Quantitative Structure-Activity Relationship (QSAR) Paradigm – Hansch Era to New Millennium

Asim Kumar Debnath*

Lindsley F. Kimball Research Institute of The New York Blood Center, 310 E 67th Street, New York, NY 10021, U.S.A.

Abstract: The analysis of structure-activity relationships started probably more than hundred years ago but the concept of quantitatively correlating physicochemical properties of molecules with their biological activities, termed as quantitative structure-activity relationship (QSAR), was initiated by Corwin Hansch and his groups in early 1960. Many new methods have emerged since then. The concept evolved from 2D QSAR to 3D QSAR and lately another dimension (4D QSAR) has been added. This evolution is briefly reviewed here.

INTRODUCTION

Understanding the role of chemical structure in influencing biological activity is critical. Though the quest for structure-activity relationship studies started in late 19th century, only the work of Corwin Hansch in the early 1960s put forth a mathematical model to correlate biological activity with chemical structure and revolutionized the field of drug research. For the last forty years, the field has progressed immensely and several review articles covering different aspects of this field have been published [1-7]. Due to the limited scope and space for this mini-review, the author will only focus on the evolution of different QSAR methods for drug design. No attempt will be made to describe any descriptors (parameters) or statistical techniques necessary to develop QSAR models.

CLASSICAL 2D QSAR

Hansch's Method and Related Approaches

The first application of QSAR came from Hansch et al. in 1962 when they correlated the plant growth regulatory activity of phenoxyacetic acids to Hammett constants and partition coefficients [8]. The major breakthrough in QSAR occurred later in 1964, when Hansch et al., showed that the biological activity could be correlated linearly by free-energy related terms (different physicochemical parameters) [9]. This approach was originally coined as Linear Free Energy Relationships (LFER) and later changed, more appropriately, to extra thermodynamic approach and expressed by the following equation:

$$\log 1/C = a\pi + b\sigma + cE_S + \text{.....} + \text{constant}$$

where C is the molar concentration of the compound to produce a defined biological response, π is the hydrophobic contribution of the substituent and represented by log P_X/P_H, σ is the Hammett electronic descriptor of the substituents [10], represented by log K_X/K_H, E_S is Taft's steric parameter [11] and a, b and c are the appropriate coefficients. In these expressions P_X and P_H are the octanol/water partition coefficients of the substituted and unsubstituted compounds, respectively, and K_X and K_H are the ionization constants of the meta- or para-substituted and unsubstituted benzoic acids at 25 °C, respectively.

Correlation models can be generated by using a single parameter or a combination of parameters. The parameters can be from experimental values (e.g., π, log P, σ, etc.) or from theoretically calculated values [e.g., ClogP, energies of lowest unoccupied molecular orbitals (LUMO) and highest occupied molecular orbitals (HOMO), charge, etc].

Hansch and co-workers introduced the parabolic model in QSAR analysis after realizing that biological activity of hydrophobic drugs started to level of or decrease after reaching the optimum value. This was attributed to the entrapment of the drugs in the lipid phase of the transport process [12,13]. They analyzed a series of datasets and proposed a second-order relationship of hydrophobicity (log P) with biological activity as follows:

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

where a and b are the coefficients of the log P and $(\log P)^2$ terms, respectively, and c is a constant term.

Kubinyi made an important contribution in QSAR methodology in 1977 when he first proposed the bilinear model [14,15]. In a large number of cases, it has been found that the biological activity increases with hydrophobicity linearly up to a certain point and then decreases in a linear fashion. The differences between observed and calculated biological activities were found to be high if parabolic models were used. The bilinear model to describe this non-linear dependence of biological activity of drugs on hydrophobicity was expressed as:

$$\log 1/C = a \log P - b \log (\beta P + 1) + c$$

*Address correspondence to this author at the Lindsley F. Kimball Research Institute of The New York Blood Center, 310 E 67th Street, New York, NY 10021, U.S.A.; Tel. (212) 570-3373; Fax. (212) 570-3299; E-mail: asim_debnath@nybc.org

1389-5575/01 $20.00+.00 © 2001 Bentham Science Publishers, Ltd.

The terms a, b and c are linear in nature and can be calculated by multiple regression analysis whereas β is a non-linear term and must be calculated by an iterative method.

The introduction of Hansch's simple linear and parabolic models and Kubinyi's bilinear model made a considerable impact on our understanding of how chemical structures influence biological activity. The application of these methods surged dramatically, as reflected in a sharp increase in the number of publications. During 1962-1969, just Hansch's group published forty-three papers related to QSAR (personal communications), which covered a wide variety of areas e.g., plant growth regulation, enzyme inhibition, tissue distribution, sweetness, drug-protein interactions, hapten-antibody interactions, metabolism to name a few. In period from 1970-1979, 113 papers were published by Hansch's group covering the areas already mentioned along with applications in pesticide design, antibacterial, antifungal, anticancer and antimalarial drugs, immunochemistry, fibrinolytics, etc. During 1980-1989 Hansch's group published 80 papers on QSAR, many of which dealt with cytotoxicity, mutagenicity and carcinogenicity issues. Hansch's group also incorporated the three-dimensional structure of receptors in understanding the mechanistic aspects of ligand-receptor interactions [16-19]. This trend continued with 58 publications by this group during 1990-1999 and many of these articles emphasized comparative QSAR studies [20-24].

Hansch's era in QSAR, which started some 40 years ago is still continuing in full force by contributing QSAR related publications and by generating software and databases to lay the foundation, in his words, for 'a science of chemico-biological interactions'. Towards this end, for the last 35 years Hansch's group at Pomona College has been collecting data from literature and created a computer database known as C-QSAR. The current database contains QSAR equations from both physicochemical data (7700) and Biological data (6300) (source- CQSAR.com web site). From these data 14000 QSAR equations have been developed so far, which are part of this database. The greatest advantage of these databases is that they provide the opportunity to derive comparative QSAR from many points of view and to facilitate lateral validation of any new QSAR with already developed QSAR in the database.

BEYOND HANSCH

Another QSAR approach was proposed by Free and Wilson in 1964 based on a simplistic mathematical approach for structure-activity studies in a congeneric series [25]. The method was based on the additivity principle. According to this principle, substituent groups contribute in constant amounts and in an additive manner towards biological activity and do not depend on any other structural changes in the compound. In their original publication [25], Free and Wilson put forward a generalized mathematical model, as follows:

$$A = \Sigma a_i x_i + \mu$$

where A is the activity data, a_i is the contribution of the *i*th substituent and x_i takes a value of 1 when the substituent is present and 0 when it is absent; μ is the average contribution of the parent molecule and is a constant term.

The mathematical model also includes symmetry equations (often called restriction equations), which assume that the summation of contributions of all substituents in a particular position is 0. According to this assumption, $\Sigma a_i x_i = 0$. Several structure-activity studies have used this method to develop models [26-30].

The major drawbacks of the Free-Wilson method are: (1) the activity contribution of all substituents including H has to be considered; (2) the summation of the group contributions at each position, the so called symmetry assumption, has to be zero; (3) the constant term (μ) should be an over-all average of the biological activity of all the compounds used to develop the QSAR model.

To circumvent the major drawbacks of the Free-Wilson method, Fujita and Ban in 1971, proposed a modified mathematical equation [31], using the logarithm of activity, which is a free energy related term and additive in nature, represented as follows:

$$\log A/A_0 = \Sigma G_i X_i$$

where A and A_0 are the activity data of the substituted and unsubstituted compounds, respectively, and G_i is the logarithm of the activity contribution of the *i*th substituent, and X_i has a value of 1 or 0 depending on the presence of the substituent or its absence, respectively.

For a set of substituents, the equation takes the following form:

$$\log A = \Sigma G_i X_i + \mu$$

where μ is a constant.

The major advantages of this modified method are: (1) the structural matrix does not need to be transformed; (2) no restriction equation is necessary; (3) the group contribution at each position is based on the parent compound (i.e., H); (4) the constant term (μ) is calculated by the least square method and is the theoretically predicted value for the unsubstituted compound. The addition or omission of a compound does not affect markedly the value of group contributions. These advantages make the Fujita-Ban method preferable over the Free-Wilson method. Several applications of this method have been reported [32-35].

Kier and Hall have introduced a method for decoding structural features such as size, branching, unsaturation, cyclicity, heteroatom content, etc. in quantitative terms and designated molecular connectivity [36-38]. These structural indices were used as parameters to correlate structure with activity (property). The indices are calculated based on the hydrogen suppressed molecular structure or graph. Simple connectivity indices ($^0\chi$, $^1\chi$, $^2\chi$), to more complex valence connectivity indices ($^0\chi_v$, $^1\chi_v$, etc.), Kappa indices ($^1\kappa$, $^2\kappa$, etc.) and lately electrotopological state (E-state values) and Molconn-Z parameters have been used in numerous structure-activity applications [39-42]. Several other

topological indices proposed by Randic [38,43-45], Basak [46-50] and Balaban [48,51,52] have also been used in toxicity prediction and many biological (QSAR) and quantitative structure property relationship (QSPR) studies. These indices have been used in diversity analysis [53] and in analyzing drug likeness of molecules in databases [54].

Klopman and his coworkers developed a new generation Computer Automated Structure Evaluation (CASE and multi-CASE) program useful in drug design [55,56]. The program generates fragments consisting of 2-10 heavy atoms of all possible chains from the input structure. These fragments are considered as structural descriptors and used to derive QSAR models for prediction of biological activity. Fragments that are responsible for activity (biophore), and those detrimental for activity (biophobe) are identified by statistical evaluation based on a binomial distribution. For the prediction of biological activity of a new compound, the program uses the information from the learning set and assigns probabilities for the compound to be active or inactive depending on the presence of biophores or biophobes.

Klopman and Ptchelintsev recently applied the Multi-CASE method to a series of 71 triazole alcohols to derive structure-antifungal, structure-teratogenicity and structure-therapeutic index relationships [57]. In a more recent study, Klopman and Tu have selected a set of 1819 chemicals out of 14,156 tested by the National Cancer Institute (NCI) and identified 74 fragments that could explain the anti-HIV activity of these compounds using this method [58].

Neural network (NN) techniques have been used successfully in QSAR [59-66]. This method is generally used when the data set size is large and the data cannot be interpreted easily by linear functions. This method is generally used in QSARs to describe a model with a non-linear hypersurface.

Although there are different ways of constructing neural networks, the multilayer feed-forward network with back-propagation is the one primarily used in drug design. In such a model the units are organized in layers starting with input units that are connected to the output unit through some layers of hidden units. Signals of representative input information about the drugs (parameters) are propagated forward using connecting weights, from the input layer to the output layer via the hidden units, and the output signals represent the predicted activity. Differences between the predictions and known activity are then used to adjust the weights "backward" until those differences become small. The major step in a neural network is to train the network using a representative training set. The design aspect of the neural network is very critical. If a network is trained with a large number of parameters it may over train the model and generate unreliable prediction results due to overfitting. On the other hand, an under trained neural network with too few parameters may generate poor results for new predictions. The quality of the fit can be validated by "leave-one-out" cross validation whereby data are removed systematically, a neural network is trained and a prediction of the removed data is made based on the trained network. The residual root mean squared error (RmsE) or correlation coefficient (R) values are also calculated to validate the model. The Bayesian regularization algorithm has also been used to eliminate the need for a test model since it minimizes a linear combination of errors and weights [64,67].

Breindl et al. recently reported the use of a back-propagation artificial neural network for predicting the octanol/water partition coefficient (log P) of a large number of organic chemicals [68].

The major advantages of the neural networks are that they are non-parametric and non-linear and few statistical assumptions are required to build the model. Their major disadvantage is that the models cannot be easily interpreted, especially in physicochemical terms.

NEW QSAR METHODS

HQSAR

Hologram QSAR (HQSAR) is a relatively new technique, which does not require any physicochemical descriptors or 3D structure to generate structure-activity models [69]. The method is based on the input of 2D structures and biological activity. The structures are converted to all possible linear, branched and overlapping fragments. These fragments are then assigned a specific integer by using a cyclic redundancy check (CRC) algorithm. These integers are then hashed to a bin in an integer array of fixed length. These arrays are known as the molecular hologram and the bin occupancies of the molecular holograms are used as the descriptors. These descriptors are expected to encode the chemical and topological information of molecules. The QSAR model is developed by using the partial least squares (PLS) regression technique and validated by the "leave-one-out" cross-validation technique. Once the final model is obtained, PLS yields the following equation correlating hologram bins with activity:

$$A_i = C + \sum_{l=1}^{L} X_{il} C_{il}$$

In this expression, A_i is the activity of compound i, C is a constant, X_{il} is the hologram occupancy value at position i or bin l and C_{il} is the coefficient for the corresponding bin from the PLS run, L is the hologram length.

The fundamental difference between the HQSAR method and other fragment based methods, e.g., the Free-Wilson method, the multi-CASE method, is that HQSAR encodes all possible fragments including overlapping fragments. The method is very fast and can also be used to predict physicochemical properties, e.g. ClogP. Several applications of the HQSAR method were reported recently [69-71].

Inverse QSAR

An inverse QSAR method implemented in a new library design technique, known as Focus-2D, has been recently reported by Cho et al. to rationally design a virtual peptide

combinatorial library [72,73]. A preconstructed QSAR was used, as one of the methods, to select compounds with high-predicted activity in a virtual library. The method was validated by developing a QSAR equation using GA-PLS (Genetic Algorithm-Partial Least Squares) method from a training set of 28 bradykinin-potentiating (BK) pentapeptides and predicting the activity of the two most active peptides from the equation. Topological descriptors, calculated by the Molconn-X program; several amino acid based descriptors (Z_1, Z_2 and Z_3) [74] related to hydrophobicity, bulk and electronic properties, respectively; isotropic surface area (ISA) and electronic charge index (ECI) were used [75]. Significant cross-validated correlation coefficients and low standard errors of predictions were achieved with both the topology-based study and the amino acid descriptor-based study. The method suggested a number of amino acids as the preferred building blocks. These amino acids were also present most frequently in the known active BK peptides. The results obtained from the training set of 28 pentapeptides were used to extrapolate on all the theoretically possible pentapeptides and comparable results were obtained. As the training set population was very small compared to the number of theoretically possible peptides, a modified "degree of fit" condition was used to control the degree of extrapolation and not to predict peptides that are structurally too distant from the training peptides.

Binary QSAR

Introduction of combinatorial chemistry for designing large libraries compelled researchers to discover rapid robotic methods for assaying millions of compounds in a short period of time. This rapid method is referred to as high throughput screening (HTS). Often, this method just generates yes/or no (active/or inactive; pass/ or fail) data and the results are prone to error. Current QSAR methodologies require less heterogeneous compounds with continuous activity data and lower error margins to have any predictive value. To overcome methodological problems in current QSAR techniques and to handle such a huge amount of binary data from HTS, Labute introduced a method termed 'binary QSAR' to handle binary measurement data from HTS [76]. Two recent reports described the successful use of this method for analyzing large sets of binary data [77,78]. This method is expected to help in extracting important structural information required for biological activity and to design more focused libraries for drug discovery. For methodological details readers are referred to recent articles of Labute's [76] and Gao et al. [78].

The performance of the QSAR model is measured by evaluating three levels of prediction from the model. If m_0 is the number of active compounds, m_1 is the number of inactive compounds and c_0 is the number of correctly predicted active and c_1 is the number of correctly predicted inactive compounds by the model, then

$100(c_0/m_0)$ represents the percentage accuracy for active compounds;

$100(c_1/m_1)$ represents the percentage accuracy for inactive compounds; and

$100(c_0 + c_1)/ (m_0 + m_1)$ represents the percentage of overall accuracy for all compounds.

The advantages and disadvantages of the binary QSAR method have also been addressed by Labute [76]. The method should be useful for (a) selecting (prioritizing) compounds for HTS, (b) designing focused combinatorial libraries, and (c) screening and synthesizing virtual libraries. One major drawback of this method is that the interpretation of the importance of descriptors in developing the model is not easy.

The first report concerning the application of binary QSAR to a drug discovery problem involving the QSAR analysis of estrogen receptor ligand was published recently by Gao et al. [78].

3D QSAR METHODS

The QSAR approaches by Hansch's group and others mentioned above have provided us with the tools to quantitate the relationships between the structural parameters associated with the change in structure and physical or biological activities. These methods have helped us to delineate drug-receptor interactions and to develop several commercial drugs and pesticides [7]. However, these methods have some limitations, the most serious among them is the lack of availability of numerical descriptions, for new or unusual substituents, which are required to develop any meaningful model. The other limitations are that one needs considerable knowledge in physical organic chemistry to design a molecule based on the prediction made by any QSAR equation and the results cannot be interpreted graphically to understand the interaction pattern. The proponents of 3D QSAR methods believe that most traditional QSAR limitations can be overcome by 3D QSAR methods. Several 3D QSAR modeling approaches have emerged in 1980s such as, active analog approach, molecular shape analysis, distance geometry and CoMFA. Several new methods have been proposed in the 1990s. This field has grown so much during the last decades that it is beyond the scope of this review to go over the subject matter in detail instead the author will touch upon some of the most widely used and newer 3D QSAR methods and their applications for drug design.

Hopfinger et al. in 1980 incorporated the molecular shape of molecules in the QSAR analysis and they designated the method as molecular shape analysis (MSA) [79]. The primary goal of this method is to incorporate conformational analysis in QSAR. The few basic steps to construct 3D QSAR using this method are as follows: (1) Conformational analysis; (2) Identification of the most likely biologically active conformation of a set of molecules under study; (3) Selection of a candidate shape reference compound; (4) Superposition of each active conformation onto the reference molecule: (5) Determination of molecular shape descriptors; (6) Determination of possible other Hansch type descriptors; (7) Construction of trial MSA 3D QSAR

Several molecular descriptors such as the common overlap steric volume V_{ov}, nonoverlap volume V_{non}, etc.

have been calculated as a quantitative measure of molecular shape. These 3D descriptors along with standard physicochemical properties used in Hansch analysis are used to generate trial 3D QSAR models. Applications of this MSA 3D QSAR method have been described [79-82].

Golender et al. in 1983 introduced an expert system, known as Apex-3D that simulates the intelligence of a drug researcher in identifying the pharmacophores responsible for activity or inactivity from a set of molecules and their biological activity [83,84]. Therefore, this method can be used to automatically deduce the biophore (pharmacophore) from three-dimensional structures and biological activity data. This biophore can be used as a reference for optimizing the superposition of ligands to build the 3D QSAR models, which can be used to predict the activity of new compounds to be synthesized. A few applications of this method have been reported [85-87].

The major advances in 3D QSAR methodology came with the introduction of the Comparative Molecular Field Analysis (CoMFA) technique by Crammer et al. in 1988 [88]. They envisioned that the changes in biological activity (or binding affinities) of any molecule could be correlated with their steric and electrostatic energies. The important steps in the CoMFA method are: (1) determination of the bioactive conformation of each molecule; (2) alignment of the molecules using either manual or automated methods in a manner that best represents the interaction of the molecule with the target receptor; (3) sampling of steric and electrostatic fields with a probe (e.g. sp^3 with a 1+ charge) placed in all intersections of a regularly placed grid with a grid size of 1-2 Å that encompasses all the molecules; (4) establishment of quantitative relationships of these interaction energies with biological activity by the partial least squares (PLS) technique and cross-validation; (5) displaying results through contour plots for visual understanding.

Since the first publication of this method and its inception in the computer program SYBYL [89] a surge in applications of CoMFA occurred. For an excellent compilation of references till 1997 see Kim [90,91]. Despite the overwhelming success of CoMFA, this method, like any other methods, has some limitations [92]. To overcome these limitations several new techniques were developed. One of the most difficult problems often encountered in developing CoMFA models is the alignment of molecules. The alignment basically represents the orientation of a molecule as it binds to a receptor site. Any miscue in this crucial stage may produce erroneous and misleading predictions of the interaction sites. To avoid the alignment problem, some new descriptors such as CoMMA [93], EVA [94,95], MS-WHIM [96], CoMSIA [97,98] have been proposed. Their properties do not depend on the orientation of the molecule and also no alignment is necessary to develop these descriptors. Several new fields, besides CoMFA fields (steric and electrostatic), have also been proposed e.g., hydrophobic fields such as hydrophobic interactions (HINT) [99], molecular lipophilicity potentials (MLP) [100]; H-bonding field using GRID [101]; molecular orbital fields such as HOMO and LUMO [102,103]; electrotopological state (E-state) fields [39]; desolvation energy fields using the DELPHI program [102]. The steric field based on indicator variables has also been used [104]. Due to a large number of variables in CoMFA type studies, it became evident that perhaps improved as well as newer methods were needed to discriminate between important and less important variables. A number of methods have been proposed and applied such as GOLPE (generating optimal linear PLS estimation) [105,106] and q^2-GRS (cross-validated (q^2) q2-guided region selection) [107].

In 1988, Doweyko introduced another 3D QSAR method known as the hypothetical active-site lattice (HASL) [108]. The method involves the creation of a four-dimensional (4D) molecular lattice space based on the three-dimensional Cartesian coordinates of an arbitrarily selected compound from a set of compounds to be studied and assignment of parameters corresponding to the physicochemical properties of the atom in that space (representing the fourth dimension). A second molecule is chosen and the lattice is generated as before and then compared with the reference molecule. The degree of correspondence between these molecules is determined by using a FIT function. The FIT function, which is a convenient measure of matching, can be represented as a sum of the fraction of the lattice points of a molecule and the reference lattices that are found to be common and represented by the following simple expression:

$$FIT = L\,(common) / L\,(ref) + L\,(common) / L\,(molecule)$$

When the molecular lattices are identical to the reference lattice then the value of the Fit function will be 2. This provides an efficient way of matching two molecules. This information and the binding data are then merged to create the HASL, which is expected to effectively capture the shape and binding properties of the receptor site. The HASL can be used to predict the binding and orientation of an inhibitor. This method has been applied to derive QSAR in several cases [109,110].

In 1989, Ghose et al. proposed a new method of modeling (REMOTEDISC) the binding site cavity based on binding data for a series of molecules [111]. The method utilizes the three-dimensional structures and the physicochemical properties of molecules (calculated using atom-based methods) to model the binding site cavity. The method assigns different weights to the physicochemical properties of the molecule at different parts of the binding cavity that determines the binding of the molecule. This method has also been applied to model binding site cavities [112,113].

Walters et al. in 1994 introduced a method to construct three-dimensional atom-based models of receptor sites using only the structure-activity information of a small set of compounds. The method was termed as genetically evolved receptor models (GERM) [114]. A genetic algorithm is used in this program to construct and improve the model so that the binding energy correlates with the experimentally determined biological activity of the test series. The activities of the yet to be synthesized compounds can be predicted by docking them onto the receptor model. The

method has been used in generating 3D QSAR models [115].

In 1994, Jain et al. reported 3D QSAR method known as COMPASS that utilizes the molecular surface properties to predict biological activity [116]. This method differs from other methods such as CoMFA in that it utilizes the physical properties, e.g., electrostatic field only at the interface of the ligand and receptor. It uses non-linear rather than linear statistics (such as MLR, PLS), and it automatically selects the most probable bioactive conformation and alignment. The method has been successfully applied [117].

Hahn in 1995 proposed a receptor site modeling method known as receptor surface models (RSM) [118,119]. A series of aligned molecules with binding data are used to generate the receptor surface models. These models can be used to visualize the receptor-ligand interaction in a qualitative and intuitive manner. These models can also be used to generate descriptors that truly represent the three-dimensional nature of the receptor site and these descriptors alone or in combinations with 2D descriptors can be used to derive QSAR. This method was also applied to the steroids binding data [120,121] used by many as benchmark for method validation and yielded somewhat better results than were those obtained by either CoMFA or COMPASS methods.

In 1997, So and Karplus put forward a 3D QSAR method using molecular similarity matrices (SM) and a genetic neural network (GNN) to derive the predictive model, thus the name SM/GNN method [122]. Different types of similarity matrices such as electrostatic similarity matrix (ESM), shape similarity matrix (SSM) and van der Waals (vdW) similarity matrix (VSM) were used as molecular field parameters. A genetic algorithm was used to select the variables and the biological activity was correlated with these descriptors by a neural network technique. The method was validated by the most commonly used corticosteroid-binding globulin (CBG) data [122]. Unlike CoMFA and CoMSIA, this method does not have any visual analysis option; therefore it is difficult to interpret. This method is also heavily dependent on the alignment.

Recently (2000), Polanski and Walczak reported the use of molecular electrostatic potential (MEP) calculated for specific areas on the molecular surface in 3D QSAR analyses [123]. This method is known as COMSA. The method has been applied to the steroid dataset [120,121] that was used by Crammer et al. for the CoMFA study and obtained somewhat improved results. A kohonen self-organizing neural network and the partial least squares (PLS) methods were used to develop the models. The method has been applied to derive 3D QSAR models with data on both biological activities and on physicochemical properties [123].

4D QSAR Analyses in Developing 3D QSAR Models

Hopfinger et al. recently (1997) introduced a fourth-dimension to the 3D QSAR modeling and termed it as 4D QSAR analysis [124]. The fundamental difference of this method from CoMFA is that it incorporates conformational and alignment freedom to the model development from a set of structures with biological activity data by performing conformational ensemble sampling. The fourth-dimension of the 4D QSAR analysis comes from ensemble averaging. According to the method, the active conformation, unlike that of many other 3D QSAR methods, is not the minimum energy conformation but the "active" conformation that optimizes the 3D QSAR models. Instead of using systematic conformational search, a multi-temperature molecular dynamics (MDS) is used to generate the conformational ensemble profiles (CEF). Sampling all the conformations that are within 2 Kcal/mole of the minimum energy conformation identifies the best "active" conformation of each compound in the training set. This set of low energy conformations is evaluated individually in the best 3D-QSAR models. The alignment problem is resolved by a similar sampling and evaluation technique. The conformations of each compound are placed in a predefined grid cell (similar to CoMFA) according to the trial alignment, and 3D descriptors are computed and then correlated with biological activity by the PLS technique. The 3D QSAR models are generated from the PLS run. The method is repeated until all the trial alignments are included to develop the models. The active conformation is selected from the optimum 3D QSAR model that predicts the observed activity best. The PLS weighting for the descriptors can be used as in CoMFA to generate graphical contours. This method has been applied to develop 3D QSAR models [124,125].

Vedani et al. recently have introduced a 4D QSAR approach implemented in software called Quasar [126]. According to these authors, this software uses 4D QSAR concepts by incorporating an ensemble of multiple conformations, orientations or protonation states of each molecule as the fourth dimension to the model and thereby reduces the bias resulting from selecting any single conformation as the bioactive conformation and the alignment resulting from those conformations. This method is especially useful when there is no receptor structure available to estimate the free energy of binding of the ligand to the receptor. The method has been successfully applied to a large series of NK-1 antagonists to develop 4D QSAR models [126].

ACKNOWLEDGEMENT

I thank Drs. A. Robert Neurath and Cathy Falk for thoroughly reading the manuscript and their critical comments. This work was supported by grants from NIH (AI 46221), Phillip Morris, Inc, Hugoton Foundation and Johnson & Johnson.

REFERENCES

[1] Topliss, D. J. Quantitative Structure-Activity Relationships of Drugs, Academic Press: New York, 1983.

[2] Hansch, C.; Fujita, T. *Classical and Three-Dimensional QSAR in Agrochemistry*, American Chemical Society: Washington, DC, 1994.

[3] Fujita, T. QSAR and Drug Design: New Developments and Applications, Elsevier: Amsterdam, 1995.

[4] Devillers, J. *Comparative QSAR*, Taylor & Francis: Washington, DC, 1997.

[5] Parrill, A. L.; Reddy, M. R. *Rational Drug Design - Novel Methods and Practical Applications*, American Chemical Society: Washington, DC, 1999.

[6] Hansch, C.; Leo, A. *Exploring QSAR Fundamentals and applications in chemistry and biology*, American Chemical Society: Washington, DC, 1995.

[7] Debnath, A. K. Quantitative Structure-Activity Relationship (QSAR) - A versatile tool in drug design, In *Combinatorial Library Design and Evaluation: Principles, Software Tools, and Applications*, Ghose, A. K.; Viswanadhan, V. N., Eds.; Mercel Dekker, Inc.: New York, **In Press**.

[8] Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. *Nature* **1962,** *194,* 178.

[9] Hansch, C.; Fujita, T. *J. Am. Chem. Soc.* **1964,** *86,* 1616.

[10] Hammett, L. P. *Physical Organic Chemistry*, McGraw-Hill: New York, 1970.

[11] Taft, R. W. *Steric Effects in Organic Chemistry*, Wiley: New York, 1956.

[12] Hansch, C.; Steward, A. R.; Anderson, S. M.; Bentley, D. L. *J. Med. Chem.* **1968,** *11,* 1.

[13] Hansch, C.; Clayton, J. M. *J. Pharm. Sci.* **1973,** *62,* 1.

[14] Kubinyi, H. *J. Med. Chem.* **1977,** *20,* 625.

[15] Kubinyi, H. *Farmaco [Sci]* **1979,** *34,* 248.

[16] Carotti, A.; Smith, R. N.; Wong, S.; Hansch, C.; Blaney, J. M.; Langridge, R. *Arch. Biochem. Biophys.* **1984,** *229,* 112.

[17] Hansch, C. *Drug Intell. Clin. Pharm.* **1982,** *16,* 391.

[18] Hansch, C.; McClarin, J.; Klein, T.; Langridge, R. *Mol. Pharmacol.* **1985,** *27,* 493.

[19] Selassie, C. D.; Fang, Z. X.; Li, R. L.; Hansch, C.; Debnath, G.; Klein, T. E.; Langridge, R.; Kaufman, B. T. *J. Med. Chem.* **1989,** *32,* 1895.

[20] Gao, H.; Katzenellenbogen, J. A.; Garg, R.; Hansch, C. *Chemical Reviews* **1999,** *99,* 723.

[21] Garg, R.; Gupta, S. P.; Gao, H.; Babu, M. S.; Debnath, A. K.; Hansch, C. *Chemical Reviews* **1999,** *99,* 3525.

[22] Hansch, C.; Hoekman, D.; Gao, H. *Chemical Reviews* **1999,** *96,* 1045.

[23] Garg, R.; Denny, W. A.; Hansch, C. *Bioorg. Med. Chem.* **2000,** *8,* 1835.

[24] Hansch, C.; McKarns, S. C.; Smith, C. J.; Doolittle, D. J. *Chem. Biol. Interact.* **2000,** *127,* 61.

[25] Free, S. M.; Wilson, J. W. *J. Med. Chem.* **1964,** *7,* 395.

[26] Borea, P. A. *Boll. Soc. Ital. Biol. Sper.* **1981,** *57,* 633.

[27] Cativiela, C.; García, J. I.; Dib, L.; Ballesti, P. *Prog. Clin. Biol. Res.* **1989,** *291,* 395.

[28] Craig, P. N. *J. Med. Chem.* **1972,** *15,* 144.

[29] Kubinyi, H. The Free-Wilson method and its relationship to the extrathermodynamic approach, In *Comprehensive Medicinal Chemistry Volume 4: Quantitative Drug Design*, Hansch, C.; Sammes, P. G.; Taylor, J. B.; Ramsden, C. A., Eds.; Pergamon Press: Oxford, **1990**; Vol. 4 , pp. 589-643.

[30] Zahradnik, P.; Foltinova, P.; Halgas, J. *SAR QSAR Environ. Res.* **1996,** *5,* 51.

[31] Fujita, T.; Ban, T. *J. Med. Chem.* **1971,** *14,* 148.

[32] Gombar, V. *Arzneimittelforschung* **1986,** *36,* 1014.

[33] Hernandez-Gallegos, Z.; Lehmann, P. A. *J. Med. Chem.* **1990,** *33,* 2813.

[34] Higginbottom, M.; Kneen, C.; Ratcliffe, G. S. *J. Med. Chem.* **1992,** *35,* 1572.

[35] Kubinyi, H. *Arzneimittelforschung* **1977,** *27,* 750.

[36] Hall, L. H.; Kier, L. B.; Murray, W. J. *J. Pharm. Sci.* **1975,** *64,* 1974.

[37] Hall, L. H.; Kier, L. B. *J. Pharm. Sci.* **1977,** *66,* 642.

[38] Kier, L. B.; Hall, L. H.; Murray, W. J.; Randic, M. *J. Pharm. Sci.* **1975,** *64,* 1971.

[39] Kellogg, G. E.; Kier, L. B.; Gaillard, P.; Hall, L. H. *J. Comput. Aided Mol. Des.* **1996,** *10,* 513.

[40] Kier, L. B.; Murray, W. J.; Hall, L. H. *J. Med. Chem.* **1975,** *18,* 1272.

[41] Kier, L. B.; Hall, L. H. *J. Pharm. Sci.* **1983,** *72,* 1170.

[42] Kier, L. B.; Hall, L. H. The electrotopological state index: An atom-centered index for QSAR, In *Advances in Drug Design, Vol. 22*, Testa, B., Ed.; Academic Press: **1992**.

[43] Pisanski, T.; Plavsic, D.; Randic, M. *J. Chem. Inf. Comput. Sci.* **2000,** *40,* 520.

[44] Randic, M. *J. Am. Chem. Soc.* **1975,** *97,* 6609.

[45] Randic, M. Topological indices, In *The Encyclopedia of Computational Chemistry*, Schleyer, P. v. R.; Allinger, N. L.; Clark, T.; Gasteiger, J.; Kollman, P. A.; Schaefer, H. F., III; Schreiner, P. R., Eds.; Wiley: Chichester, U.K., **1999**; pp. 3018-3032.

[46] Basak, S. C.; Magnuson, V. R. *Arzneimittelforschung* **1983,** *33,* 501.

[47] Basak, S. C.; Bertelsen, S.; Grunwald, G. D. *Toxicol. Lett.* **1995,** *79,* 239.

[48] Basak, S. C.; Grunwald, G. D.; Niemi, G. J. Use of graph theoretic and geometrical molecular descriptors in structure-activity relationships, In *From Chemical*

Topology to Three-Dimensional Geometry, Balaban, A. T., Ed.; Plenum Press: New York, **1997**; pp. 73-116.

[49] Basak, S. C.; Gute, B. D.; Ghatak, S. *J. Chem. Inf. Comput. Sci.* **1999,** *39,* 255.

[50] Gute, B. D.; Grunwald, G. D.; Basak, S. C. *SAR QSAR Environ. Res.* **1999,** *10,* 1.

[51] Balaban, A. T. *Chem. Phys. Lett.* **1982,** *89,* 399.

[52] Balaban, A. T.; Mills, D.; Basak, S. C. *J. Chem. Inf. Comput. Sci.* **1999,** *39,* 758.

[53] Lewis, R. A.; Mason, J. S.; McLay, I. M. *J. Chem. Inf. Comput. Sci.* **1997,** *37,* 599.

[54] Dixon, S. L.; Villar, H. O. *J. Comput. Aided Mol. Des.* **1999,** *13,* 533.

[55] Klopman, G.; Macina, O. T. *J. Theor. Biol.* **1985,** *113,* 637.

[56] Klopman, G. *Quant. Struct.-Act. Relat.* **1992,** *11,* 176.

[57] Klopman, G.; Ptchelintsev, D. *J. Comput. Aided Mol. Des.* **1993,** *7,* 349.

[58] Klopman, G.; Tu, M. *J. Med. Chem.* **1999,** *42,* 992.

[59] Hatrik, S.; Zahradnik, P. *J Chem Inf.Comput. Sci.* **1996,** *36,* 992.

[60] So, S. S.; Karplus, M. *J. Med. Chem.* **1996,** *39,* 5246.

[61] Tmej, C.; Chiba, P.; Schaper, K. J.; Ecker, G.; Fleischhacker, W. *Adv. Exp. Med. Biol.* **1999,** *457,* 95.

[62] Ajay *J. Med. Chem.* **1993,** *36,* 3565.

[63] Aoyama, T.; Suzuki, Y.; Ichikawa, H. *J. Med. Chem.* **1990,** *33,* 2583.

[64] Burden, F. R.; Winkler, D. A. *J. Med. Chem.* **1999,** *42,* 3183.

[65] Duprat, A. F.; Huynh, T.; Dreyfus, G. *J. Chem. Inf. Comput. Sci.* **1998,** *38,* 586.

[66] Gobburu, J. V.; Shelver, W. H. *J. Pharm. Sci.* **1995,** *84,* 862.

[67] Burden, F. R.; Winkler, D. A. *J. Chem. Inf. Comput. Sci.* **1999,** *39,* 236.

[68] Breindl, A.; Beck, B.; Clark, T.; Glen, R. C. *J. Mol. Model* **1997,** *3,* 142.

[69] Naumann, T.; Lowis, D. R. First International Electronic Conference on Synthetic Organic Chemistry (ECSOC-1) **1997**.

[70] Heritage, T. W.; Lowis, D. R. Molecular hologram QSAR, In *Rational Drug Design Novel Methodology and Practical Applications*, Parrill, A. L.; Reddy, M. R., Eds.; American Chemical Society: Washington D.C., **1999**; pp. 212-225.

[71] Tong, W.; Lowis, D. R.; Perkins, R.; Chen, Y.; Welsh, W. J.; Goddette, D. W.; Heritage, T. W.; Sheehan, D. M. *J. Chem. Inf. Comput. Sci.* **1998,** *38,* 669.

[72] Cho, S. J.; Zheng, W.; Tropsha, A. *Pac. Symp. Biocomput.* **1998,** 305.

[73] Cho, S. J.; Zheng, W.; Tropsha, A. *J. Chem. Inf. Comput. Sci.* **1998,** *38,* 259.

[74] Hellberg, S.; Sjöström, M.; Skagerberg, B.; Wold, S. *J. Med. Chem.* **1987,** *30,* 1126.

[75] Collantes, E. R.; Dunn, W. J. *J. Med. Chem.* **1995,** *38,* 2705.

[76] Labute, P. *Pac. Symp. Biocomput.* **1999,** 444.

[77] Gao, H.; Bajorath, J. *Mol. Divers.* **1999,** *4,* 115.

[78] Gao, H.; Williams, C.; Labute, P.; Bajorath, J. *J. Chem. Inf. Comput. Sci.* **1999,** *39,* 164.

[79] Hopfinger, A. J. *J. Am. Chem. Soc.* **1980,** *102,* 7196.

[80] Hopfinger, A. J. *Arch. Biochem. Biophys.* **1981,** *206,* 153.

[81] Tokarski, J. S.; Hopfinger, A. J. *J. Med. Chem.* **1994,** *37,* 3639.

[82] Tokarski, J. S.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **1997,** *37,* 792.

[83] Golender, V. E.; Rozenblit, A. B. *Logical and combinatorial algorithm for drug design,* Research Studies Press, Wiley: Letchworth, 1983.

[84] Golender, V. E.; Vorpagel, E. R. Computer-assisted pharmacophore identification, In *3D QSAR in Drug Design: Theory, Methods and Applications*, Kubinyi, H., Ed.; ESCOM: Leiden, **1993**; pp. 137-149.

[85] Kurogi, Y. *Drug Des Discov.* **1999,** *16,* 109.

[86] Talele, T. T.; Kulkarni, V. M. *J Chem Inf.Comput.Sci.* **1999,** *39,* 204.

[87] Hariprasad, V.; Kulkarni, V. M. *J. Comput. Aided Mol. Des.* **1996,** *10,* 284.

[88] Crammer III, R. D.; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988,** *110,* 5959.

[89] *SYBYL*, Tripos Inc.: 1699 South Hanley Road, St. Louis, Missouri 63144, 2000

[90] Kim, K. H. List of references, 1997, In *3D QSAR in Drug Design: Recent Advances*, Kubinyi, H.; Folkers, G.; Martin, Y. C., Eds.; Kluwer/ESCOm: **1998**; pp. 334-338.

[91] Kim, K. H. List of CoMFA references, 1993-1996, In *3D QSAR in Drug design: Recent Advances*, Kubinyi, H.; Folkers, G.; Martin, Y. C., Eds.; Kluwer/ESCOM: **1998**; pp. 317-333.

[92] Folkers, G.; Merz, A.; Rognan, D. CoMFA: Scope and Limitations, In *3D QSAR in Drug Design: Theory, Methods and Applications*, Kubinyi, H., Ed.; ESCOM: Leiden, **1993**; pp. 583-618.

[93] Silverman, B. D.; Platt, D. E. *J. Med. Chem.* **1996,** *39,* 2129.

[94] Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. *J. Comput. Aided Mol. Des.* **1997,** *11,* 409.

[95] Turner, D. B.; Willett, P. *Eur. J. Med. Chem.* **2000,** *35,* 367.

[96] Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A. *J. Comput. Aided Mol. Des.* **1997,** *11,* 79.

[97] Klebe, G.; Abraham, U.; Mietzner, T. *J. Med. Chem.* **1994,** *37,* 4130.

[98] Klebe, G.; Abraham, U. *J Comput.Aided Mol.Des* **1999,** *13,* 1.

[99] Kellogg, G. E.; Semus, S. F.; Abraham, D. J. *J. Comput. Aided Mol. Des.* **1991,** *5,* 545.

[100] Fauchère, J. L.; Quarendon, P.; Kaettere, L. J. *J. Mol. Graph* **1998,** *8,* 202.

[101] Kim, K. H.; Greco, G.; Novellino, E.; Silipo, C.; Vittoria, A. *J Comput.Aided Mol Des* **1993,** *7,* 263.

[102] Waller, C. L.; Marshall, G. R. *J. Med. Chem.* **1993,** *36,* 2390.

[103] Poso, A.; Tuppurainen, K.; Gynther, J. *J. Mol. Struc. (Theochem)* **1994,** *304,* 255.

[104] Kroemer, R. T.; Hecht, P. *J. Comput. Aided Mol. Des.* **1995,** *9,* 205.

[105] Baroni, M.; Constantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. *Quant. Struct.-Act. Relat.* **1993,** *12,* 9.

[106] Cruciani, G.; Watson, K. A. *J. Med. Chem.* **1994,** *37,* 2589.

[107] Cho, S. J.; Tropsha, A. *J. Med. Chem.* **1995,** *38,* 1060.

[108] Doweyko, A. M. *J. Med. Chem.* **1988,** *31,* 1396.

[109] Doweyko, A. M.; Mattes, W. B. *Biochemistry* **1992,** *31,* 9388.

[110] Guccione, S.; Doweyko, A. M.; Chen, H.; Barretta, G. U.; Balzano, F. *J. Comput. Aided Mol. Des.* **2000,** *14,* 647.

[111] Ghose, A. K.; Crippen, G. M.; Revankar, G. R.; McKernan, P. A.; Smee, D. F.; Robins, R. K. *J. Med. Chem.* **1989,** *32,* 746.

[112] Viswanadhan, V. N.; Ghose, A. K.; Weinstein, J. N. *Biochim. Biophys. Acta* **1990,** *1039,* 356.

[113] Viswanadhan, V. N.; Ghose, A. K.; Hanna, N. B.; Matsumoto, S. S.; Avery, T. L.; Revankar, G. R.; Robins, R. K. *J. Med. Chem.* **1991,** *34,* 526.

[114] Walters, D. E.; Hinds, R. M. *J. Med. Chem.* **1994,** *37,* 2527.

[115] Walters, D. E. Genetically evolved receptor models (GERM) as a 3D QSAR Tool, In *3D QSAR in Drug design: Recent Advances*, Kubinyi, H.; Folkers, G.; Martin, Y. C., Eds.; Klnuwer/ESCOM: Dordrecht, Netherlands, **1998**; Vol. 12/13/14, pp. 159-166.

[116] Jain, A. N.; Koile, K.; Chapman, D. *J. Med. Chem.* **1994,** *37,* 2315.

[117] Jain, A. N.; Harris, N. L.; Park, J. Y. *J. Med. Chem.* **1995,** *38,* 1295.

[118] Hahn, M. *J. Med. Chem.* **1995,** *38,* 2080.

[119] Hahn, M.; Rogers, D. *J. Med. Chem.* **1995,** *38,* 2091.

[120] Dunn, J. F.; Nisula, B. C.; Rodbard, D. *J. Clin. Endocrinol. Metab.* **1981,** *53,* 58.

[121] Mickelson, K. E.; Forsthoefel, J.; Westphal, U. *Biochemistry* **1981,** *20,* 6211.

[122] So, S. S.; Karplus, M. *J. Med. Chem.* **1997,** *40,* 4347.

[123] Polanski, J.; Walczak, B. *Comput. Chem.* **2000,** *24,* 615.

[124] Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraswami, C. *J. Am. Chem. Soc.* **1997,** *119,* 10509.

[125] Albuquerque, M. G.; Hopfinger, A. J.; Barreiro, E. J.; de Alencastro, R. B. *J. Chem. Inf. Comput. Sci.* **1998,** *38,* 925.

[126] Vedani, A.; Briem, H.; Dobler, M.; Dollinger, H.; McMasters, D. R. *J. Med. Chem.* **2000,** *43,* 4416.