each molecule in the data set to a fixed valence geometry intramolecular conformational analysis with a scan at 30° increments for all torsion angles except for amide N−(C=O) torsion which is scanned at 180° increment. The conformational energies are estimated using a fixed valence geometry molecular mechanics force-field consisting of a dispersion/steric, electrostatic, and, if applicable, hydrogen bonding contributions. For each compound, all apparent intramolecular energy minima are identified and recorded, each of which are then used as starting points in rigorous fixed valence geometry energy minimizations. Both apparently as well as rigorously minimized energy conformations are aspirants for the 'active' conformation of each analog in the ensuing steps. To identify the active conformation of each analog, the LBA-LCS (loss in biological activity-loss in conformational stability) approach is used; this is based on the identification of stable low-energy intramolecular conformer states common to the active analogs, which is a high-energy, unstable state for the inactive analogs.

A shape reference structure is selected as the mutant shape generated by the common and difference volume combinations realized by multiple compound alignments or active conformations. The potential active conformation of each compound in the data set is pair-wise compared and aligned with the shape reference. This is followed by the calculation of various descriptors which measure relative molecular shape similarity. One of the important shape variables is the common overlap steric volume (COSV) between pairs of molecules as a function of conformation and relative intermolecular geometry. It actually measures how much steric space a pair of molecules share under a prescribed intermolecular relationship. Two other descriptors are also arbitrarily defined as alternate mathematical representations of COSV that can advantageously be used in developing empirical QSARs; one has the dimensions of area but is not a physical measure of common atomic surface areas between two molecules, and another has the dimensions of length but is not a cumulative measure of distances between the molecules [43]. These pair-wise shape variables can also be amalgamated with the non-shape ther-modynamic and electronic descriptors including the terms from the Hansch equation (π, E_s, σ) in developing a MSA 3D-QSAR model.

The shape similarity descriptors along with the non-shape variables are eventually correlated with the biological activities of the molecules using the MLR technique, however, other chemometric methods like PLS and GA can also be employed. The MSA results can be graphically represented as a picture of the most active analog placed in its active conformation or as the superimposition of shape descriptors onto the molecular geometry of the most active molecule. Some of the recent successful applications of MSA include the generation of useful 3D-QSAR models of the allosteric modulators of muscarinic receptors [44], anticoccidial triazines [45], cholecystokinin-A receptor antagonists [46], and indanone-benzylpiperidine inhibitors of acetylcholinesterase [47]. MSA is being provided in the Cerius2 software [48] from Accelrys Inc.

GRID

GRID was the first tailor-made program designed for the medicinal chemist as an alternative to the original CoMFA approach. It calculates the interaction energy fields in molecular field analysis and determines the energetically favorable binding sites on molecules of known structure. Though the approach is similar to CoMFA in that it too computes explicit non-bonded (or non-covalent) interactions between a molecule of known three-dimensional structure and a probe (*i.e.,* a small chemical group with user-defined properties) located at the sample positions on a lattice throughout and around the macromolecule, it offers two

distinct advantages; first is the use of a 6-4 potential function for calculating the interaction energies, which is smoother than the 6-12 form of the Lennard-Jones type in CoMFA, and second is, the availability of different types of probes [49]. The program in addition of computing the regular steric and electrostatic potentials, also calculates the hydrogen bonding potential using a hydrogen bond donor and acceptor, and the hydrophobic potential using a "DRY probe". Later on a water probe was added to calculate hydrophobic interactions [41, 50]. Since the water probe is not only electrically neutral but can also donate and accept a hydrogen bond, the energies determined using this probe are supposed to embrace steric and hydrogen bonding interactions also, besides representing the hydrophobic interaction energy like log P due to its molecular surface area. In addition to the water and DRY probes, other probes which are usually used singly, include the methyl group, the amine NH_2 group, the carboxylate group and the hydroxyl group. Contour surfaces are calculated at various energy levels for each probe for every point on the grid and are displayed graphically along with the protein structure. While negative energy levels of the contours describe regions at which ligand binding should be favored, positive energy levels normally characterize the shape of the target. Some of the recent applications of the GRID method include determining energetically favorable binding sites and characterization of their surface properties [51], building predictive pharmacophore models [52], classification and comparison of ligand-binding sites [53], and rational design of potent inhibitors of influenza virus sialidase [54]. Many times GRID maps are also used as input descriptors in CoMFA, GOLPE or SIMCA for QSAR or 3D-QSAR analyses [55]. It is possible to use these interaction energies in a statistical technique to relate to the biological activity in a quantitative manner. The GRID software is supplied by Molecular Discovery Ltd [56].

HASL

The Hypothetical Active Site Lattice (HASL) method is an inverse grid-based approach which represents the shapes of the molecules inside an active site as a collection of grid-points [57]. The methodology begins with the intermediate conversion of the Cartesian coordinates (x, y, z) of a superposed set of molecules to a 3D-grid consisting of the regularly-spaced points that are:

- arranged orthogonally to each other
- separated by a particular distance termed as the resolution (which determines the number of grid points representing a molecule)
- all sprawl within the van der Waals radii of the atoms constituting the molecule

The resulting framework of points is referred to as the molecular lattice and represents the receptor active site map. The overall lattice dimensions are dependent on the size of the molecules and the resolution chosen.

Typically a reference molecule is selected arbitrarily and its user-defined conformations similar in shape and that have been energy-minimized, are used to generate the HASL. The selected conformation of the reference molecule is centered about the origin of a Cartesian coordinate system, and a regular grid with a chosen resolution is then laid over the molecule. All grid points lying within the van der Waals radii of the atoms of the molecule are designated as 'occupied' and form the molecular lattice. The electronic properties of the occupying atoms are distinctly represented by assigning the lattice points a 'HASL-type' value based on the electron density of the atoms, which constitute the fourth dimension of the molecular lattice. The values of +1, -1 and 0 are assigned to the electron-rich (*e.g.* O, N), electron-poor (*e.g.* C in C=O) and neutral atoms/substituents, which roughly represent H-bond acceptors, donors, and lipophilic atom types, respectively. Such internal atom type designations allow apparently different structures to be overlaid with an equivalent electronic "sense", (*i.e.*, similar atomic characteristics of two molecules can be superimposed in 3D-space), to obtain a maximum complementarity of space and physiochemical character within that space. Similar to electron density, other user-selected physio-chemical properties such as hydrophobicity, can also be employed as the fourth dimension. A second molecule is subsequently selected and subjected to the same routine, generating its 4D lattice which is then compared and consequently aligned to that of the stationary reference molecule. In order to optimally align the molecules based on their lattices, a systematic search is performed which involves a stepped progression of translational and rotational movements, with an intermediate lattice generated at each step, until a perfect match is obtained. The extent of similarity between the two molecules is calculated according to a fitting function, based upon the degree of corres-pondence between the points of the two lattices, *i.e*, on the number of points the two lattices have in common. Once the best possible alignment between the two molecular lattices is obtained, those lattice points of the fitted molecule which are not yet in common with the reference molecule are added to create a new composite construct or larger reference lattice containing the information from both the molecules. This fitting and merging process is then repeated to include all the molecules of the training set in the growing HASL, resulting in a reference lattice entailing every point from all the molecular lattices.

In order to determine the activity contributions from different lattice points, initially the experimental activity value of a molecule is homogeneously divided among its all lattice points. For lattice points which are shared by more than one molecule, the partial bioactivity values are, at first, averaged over these points and, afterwards adjusted by an iterative protocol to fit the experimental activity data of the entire training set. This iteratively optimized HASL is then used as a standardized model to predict the activities of untested molecules. The bioactivity of a specific compound is forecasted by summing all the partial activity values at points in common with the composite reference lattice. Some of the successful applications of HASL approach include the analysis of the *in vitro* antimalarial activity of artemisinin analogs [58], *in vitro* biochemical and *in vivo* gastric antisecretory activity of substituted imidazo[1,2-a]pyridines [59], sequence specificity of DNA alkylation by uracil mustard [60], and the generation of putative pharmacophoric models of the HIV-1 protease inhibitors [61]. HASL is a copyrighted program of Hypothesis Software and eduSoft

LC [62], and also comes as one of the modules in Sybyl Software [33].

CoMSIA

Comparative Molecular Similarity Indices Analysis (CoMSIA) was developed to overcome 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. For computing similarity at all grid points, the mutual distances between the probe atom and the atoms of the molecules in the aligned dataset are also taken into account. To describe this distance-dependence and calculate the molecular properties, Gaussian-type functions are employed. Since the underlying Gaussian-type functional forms are 'smooth' with no singularities, their slopes are not as steep as the Coulombic and Lennard-Jones potentials in CoMFA; therefore, no arbitrary cut-off limits are required to be defined. These functions tend to produce values within a reasonable range, even in the case of overlapping atoms. Despite the fact that CoMSIA also suffer from most of the limitations of CoMFA, it offers following distinctive advantages:

- Use of the Gaussian distribution of similarity indices, which avoids the abrupt changes in grid-based probe–atom interactions
- The choice of similarity probe, is not limited to either steric or electrostatic potential fields but also include hydrophobic and hydrogen bonding (hydrogen bond acceptors and donors) fields
- Effect of the solvent entropic terms can also be included by using a hydrophobic probe
- The standard CoMFA contours highlights those regions in space where the aligned molecules would favorably or unfavorably interact with a possible receptor environment. On the other hand, 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. This relationship between the required properties and a probable ligand shape is a more direct guide to substantiate whether all features imperative for activity are present in the structures being considered.

Some of the recent applications of CoMSIA include generation of predictive 3D-QSAR models of boron-containing dipeptides as proteasome inhibitors [63], hydroxamic acid derivatives as urease inhibitors [64], thiazolidin-4-one derivatives as anti-HIV-1 agents [65], and thiazolidinediones derivatives as aldose reductase inhibitors [66]. CoMSIA is provided by Tripos Inc. in the Sybyl software [33], along with CoMFA.

GERM

Genetically Evolved Receptor Models (GERM) is a technique for 3D-QSAR and for constructing useful three-dimensional models of macromolecular binding sites in the absence of a crystallographically-determined or homology-modeled structure of the target receptor [67]. The primary requirement for GERM is a structure–activity series for which a sensible alignment of realistic conformers has been determined. The methodology consists of enclosing the superimposed set of molecules in a shell of atoms (analogous to the first layer of atoms in the active site) and allocating these atoms with explicit atom types (aliphatic H, polar H, *etc*. to match the types of atoms usually found in the proteins). Aliphatic carbon atoms are disseminated uniformly over a sphere surrounding the training set of aligned ligands, and their positions are adjusted to obtain maximum van der Waals interaction between the model carbon atoms and the ligand molecules. Once the positions of the carbons have been recognized, they can be occupied by any of the atom types, including no atom at all. One practical problem arises when the number of shell atoms and their atom types increases, since the number of possible combinations rises to a huge value thereby rendering it impossible to systematically find a best possible model. The method therefore makes use of the genetic algorithms (GA) to solve this highly multi-dimensional search problem. The ligands in the training set are then docked into a GA generated receptor active site model, one at a time, and the intermolecular non-bonded interaction energies (van der Waals and electrostatic terms) are computed using a CHARMm molecular mechanics force field. Finally these calculated interaction energies are correlated with the biological activities of the molecules. The affirmative feature of this method is that the model is presented as a 3D-display of the receptor properties in space. The limitation of GERM methodology is that it considers only a single conformation of each ligand in the training set, as well as its single orientation in the binding site. Since this method is based on the computation of interaction energies with the hypothetical receptor, it is subjected to all the limitations of such methods including the alignment problem. However, if all the molecules of the set do bind in a manner that doesn't alter the binding site too much; GERM could be a good approach. The method has been applied profitably on a series of sweeteners, correlating their bioactivities with the calculated intermolecular energy [68]. The methodology has a fair potential for application in screening 3D-structural databases to find new leads, or in combination with *de novo* ligand-design programs. The program GERM is available from D. Eric Walters [69], Associate Professor, Finch University of Health Sciences, North Chicago, USA.

COMBINE

Comparative Binding Energy Analysis (COMBINE) method was developed to take advantage of the structural data from ligand-macromolecule complexes, in a 3D-QSAR paradigm. The technique is based upon the hypothesis that the free energy of binding can be correlated with a subset of

energy components calculated from the structures of receptors and ligands in bound and unbound forms [70, 71]. The ligands are divided into fragments and the same number of fragments is allocated to all the compounds, adding "dummy" fragments to the ligands lacking a particular fragment. The non-bonded (van der Waals and electrostatic) interaction energies are computed between each residue of the receptor and every fragment of the ligand, using a molecular mechanics force field. The energies are also calculated between all pairs of residues/fragments for the complexes and for the free ligands and receptor. The electrostatic interactions are computed using a distance-dependent dielectric constant, and no cutoff limits are employed for the non-bonded interactions. The insignificant descriptors are then eliminated from the data using the variable selection utility in GOLPE program, and finally the biological activities of the molecules are correlated with the interaction energy values by employing PLS technique. Like all other interaction energy based 3D-QSAR approaches, COMBINE also suffers from the inherent errors involved in the computation of these energies. Also, the predictive ability of the method can be enhanced by making improvements in various aspects like the description of the electrostatic term, the inclusion of suitable descriptors for solvation and entropic effects, and the optimization of particular facets of the methodology, such as the choice of ligand fragment definitions and the details of the variable selection protocol. Recently COMBINE methodology has been utilized to build 3D-QSAR models to determine the selectivity and specificity of Ras proteins [72], predict binding affinity of non-peptide inhibitors of HIV-1 protease [73], and to identify amino acid residues in haloalkane dehalogenase LinB that modulate its substrate specificity [74].

CoMMA

Comparative Molecular Moment Analysis (CoMMA) is one of the unique alignment-independent 3D-QSAR methods, which involves the computation of molecular similarity descriptors based on the spatial moments of molecular mass (shape) and charge distributions up to and including second order as well as related quantities [75]. With respect to each molecular structure, two Cartesian reference frames are then defined. One frame is the principal inertial axes calculated with respect to the center-of-mass. For neutral molecular species, the other reference frame is the principal quadrupolar axes calculated with respect to the molecular "center-of-dipole". Dipolar, quadrupolar, and displacement descriptors are then calculated with reference to the principal inertial axes translated such that its origin is superposed on the center-of-dipole. It is noteworthy that these descriptors are obtained after translation to the center of mass as well as the center of dipole for each molecule, to keep the system alignment-independent. Finally these mole-cular moment descriptors are correlated with the biological activities of molecules using the PLS technique. Literature reports suggest that CoMMA descriptors are sensitive to molecular conformations, but less sensitive than CoMFA field parameters. The authors propose that the CoMMA descriptors have a potential role in addressing the issues like large scale screening and molecular diversity. The method has been used to build robust 3D-QSAR models to comprehend the structure-activity relationships of the benchmark steroid data set [75], and to develop combinatorial QSAR of ambergris fragrance compounds [76]. A web version of the CoMMA program is provided by the IBM informatics group [77]. A slight variant of this approach, termed as CoMMA2, has also been developed by the author [78].

CoMSA

Comparative Molecular Surface Analysis (CoMSA) is a non-grid 3D-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 [79, 80]. The methodology proceeds by subjecting the molecules in the data set to geometry optimization and assigning them with partial atomic charges. The Kohonen's self-organizing maps (SOM, a type of neural network) are then employed to transform the three-dimensional surface of the molecules into two-dimensional topographical maps, by extracting the signals from the Cartesian coordinates of the points sampled randomly at the van der Waals surface of the molecules. The partial atomic charges of the atomic molecular represen-tations are also projected onto the 2D-topographical maps. The molecular electrostatic potentials (MEPs) are calculated at the surface points and a mean value of the potential analogous to the respective points found in each grid cell (of CoMFA like methods) is utilized to explain this 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 ensuing comparative matrix of the mean electrostatic potentials (transformed into vectors) is finally used to develop a 3D-QSAR model using the PLS technique. The distinctive feature of CoMSA is that, in contrast to CoMFA and related approaches, 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. Recently a receptor-dependent CoMSA model, using multi-pose molecular docking and iterative variable elimination PLS (IVE-PLS), has been developed and applied on sulforaphane compounds as activators of quinone reduc-tase [81]. Other recent applications of CoMSA include the modeling of pKa values of benzoic acids [82], and hypolipidemic asarones [83], virtual combinatorial library screening of styrylquinoline HIV-1 blocking agents [84], and determination of the binding mode for a series of benzoxazine oxytocin antagonists using docking and 3D-QSAR studies [85].

AFMoC

Adaptation of Fields for Molecular Comparison (AFMoC) is a 3D-QSAR method involving fields derived from the protein environments (and not from the superimposed ligands as in CoMFA), therefore it is also known as a 'reverse' CoMFA (=AFMoC) approach [86]. The methodology begins by placing a regularly-spaced grid into the receptor binding site, followed by mapping of the knowledge based pair-potentials between protein atoms and ligand atom probes onto the grid intersections resulting in

the potential fields. Based on these potential fields, interaction fields are generated by multiplying distance-dependent atom-type properties of actual ligands docked into the active site with the neighboring grid values. These atom-type specific interaction fields are then correlated with the binding affinities of the molecules using PLS technique, which assigns individual weighting factors to each field value. Finally the results are displayed graphically by using contribution maps, and binding affinities of novel ligands are predicted by applying the derived 3D-QSAR equation. The distinctive features of this approach include:

- A tailor-made scoring function is combined with a protein-based CoMFA approach, thereby overcoming the prerequisite to involve complete ligand training sets
- The gradual shift from generally valid knowledge-based potentials to protein-specific pair-potentials, reflects the amount and the degree of structural diversity existing in the ligand training data
- Atom-type specific interaction fields are used which are mutually orthogonal in nature and thus eases the interpretation of PLS results
- In addition of the enthalpic contribution, the methodology is also expected to include the entropic effects resulting from (de-)solvation, since structural knowledge from experimentally determined complexes is converted into statistical pair potentials

Some of the thriving applications of AFMoC include building predictive 3D-QSAR models for 1-deoxyxylulose-5-phosphate (DOXP)-reductoisomerase inhibitors [87], 3-oxybenzamides as potent inhibitors of the coagulation protease factor Xa [88], thermolysin and glycogen phosphorylase b inhibitors [86], and for analyzing selectivity- and affinity-determining features of carbonic anhydrase isozymes [89]. Recently the methodology has been modified to account for the multiple ligand conformations in an ensemble of protein configurations. The improved method has been termed as consensus AFMoC (AFMoCcon), and was validated on the thrombin inhibitors [90].

CoRIA AND ITS VARIANTS

Comparative Residue Interaction Analysis (CoRIA) is a 3D-QSAR approach which uses the descriptors that describe the thermodynamic events involved in ligand binding, to explore both the qualitative as well as the quantitative facets of the ligand–receptor recognition process. Initial CoRIA methodology simply consisted of calculating the non-bonded (van der Waals and Coulombic) interaction energies between the ligand and the individual active site residues of the receptor that are involved in interaction with the ligand [91-93]. Employing the genetic version of PLS technique (G/PLS), these energies were then correlated with the biological activities of molecules, along with the other physiochemical variables describing the thermodynamics of binding like, lipophilicity, molar refractivity, surface area, molecular volume, Jurs descriptors, strain energy *etc*.

Later on to deal with the problems of peptide QSAR, this approach was further extended and modified to develop two new variants of CoRIA: *reverse*-CoRIA (*r*CoRIA) and *mixed*-CoRIA (*m*CoRIA). In these methodologies, the peptide (ligand) is fragmented into individual amino acids and the interaction energies (van der Waals, Coulombic and hydrophobic interactions) of each amino acid in the peptide with the receptor as a whole (*r*CoRIA) and with individual active site residues in the receptor (*m*CoRIA) are calculated, which along with other thermodynamic descriptors (like free energy of solvation, entropy loss on binding, strain energy, and solvent assessable surface area) are used as independent variables that are correlated to the biological activity by G/PLS chemometric method [94].

CoRIA methodologies makes full use of the wealth of knowledge contained in the ligand-receptor complexes and extract crucial information regarding the nature and type of important interactions at the level of both the receptor and the ligand, which can be directly employed in the design of new molecules and receptors. The approaches have the ability to forecast modifications in both the ligand as well as the receptor, provided structures of some ligand–receptor complexes are available. The methodology has been successfully applied to study the interactions of inhibitors with Cyclooxygenase-2 [91], MurF Enzyme of *Strepto-coccus pneumonia* [92], HIV-1 integrase [93], and peptides binding to MHC class-I molecule HLA-A*0201 [94]. However, these methods are difficult to be applied on small organic molecules, because unlike peptides there is no logical or universally accepted protocol for fragmenting small molecules. Also the methodologies can be further improved by solvating entire ligand-protein complexes, extensive conformational sampling by molecular dynamics, inclusion of other important interactions like hydrogen bonding *etc*.

OTHER 3D-QSAR METHODOLOGIES

In addition of the above mentioned formalisms, several other 3D-QSAR methodologies have been developed. Some of them are as follows:

Method	Steps
Compass	• Conformational analysis is carried out to determine the probable bioactive conformation of each ligand • descriptors measuring surface shape or polar functionality of each ligand's pose in a specific alignment in the vicinity of a particular point in space are then computed • a neural network is constructed and models built, realignment of molecules is continuously carried out to achieve the best fit to the binding site with improvements in the neural network model • the final model is developed from these improved and realigned molecular poses [95]

Method	Steps
RSA/RSM/CoRSA (Receptor Surface Analysis/Modeling, Comparative Receptor Surface Analysis)	• The structures of molecules are optimized and superimposed in their bioactive conformation • a receptor-complementary surface is generated using shape fields (defined by some distance-dependent function) that encloses a volume common to all the aligned molecules and which represents their aggregate molecular shape • the putative chemical properties of the receptor at every surface point are computed • PLS models are developed that correlate surface properties with molecular activities [96, 97]
VFA (Voronoi Field Analysis)	• A conformational analysis, minimization and superimposition of all the molecules is first carried out. • the volume occupied by superimposed set of molecules is divided into subspaces referred to as Voronoi polyhedral, each including a reference point (an atom) with certain coordinates as explained in the following steps • first a template (the simplest) molecule is selected and all the atoms of the template molecule are allocated as initial reference points • next the largest molecule in the dataset is superimposed on the template in terms of the number of atoms and new reference points are designated if this point is greater than 1 Å distance of the reference points identified in the above step. • the above two steps are repeated with superimposition of other molecules in decreasing order of their size, each time defining isolated atoms as new reference points by the criteria stated above, until all compounds are superimposed • a cuboid with six tangential planes divided into a 3D-lattice with a spacing of 0.3 Å, surrounding the union volume of the superposed set of molecules is constructed. This gives the Voronoi polyhedral. • the potential and electrostatic energy indices at each lattice point is computed according to the 'hard-sphere potential' model and Coulomb's law respectively • the PLS algorithm is then applied to correlate independent steric and electrostatic latent variables with the activity index [98]
PARM (Pseudo Atomic Receptor Model)	• fifteen types of pseudo receptor atoms types possibly found in a protein are selected • the molecules are superimposed and a 3D-grid around their common surface is generated • pre-defined atom types and formal charge at these grid points are assigned using a genetic algorithm; this is based on the charge of the ligand atom closest to the grid point • a GA-based initial population of individuals or receptor models are generated • van der Waals and electrostatic interaction energies between each ligand and the receptor model are computed and are correlated to their molecular activities using a linear regression technique [99]
SOMFA (Self-Organizing Molecular Field Analysis)	• Firstly the mean activity of training set is subtracted from the activity of each molecule to obtain their mean centered activity values • a 3D-grid around the molecules with values at the grid points signifying the shape or electrostatic potential is generated • the shape or electrostatic potential value at every grid point for each molecule is multiplied by its mean centered activity • the grid values for each molecule are summed up to give the master grids for each property • the so called $SOMFA_{property,i}$ descriptors from the master grid values are then calculated and correlated with the log-transformed molecular activities [100]
FLUFF-BALL	• A semiautomatic superimposition of the molecules based on a novel field-fitting procedure called Flexible Ligand Unified Force Field (FLUFF) is carried out; this is a MMFF94 force field that is customized to impart flexibility to the ligand to maximize adaptation/similarity between the steric and electrostatic field volumes of the ligand and the template • the internal coordinate system is attached to the template molecule by placing the vertices of the local grid at the atomic centers of the template, using the Boundless Adaptive Localized Ligand (BALL) approach, thus rendering the system grid-independent • 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 [101]
CoMASA (Comparative Molecular Active Site Analysis)	• The molecules are first superimposed and their interatomic distances calculated • then is extracted the co-ordinates of the molecular representation (instead of the lattice points as in CoMFA) by continuously removing the atoms that are closer to each other, and replacing them with pseudo atoms (created from their weighted average), until the distances between all the atoms/pseudo atoms are greater than the threshold value of 0.75 Å • the interaction energies (steric, electrostatic and hydrophobic properties) are then computed for each molecule at these points by different evaluation functions and finally these are correlated with their molecular activities using PLS [102]

Method	Steps
CoMPIA (Comparative Molecule/Pseudo receptor Interaction Analysis)	• The geometry of the molecules is optimized which is followed by their superimposition based on a common template molecule • 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 between different probes and every molecule in the set are computed and then correlated with the biological activities using PLS [103]

STATISTICAL METHODS USED FOR BUILDING QSAR MODELS

Statistical or chemometric techniques form the mathematical foundation for building a QSAR model. Some of these methods are briefly described below:

Table 2. Statistical Techniques for Building QSAR Models

Linear Regression Analysis (RA)
Simple linear regression
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)

Among the increasing pool of various statistical methods available in the literature, **Linear Regression analyses** are considered as an easily interpretable methods indicated for QSAR analysis [104]. These regression techniques construct a statistical model to represent the correlation of one or more independent variables (x) with a dependent explicative variable (y). The model can be utilized to predict y from the knowledge of x variables, which can be either quantitative or qualitative. Simple linear regression, multiple linear regression, and stepwise multiple linear regression are some of its variants.

Simple linear regression method performs a standard linear regression calculation to generate a set of QSAR equations that include a single independent descriptor **x** and a dependent variable **y** [104]. Thus, a one-term linear equation is produced separately for each independent variable from the descriptor set**.** This technique is suitable for generating simple relationships between structure and activity exploring some of the most important descriptors governing the activity. However, the interaction of multiple descriptors is ignored. The simple linear regression can be expressed by the equation:

$$\mathbf{y} = a + b\mathbf{x}$$

where the dependent variable **y** is expressed in terms of the independent variable **x** by means of two parameters: the constant *a*, also referred to as the intercept and the regression coefficient *b*.

Multiple linear regression (MLR) also referred to as the linear free-energy relationship (LFER) method, is an extension of the simple regression analysis to more than one dimension [105]. MLR generates QSAR equations by performing standard multivariable regression calculations to identify the dependence of a drug property on any or all of the descriptors under investigation. The possibility of chance correlation is checked through the values of multiple correlation coefficient (*r*), Student's *t*-value; Fisher's F ratio, standard deviation (*s*), and through independent tests like the leave-one-out (LOO) method. The significance of correlation can be judged through cross-validated correlation coefficient (r^2_{cv} or q^2) values and also by the y-scrambling technique. MLR assumes that all variable are independent, and not correlated. However, in the multivariate case, *i.e.*, MLR analysis involving more than one independent variable, the relationship is expressed with the following single multiple-term linear equation:

$$\mathbf{y} = b_0 + b_1\,\mathbf{x_1} + b_2\mathbf{x_2} + \ldots\ldots\ldots\ldots + b_m\,\mathbf{x_m} + \mathbf{e}$$

the MLR analysis estimates the regression coefficients **(**b_i), by minimizing the residual error (e), which quantify the deviation of a particular point from the regression line, as in the case of simple linear regression.

Stepwise multiple linear regression is a commonly used variant of MLR which also creates a multiple-term linear equation, but not all the independent variables are used [106]. In contrast to MLR, each independent variable is sequentially added to the equation and new regression is performed every time. The new term is preserved only if the model passes a test for significance. This regression technique is especially useful when the number of descriptors is large and the key descriptors are unknown.

The methods described above have now been replaced by **multivariate chemometric methods** which try to explain an extended set of variables by means of a reduced number of new latent variables possessing the maximum amount of information relevant to the problem. These techniques

project multivariate data into a space of lower dimensions providing insight to visualize, classify, and model large data sets. These latent variables are orthogonal and hence can be used in multiple linear regressions.

Partial least squares (PLS) is an iterative regression procedure 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 [107]. 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. Thus, PLS is able to analyze complex structure-activity data in a more realistic way, and effectively interpret the influence of molecular structure on biological activity. This is one of the standard statistical methods used for the development of predictive 3D-QSAR models.

Principal components analysis (PCA) is another data reduction technique that does not generate a QSAR model but seeks for relationships among independent variables [108]. It then creates 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. Consequently, PCA reduces dimensionality of a multivariate data set of descriptors to the actual amount of data available. When principal components are employed as the independent variables to perform a linear regression, the method is termed as the **principal components regression (PCR)**. In other words, PCR applies the scores from PCA decomposition as regressors in the QSAR model, to generate a multiple-term linear equation [109].

Genetic function approximation (GFA) serves as an alternative to standard regression analysis for building QSAR equations [110]. It employs the natural principles of evolution of species which leads to improvements by recombination (mutation and crossover) of independent variables. This method results in multiple models generated by evolving random initial models using genetic algorithm. The method is suitable for obtaining QSAR equations when dealing with a larger number of independent variables. It can build linear as well as higher-order nonlinear equations, perform automatic outlier removal and classification by utilizing spline-based terms. **Genetic partial least squares (G/PLS or GA-PLS)** is a valuable analytical tool that has evolved by combining the best features of GFA and PLS [111], and has been widely preferred by the researchers [91-94].

In recent years, other methods to perform qualitative or classification studies have been spurred in the field of QSAR. The so-called pattern recognition methods based on the principle of analogy are used for the detection of the distance or closeness within the large amount of multivariate data [112]. It searches for structural features such as the presence (or absence) of certain groups, number of a certain type of atom, or mass spectral-fragmentation so that new compounds can be classified as similar or dissimilar to the members of the existing classes.

Cluster analysis is a statistical pattern recognition method used to investigate the relationship between observations associated with several properties and to partition the data set into categories consisting of similar elements [113]. It allows for the consideration of the inactive compounds in the analysis and can be used to study a large set of substituents to identify which of the subsets share similar physical properties.

The technique of **artificial neural networks (ANNs)** has its origin from the real neurons 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 [114]. The first layer is termed the input layer, and each of its neurons receives data from outside/user, corresponding to one of the independent variables used as inputs in QSAR. Subsequent to the input layer, there are one or many layers of neurons, collectively termed as the hidden layers. The last layer is the output layer, and its neurons handle the output from the network. Each layer may make its independent computations and may pass the results yet to another layer. The working of ANNs is given below:

- Each input descriptor value is multiplied by the connection weight, as per its significance
- The weighted inputs are summed up and supplied to the hidden layers, where a nonlinear transfer function does all the required processing
- The results of the transfer function are communicated to the neurons in the output layer, where the results are interpreted and finally presented to the user.

The ***k*-Nearest Neighbor (*k*NN)** method is one of the simplest machine learning algorithms, most commonly used for classifying a new pattern (*e.g.* a molecule). The technique is based on a simple distance learning approach whereby an unknown/new molecule is classified according to the majority of its *k*-nearest neighbors in the training set [115]. The nearness is determined by a Euclidean distance metric (*e.g.* a similarity measure computed using the structural descriptors of the molecules). Typically, the *k*NN approach is executed as follows:

- Euclidean distances between an unknown object (u) and all the objects in the training set are computed
- Based on the calculated distances, *k* objects from the training set most similar to object u are selected
- Object u is assigned to the group to which the majority of the *k* objects belong
- An optimal *k* value is selected by optimization through the categorization of a test set of samples or by leave-one out cross-validation.

VALIDATION OF 3D-QSAR MODELS

Validation is a crucial element of any QSAR analysis. The reliability of a 3D-QSAR model depends on how well the model can predict the activity of compounds outside the training set rather than how well the model reproduces the biological activity of compounds included in the model. Various approaches used for this purpose are described below:

The correlation coefficient, *r* is a measure of the degree of linearity of the relationship. It signifies the quality of fit of the model and quantifies the variance in the data [116]. In an ideal situation the correlation coefficient must be equal to or approach 1, but in reality due the complexity of biological data, any value above 0.9 is appreciable. Correlation coefficients for the variables in a dataset are compiled in a correlation matrix, which shows the relationship of one descriptor with another. The correlation matrix ensures that variables of significance are orthogonal to each other. The addition of every new variable to the model always increases the *r*, unless the new variable is a constant of a linear combination of other variables, which would not produce any effect. The increase in *r* caused by adding new variable signifies over-fitting of the data.

The coefficient of multiple determinations also called **Pearson's correlation coefficient, r^2** is the squared correlation coefficient which informs about how well the model reproduces the experimental data [116]. It is a quantitative measure of the precision of adjustment for the fitted values to the observed ones. The closer it approaches to the unity, the more similar are the adjusted values to the experimental ones, suggesting that the model fits the data unerringly. However, an r^2 close to 1 does not mean that the model is perfect; the addition of any new descriptor to the model induces an ever-increasing of r^2, even if the newly added descriptor does not contribute to the model. Thus, other measures are required to determine the predictive capacity of the model.

Cross-validation (CV) is one of the most extensively employed methods for the internal validation of a statistical model [117]. In cross-validation, the predictive ability of a model is estimated using a reduced set of structural data. Usually, one element of the set is extracted each time, and a new model is derived based on the reduced dataset, which is then employed to predict the activity of the excluded molecule. The procedure is repeated *n* number of times until all compounds have been excluded and predicted once. This is the so-called **leave-one-out (LOO**) method [38]. Analogously, leaving out more than one molecule of the dataset at a time is termed as **leave-n-out** or **leave-many-out** CV method [38]. The outcome of LOO procedure is a cross-validated correlation coefficient r^2_{cv} (or q^2) which is a criterion of both robustness and predictive ability of the model:

$$r^2_{cv} = (PRESS_0 - PRESS) / (PRESS_0)$$

where $PRESS_0$ is the mean of the observed biological activity while PRESS is the sum of the squares of the differences between the predicted and the observed activity values [118]. Many researchers consider high q^2 as the ultimate proof of high predictive power of the QSAR model which is incorrect. It has been established that, in cases where test sets with known values of biological activities were available for prediction, there existed no correlation between the q^2 and r^2. Therefore, q^2 should be regarded as a measure of internal consistency of the derived model rather than as a true indicator of the predictability. It should be noted that, since it is easier to fit the experimental data than to predict them from the QSAR model, r^2 of the model is always higher than q^2. Cross-validation is not foolproof. In highly redundant data sets with fewer degrees of freedom, it can give an over-optimistic result. It may also improperly indicate a lack of correlation if all the compounds in the dataset are unique. Therefore, we can conclude that despite its wide acceptance, a high value of q^2 alone is an insufficient criterion for a QSAR model to be highly predictive.

Bootstrapping is another technique that can be used along with cross-validation to evaluate the robustness and the statistical confidence of the QSAR model. It involves simulating a large number of datasets which are of the same size as original and are produced by randomly selecting samples from the original dataset [119]. In each PLS run some objects may be excluded while some others might be sampled more than once. The statistical calculation is run on each of these bootstrap samplings. The difference between the parameters calculated from the original dataset and the mean of the parameters calculated from many bootstrap samplings is a measure of the biasness of the original calculations. Since it demands heavy computation with relatively smaller gains compared to cross-validation, the technique is not very attractive.

A rigorous alternative to cross-validation and bootstrapping is **randomization or y-scrambling** in which the biological activity values are re-assigned arbitrarily to different molecules in the same data set, and a new regression is performed [120]. Only if the results from a PLS model, using the original sequence of the biological data, is significantly better than the results from the 'scrambled' models, can one be sure that significant correlation indeed exists between the biological data and the independent variables, and it has not been resulted from a chance correlation. The randomization test analyses the ability of the statistical model to derive real structure-activity relationships.

Predictive ability of the model can also be evaluated by forecasting the activity of an external test set of molecules using the models derived from the training set. **Predictive correlation coefficient (r^2_{pred})**, which is analogous to cross-validated r^2 (or q^2) is a measure of the predictive ability of the derived QSAR model and is calculated by the following formula [121]:

$$r^2_{pred} = (SD - PRESS) / SD$$

where SD is the sum of squared deviations between the biological activities of the test set molecules and the mean activity of the training set molecules, while PRESS is the sum of squared deviations between the observed and the predicted activities of the test set molecules.

The **Fischer statistic** (F value) parameter is one of the several variance-related parameters that can be used as a measure of the level of statistical significance of the regression model [122]. A higher F value implies that a more significant correlation has been reached. It is used as a criterion to determine whether a more complex model is significantly better than a less complex one.

GUIDELINES FOR DEVELOPING A GOOD QSAR MODEL

Various guidelines have been proposed for developing a valid and universally acceptable QSAR model, some of which are given below:

SETUBAL / OECD PRINCIPLES

These principles were agreed by OECD (Organization for Economic Cooperation and Development) member countries at the 37th Joint Meeting of the Chemicals Committee and Working Party on Chemicals, Pesticides and Biotechnology in November 2004 at Setubal in Portugal [123]. According to these principles, a QSAR model should be associated with the following information so as to be applicable for regulatory purposes:

- a defined biological/biochemical/pharmaco-toxicological endpoint or intent which it serves to predict
- an unambiguous and easily applicable algorithm for predicting the endpoint
- a clearly defined domain of its applicability
- an appropriate measures of goodness-of–fit, robustness and predictivity assessed by internal and external validation
- a clear mechanistic basis, if possible apparent

RECOMMENDATIONS FROM TOPLISS & COSTELLO AND UNGER & HANSCH

A functional QSAR model must comply with the following general characteristics proposed by Topliss & Costello [124] and Unger & Hansch [125], for the proper derivation of extrathermodynamic equations:

- Sufficient examples must be present in the training set to cover the range of properties required to be predicted by the model. Generally this includes several log orders of magnitude of the end point being predicted.
- A wide range of different, meaningful and essentially independent/non-correlated descriptors, biophysically related to the property being predicted, should be selected. All "logical" parameters should be pruned and validated by appropriate statistical method like stepwise regression, F-value, cross-validation, Y-scrambling *etc*. Ideally while using MLR, the selected descriptors should be far less numerous than the number of compounds in the training set (at least 5 - 10 fold).
- In case of multiple models, one should follow the principle of parsimony (Occam's Razor), which states that if all things are (approximately) equal, the simplest model should preferably be selected.
- For an intermediate size data set, one should have at least 5 - 6 data points per variable (or per component) to circumvent chance correlations.
- In order to evade chance correlations, it is important to have a qualitative model which is consistent with the known physical-organic and biomedicinal chemistry of the process under consideration.
- The model should be supported by a number of statistical parameters to test its internal predictivity. A separate test set should be used to validate external predictivity of the model.

OTHER IMPORTANT SUGGESTIONS

Statistical method should be chosen in such a way to give a clear correlation between the descriptors and the biological activity. This can be made sure by additional tests like scrambling of activity data, boot-strapping, and cross-validation.

- Since prediction is the main goal of QSAR, a model must be evaluated for its predictive power both internally (via cross-validation) and externally (using a separate test set). The external testing becomes more significant in light of the 'Kubinyi Paradox' which states that high internal predictivity may often result in low external predictivity and vice versa. This probably might be due to the fact that the overall error of the prediction is compounded when errors inherent in the model are coupled with experimental errors in the data from external compounds.
- The QSAR model should be explanatory and interpretable. It should help in understanding the mode of action for active compounds originating from different data sets. Whenever structural data is available, the results should be validated with receptor information.
- The descriptors selected for model development must be pharmacologically or mechanistically relevant to the biological endpoint being examined.
- As far as possible, the descriptors which are simple and easier to interpret should be chosen for better understanding of the modeled system.
- QSAR should be applied only to pure compounds. Its application on mixtures should be avoided.
- The QSAR should not be applied outside of its domain of validity, *i.e.*, outside of the parameter space covered by the training set.

CONCLUDING REMARKS

Despite of all the pitfalls and caveats, it has now been globally apprehended by the contemporary drug discovery community that QSAR, based on well-established principles of statistics, is intrinsically a valuable and viable medicinal chemistry tool whose application domain range from explaining the structure-activity relationships quantitatively and retrospectively, to endowing synthetic guidance leading to logical and experimentally testable hypotheses. Ever-increasing information from structural biology will present valuable feedback to the assumptions that form the basis of 3D-QSAR methods. Before applying the predictive models to real-life situations, one must look into the technicalities of the underlying QSAR methodologies, in order to circumvent their inappropriate use and misinterpretation. More specifically, the problems associated with alignment-dependency and conformational sensitivity must be taken into consideration. In this regard, approaches like COMPASS, CoMMA, 4-way PLS *etc*. are helpful. Similarly whenever information about the target receptor is available, it must be utilized in building the models or validating the approaches developed solely on the basis of ligand data. In other words,

receptor-based models like COMBINE, AFMoC, CoRIA *etc*. are more enlightening and extrapolative than the conventional ligand-based methods like CoMFA, CoMSIA *etc*. Also when the inherent relationship between the descriptors and the biological endpoint to be modeled, is not expected to be linear in nature (*e.g.*, in case of ADMET properties), a non-linear chemometric method like neural networks, k-nearest neighbors *etc*. should be preferred. Finally, since the reliability of QSAR models depends on their statistical significance and the ability to predict accurately the activity of compounds not included in the training set, the models must be thoroughly validated both internally as well as externally using rigorous cross-validation techniques. Moreover, QSAR results should be accepted as a functional hypothesis that must be supported by pertinent statistical analysis as well as justified by further synthesis and biological testing for its approval or disapproval. 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. To end with a positive and motivating note, QSAR models, if used impeccably within their application domains and without unjustifiable extrapolations, will continue to impact the *in silico* drug discovery research.

REFERENCES

[1] Crum-Brown, A.; Fraser, T.R. On the connection between chemical constitution and physiological action. Part 1. On the physiological action of the ammonium bases, derived from Strychia, Brucia, Thebaia, Codeia, Morphia and Nicotia. *Trans. R. Soc. Edinburgh*, **1868**, *25*, 151-203.

[2] Richardson, B.J. Physiological research on alcohols. *Med. Times Gaz.*, **1868**, *2*, 703-706.

[3] Mills, E.J. On melting point and boiling point as related to composition. *Philos. Mag.*, **1884**, *17*, 173-187.

[4] Richet, C. On the relationship between the toxicity and the physical properties of substances. *Compt. Rendus Seances Soc. Biol.*, **1893**, *9*, 775-776.

[5] Overton, E. Osmotic properties of cells in the bearing on toxicology and pharmacology. *Z. Physik. Chem.*, **1897**, *22*, 189-209.

[6] Meyer, H. On the theory of alcohol narcosis I. Which property of anesthetics gives them their narcotic activity? *Arch. Exp. Pathol. Pharmakol.*, **1899**, *42*, 109-118.

[7] Hammett, L.P. Some relations between reaction rates and equilibrium constants. *Chem. Rev.*, **1935**, *17*, 125-136.

[8] Hammett, L.P. The effect of structure upon the reactions of organic compounds. benzene derivatives. *J. Am. Chem. Soc.*, **1937**, *59*, 96-103.

[9] Ferguson, J. The Use of Chemical Potentials as Indices of Toxicity. *Proc. R. Soc. Lond. B*, **1939**, *127*, 387-404.

[10] Bell, P.H.; Roblin, R.O. Studies in chemotherapy. vii. a theory of the relation of structure to activity of sulfanilamide type compounds1. *J. Am. Chem. Soc.*, **1942**, *64*, 2905-2917.

[11] Albert, A.; Goldacre, R.; Phillips, J. The strength of heterocyclic bases. *J. Chem. Soc.*, **1948**, 2240-2249.

[12] Taft, R.W. Polar and steric substituent constants for aliphatic and o-Benzoate groups from rates of esterification and hydrolysis of esters1. *J. Am. Chem. Soc.*, **1952**, *74*, 3120-3128.

[13] Hansch, C.; Maloney, P.P.; Fujita, T.; Muir, R.M. Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients. *Nature*, **1962**, *194*, 178-180.

[14] Hansch, C.; Fujita, T. p-σ-π Analysis. A Method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.*, **1964**, *86*, 1616-1626.

[15] Hansch, C. Quantitative approach to biochemical structure-activity relationships. *Acc. Chem. Res.*, **1969**, *2*, 232-239.

[16] Free, S.M., Jr.; Wilson, J.W. A Mathematical contribution to structure-activity studies. *J. Med. Chem.*, **1964**, *7*, 395-399.

[17] Fujita, T.; Ban, T. Structure-activity study of phenethylamines as substrates of biosynthetic enzymes of sympathetic transmitters. *J. Med. Chem.*, **1971**, *14*, 148-152.

[18] Kubinyi, H. Quantitative structure-activity relationships. IV. Non-linear dependence of biological activity on hydrophobic character: a new model. *Arzneimittelforschung*, **1976**, *26*, 1991-1997.

[19] Hansch, C.; Gao, H. Comparative QSAR: Radical reactions of benzene derivatives in chemistry and biology. *Chem. Rev.*, **1997**, *97*, 2995-3060.

[20] Hurst, T.; Heritage, T. *HQSAR - A Highly Predictive QSAR Technique Based on Molecular Holograms*. In: 213th ACS Natl. Meeting, San Francisco, CA, **1997**.

[21] Lowis, D.R. HQSAR: A New, Highly Predictive QSAR Technique. In: *Tripos Technical Notes;* Tripos Inc.: USA, Vol. *1*,**1997**.

[22] Cho, S.J.; Zheng, W.; Tropsha, A. Rational combinatorial library design. 2. Rational design of targeted combinatorial peptide libraries using chemical similarity probe and the inverse QSAR approaches. *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 259-268.

[23] Labute, P. Binary QSAR: a new method for the determination of quantitative structure activity relationships. *Pac. Symp. Biocomput.*, **1999**, 444-455.

[24] Kubinyi, H. 2D QSAR Models: Hansch and Free-Wilson Analyses. In: *Comput. Med. Chem. Drug Discov*., Bultinck, P., Winter, H.D., Langenaeker, W., Tollenaere, J.P., Eds.; Marcel Dekker, Inc: New York, USA, **2004**, pp. 539-570.

[25] Akamatsu, M. Current state and perspectives of 3D-QSAR. *Curr. Top. Med. Chem.*, **2002**, *2*, 1381-1394.

[26] Hopfinger, A.J.; Tokarski, J.S. Three-Dimensional Quantitative Structure-Activity Relationship Analysis. In: *Practical Application of Computer-Aided Drug Design;* Charifson, P.S., Ed.; Marcel Dekker, Inc.: New York, USA, **1997**; pp. 105-164.

[27] Martin, Y.C. 3D QSAR: Current State, Scope, and Limitations. In: *3D QSAR in Drug Design - Recent Advances;* Kubinyi, H., Folkers, G., Martin, Y.C., Eds.; Kluwer Academic Publishers: New York, USA, **1998**, Vol. *3*, pp. 3-23.

[28] Matyus, P.; Borosy, A.P. Three dimensional structure-activity relationships. *Acta Pharm. Hung.*, **1998**, *68*, 33-38.

[29] Oprea, T.I. 3D QSAR Modeling in Drug Design. In: *Computational Medicinal Chemistry for Drug Discovery;* Bultinck, P., Winter, H.D., Langenaeker, W., Tollenaere, J.P., Eds.; Marcel Dekker, Inc.: New York, USA, **2004**, pp. 571-616.

[30] Wise, M.; Cramer, R.D.; Smith, D.; Exman, I. Progress in Three-Dimensional Drug Design: the use of Real Time Colour Graphics and Computer Postulation of Bioactive Molecules in DYLOMMS. In: *Quantitative Approaches to Drug Design;* Dearden, J., Ed.; Elsevier: Amsterdam, UK, **1983**, pp. 145-146.

[31] Cramer III, R.D.; Patterson, D.E.; Bunce, J.D. Comparative molecular field analysis (comfa). i. effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.*, **1988**, *110*, 5959-5967.

[32] Podlogar, B.L.; Ferguson, D.M. QSAR and CoMFA: a perspective on the practical application to drug discovery. *Drug Des. Discov.*, **2000**, *17*, 4-12.

[33] *Sybyl*, version 7.1; Tripos Associates Inc.: 1699 S Hanley Rd., St. Louis, MO 631444, USA, **2005**.

[34] Kim, K.H. Comparative molecular field analysis (CoMFA). In: *Molecular Similarity in Drug Design;* Dean, P.M., Ed.; Blackie Academic & Professional: Glasgow, UK, **1995**, pp. 291-331.

[35] Allen, F. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr. B*, **2002**, *58*, 380-388.

[36] Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The protein data bank. *Nucleic Acids Res.*, **2000**, *28*, 235-242.

[37] Norinder, U. Recent progress in CoMFA Methodology and Related Techniques. In: *3D QSAR in Drug Design - Recent Advances;* Kubinyi, H., Folkers, G., Martin, Y.C., Eds.; Kluwer Academic Publishers: New York, USA, **1998**, Vol. *3*, pp. 24-39.

[38] Richard, D.; Cramer III, R.D.; Bunce, J.D.; Patterson, D.E.; Frank, I.E. Crossvalidation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies. *Quant. Struct.-Act. Relat.*, **1988**, *7*, 18-25.

[39] Kim, K.H. List of CoMFA References. In: *3D QSAR in Drug Design - Recent Advances;* Kubinyi, H.; Folkers, G.; Martin, Y.C.; Eds.; Kluwer Academic Publishers: New York, USA, **1998**, Vol. *3*, pp. 316-338.

[40] Coats, E.A. The CoMFA Steroids as a Benchmark Dataset for Development of 3D QSAR Methods. In: *3D QSAR in Drug Design - Recent Advances;* Kubinyi, H., Folkers, G., Martin, Y.C., Eds.; Kluwer Academic Publishers: New York, USA, **1998**, Vol. *3*, pp. 199-213.

[41] Kim, K.H.; Greco, G.; Novellino, E. A Critical Review of Recent CoMFA Applications. In: *3D QSAR in Drug Design - Recent Advances;* Kubinyi, H., Folkers, G., Martin, Y.C., Eds.; Kluwer Academic Publishers: New York, USA, **1998**, Vol. *3*, pp. 257-315.

[42] Bordas, B.; Komives, T.; Lopata, A. Ligand-based computer-aided pesticide design. A review of applications of the CoMFA and CoMSIA methodologies. *Pest Manag. Sci.,* **2003**, *59*, 393-400.

[43] Hopfinger, A.J. A QSAR investigation of dihydrofolate reductase inhibition by Baker triazines based upon molecular shape analysis. *J. Am. Chem. Soc.,* **1980**, *102*, 7196-7206.

[44] Holzgrabe, U.; Hopfinger, A.J. Conformational analysis, molecular shape comparison, and pharmacophore identification of different allosteric modulators of muscarinic receptors. *J. Chem. Inf. Comput. Sci.,* **1996**, *36*, 1018-1024.

[45] Rhyu, K.B.; Patel, H.C.; Hopfinger, A.J. A 3D-QSAR study of anticoccidial triazines using molecular shape analysis. *J. Chem. Inf. Comput. Sci.,* **1995**, *35*, 771-778.

[46] Tokarski, J.S.; Hopfinger, A.J. Three-dimensional molecular shape analysis-quantitative structure-activity relationship of a series of cholecystokinin-A receptor antagonists. *J. Med. Chem.,* **1994**, *37*, 3639-3654.

[47] Cardozo, M.G.; Iimura, Y.; Sugimoto, H.; Yamanishi, Y.; Hopfinger, A.J. QSAR analyses of the substituted indanone and benzylpiperidine rings of a series of indanone-benzylpiperidine inhibitors of acetylcholinesterase. *J. Med. Chem.,* **1992**, *35*, 584-589.

[48] *Cerius2*, version 4.8; Accelrys Inc.: San Diego, CA, USA, **1998**.

[49] Goodford, P.J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.,* **1985**, *28*, 849-857.

[50] Kim, K.H. Thermodynamic aspects of hydrophobicity and biological QSAR. *J. Comput. Aided Mol. Des.,* **2001**, *15*, 367-380.

[51] Caron, G.; Nurisso, A.; Ermondi, G. How to extend the use of grid-based interaction energy maps from chemistry to biotopics. *ChemMedChem,* **2009**, *4*, 29-36.

[52] Ortuso, F.; Langer, T.; Alcaro, S. GBPM: GRID-based pharmacophore model: concept and application studies to protein-protein recognition. *Bioinformatics,* **2006**, *22*, 1449-1455.

[53] Hoppe, C.; Steinbeck, C.; Wohlfahrt, G. Classification and comparison of ligand-binding sites derived from grid-mapped knowledge-based potentials. *J. Mol. Graph. Model.,* **2006**, *24*, 328-340.

[54] von Itzstein, M.; Wu, W.Y.; Kok, G.B.; Pegg, M.S.; Dyason, J.C.; Jin, B.; Van Phan, T.; Smythe, M.L.; White, H.F.; Oliver, S.W.; Colman, P.A.; Varghese, J.N.; Ryan, DM.; Woods, J.M.; Bethell, R.C.; Hotham, V.J.; Cameron, J.M.; Penn, C.R. Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature,* **1993**, *363*, 418-423.

[55] Pastor, M.; Cruciani, G.; Watson, K.A. A strategy for the incorporation of water molecules present in a ligand binding site into a three-dimensional quantitative structure--activity relationship analysis. *J. Med. Chem.,* **1997**, *40*, 4089-4102.

[56] GRID. Molecular Discovery Ltd. http://www.moldiscovery.com/soft_grid.php [Accessed on 1st July, **2009**].

[57] Doweyko, A.M. The hypothetical active site lattice. An approach to modelling active sites from data on inhibitor molecules. *J. Med. Chem.,* **1988**, *31*, 1396-1406.

[58] Woolfrey, J.R.; Avery, M.A.; Doweyko, A.M. Comparison of 3D quantitative structure-activity relationship methods: analysis of the *in vitro* antimalarial activity of 154 artemisinin analogues by hypothetical active-site lattice and comparative molecular field analysis. *J. Comput. Aided Mol. Des.,* **1998**, *12*, 165-181.

[59] Kaminski, J.J.; Doweyko, A.M. Antiulcer agents. 6. Analysis of the *in vitro* biochemical and *in vivo* gastric antisecretory activity of substituted imidazo[1,2-a]pyridines and related analogues using comparative molecular field analysis and hypothetical active site lattice methodologies. *J. Med. Chem.,* **1997**, *40*, 427-436.

[60] Doweyko, A.M.; Mattes, W.B. An application of 3D-QSAR to the analysis of the sequence specificity of DNA alkylation by uracil mustard. *Biochemistry,* **1992**, *31*, 9388-9392.

[61] Doweyko, A.M. Three-dimensional pharmacophores from binding data. *J. Med. Chem.,* **1994**, *37*, 1769-1778.

[62] HASL. eduSoft. http://www.edusoft-lc.com/hasl/ [Accessed on 1st July, **2009**].

[63] Zhu, Y.Q.; Lei, M.; Lu, A.J.; Zhao, X.; Yin, X.J.; Gao, Q.Z. 3D-QSAR studies of boron-containing dipeptides as proteasome inhibitors with CoMFA and CoMSIA methods. *Eur. J. Med. Chem.,* **2009**, *44*, 1486-1499.

[64] Ul-Haq, Z.; Wadood, A.; Uddin, R. CoMFA and CoMSIA 3D-QSAR analysis on hydroxamic acid derivatives as urease inhibitors. *J. Enzyme Inhib. Med. Chem.,* **2009**, *24*, 272-278.

[65] Murugesan, V.; Prabhakar, Y.S.; Katti, S.B. CoMFA and CoMSIA studies on thiazolidin-4-one as anti-HIV-1 agents. *J. Mol. Graph. Model.,* **2009**, *27*, 735-743.

[66] Liu, H.Y.; Liu, S.S.; Qin, L.T.; Mo, L.Y. CoMFA and CoMSIA analysis of 2,4-thiazolidinediones derivatives as aldose reductase inhibitors. *J. Mol. Model.,* **2009**.

[67] Walters, D.E.; Hinds, R.M. Genetically evolved receptor models: a computational approach to construction of receptor models. *J. Med. Chem.,* **1994**, *37*, 2527-2536.

[68] Walters, D.E.; Muhammad, T.D. Genetically Evolved Receptor Models (GERM): A Procedure for Construction of Atomic-level Receptor Site Models in the Absence of a Receptor Crystal Structure. In: *Genetic Algorithms in Molecular Modeling;* Devillers, J., Ed.; Academic Press: London, **1996**, pp. 193-210.

[69] GERM. Walters, D. E. http://www.finchcms.edu/biochem/Walters/germ.html [Accessed on 1st April, **2009**].

[70] Ortiz, A.R.; Pisabarro, M.T.; Gago, F.; Wade, R.C. Prediction of drug binding affinities by comparative binding energy analysis. *J. Med. Chem.,* **1995**, *38*, 2681-2691.

[71] Lushington, G.H.; Guo, J.X.; Wang, J.L. Whither combine? New opportunities for receptor-based QSAR. *Curr. Med. Chem.,* **2007**, *14*, 1863-1877.

[72] Tomic, S.; Bertosa, B.; Wang, T.; Wade, R.C. COMBINE analysis of the specificity of binding of Ras proteins to their effectors. *Proteins,* **2007**, *67*, 435-447.

[73] Nakamura, S.; Nakanishi, I.; Kitaura, K. Binding affinity prediction of non-peptide inhibitors of HIV-1 protease using COMBINE model introduced from peptide inhibitors. *Bioorg. Med. Chem. Lett.,* **2006**, *16*, 6334-6337.

[74] Kmunicek, J.; Hynkova, K.; Jedlicka, T.; Nagata, Y.; Negri, A.; Gago, F.; Wade, R.C.; Damborsky, J. Quantitative analysis of substrate specificity of haloalkane dehalogenase LinB from Sphingomonas paucimobilis UT26. *Biochemistry,* **2005**, *44*, 3390-3401.

[75] Silverman, B.D.; Platt, D.E. Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. *J. Med. Chem.,* **1996**, *39*, 2129-2140.

[76] Kovatcheva, A.; Golbraikh, A.; Oloff, S.; Xiao, Y.D.; Zheng, W.; Wolschann, P.; Buchbauer, G.; Tropsha, A. Combinatorial QSAR of ambergris fragrance compounds. *J. Chem. Inf. Comput. Sci.,* **2004**, *44*, 582-595.

[77] CoMMA. IBM Bioinformatics Group. http://cbcsrv.watson.ibm.com/Tco.html [Accessed on 1st July, **2009**].

[78] Silverman, B.D. Three-dimensional moments of molecular property fields. *J. Chem. Inf. Comput. Sci.,* **2000**, *40*, 1470-1476.

[79] Polanski, J.; Walczak, B. The comparative molecular surface analysis (COMSA): a novel tool for molecular design. *Comput. Chem.,* **2000**, *24*, 615-625.

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

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

[82] Gieleciak, R.; Polanski, J. Modeling robust QSAR. 2. iterative variable elimination schemes for CoMSA: application for modeling benzoic acid pKa values. *J. Chem. Inf. Model.,* **2007**, *47*, 547-556.

[83] Magdziarz, T.; Lozowicka, B.; Gieleciak, R.; Bak, A.; Polanski, J.; Chilmonczyk, Z. 3D QSAR study of hypolipidemic asarones by comparative molecular surface analysis. *Bioorg. Med. Chem.,* **2006**, *14*, 1630-1643.

[84] Niedbala, H.; Polanski, J.; Gieleciak, R.; Musiol, R.; Tabak, D.; Podeszwa, B.; Bak, A.; Palka, A.; Mouscadet, J.F.; Gasteiger, J.; Le Bret, M. Comparative molecular surface analysis (CoMSA) for virtual combinatorial library screening of styrylquinoline HIV-1 blocking agents. *Comb. Chem. High Throughput Screen.*, **2006**, *9*, 753-770.

[85] Jojart, B.; Martinek, T.A.; Marki, A. 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.*, **2005**, *19*, 341-356.

[86] Gohlke, H.; Klebe, G. DrugScore meets CoMFA: adaptation of fields for molecular comparison (AFMoC) or how to tailor knowledge-based pair-potentials to a particular protein. *J. Med. Chem.*, **2002**, *45*, 4153-4170.

[87] Silber, K.; Heidler, P.; Kurz, T.; Klebe, G. AFMoC enhances predictivity of 3D QSAR: a case study with DOXP-reducto-isomerase. *J. Med. Chem.*, **2005**, *48*, 3547-3563.

[88] Matter, H.; Will, D.W.; Nazare, M.; Schreuder, H.; Laux, V.; Wehner, V. Structural requirements for factor Xa inhibition by 3-oxybenzamides with neutral P1 substituents: combining X-ray crystallography, 3D-QSAR, and tailored scoring functions. *J. Med. Chem.*, **2005**, *48*, 3290-3312.

[89] Hillebrecht, A.; Supuran, C.T.; Klebe, G. Integrated approach using protein and ligand information to analyze selectivity- and affinity-determining features of carbonic anhydrase isozymes. *ChemMedChem*, **2006**, *1*, 839-853.

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

[91] Datar, P.A.; Khedkar, S.A.; Malde, A.K.; Coutinho, E.C. 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.*, **2006**, *20*, 343-360.

[92] Khedkar, S.A.; Malde, A.K.; Coutinho, E.C. Design of inhibitors of the MurF enzyme of *Streptococcus pneumoniae* using docking, 3D-QSAR, and de novo design. *J. Chem. Inf. Model.*, **2007**, *47*, 1839-1846.

[93] Dhaked, D.K.; Verma, J.; Saran, A.; Coutinho, E.C. Exploring the binding of HIV-1 integrase inhibitors by comparative residue interaction analysis (CoRIA). *J. Mol. Model.*, **2009**, *15*, 233-245.

[94] Verma, J.; Khedkar, V.M.; Prabhu, A.S.; Khedkar, S.A.; Malde, A.K.; Coutinho, E.C. A comprehensive analysis of the thermodynamic events involved in ligand-receptor binding using CoRIA and its variants. *J. Comput. Aided Mol. Des.*, **2008**, *22*, 91-104.

[95] Jain, A.N.; Koile, K.; Chapman, D. Compass: predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. *J. Med. Chem.*, **1994**, *37*, 2315-2327.

[96] Hahn, M. Receptor surface models. 1. Definition and construction. *J. Med. Chem.*, **1995**, *38*, 2080-2090.

[97] Ivanciuc, O.; Ivanciuc, T.; Cabrol-Bass, D. 3D quantitative structure activity relationships with CoRSA. Comparative receptor surface analysis. Application to calcium channel agonists. *Analysis*, **2000**, *28*, 637-642.

[98] Chuman, H.; Karasawa, M.; Fujita, T. A novel three-dimensional QSAR procedure: voronoi field analysis. *Quant. Struct.-Act. Relat.*, **1998**, *17*, 313-326.

[99] Chen, H.; Zhou, J.; Xie, G. PARM: a genetic evolved algorithm to predict bioactivity. *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 243-250.

[100] Robinson, D.D.; Winn, P.J.; Lyne, P.D.; Richards, W.G. Self-organizing molecular field analysis: a tool for structure-activity studies. *J. Med. Chem.*, **1999**, *42*, 573-583.

[101] Korhonen, S.P.; Tuppurainen, K.; Laatikainen, R.; Perakyla, M. FLUFF-BALL, a template-based grid-independent superposition and QSAR technique: validation using a benchmark steroid data set. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 1780-1793.

[102] Kotani, T.; Higashiura, K. Comparative molecular active site analysis (CoMASA). 1. An approach to rapid evaluation of 3D QSAR. *J. Med. Chem.*, **2004**, *47*, 2732-2742.

[103] Zhou, P.; Tong, J.; Tian, F.; Li, Z. A novel comparative molecule/ pseudo receptor interaction analysis. *Chin. Sci. Bull.*, **2006**, *51*, 1824-1829.

[104] Berk, R.A. Simple Linear Regression. In: *Regression Analysis: A Constructive Critique;* Berk, R.A., Ed.; SAGE Publications Ltd: London, **2003**, pp. 21-38.

[105] Berk, R.A. The Formalities of Multiple Regression. In: *Regression Analysis: A Constructive Critique;* Berk, R.A., Ed.; SAGE Publications Ltd: London, **2003**, pp. 103-110.

[106] Berk, R.A. Some Popular Extensions of Multiple Regression. In: *Regression Analysis: A Constructive Critique;* Berk, R.A., Ed.; SAGE Publications Ltd: London, **2003**, pp. 125-150.

[107] Wold, S.; Johansson, E.; Cocchi, M. PLS : Partial Least Squares Projections to Latent Structures. In: *3D QSAR in Drug Design: Theory, Methods and Applications;* Kubinyi, H., Ed.; ESCOM Science Publishers: Leiden, **1993**, pp. 523-550.

[108] Dunteman, G.H. Basic Concepts of Principal Components Analysis. In: *Principal Components Analysis;* Dunteman, G.H., Ed.; SAGE Publications Ltd: London, **1989**, pp. 15-22.

[109] Dunteman, G.H. Uses of Principal Components in Regression Analysis. In: *Principal Components Analysis;* Dunteman, G.H., Ed.; SAGE Publications Ltd: London, **1989**, pp. 65-74.

[110] Rogers, D.; Hopfinger, A.J. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 854-866.

[111] Dunn III, W.J.; Rogers, D. Genetic partial least squares in QSAR. In: *Genetic algorithms in molecular modeling;* Devillers, J., Ed.; Academic Press: London, **1996**, pp. 109-130.

[112] Hyde, R.M.; Livingstone, D.J. Perspectives in QSAR: computer chemistry and pattern recognition. *J. Comput. Aided Mol. Des.*, **1988**, *2*, 145-155.

[113] Aldenderfer, M.S.; Blashfield, R.K. A Review of Clustering Methods. In: *Cluster Analysis;* Aldenderfer, M.S., Blashfield, R.K., Eds.; SAGE Publications Ltd: London, **1984**, pp. 33-61.

[114] Baskin, II; Palyulin, V.A.; Zefirov, N.S. Neural networks in building QSAR models. *Methods Mol. Biol.*, **2008**, *458*, 137-158.

[115] Ajmani, S.; Jadhav, K.; Kulkarni, S.A. Three-dimensional QSAR using the k-nearest neighbor method and its interpretation. *J. Chem. Inf. Model.*, **2006**, *46*, 24-31.

[116] Archdeacon, T.J. Regression and explained variance. In: *Correlation and Regression Analysis: a Historian's Guide;* Archdeacon, T.J., Ed.; Univ of Wisconsin Press: USA, **1994**, pp. 178-196.

[117] Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. Stat. Soc. B*, **1974**, *36*, 111-147.

[118] Deep, R. Regression. In: *Probability and Statistics;* Deep, R., Ed.; Academic Press: UK, **2006**, pp. 455-515.

[119] Shao, J. Bootstrap model selection. *J. Am. Stat. Assoc.*, **1996**, *91*, 655-665.

[120] Rucker, C.; Rucker, G.; Meringer, M. y-Randomization and its variants in QSPR/QSAR. *J. Chem. Inf. Model.*, **2007**, *47*, 2345-2357.

[121] Marshall, G.R. Binding-Site Modeling of Unknown Receptors. In: *3D QSAR in Drug Design: Theory Methods and Applications;* Kubinyi, H.; Martin, Y.C.; Folkers, G.; Eds.; Springer Publications: London, UK, **1998**, pp. 80-116.

[122] Archdeacon, T.J. Evaluating the Regression Equation. In: *Correlation and Regression Analysis: a historian's Guide;* Archdeacon, T.J., Ed.; Univ of Wisconsin Press: USA, **1994**, pp. 160-177.

[123] Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.*, **2007**, *26*, 694-701.

[124] Topliss, J.G.; Costello, R.J. Chance correlations in structure-activity studies using multiple regression analysis. *J. Med. Chem.*, **1972**, *15*, 1066-1068.

[125] Unger, S.H.; Hansch, C. On model building in structure-activity relationships. A reexamination of adrenergic blocking activity of beta-halo-beta-arylalkylamines. *J. Med. Chem.*, **1973**, *16*, 745-749.

Received: May 6, 2009 Revised: September 20, 2009