International Journal of Chemical Modeling
Volume 1, Number 2, pp. 85-107
ISSN 1941-3955

METHODOLOGIES QSAR/QSPR/QSTR: CURRENT STATE AND PERSPECTIVES

Fabiana Paola Maguna[1], María Beatriz Nuñez[1], Nora Beatriz Okulik[1] and Eduardo Alberto Castro[2,*]

[1] Facultad de Agroindustrias, Universidad Nacional del Nordeste, Cdte. Fernández 755, Sáenz Peña 3700, Chaco, Argentina.

[2] INIFTA, División Química Teórica, Departamento de Química, Facultad de Ciencias Exactas, UNLP, Diag. 113 y 64, Suc. 4, C.C. 16, La Plata 1900, Buenos Aires, Argentina

ABSTRACT

Quantitative structure–activity/property/toxicity relationships (QSAR, QSPR, QSTR) modeling is a well known and firmly established discipline, where physicochemical and molecular descriptors are correlated with bioassay the drug eliciting a standard pharmacological response. These relationships have since long been considered a vital component of drug discovery and development, providing significant insights into the role of molecular properties in the biological activity of similar and unrelated compounds.

The objective of this article is to describe the concepts and theories that form the foundation of the traditional and contemporary approaches to QSAR/QSPR/QSTR modeling. Relevant descriptors, observations and examples are presented, pointing out problems that may be encountered, suggesting alternative ways for avoiding pitfalls, and citing pertinent references that should be consulted for more details on specific applications. Consideration of the concepts outlined in this paper may facilitate future advancement of these methodologies and encourage the participation of scientists not presently engaged in this line of research, which is by necessity multi-disciplinary in nature.

Keywords: QSAR, Molecular Descriptors, 3D-Molecular Analysis, Drug development, Receptor binding.

1. INTRODUCTION TO THE QSAR STUDY

Drug design is an iterative process which begins with a compound that displays an interesting biological profile and ends with optimizing both the activity profile for the molecule and its chemical synthesis [1].

In the last decade, much attention has been paid for combinatorial chemistry and high throughput screening in drug discovery setting. Such new technologies increased the

[*] E-mail address: castro@quimica.unlp.edu.ar. (correspondent author)

possibility of finding new lead compounds at much shorter time periods than conventional medicinal chemistry. However, in the biological field, too much promising drug candidates often fail because of unsatisfactory adsorption, distribution, metabolism, elimination, and toxicity (ADMET) properties [2].

The value of computational tools arises from their applicability at an early stage in the development of the synthetic process. At a stage when chemical series are initially screened concerning undesired activities, information on possible adverse properties should be obtained by globally valid computational tools. An excellent correlation with 'wet-lab' data, which is, high sensitivity, as well as high specificity, an easy to use and easy to interpret in silico model are key requirements for its usefulness. As a non-expert tool it should be available to the medicinal chemist via computer networks [3].

In drug discovery, QSAR are widely used to identify ligands with high affinity for a given macromolecular target. More recently, the technology has been extended to predict ADMET properties or the oral bioavailability of compounds. Originally based on the idea that compounds with similar physico-chemical properties trigger similar biological effects, QSAR are often employed to establish a correlation between structural and electronic properties of potential drug candidates and their binding affinity towards a common macromolecular target [4].

The great advantage of QSAR models is that predicting activity instead of measuring it allows to save time and money in chemical management and to speed up managerial decision. In addition, QSAR models can be used to reduce, refine, or replace the use of animals for an experimental purpose [5].

The alternative to this intensive labor approach to compound optimization is to develop a theory that quantitatively relates variations in biological activity to changes in molecular descriptors which can easily be obtained for each compound. A QSAR can then be utilized to help and guide chemical synthesis [1].

The importance of optimizing molecules during early drug development not only for efficacy but also in parallel with regard to their pharmacokinetic and toxicological properties is now widely recognized. It is the fine balance of target potency, selectivity, favorable ADME and (pre)-clinical safety properties that will ultimately lead to the selection and clinical development of a potential new drug [3].

One of the most attractive applications of computer-aided techniques in molecular modeling stands on the possibility of assessing certain molecular properties before the molecule is synthesized. The field of QSAR/QSPR has demonstrated that the biological activity and the physical properties of a set of compounds can be mathematically related to some "simple" molecular structure parameters [6]. When the relationships are applied to modelling of toxicological data, it is termed quantitative structure-toxicity relationships (QSTR).

Thus, the (Q)SARs are being increasingly used as tools for the prediction of environmental and human health endpoints [7].

In recent years, SAR and QSAR methods have received considerable attention because they are capable of identifying similar structural alerts associated with toxicity in a test compound [8].

QSAR predictive software offers a rapid, reliable, and cost-effective method of identifying the potential risk of chemicals that are well represented in QSAR training data sets, even when experimental data are limited or lacking [8].

In the pharmaceutical industry, computational toxicology is now being used as a sentinel tool for the early assessment of the toxicological potential of candidate molecules in lead selection and drug discovery [9].

While early QSAR studies were typically based on a single physico-chemical property, such as the solubility or the pKa value, to explain the biological effect of a molecule (1D-QSAR), Hansch, Fujita, Free and Wilson implicitly included the connectivity of a compound by considering physico-chemical properties of single atoms and functional groups and their contribution to biological activity (2D-QSAR). Nowadays, Hansch-Fujita like QSAR models can also contain 3D-structural descriptors such as the length or width of a substituent [4].

The introduction of comparative molecular field analysis (CoMFA) in 1988 represents another milestone in QSAR since for the first time, such structure–activity relationships were based on the three-dimensional structure of the ligand molecules (3D-QSAR). In 3D-QSAR the ligands interaction with chemical probes is mapped onto a surface or grid surrounding a series of compounds (superimposed in 3D space). This surface or grid represents a surrogate of the binding site of the true biological receptor. The quality of the 3D-QSAR model depends critically on the correct superposition of the ligands, the identification of which is almost impossible in the absence of structural information for the target protein. While this problem has long been recognized, only recently developed 4D-QSAR technologies would seem to provide decent solutions [4].

Thus, 4D-QSAR concepts approach the alignment issue by incorporating molecular and spatial variety by representing each molecule in different conformations, orientations, tautomers, stereo-isomers or protonation states. 4D-QSAR can be interpreted as a feasible extension of 3D-QSAR to address the uncertainties during the alignment process [4].

Additional dimensions further hone predictive powers. 5D-QSAR allows the model to include induced fit, which occurs during the binding of many ligands. In this case, the ligand changes the protein's shape, bringing the active parts into proximity with the substrate. One pilot project use 6D-QSAR, the additional dimension simultaneously considers various salvation models, which is when solute and solvent molecules combine using relatively weak covalent bonds, to screen for adverse effects *in silico* [4].

This paper introduces mainly QSAR/QSPR/QSTR methods, and discusses several important developments in the field that have taken place in the last decade. This work also exposes on chemical descriptors and the choice of descriptors that enter to QSAR different models, focusing in the 3D-QSAR analysis. Examples of studies about the development of drugs or the toxicology of chemical products in the health and the environment are also discussed.

2. Materials and Methods

2.1. Molecular Descriptors

The different classes of descriptor encode different classes of information, and main classes of physicochemical properties are used to characterize chemical structures. The primary purpose of this characterization is “molecular design”, that is to say, the discovery of “performance” chemicals for use as pharmaceuticals, agrochemicals, fragrance ingredients, dyestuffs, flavorings, and so on [10].

Todeschini and Consonni (2000) have already listed more than 3000 molecular descriptors in their Handbook of molecular descriptors [11]. Nowadays, around 8000 chemical properties can be used to describe a molecule.

The descriptors can sometimes be classified into fragment descriptors, involving properties of sections of molecules, and whole molecule descriptors, based on the properties of the molecule itself [12].

The descriptor-based QSAR approaches make use of the whole structure of the compounds by modelling the activity as a function of chemical descriptors. Those descriptors could be simply the number of such atom, as well as properties characterizing the links between atoms or the three-dimensional shape of the molecule. Descriptors are often classified according to the molecule dimensions. This way, 1D descriptor is one-dimensional linear representation of the molecule, 2D descriptors are two-dimensional planar representation of the molecule, and 3D descriptors are three-dimensional spatial representation of the molecule [5].

On the other hand, the descriptors may also be classified into one of four hierarchical categories based on level of complexity and computational demand, such as topo-structural (TS) descriptors, topochemical (TC) descriptors and quantum chemical (QC) descriptors [13].

The chemical-biological interaction, chemical structure or the composition of the biological target may determine the selection of relevant descriptors. For example, in a previous study involving QSAR modelling of tissue-air partition coefficients; it was found that the dominant descriptors varied according to tissue composition [13].

2.1.1. Two-Dimensional Molecular Descriptors

Two-dimensional descriptors rearrange mainly topological and connectivity indexes. One of the most well-known 2D- descriptor is the octanol-water partition coefficient, K_{OW} that is a measure of hydrophobicity and hydrophilicity of a substance. [5].

Molecular connectivity indices are non-empirical structure descriptions that contain information on intermolecular accessibility, while E-state indices contains information reflecting intermolecular accessibility of atoms and groups in a molecule, specifically electron accessibility [14].

Other descriptors such as the topostructural (TS) descriptors are based only on topologic features, without consideration of chemical information. Although the topochemical (TC) descriptors are also based on topology, they encode chemical information, as well, such as atom type and bond type. The TS and TC are two-dimensional descriptors collectively referred to as topological indices and can be calculated quickly without concern for conformational assumptions that are associated with higher dimensional descriptors [13].

2.1.2 Three-Dimensional Molecular Descriptors

Three-dimensional descriptors summarize the geometry, the surface and the volume, as well as other electrostatic properties of the molecule such as Highest Occupied Molecular Orbital (HOMO) or Lowest Unoccupied Molecular Orbital (LUMO). The difference of the energies of the HOMO and LUMO can sometimes serve as a measure of the excitability of the molecule [5].

The quantum chemical (QC) descriptors are based on electronic aspects of molecular structure and their calculation requires significant computational resources, which can be prohibitive in the case of QC descriptors calculated at the *ab initio* level of theory [13].

The descriptors mentioned above encode 2D representations of molecules. However, the CoMFA method surrounds 3D structures of molecules by an array of grid points. At each grid point outside of the molecule a probe atom or functional groups is used to calculate steric, electrostatic and lipophilic fields at that point. This generates a molecular field representation of molecules in a training data set. The large number of descriptors that this method generates must be dealt with by special regression techniques such as partial least squares (PLS) analysis [12].

The CoMFA analysis has become one of the most powerful tools for QSAR and drug design. It has pioneered a new paradigm of three-dimensional QSAR [15] where the shapes, properties, etc. of molecules are related to specific molecular features (substituents, etc.) and their spatial relationship.

2.2. Methodologies to Calculate Molecular Descriptors

CoMFA evaluates the electrostatic (Coulombic interactions) and steric (van der Waals interactions) fields at the binding site through mutual 3D alignment of the ligands [16].

While steric and electrostatic properties of molecules are major physicochemical properties related to biological activity, they are purely enthalpic. Efforts to include entropic properties within a CoMFA framework have been to characterize the hydrophobic nature of molecules. More recently, reactivity-based fields such as those of molecular orbitals have also been imported into CoMFA studies [17].

An alternative approach to the computation of molecular potential fields has been described as comparative molecular similarity indices analysis (CoMSIA) by Klebe *et al* [18]. The form of the distance functions in the standard Lennard-Jones and Coulomb type potentials provides for the generation of unrealistically extreme values as the surface of the molecules under examination is approached.

Among 3D-descriptors, the GRID-based one deserves particular mentioning. For their determination a molecule is placed in a box and for an orthogonal grid of points the interaction energy values between this molecule and a small probe molecule, such as water, are calculated. The GRID fields thus obtained, named Molecular Interaction Fields (MIF), characterize, for example, molecular shape, charge distribution and hydrophobicity [19]. Development and successful application of GRID based 3D-descriptors is recently summarized by Cruciani [20].

Similarly, the GRIND [21] (GRid-INdependent Descriptors) approach has been used for virtual screening [22] and 3D QSAR analyses. [23] COMparative BINding Energy (COMBINE) analysis [24] is a receptor-based 3D QSAR approach using protein-ligand interaction energy terms as independent variables thus requiring knowledge or reasonable assumptions about the binding geometries [25].

Hopefully, more and more descriptors can be computed exactly by applying new software such as ADAPT [26, 27], OASIS [28], CODESSA [29], and DRAGON [30]. That software can theoretically compute a set of descriptors for any molecule coded using a universal text representation called SMILES. The name SMILES stands for Simplified

Molecular Input Line Entry Specification. This universal text coding of a molecule was first introduced by Weininger [31]. It allows specifying the structure of chemical substances in a ASCII format that can be understood by most chemical software that compute descriptors [5].

2.3. More Used Descriptors and Their Main Features

2.3.1. Hydrophobic Descriptors

The octanol/water partition coefficient (log *P*) is the standard quantity to characterize the hydrophobicity/hydrophilicity of a molecule, a property of major importance in biomedical applications. Over the years a number of procedures have been proposed for calculating partition coefficients from the molecular structure [32].

Other archetypal hydrophobicity parameter is π, a substituent constant derived from measurements of octanol-water partition coefficients (log P). The hydrophobic constant π is related to the difference between the log of the partition coefficient of an unsubstituted molecule and that of a molecule with given substituent.

Partition coefficients and π values show to correlate with measures of biological activity in a wide variety of experimental systems. This relation is presumably because hydrophobic effects are important not only in the intermolecular interactions that occur between a drug and its target site but also in the distribution of a compound within a biosystem and its interaction with competing binding site.

2.3.2. Steric Descriptors

Steric molecular descriptors are employed to characterize the size or confirmation of an entire molecule or specific fragments, and its influence in drug transport and protein recognition is well appreciated.

The first steric descriptor to be used in QSAR studies was the E_s parameter due to Taft [33]. He assumed that E_s should reflect various types of steric influence (strain, repulsion, etc.) and that the rate constants, as well as E_s, are linearly related to the activation energies [32].

A more commonly used descriptor of steric properties, both for substituents and whole molecules, is molar refractivity, M_R. This is a measure of size and polarizability of substituents and it is derived from the Lorentz- Lorentz equation.

This descriptor correlates satisfactorily with Van der Waals´ volume, parachor (a surface-tension-weighted molar volume) and fragmental volume constants [34].

STERIMOL is an appropriate descriptor to describe the three-dimensional shape of the molecules and it was developed by Verloop [35]. Five parameters are deemed necessary to define shape: *L*, B1, B2, B3, and B4. *L* represents the length of a substituent along the axis of a bond between the parent molecule and the substituent; B1 to B4 represent four different width parameters [36].

2.3.3. Electronic Descriptors

The traditional electronic parameters, such as pKa and Hammett's σ, are frequently considered along with those available from quantum chemical calculations.

The Hammett constant is a measure of electron donating properties of substituents and was derived originally from the differences between the logarithm of the ionization constant of benzoic acid and that of substituted benzoic acids. In other words, the Hammett equation defines the electronic effect of a substituent on a reaction.

Part of the problem in the description of the electronic effects of substituents is that these effects may be transmitted through field (inductive) effects and resonance effects. It has generated a considerable number of different σ scales. A major problem in the use of electronic substituent constants, i.e. in the use of any substituent constants, lies in the definition of "parent" substituents, and even substituent positions [10].

In recent years, there has been a rapid growth in the application of quantum chemical methodology to QSAR, and direct derivation of electronic descriptors from the molecular wave functions [37]. As in other electronic parameters, QSAR models incorporating quantum chemical descriptors will include information on the nature of the intermolecular forces involved in the biological response.

Quantum chemical descriptors such as net atomic changes, HOMO and LUMO energies, and superdelocalizabilities have shown to correlate quite well with various biological activities [38].

2.3.4. Molecular Structure Descriptors

These are descriptors based only on the standard 2D representation of a chemical structure. The most known topological parameters are the molecular connectivity indices first described by Randic [39] and extensively investigated by Hall and Kier and co-workers [40-42].

Molecular connectivity indices are non-empirical structure descriptions that contain information on intermolecular accessibility [43].

Connectivity indices in their simplest form are computed from the hydrogen-suppressed skeleton of a compound by the assignment of a degree of connectivity, δ_i, to each atom (i) representing the number of atoms connected to it. For each bond in the structure, bond connectivity, C_k, can be calculated by taking the reciprocal of the square root of the product of the connectivity of the atoms at either end of the bond.

The first order index is the most simple connectivity index because it considers only individual bonds or paths of two atoms in the structure. Higher order indices may be generated by the consideration of longer paths in a molecule, and other refinements have been considered, such as valence connectivity values, path, cluster, and chain connectivity [44].

Higher molecular connectivity indices encode more complex attributes of molecular structure by considering longer paths [45].

A topological index (TI) is a numerical descriptor of the molecular structure based on certain topological features of the molecular graph, offering an effective way of measuring molecular branching, shape, size, and molecular similarity [46]. Electrotopological state (E-state) indices contain information reflecting intermolecular accessibility of atoms and groups in a molecule, specifically electron accessibility [43].

2.3.5. Spectroscopic Descriptors

The use of experimental spectroscopic properties, such as NMR chemical shifts and infrared and Raman stretching frequencies, has already been briefly mentioned and the major problem with the employment of experimental quantities, their predictability, pointed out. Spectroscopic properties can be calculated, with varying degrees of reliability, from ab initio and semiempirical quantum mechanics packages, and these may be used as a replacement for experimentally determined values [10].

Also, other approach [47], EVA (Eigen VAlues), makes use of quantities derived from quantum chemical calculations based on 3D molecular coordinates which bear some similarity to spectroscopic data. The procedure followed above is chosen to overcome a problem associate with the direct use of normal coordinate frequencies. The number of normal modes varies with the number of atoms in a molecule, and thus generally, unless each molecule contains the same number of atoms, there would be different numbers of descriptors for each compound. The normal coordinate frequencies are calculated using a semiempirical program such as MOPAC [48].

The EVA descriptor has been successfully applied to the calculation of octanol/water partition coefficients and to a number of biological datasets [49].

2.4. Three Dimensional QSAR Generation Process

The general procedure for generating a QSAR method requires the identification of a training set, creating a congeneric series of 3D structures, which can have information on the activity associated with each molecule. The 3D QSAR analysis usually needs conformational information that can be obtained by performing a conformational search by way of a variety of methods.

The exploration of data can be carried out generating graphs to depict descriptor distribution. If there are holes in descriptor sets one can resort to the choice of new compounds to fill in the holes. Also, correlation matrices assist in identification the descriptors that are highly correlated and histograms and plots help in examine the uniformity of data. Descriptive statistics are available to further characterize descriptors. Additionally, it can transform and normalize descriptors, as appropriate. The principal components analysis (PCA) and cluster analysis can carry out to further characterize the data.

The appropriate identification of the dependent and independent variables can choose from several statistical methods for generating a QSAR equation. These include multiple linear regressions (MLR), partial least squares (PLS), simple linear regression, stepwise multiple linear regression, and principal components regression (PCR). Additionally, the genetic function approximation (GFA) can perform a genetic analysis, either GFA or G/PLS, to create a QSAR equation. The validation techniques to identify outliers and leverage points and the use graphic analyses and cross-validation characterize the robustness of the QSAR [5].

The analysis of the equation can be carried out using the plot of observed vs. predicted activities and to identify outliers. The 3D plots allow visualizing the positions of important 3D-QSAR descriptors from Molecular Field Analysis (MFA) or Receptor Surface Analysis (RSA) in relation to the molecules. The use of the calculated QSAR equation allows one to predict biological activity of compounds and to derive a candidate structure.

2.5. Technical for QSAR Analysis

The most common modelling techniques in QSAR to link the descriptors to the activity response are based on regression analysis. Among those techniques, MLR is the most classical one. However, MLR is not adapted to the existing correlations between descriptors. Accordingly multivariate projection methods to a subspace of orthogonal latent variables have become more and more popular in QSAR. Partial Least Squares Regression (PLSR) and PCR are two such projection techniques increasingly used. Another alternative, close to MLR, is the Ridge Regression (RR) who imposes a penalty to the size of the coefficient in the linear regression model. Other modeling techniques such as Regression Trees (RT), k-Nearest Neighbours (kNN) and Neural Networks (NN) are also recently introduced.

2.5.1. Description of QSAR Techniques

MLR assumes that there is a link between the activity response and the standardized descriptors. This model is traditionally fitted to the data by least squares method. MLR is very popular in QSAR modelling because of its simplicity and its well-known theoretical background [50, 51].

PCR is an alternative to MLR when the explanatory variables are correlated. PCA is first applied to construct orthogonal latent variables, called the Principal Components (PCs), which can then enter a MLR model without any more co-linearity problem. The PCs are computed as linear combinations of the mean-centered descriptors. One may use all the PCs (ordinary least squares results) as they are orthogonal or one may select only a few ones. PLSR is an alternative to MLR analysis and can accommodate descriptor matrices that exhibit high co-linearity and contain more descriptors than compounds [52].

The general expression for PLS analysis is defined for any dependent variable (*y*) as an orthogonal linear combination of independent variables (*x*). The number of latent variables and the coefficients are optimized to both capture the variation in the dependent variables and enhance the ability to predict the variations in the observations [32].

The principle of PLSR is very similar to PCR as the link between the activity response and the descriptors is modeled through newly constructed latent variables. The specificity of PLSR is that those latent variables are recursively chosen to perform a simultaneous decomposition of the observed mean-centered descriptors (X^c) and the mean-centered observed responses (y^c) with the constraint that they explain as much as possible of the covariance between X^c and y^c.

Instead of selecting a variables subset and dropping descriptors from a MLR model, one can use a modification of the classical least squares MLR method, which is the RR method. The assumed model of RR is exactly the same as the MLR model. The difference stands on the least squares criteria that are penalized by a multiple of the sum of squared regression coefficients.

This RR sum of squares is equivalent to the minimization of the ordinary sum of squares under the constraint that the sum of the squared coefficients does not exceed a given size. This constraint protects the gradients of the response surface in the direction of the smallest PCs in correlated Z-space against potentially high variance [5].

The principle of RT is really different from the previous regression methods as the nature of the relationship between the activity response and the descriptors is not pre-specified. RTs are nonparametric models. The most simple technique for fitting a RT to QSAR data consists in recursively partitioning the data into successively smaller groups (called nodes) with binary splits based on a single descriptor (e.g: $x_j \leq c$ and $x_j > c$ for the j^{th} transformed descriptor and a constant c). At each step, splits for all of the descriptors are examined by an exhaustive search procedure and the best split is chosen. The idea is to define the best split at stage M as the one that minimizes the total residual sum of squares over all the current M nodes [5].

The kNN method consists in predicting the activity of a molecule as the average (or weighted average) of the observed activity values of the k nearest molecules. kNN is greatly appreciated by QSAR practitioners as it is really intuitive. Indeed, its principle reflects the ancestral idea those similar compounds reveal similar activity property. Like RT, kNN does not make any a priori assumption about the nature of the activity-structure relationship. The descriptors are not directly used to model the activity but to define the neighborhood.

NN models are nonlinear models that can be represented as a network structure. There exist different architectures of NN according to the number of hidden layers, the number of nodes for each layer and the connections existing between all the different layers. The two main forms of NN are the feed forward NN and NN using radial basis functions [6].

In other fields the applications of QSAR methods are related with structural alerts. They are defined as molecular functionalities (structural features) that are known to cause toxicity, and their presence in a molecular structure alerts the investigator to the potential toxicities of the test chemical. Predictive toxicology software can be classified as either qualitative rule-based SAR software [e.g., Derek (Lhasa), ONCOLOGIC (EPA/Logichem)] or QSAR software [e.g., MC4PC (MultiCASE), MDL-QSAR (MDL Information Systems)]. QSAR software programs generate equations (models) by statistically identifying molecular descriptors and/or sub-structural molecular attributes that are correlated with toxicity, whereas SAR programs use expert rules, developed by panels of human experts that are applied to the test chemical to yield a computer-automated prediction of toxicity.

Both SAR and QSAR methods are capable of identifying similar structural alerts associated with toxicity in a test compound [8].

2.5.2. Validation of Methods QSAR

The most important method in QSAR modelling to quantify the predictive power on the basis only of the training set is the cross-validation technique. The principle of cross-validation is to simulate predictions for new molecules not used in the fitting of the model. The training set is divided in distinct subgroups.

There are two main practices in defining the subgroups. The leave-one-out method considers each compound as a subgroup. Each compound is omitted in its turn. The model is fitted on the N-1 other observations of the training set and the activity response of the remaining compound is predicted using the obtained model. To simulate predictions for new molecules not used in the fitting of the model, the leave-one-out principle is the most intuitive idea as the most information as possible is taken into account in the training of the model by removing only one observation [13].

The alternative is to delete more than one compound at each turn by defining subgroups with up to 50% of the data set. This method is called leave-many-out. The reasonable number of compounds to be included in each subgroup depends on the sample size. Let us denote *G* the number of groups, generally selected between 2 and 10, being N/G is the size of groups. Every group in its turn is left out of the training set and the model is fitted on the N – (N/G) other compounds. This fitted model is then applied to the remaining group to predict the responses of its N/G compounds [5].

Another technique to quantify the predictive power is the *bootstrap*. N compounds are drawn at random with replacement from the original dataset. Some of the original molecules might appear more than once in the resample while other molecules might not be included. The model is fitted on this resample and applied on the remaining compounds to predict the endpoint.

2.6. QSAR for the Study of Drugs- Receptor Interaction

The central theme of molecular pharmacology, and the basis of SAR studies, has focused on the elucidation of the structure and function of drug receptors. It is generally accepted that endogenous and exogenous chemicals interact with a binding site on a specific macromolecular receptor. This interaction, which is determined by intermolecular forces, may or may not elicit a pharmacological response depending on its eventual site of action.

In connection with this, from the calculated electron distribution in the molecule, some properties such as the net atomic charges and bond polarities can be predicted, which help to characterize the nature of interactions at specific receptor sites. The electron distribution can also be used to quantitatively map the electrostatic potential generated by a molecule in all regions that surround [53].

The integration of ligand- and structure-based strategies might sensitively increase the success of the drug discovery process. In fact, ligand-based approaches are widely and successfully used to develop quantitative models able to correlate and predict the biological activities based on various molecular properties (molecular descriptors) [54].

3D-QSAR allows developing an approach to determine the pharmacophor of the molecules under study. In this method it is calculated and it correlates the biological activity of a series of molecules with their global properties in a three-dimensional space.

A pharmacophore model is the 3D arrangement of essential features that enables a molecule to exert a particular biological effect. As an important category in drug design, it has been widely used as a query for database mining or a guide for de novo design. Generally, a pharmacophore model is deduced from a set of ligands with known activities while lacking the three-dimensional structure of a receptor. Though there are already many automatic ligand based methods such as DISCO, Catalyst/HipHop, and GASP, the derivation of a pharmacophore model remains a difficult task [55].

On the other hand, the CoMFA is one of the well known 3D-QSAR descriptors which have been used regularly to produce the 3-D models to indicate the regions that affect biological activity with a change in the chemical substitution [56]. The advantages of CoMFA are the ability to predict the biological activities of the molecules and to represent the relationships between steric/electrostatic property and biological activity in the form of contour maps gives key features on not only the ligand-receptor interaction but also the topology of the receptor [57].

In this context the molecular mechanics is adapted to interpret many molecular properties. This way, starting from a reference molecule their conformation of minimum energy is determined, a mesh is built to its surroundings and the structures of the ligands are minimized in this cases what is supposed they act on the biological same reveille. Then the different selected properties (descriptors) for each one of the molecules are calculated.

The relationship between the structural descriptors (CoMFA interaction energies) and the biological activities is quantified by the PLS algorithm. PLS regression technique is especially useful in quite common cases where the number of descriptors (independent variables) is comparable to or greater than the number of compounds (data points) and/or there exist other factors leading to correlations between variables [58].

At the end of the process, the graphic representation of the correlation "structures three-dimensional-activity" is visualized in form of contour maps, where it is observed the regions of the space in those that the studied properties are favorable or unfavorable for a biological certain activity.

Recently, the generation of a pharmacophore model directly from a protein crystal structure is an alternative approach, which can reveal the key elements in protein-ligand binding more straightforwardly. With the advances in experimental techniques of X-ray crystallography and NMR, the determination of a protein structure as a drug target has become more convenient. The development of reliable and robust techniques to construct pharmacophores from a receptor structure is really important [55].

On the other hand, there are the QSAR indirect methods. These models are an indirect structure-based way, which can assess the potential of small molecules to interact with a pharmacophore of interest. This approach consists of three steps: (i) identification of chemically equivalent atoms or groups in terms of their physical or chemical properties; (ii) estimation of the relative 3D position of the possible pharmacophoric groups in allowed low energy conformations of the molecules; and (iii) weighing similar pharmacophoric groups when there are multiple choices [2].

Since chemical interactions are three-dimensional (3D) events, QSARs often depend on the 3D molecular models adopted for the chemicals under study. This certainly applies to receptor-site mapping models dealing directly with molecular shapes and fields. Correlative QSARs may also be influenced indirectly when employing electronic-quantum chemical descriptors that generally depend on 3D structure [59].

3. Discussion

The QSAR methods (and analogously QSTR and QSPR) involves a number of key steps, where the selection the best descriptors from a larger set of accessible, relevant descriptors is truly relevant. Thus, for a given study, descriptors should be selected which are relevant to the activity for the series of molecules under investigation and these parameters should have values which are obtained in a consistent manner. For that reason, here it will be commented some aspects related to the same ones and information will be contributed regarding particular cases where it is favorable its employment.

Some descriptors have to be measured in laboratories by performing experiences on the molecule, such as the acidity constant Ka. They are often referred to as empirical descriptors. These descriptors can be heavy to use in practice, especially if the goal of QSAR modelling is to speed up the drug discovery process.

One of the problems with using single descriptors is that there is no discrimination between substituents that have different shapes. Recognizing this, Verloop and co-workers devised the STERIMOL parameters based on models of compounds or substituents using standard bond length and angles [10].

However, the high degree of co-linearity between B1, B2, and B3 and the large number of training set members needed to establish the statistical validity of this group of parameters led to their demise in QSAR studies. Verloop subsequently established the adequacy of just three parameters for QSAR analysis: a slightly modified length *L*, a minimum width B1, and a maximum width B5 that is orthogonal to *L* [60]. The uses of these insightful parameters have done much to enhance correlations with biological activities.

In recent years, there has been a rapid growth in the application of quantum chemical methodology to QSAR, and direct derivation of electronic descriptors from the molecular wave functions [37]. As in other electronic descriptors, QSAR models incorporating quantum chemical descriptors must include information on the nature of the intermolecular forces involved in the biological response.

Quantum chemical descriptors such as net atomic changes, HOMO and LUMO energies, frontier orbital electron densities, and superdelocalizabilities have shown to correlate quite well with various biological activities [38, 61].

Molecular connectivity indices have been shown to be closely related to many physicochemical properties such as boiling points, molar refraction, polarizability and partition coefficients [42, 62]. Also, molecular connectivity descriptors have been shown to explain a variety of different types of biological properties including a number of applications in the environmental area [63-66].

Some descriptors can be derived from atomic or molecular properties and can encode physicochemical (e.g. octanol–water partition coefficient), topological (e.g. electronegativities) and surface properties (e.g. polarity) of molecules. These descriptors are then correlated to the extent of a biological or toxicological response, resulting in correlation curves or regression lines [67].

Topological indices have several obvious advantages when compared with geometrical, electrostatic, and quantum descriptors: they are computed only from the information contained in the molecular graph, they have a unique value for a particular chemical compound, and their calculation requires small computational resources. [46, 68].

The E-State is established as a composite index encoding both electronic and steric properties of atoms in molecules. It reflects an atom’s electronegativity, the electronegativity of proximal and distal atoms, and topological state. Extensions of this method include the HE-State, atom-type E-State, and the polarity index *Q*. Log *P* showed a strong correlation with the *Q* index of a small set (n=21) of miscellaneous compounds [69]. E-state descriptors that combine electronic, topological and valence state information represent a major descriptor category for rodent carcinogenicity, followed by connectivity parameters [70].

As computers have increased in calculation capabilities, new QSAR methodologies that take into account the 3D structure of molecules have emerged such as, for example, CoMFA analysis [71]. Numerous approaches about CoMFA analysis in drugs discovery has been recently published [25, 56, 57, 72-75].

In contrast to the fragmentary nature of standard CoMFA maps, CoMSIA derived maps are contiguous and located closer to the molecular skeletons thus providing a more direct representation of the physicochemical features localized in the design space (i.e., occupied by training set molecules) which are required for bioactivity.

3D QSAR methods, CoMFA and CoMSIA, were applied on a series of 1,4-dihydropyridines possessing antitubercular activity, and steric and electrostatic fields of the inhibitors were found to be relevant descriptors for SAR [76].

Also, a study on histone deacetylase inhibitors (HDACIs) was performed using the GRID/GOLPE combination using structure-based alignment. The 3D QSAR investigation proved to be of higher statistical value, displaying for the best global model r^2, q^2, and cross-validated SDEP (standard deviation of errors of prediction) values. A comparison of the 3D QSAR maps with the structural features of the binding site showed good correlation. These results present a 3D QSAR application on a broad molecular diversity training set of HDACIs [77].

Zhang et al. performed a virtual screen for HMG-CoA reductase inhibitors using a combination of pharmacophore filtering, docking, and CoMFA predictions. Therein, the docking procedure provided the molecular alignment rule [25].

Comparatively, in terms of numerical affinity prediction the 3D QSAR models significantly outperformed the 1D and 2D approaches. They tend to be more specific what the MDL MACCS (Molecular Design Limited, Inc.-Molecular ACCess System) keys (one kind and the public version to 2D descriptors) but at the cost of a lower sensitivity. The models are difficult to mutually rank against each other since relevance, predictive value, and applicability depends upon the specific goal of the project. Most of the descriptors to molecular fields of a ligand approximate better the concept of molecular recognition, and they should not suffer from the fact of missing connectivity information as the MACCS keys do.

Most likely, the 3D QSAR methods are much more robust in handling data sets composed of compounds with structurally rather diverse molecular skeletons [78, 79].

MLR is a technique of regression analysis very popular in QSAR modelling because of its simplicity and its well-known theoretical foundation. Some studies that involve the MLR define the possibility to find the most active groups in the compounds [80-82], and others study the relationship with the cytotoxicity [83-86].

As stated before, MLR is not adapted to the existing correlations between descriptors. PLS also does not assume that the explanatory variables are exact and 100% relevant for modelling the response. PLSR is increasingly used in QSAR modelling as it manages a great number of explanatory variables (even possibly greater than the number of available data) and the existing correlations between descriptors [87].

The autocorrelation vectors of the molecular electrostatic potential (MEP) in conjunction with PLS analysis produces descriptors which encode the three-dimensional distribution of molecular properties, but their comparative evaluation does not depend on an alignment.

Although RR, PCR and PLS are all appropriate modelling approaches when the number of descriptors exceeds the number of chemicals in the data set, a conservative reduction in the number of descriptors facilitate the model interpretation without introducing bias from extreme data reduction. For this reason, a modified Gram-Schmidt orthogonalization procedure was used to modestly reduce the size of the descriptors sets [13].

Since the co-linearity between variables is an important factor to consider, the three multiple linear regression methodologies are used comparatively in model development, i.e. RR, PCR and PLS. Each of these methods is useful when the number of descriptors exceeds the number of chemicals in the data set and when the descriptors are highly inter-correlated. Statistical theory suggests that RR is the best of the three methods and this has been generally borne out in comparative studies [13].

Recently, it has been reported that the *autocorrelation* Molecular Electrostatic Potential (*auto*MEP) vectors in combination with PLS analysis can represent an alternative 3D-QSAR tool to CoMFA. Incidentally, both CoMFA and *auto*MEP/PLS methodologies can be classified as *linear* QSAR methods considering the mathematical relationship among molecular descriptors and chemical/biological response space. Very recently, has been also presented a nonlinear method based on a response surface analysis (RSA) application in tandem with the *auto*MEP descriptors (*auto*MEP/RSA) as an alternative *nonlinear* 3D-QSAR method [54].

Other technique within the QSAR analysis area is RT which is a very flexible method. Indeed, there are various methods for growing the tree, allowing not only binary splitting, various criteria to optimize while splitting and also various tools for pruning [88]. RT is appreciated in QSAR applications as it can handle large data sets, it allows interactions and nonlinear relation between descriptors and response, and also because it produces sequences of prediction rules that are readily interpretable. Some applications of RT to QSAR modelling has been published [5, 89-91].

Artificial neural network (ANN) approaches provide an alternative to established predictive algorithms for analyzing massive chemical databases, potentially overcoming obstacles arising from variable selection, multi-co-linearity, specification of important parameters, and sensitivy to erroneous values. In most instances, ANN's have proven to be better than MLR, PCA or PLS because of their ability to handle non-linear associations [6].

In the last years there has been a growing interest in the application of NN to the development of QSAR/QSPR. The bigger advantage of ANN lies in the fact QSAR/QSPR can be developed without having a priori to specify an analytical form for the correlation model. The NN approach is especially suited for mapping complex non-linear relationships that exists between models output (physicochemical or biological properties) and input model (molecular descriptors). The NN approach could also be used to classify chemicals according to their chemical descriptors and used this information to select the most suitable indices capable of characterize the set of molecules. Existing NN based QSAR/QSPR for estimating properties of chemicals have relied primarily on back-propagation architecture [6, 92].

The cross-validation technique is among one of the most important method in QSAR modelling to quantify the predictive power. Results of the cross-validation and validates the equation offers information like the outliers, the sum of squared deviations and the predicted sum of squares (PRESS). In this sense, a cross-validation algorithm is used to determine the number of latent variables, such that the cross-validated correlation (predictive r^2 or so-called q^2) is optimal. This method has demonstrated broad applicability, combining 3D molecular modeling and QSAR/QSPR analysis, and has been further extended to include similarity indices for improving predictive performance and model interpretation [18].

Global external performance according to Doweyko et al. is only about half of the QSAR models published in the last decade made reasonable predictions about test compounds not used to create the model [93]. Tropsha et al. pointed out that the quality of QSAR models is often typically measured on the training set alone, but this approach does not necessarily generate good predictive QSAR models [94].

The most demanding way to quantify the predictive power of a QSAR model should be based on an external data set, by making predictions for an independent set of data, not used in the model calibration. This provides a more rigorous evaluation of the model predictive capability for untested chemicals than cross-validation or bootstrap on the training set. Local external performance allows one to define the predictive power of the QSAR model assessed on an external test set and it can be used to predict the activity of a new compound [5].

To simulate predictions for new molecules not used in the fitting of the model, the leave-one-out principle is the most intuitive idea as the most information as possible is taken into account in the training of the model by removing only one observation. But leave-one-out is too optimistic and tends to overestimate the predictive power for independent new compounds.

Anyway, there is disagreement among QSAR modelers with respect to the proper method of model validation. Whereas may believe that LOO is inferior to the use of a hold-out-test set, it has been shown by theoretic argument and empiric study that the LOO cross-validation approach is actually preferred to the use of a hold-out set when working with small to moderate sized chemical databases [95].

The central theme of molecular pharmacology, and the basis of SAR studies, has focused on the elucidation of the structure and function of drug receptors. A pharmacophore model is the 3D arrangement of essential features that enables a molecule to exert a particular biological effect. Generally, a pharmacophore model is deduced from a set of ligands with known activities while lacking the three-dimensional structure of a receptor. Recently, the generation of a pharmacophore model is obtained directly from a protein crystal structure or with an alternative approach of indirect methods QSAR. Different results may be given due to different strategies taken in conformational analysis and molecular superposition [55]. Some examples can be analyzed in recent studies [72, 96].

4. CONCLUSION

Quantitative structure property relationship (QSPR) and quantitative structure activity relationship (QSAR) modeling is an emerging tool in pharmaceutical industry and the actual environmental research. The great advantage of QSAR/QSPR models is that predicting activity instead of measuring it allows us to save time and money in chemical management and to speed up managerial decision.

QSAR models represent well established tools for the molecular design of new compounds with desired properties. All QSPR and QSAR models are statistically based and aimed at extracting the maximum information from experimental data on compounds of known structure.

These structure-property/activity studies require atomic and molecular descriptors to encode in a numerical form the local (atomic) or global (molecular) structure. Numerous software to QSPR or QSAR studies are used, which integrate the computation of structural descriptors with the generation of structure-property models. Afterward, the best descriptors are selected in the final structure-property model using statistical methods.

In the rational design of new drugs it is necessary to keep in mind that 3D QSAR methods comprises several subsequent steps: conformational analyses, alignment of the molecules, generation of molecular descriptors and regression analysis. Optionally, one or more biological response(s) can be used as the independent variable(s). However, although 2D-QSAR methods do not give us information for rational drug design like 3D-QSAR methods do, they would enable an exhaustive screening of a combinatorial library of compounds because its simplicity of computation.

In this topic the bioactive conformation represents the crucial starting point of all 3D-QSAR strategies such as Comparative Molecular Field Analysis (CoMFA) or 3D Pharmacophore search. As anticipated, 3D-QSAR methods require the knowledge of the conformational properties of the molecules in order to calculate their structural or property descriptors. CoMFA is probably one of the most successful 3D-QSAR methods used in medicinal chemistry in the last two decades.

On the other hand, one attractive application of receptor-based pharmacophore model is to discover new binding spots that have not yet been occupied by known ligands so as to guide the improvement of binding affinity and/or maximizing selectivity. Reproducing known ligand geometries is insufficient because they represent an extremely limited and biased sampling of all bound ligand conformations. Therefore, the methods generally derive the probability that each feature is important and that each conformation is the active conformation, starting with no knowledge of important features or binding conformations.

The present work lets to know the QSPR and QSAR models developed and the molecular graph descriptors and topological indices used with success to model various properties and demonstrates that they are valuable descriptors of chemical structure. A final important remark is that QSAR models have proven their utility, from both the pharmaceutical and toxicological perspectives, for identification of chemicals that might interact with nuclear receptors. While their primary function in the pharmaceutical enterprise is lead discovery and optimization, QSAR models have played an essential role in toxicology as a priority setting tool for risk assessment. Basically, QSAR models employ quantitative regression methods to correlate, and rationalize, variations in the biological activity of a structurally related series of chemicals with variations in their molecular structures as encoded in pre-selected quantities commonly known as molecular descriptors. Reflecting the availability of relevant biological data, the largest number of QSAR models in the published literature is associated with protein and estrogen receptors binding.

The threshold of innovation has increased dramatically in recent years, and there is a great need for methods that can provide access to uncharted chemical space. One such method is de novo drug design which is a automated, computer-assisted construction of new molecules that satisfy a set of desired constraints, such as shape and electrostatic complementarity to a protein binding site. Although de novo design programs tend to generate a large number of candidate ligands, only a small fraction of them can be physically synthesized and tested in a biological context and for that additional investigations are necessary. In the same way, other QSAR techniques (other statistical methods, 4D, 5D, 6D approaches) need to be done to collect more experience on the scope and limitations of QSAR methods for database screening.

Acknowledgements

Financial support from SGCyT-UNNE is gratefully acknowledged. E.A.C. and N.B.O. are members of the Research Career of CONICET, Argentine.

REFERENCES

[1] A.B. Richon and S.S. Young. *An Introduction to QSAR Methodology*. Available online at: http://www.netsci.org/Science/Compchem/feature19.html (accessed 15 August 2007).

[2] F. Yamashita, S. Fujiwara, S. Wanchana and M. Hashida, Quantitative structure/activity relationship modelling of pharmacokinetic properties using genetic algorithm-combined partial least squares method, *Journal of Drug Targeting*, **14** (2006), pp. 496-504.

[3] W. Muster, A. Breidenbach, H. Fischer, S. Kirchner, L. Müller and A. Pähler, *Computational toxicology in drug development,* Drug Discov. Today, in press. (2008). doi:10.1016/j.drudis.2007.12.007

[4] M.A. Lill, Multi-dimensional QSAR in drug discovery, *Drug Discov. Today*, **12** (2007), pp. 1013-1017.

[5] C. Le Bailly de Tilleghem and B. Govaerts. *A review of Quantitative Structure-Activity Relationship (QSAR) models.* Available online at: http://www.stat.ucl.ac.be/IAP (accessed 27 February 2008).

[6] G. Espinosa Porragas, *Modelos QSPR/QSAR/QSTR basados en sistemas neuronales cognitivos* Thesis Doctoral. UNIVERSITAT ROVIRA I VIRGILI. Tarragona, España. (2002).

[7] Gallegos Saliner and X. Gironés, Topological quantum similarity measures: applications in QSAR, *Journal of Molecular Structure: THEOCHEM* **727** (2005), pp. 97–106

[8] N.L. Kruhlak, J.F. Contrera, D.R. Benz and E.J. Matthews, Progress in QSAR toxicity screening of pharmaceutical impurities and other FDA regulated products, *Advanced Drug Delivery Reviews* **59** (2007), pp. 43–55.

[9] G.M. Pearl, S. Livingston-Carr and S.K. Durham, Integration of computational analysis as a sentinel tool in toxicological assessments, *Curr. Top. Med. Chem.* **1** (2001); pp. 247–255.

[10] D.J. Livingstone, The Characterization of Chemical Structures using Molecular Properties. A Survey, *J. Chem. Inf. Comput. Sci.*, **40** (2000), pp. 195-209.

[11] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*, Wiley–VCH, New York, 2000.

[12] D.A. Winkler, The role of quantitative structure-activity relationships (QSAR) in biomolecular discovery, *Briefings in Bioinformatics*, **3** (2002), pp. 79-86.

[13] S.C. Basak; D. Mills and M.M. Mumtaz, A quantitative structure-activity relationship (QSAR) study of dermal absorption using theoretical molecular descriptors, *SAR QSAR Environ. Res.,,* **18** (2007), pp. 45-55.

[14] L.B. Kier and L.H. Hall, Intermolecular accessibility: The meaning of molecular connectivity. *J. Chem. Inf. Comput. Sci.,***40** (2000), pp.792–795.

[15] H. Kubyini, ed. *3D QSAR in Drug Design, Theory, Methods, and Applications*, ESCOM Science Publishers, B.V., Leiden, 1993.

[16] R.D. Cramer III, D.E. Patterson and J.D. Bruce, Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins, *J. Am. Chem. Soc*. **110** (1988) pp. 5959-5967.

[17] C.L. Waller and G.E. Kellogg, *Adding Chemical Information to CoMFA Models with Alternative 3D QSAR Fields*, NetSci Web (1996). Available online at: http://www.netsci.org/Science/Compchem/ feature10.html (accessed february 2008)

[18] G. Klebe, U. Abraham and T. Mietzner, Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity *J. Med. Chem*. **37** (1994) pp. 4130-4146.

[19] S. Sciabola, E. Carosati, L. Cucurull-Sanchez, M. Baroni and R. Mannhold, Novel TOPP descriptors in 3D-QSAR analysis of apoptosis inducing 4-aryl-4H-chromenes: comparison versus other 2D- and 3D-descriptors, *Bioorg. Med. Chem*., **15** (2007) pp. 6450–6462.

[20] G. Cruciani, *Molecular Interaction Fields. In Methods and Principles in Medicinal Chemistry*. R. Mannhold, H. Kubinyi and G. Folkers, Eds. Wiley-VCH; 2006; Vol. 27.

[21] M. Pastor, G. Cruciani, I. McLay, S. Pickett and S. Clementi, GRid-INdependent descriptors (GRIND): a novel class of alignment independent three-dimensional molecular descriptors, *J. Med. Chem*, **43** (2000), pp. 3233-3243.

[22] E. Carosati, R. Mannhold, P. Wahl, J. Hansen, T. Fremming, I. Zamora, G. Cianchetta and M. Baroni, Virtual screening for novel openers of pancreatic K(ATP) channels, J. *Med. Chem*., **50** (2007), pp. 2117-2126.

[23] P. Benedetti, R. Mannhold, G. Cruciani and G. Ottaviani, GRIND/ALMOND investigations on CysLT1 receptor antagonists of the quinolinyl(bridged)aryl type, *Bioorg. Med. Chem.*, **12** (2004), pp. 3607-3617.

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

[25] Q.Y. Zhang, J. Wan, X. Xu, G.F. Yang, Y.L. Ren, J.J. Liu, H. Wang and Y. Guo, Structure-based rational quest for potential novel inhibitors of human HMG-CoA reductase by combining CoMFA 3D QSAR modeling and virtual screening, *J. Comb. Chem.,* **9** (2007), pp. 131-138.

[26] P.C. Jurs, *ADAPT (Automated Data Analysis and Pattern Recognition Toolkit),* University Park, PA: Pennsylvania State University, (2003). Available online at: http://research.chem.psu.edu=pcjgroup=ADAPT.html (accessed 23 April 2002).

[27] A.J. Stuper and P.C. Jurs, *ADAPT:* A computer system for automated data analysis using pattern recognition techniques, *J. Chem. Inf. Comput. Sci.,* **16** (1976), pp. 99-105.

[28] O. Mekenyan and D. Bonchev, OASIS: method for predicting biological activity of chemical compounds, *Acta Pharmaceutica Jugoslavica*, **36** (1986), pp. 225-237.

[29] A.R. Katritzky, V.S. Lobanov and M. Karelson, *CODESSA, Reference Manual*, Gainesville, FL University of Florida (1994). Available online at: http://www.semichem.com=codessarefs.html (accessed 19 April 2007).

[30] V. Consonni, A. Mauri and M. Pavan, DRAGON : software for the calculation of molecular descriptors. Version 5 (2005), Milano, Italy. Available online at: http://www.talete.mi.it=mainexp.htm (accessed 20 May 2006)

[31] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *Journal of Chemical Information and Computer Sciences,* **28** (1988), pp. 31-36.

[32] D.E. Mager, Quantitative structure–pharmacokinetic/pharmacodynamic relationships, *Adv. Drug Deliv. Rev*., **58** (2006), pp. 1326-1356.

[33] R.W. Taft, Polar and steric substituent constants for aliphatic and o-benzoate groups from rates of esterification and hydrolysis of esters, *J. Am. Chem. Soc.* **74** (1952) 3120–3128.

[34] H. Van de Waterbeemd and B. Testa, The parametrization of lipophilicity and other structural properties in drug design*, Adv. Drug Res*., **16** (1987) pp. 85-225.

[35] Verloop, W. Hoogenstraaten, and J. Tipker in E.J. Ariens, Ed., *Drug Design*, Vol. VII, Academic Press, New York, 1976

[36] C.D. Selassie, Chapter One: History of Quantitative Structure-Activity Relationships, in *Burger´s Medicinal Chemistry and Drug Discover,* D. Abraham, ed., John Wiley & Sons, Inc. Publishers, California, 2003, pp. 1-48.

[37] M. Karelson, V.S. Lobanov and A.R. Katritzky, Quantum-Chemical Descriptors in QSAR/QSPR Studies, *Chem. Rev*., (1996) 96, pp. 1027-1043.

[38] S.P. Gupta, Quantitative structure-activity relationship studies of local anesthetics, *Chem. Rev*., **91** (1991) pp 1109–1119.

[39] M. Randic, On characterization of molecular branching, *J. Am. Chem. Soc*., **97** (1975), pp. 6609-6615.

[40] L. B. Kier, L.H. Hall, W.J. Murray and M. Randic, Molecular connectivity I. Relationship tonon-specific local anaesthesia*, J. Pharm. Sci.* **64** (1975) pp. 1971-1974.

[41] L.B. Kier, W.J. Murray and L.H. Hall, Molecular connectivity. 4. Relations to biological activities *J. Med. Chem*., **18** (1975) pp. 1272-1274.

[42] L.B. Kier and L.H. Hall, *Molecular Connectivity in Chemistry and Drug Research.* Academic Press, New York, 1976.

[43] L.B. Kier and L.H Hall, Intermolecular Accessibility: The Meaning of Molecular Connectivity*, J. Chem., Inf. Comput. Sci*., **40** (2000), pp. 792-795.

[44] L.B. Kier and L.H. Hall, *Molecular connectivity in structure – activity analysis;* Wiley: New York, 1976.

[45] L.B. Kier and M.H. Hall, General Definition of Valence Delta Values for Molecular Connectivity*, J. Pharm. Sci*., **72** (1983) pp. 1170-1181.

[46] O. Ivanciuc, T. Ivanciuc, D. Cabrol-Bass and A.T. Balaban, Evaluation in Quantitative Structure-Property Relationship Models of Structural Descriptors Derived from Information-Theory Operators*, J. Chem. Inf. Comput. Sci.,* **40** (2000) pp. 631-643.

[47] M. R. Ginn, D.B. Turner, P. Willett, A.M. Ferguson and T. Heritage, Similarity Searching in Files of Three-Dimensional Chemical Structures: Evaluation of the EVA Descriptor and Combination of Rankings Using Data Fusion*, J. Chem. Inf. Comput. Sci.*, **37** (1997), pp. 23-37.

[48] MOPAC (Molecular Orbital PACkage) is a semiempirical quantum chemistry program based on Dewar and Thiel's NDDO approximation. Stewart Computational Chemistry. Available online at: http://openmopac.net/ (accessed 12 agost, 2007)

[49] A.M. Ferguson, V. Heritage, P. Jonathon, S.E. Pack, L. Phillips, J. Rogan and P.J. Snaith, EVA: A new theoretically based molecular descriptor for use in QSAR/QSPR analysis*, J. Comput.-Aided Mol. Des.,* **11** (1997) pp. 143-152.

[50] M.T.D. Cronin and T.W. Schultz, Pitfalls in QSAR, *Journal of Molecular Structure*, **622** (2003), pp. 39-51.

[51] T.W. Schultz and M.T.D. Cronin, Essential and desirable characteristics of ecotoxicity quantitative structure-activity relationships, *Environmental Toxicological and Chemistry*, **22** (2003), pp. 599-607.

[52] P. Geladi and B. Kowalski, Partial least-squares regression: a tutorial, *Anal. Chim. Acta* **185** (1986), pp. 1-17.

[53] S. Srebenik, M. Weinstein and R. Pauncz, *Chem. Phys. Lett.* **20** (1973) pp. 419, in F.D. Suvire, M. Sortino, V.V. Kouznetsov, L.Y. Vargas, S.A. Zacchino, U. Mora Cruz and R.D. Enriz, Structure–activity relationship study of homoallylamines and related derivatives acting as antifungal agents, *Bioorg. Med. Chem.* **14** (2006) 1851–1862.

[54] L. Michielan, M. Bacilieri, A. Schiesaro, C. Bolcato, G. Pastorin, G. Spalluto, B. Cacciari, K. Norbet Klotz, C. Kaseda and S. Moro, Linear and Nonlinear 3D-QSAR Approaches in Tandem with Ligand-Based Homology Modeling as a Computational Strategy To Depict the Pyrazolo-Triazolo-Pyrimidine Antagonists Binding Site of the Human Adenosine A_{2A} Receptor, *J. Chem. Inf. Model.*, **48** (2008), pp. 350-363.

[55] J. Chen and L. Lai, Pocket v.2: Further Developments on Receptor-Based Pharmacophore Modeling, *J. Chem. Inf. Model.* **46** (2006), pp. 2684-2669.

[56] M. Huang, D.Y. Yang, Z. Shang, J. Zou and Q. Yu, 3D-QSAR Studies on 4-Hydroxy phenylpyruvate Dioxygenase Inhibitors by Comparative Molecular Field Analysis (CoMFA), *Bioorg. Med. Chem. Lett.*, **12** (2002) pp. 2271-2275.

[57] T.A. Ozlem, T.G. Betul, I. Yildiz, A.S. Esin and Y. Ismail, 3D-QSAR analysis on benzazole derivatives as eukaryotic topoisomerase II inhibitors by using comparative molecular field analysis method, *Bioorg. Med. Chem.*, **13** (2005), pp.6354-6359.

[58] Hokuldsson, PLS regression methods, *J. Chemometrics*, **2** (1988) pp. 211-228.

[59] O. Mekenyan, N. Nikolova and P. Schmieder, Dynamic 3D QSAR techniques: applications in toxicology, *Journal of Molecular Structure (Theochem)* **622** (2003), pp. 147–165.

[60] Verloop, *The STERIMOL Approach to Drug Design*, Marcel Dekker, New York, 1987.

[61] J.O. Morley and T.P. Matthews, Structure-activity relationships in nitrothiophenes, *Bioorg. Med. Chem.*, **14** (2006) pp. 8099–8108.

[62] W.J. Murray, L.H. Hall and L.B. Kier, Molecular connectivity III: Relationship to partition coefficients, *J. Pharm. Sci.*, **64** (1975) pp 1978-1981

[63] S.C. Basak and V.R. Magnuson, Molecular topology and narcosis. A quantitative structure-activity relationship (QSAR) study of alcohols using complementary information content (CIC*), Arzneim.-Forsch./Drug Res.* **33** (1983), pp. 501-503.

[64] G.J. Niemi, R.R. Regal and G.D. Veith, *Applications of molecular connectivity indices and multivariate analysis in environmental chemistry,* in *Environmental Applications of Chemometrics,* J. Breen and P. Robinson, eds., American Chemical Society, Washington, 1985, pp. 148-159.

[65] N. Nirmalakhandan and R.E. Speece, Structure-activity Relationships, quantitative techniques for predicting the behavior of chemicals in the ecosystem, *Environ. Sci. Technol.*, **22** (1988), pp. 606-615.

[66] R.W. Okey and H.D. Stensel, A QSBR development procedure for aromatic xenobiotic degradation by unacclimated bacteria, *Water Environ. Res.*, **65** (1993) pp. 772-780.

[67] Simon-Hettich, A. Rothfuss and T. Steger-Hartmann, Use of computer-assisted prediction of toxic effects of chemical substances, *Toxicology*, **224** (2006) pp. 156–162.

[68] R. Biye, Atomic-Level-Based AI Topological Descriptors for Structure-Property Correlations, *J. Chem. Inf. Comput. Sci*., **43** (2003) pp. 161-169.

[69] L.B. Kier and L.H. Hall, *Molecular Structure Description. The Electrotopological State*. Academic Press, San Diego, CA, 1999.

[70] J.F. Contrera, E.J. Matthews and R.D. Benz, Predicting the carcinogenic potential of pharmaceuticals in rodents using molecular structural similarity and E-state indices, *Regul. Toxicol. Pharmacol*., **38** (2003) 243-259.

[71] J.L. Melville and J.D. Hirst, *TMACC:* Interpretable Correlation Descriptors for Quantitative Structure-Activity Relationships, *J. Chem. Inf. Model*., **47** (2007) pp. 626-634.

[72] Z. Yang and P. Sun, 3D-QSAR Study of Potent Inhibitors of Phosphodiesterase-4 Using a CoMFA Approach, *Int. J. Mol. Sci.,* **8** (2007) pp. 714-722. Avalaible online at: http://www.mdpi.org/ijms/

[73] R. Katritzky, L.M. Pacureanu, S. Slavov, D.A. Dobchev and M. Karelson, QSAR study of antiplatelet agents, *Bioorg. Med. Chem*. **14** (2006), pp. 7490–7500.

[74] S. Holder, M. Lilly and M.L. Brown, *Comparative molecular field analysis of flavonoid inhibitors of the PIM-1 kinase,* Bioorg. Med. Chem. 15 (2007), pp. 6463–6473.

[75] Bak and J. Polanski, A 4D-QSAR study on anti-HIV HEPT analogues, *Bioorg. Med. Chem*. **14** (2006), pp. 273–279.

[76] P.S. Kharkar, B. Desai, H. Gaveria, B. Varu, R. Loriya, Y. Naliapara, A. Shah and V.M. Kulkarni, Three-Dimensional Quantitative Structure-Activity Relationship of 1,4-Dihydropyridines As Antitubercular Agents, *J. Med. Chem*., **45** (2002), pp. 4858-4867.

[77] R. Ragno, S. Simeoni, S. Valente, S. Massa, and A. Mai, 3-D QSAR Studies on Histone Deacetylase Inhibitors. A GOLPE/GRID Approach on Different Series of Compounds, *J. Chem. Inf. Model*., **46** (2006), pp. 1420 -1430.

[78] P.R. Duchowicz, M.G. Vitale, E.A. Castro, Fernández and M. Caballero, J. QSAR analysis for heterocyclic antifungals, *Bioorg. Med. Chem*. 15 (2007) pp. 2680–2689.

[79] Hillebrecht and G. Klebe, Use of 3D QSAR Models for Database Screening: A Feasibility Study, *J. Chem. Inf. Model*. **48** (2008), pp. 384-396.

[80] T. Abhilash, S. Vishwakarma and M. Thakur, QSAR study of flavonoid derivatives as p56lck tyrosinkinase Inhibitors, *Bioorg. Med. Chem*. **12** (2004), pp. 1209–1214.

[81] J. Gálvez, J.V. De Julián-Ortiz and R. García-Doménech, Diseño y desarrollo de nuevos fármacos contra la malaria, *Enf. Emerg*., **7** (2005), pp. 44-51.

[82] J. Liu, G. Wu, G. Cui, W.X. Wang, M. Zhao, C. Wang, C. Zhang and S. Peng, A new class of anti-thrombosis hexahydropirazyno-[1,2:1,6]pyrido-[3,4-b]-indole-1,4-dions:Design, synthesis, log K determination, and QSAR analysis. *Bioorg. Med. Chem.,* **15** (2007), pp. 5672–5693.

[83] R. Martínez, M.A. Peña-Montiel and J.G. Avila-Zárraga, Estudio cuantitativo de la relación estructura-actividad (QSAR) de Bis(acridin-4-carboxamidas) con actividad citotóxica, *Rev. Soc. Quím. Méx*., **44** (2000), pp. 62-66.

[84] J.A. Spiecer, S.A. Gamage, G.J. Atwell, G.I Finlay, B.C. Baguley and W.A. Denny, Structure-Activity Relationships for Acridine-Substituted Analogues of the Mixed Topoisomerase I/II Inhibitor N-[2-(Dimethylamino)ethyl]acridine-4-carboxamide, *J. Med. Chem*. **40** (1997), pp. 1919-1929.

[85] M.T. Scotti, M.B. Fernandes, M.J.P. Ferreira and V.P. Emerenciano, Quantitative structure–activity relationship of sesquiterpene lactones with cytotoxic activity, *Bioorg. Med. Chem.* **15** (2007) pp. 2927–2934.

[86] Y. Shao, H. Ding, W. Tang, L. Lou, and L. Hu, Synthesis and structure–activity relationships study of novel anti-tumor carbamate anhydrovinblastine analogues, *Bioorg. Med. Chem.* **15** (2007), pp. 5061–5075.

[87] L. Eriksson, E. Johansson, N. Kettaneh-Wold and S. Wold, *Multi-and megavariate data analysis - principles and applications*, Umetrics AB. (2001).

[88] T. Hastie, R. Tibshirani and J. Friedman, *The elements of statistical learning. Data Mining, inference and predictions*, Springer, 2001.

[89] S. Izrailev and D. Agrafiotis, *Variable selection for QSAR by artificial ant colony system*, SAR QSAR Environ. Res.,, in press 2002.

[90] S. Izrailev and D. Agrafiotis, A novel method of building regression tree models for QSAR based on artificial ants, *J. Chem. Inf. Comput. Sci.,* **41** (2001), pp. 176-180.

[91] H. Blockeel, S. Dzeroski, B. Kompare, S. Kramer, B. Pfahringer and W. Van Laer, Experiments In Predicting Biodegradability, *Applied Artificial Intelligence*, **18** (2004), pp. 157-181.

[92] M. Fernández and J. Caballero, QSAR modeling of matrix metalloproteinase inhibition by N-hydroxy-a-phenylsulfonylacetamide derivatives, *Bioorg. Med. Chem*., **15** (2007), pp. 6298–6310.

[93] A.M. Doweyko, 3D-QSAR illusions, *J. Comput. Aided Mol. Des*., **18** (2004) pp. 587-596.

[94] Tropsha, P. Gramatica and V.K. Gombar, The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models, *QSAR Comb. Sci*., **22** (2003), pp. 69-77.

[95] M. Hawkins, S. C. Basak and D. Mills, Assessing Model Fit by Cross-Validation, *J. Chem. Inf. Comput. Sci*. **43** (2003), pp 579-586.

[96] F.D. Suvire, M. Sortino, V.V. Kouznetsov, L.Y. Vargas, S.A. Zacchino, U. Mora Cruz and R.D. Enriz, Structure–activity relationship study of homoallylamines and related derivatives acting as antifungal agents, *Bioorg. Med. Chem.* **14** (2006), pp. 1851–1862.