

Research Signpost
37/661 (2), Fort P.O., Trivandrum-695 023, Kerala, India



Current Methods in Medicinal Chemistry and Biological Physics 2007, Vol.1: 61-90 ISBN: 81-308-0141-8
Editors: Carlton A. Taft and Carlos H. T. P. Silva

4

3D quantitative structure-activity relationships: The three-dimensional road to lead design

Andrei Leitão¹, Adriano A. Andricopulo² and Carlos A. Montanari³

¹Núcleo de Estudos em Química Medicinal, Departamento de Química, Universidade Federal de Minas Gerais, Av. Antônio Carlos, 6627, 31270-901 Belo Horizonte/MG, Brazil

²Laboratório de Química Medicinal Computacional, Centro de Biotecnologia Molecular Estrutural, Instituto de Física de São Carlos, Universidade de São Paulo, Av. Trabalhador Sancarlenense, 400, 13560-970, São Carlos/SP, Brazil; ³Grupo de Estudos em Química Medicinal de Produtos Naturais, Instituto de Química de São Carlos, Universidade de São Paulo, Av. Trabalhador Sancarlenense, 400, 13560-970, São Carlos/SP, Brazil

Introduction

The drug design process does not start with the identification and validation of the target, but rather with the hit-to-lead endeavor [1]. However, the identification and validation of the target biomacromolecule for a desired disease state are very important tasks for any research project aimed at discovering new chemical entities (NCE) with the bioactivity profile well defined and well characterized. In order to do so, the field of bioinformatics [2] (the application of computer

technology to the management of biological information), molecular biology [3] (a branch of biological science that studies the biology of a cell at the molecular level) and proteomics [4] (the study of proteins and their interactions) have to be integrated to the discovery and development processes to facilitate the finding of good complementary interactions between a small molecule and its target (or targets). This can be easily justified, for instance, by the fact that at present the “blockbuster” [5,6] Lipitor[®], which sold 12.2 bi USD in 2005, is an inhibitor of HMG-CoA (3-hydroxy-3-methylglutaryl-coenzyme A) reductase (HMGR) that catalyzes the committed step in cholesterol biosynthesis [7].

The identification of a valid ligand is a course of action that is incorporated in the lead discovery that is the process of identifying active NCE, which by subsequent modification may be transformed into a clinically useful drug. Natural [8] and combinatorial chemistry products [9] can serve as chemical sources for building virtual library of real molecules. By employing chemoinformatic tools, *in silico* models can be generated for affinity, potency and the disclosing of pharmacokinetics (PK) by accessing absorption, distribution, metabolism and excretion (ADME) properties [10]. The synthetic viability [11] leads to the possibility of establishing structure-activity and – property relationships (SAR and SPR, respectively), from which the generation of the lead can be fulfilled. These strategies are used to identify compounds that possess a desired but non-optimized biological activity. In the lead optimization process, the synthetic modification of a biologically active compound to fulfill all stereoelectronic, physicochemical, pharmacokinetic and toxicological studies required for clinical usefulness, is then accomplished right in the beginning of the drug design pipeline.

Computer-aided molecular design (CAMD) is now being thoroughly applied in the drug and development processes to allow for the value aggregation of affinity, potency, selectivity, solubility, metabolism, permeability issues and so forth. However, there is a challenge in describing metabolism. In this sense, the parallel optimization of properties and activity through *in vitro* and subsequently *in vivo* experiments represent a good deal of efforts to incorporate appropriate information into the molecular panels before going to the next step. Thus far, the actual task of medicinal chemistry is to accomplish with the integration of experimental and CAMD methods. In this way, only molecules with desired and well characterized properties will reach the status of being promoted to the next step of the pharmaceutical pipeline.

The chemical abstract has ca. 29.10^6 molecules, and this constitutes the actual universe of putative bioactive entities. However, only 8.10^4 are described to have either a bioactive property or being a drug. Nonetheless, ca. 2.000 molecules are really in therapeutically areas and are marketed throughout the world. If we take a molecule with molar mass of less than 500 Da and atoms like H, C, N, O, P, S, F, Cl and Br (just to follow some of the Lipinski's rules), we reach the amazing number of 10^{62-64} molecules to be synthesized and tested to find out a single NCE! The best estimates state that mankind can accomplish with this by only synthesizing around 10^{40} molecules, and even though not all of them will have any chance of becoming a drug molecule. CAMD is then a tool for helping the description of the chemical-biological space where the islands of bioactive molecules are widespread over the sea of small molecules with great potential of becoming a drug [12].

If a biological system (an enzyme or a receptor) interacts with a small molecule and from this interaction an action or perturbation occurs, CAMD can be of utmost importance to describe properly the process. If the description of the biological activity is done correctly and its quantitative measurement is assured, quantitative structure activity relationships (QSAR) can be generated. Since it is necessary to describe small molecules through all their physicochemical parameters needed to be correlated with the biological responses, quantitative structure property relationships (QSPR) can also be valuable to incorporate them in the appropriate chemical-biological space.

In this chapter, we will focus on some of the technologies that are needed to generate QSAR based on the three-dimensional description of the ligands, termed 3D QSAR. Nevertheless, the chapter will mostly take account of the models describes for when the 3D structure of the receptor is known. We do so because of the understanding that in the near future the proteomics will identify all the therapeutically relevant targets and their 3D structures will be available. However, since this is not a necessary condition for developing a 3D QSAR model, we will provide a deal of information regarding the description of the 3D structure of the small molecule necessitated for developing a model.

Strategies in drug design

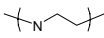
Out of many different ways of applying CAMD to the drug design effort, two of them are of importance here. The first one concerns with the fact that the ligand structures are known but the target macromolecular structure is unknown. In this case, pharmacophores [13], 3D similarity and 3D QSAR play a role in describing the putative intermolecular interactions. It is possible to incorporate information of the “virtual receptor site” nature by means of the interacting groups we can place in the ligand structures. This method is known as ligand based drug design (LBDD). The 3D similarity is dealt with by the knowledge of the 3D structure of the target and ligands. Here, the structure based drug design (SBDD) [14,15] can be brought to the scene. Moreover, by mapping the receptor and docking small molecules into the receptor site through a proper virtual screening [16] process, new classes of unknown ligands to the desired target can be found to be of valuable interest in the hit-to-lead process.

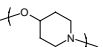

In this chapter we will refer to the interaction site through the receptor target that is an enzyme. However, readers are directed to more broad definitions of the terms and targets used [17].

Let us first suppose we have the 3D structure of the receptor target. If a ligand one interacts with the target in the receptor site, it may disclose complementary intermolecular interactions like hydrogen bond donor-acceptor, hydrophobic and π - π stacking. In the receptor site, these kinds of interactions are all located in the appropriate positions to accommodate the ligand and keep the most favorable energy for the co-complex. This is pictured so just to help the comparative description of many ligands interacting in the same receptor site. The alleged properties do not account for any kind of moiety that is needed to form the co-complex. For instance, the ligand one can have a hydrogen bond donor site along with hydrogen acceptor site provided the receptor has the opposite complementary sites. But, nothing can be told about the kind of sub-structure is needed to establish these interactions. In this context, a putative ligand two could bear $-OH$ and $=S$ groups instead of $-NH_2$ and $=O$ as found in ligand one. Ligand

three, on the other hand would have a $-(\text{NH})\text{NH}_2$ moiety as hydrogen bond donor. Hydrogen bonds are well dependent on their geometries. So when finding affinity, each of these ligands would contribute differently to the overall affinity interactions that account for the formation of the co-complex ligand-enzyme. That is, enthalpies of enzyme-ligand themselves, enzyme and ligand solvations, besides the vibration are the most important within entropic contributions.

Provided the description of the complementary intermolecular interactions are fulfilled for all small molecules that can fit in the receptor site, the question that arises is how such small molecules can be placed to elicit the desired biological response? How small can high-activity molecules be? For instance, histamine is a small endogenous molecule with a molar mass of 111 Da, partition coefficient (by means of its $\log P$) of -0.7 and has an affinity constant of as low as 8.2 nM for its H_3 receptor. One strategy to go for is related to the definition of the chemical-biological space of drug-like molecules as proposed by Lipinski and followed by others by including ligand-similar and lead-like properties. In the so called “rule-of-five” (Ro5) [18], drugs being administered by oral rout should possess molar mass lower than 500 Da, $\log P$ less than 5, H-bond donor of 5 and H-bond acceptors of 10 (N+O). By extending this, Vieth *et al.* have included number of atoms, rings, rotating bonds, polar surface areas, H-bond donors and receptors, number of halogen atoms, in addition to the numbers of O and N and OH and NH, for 1729 therapeutically used drugs [19,20]. Others, like Carr *et al.* have introduced the “rule-of-three” (Ro3) [21,22] in the search for fragment-like [23] molecules to interact with receptor sites. This would allow for better incorporation of fragments that would help in getting molecules with drug-like properties by adding balanced polar and apolar groups to end with the appropriate physicochemical properties desired for a drug molecule.

Different ligands can interact with the same target, and the same ligand can interact with different targets [24,25]. The important thing that comes out from this is the existence of privileged fragments, sub-structures and molecular anchors, as proposed by Müller [26]. For instance, comparison of most frequent scaffolds in 1729 therapeutically used drugs reveals that 112  groups are to be found in oral drug molecules, but

only 4 are found for . For oral drug side chains, 241  are registered [19].

The pharmacophore identification is of chief importance in order to describe the putative receptor site interactions. 2D and 3D pharmacophore perceptions might be envisaged. However, they have to comprise different tautomers that can be found in the small molecule. For instance, histamine has 3 different 2D pharmacophore hypotheses depending on its tautomers of singly extended Acceptor, Donor and Protonated states: A-D-P3, D-A-P3, P-P-P3, and other 3 similar for the “cycled” conformation, Figure 1 [27].

The integration of 3D QSAR in the drug design arena is carried out in a cyclic process. Because of this, its start point depends on the amount and quality of data that are in hands. The biological responses obtained for a set of compounds can be treated in the CAMD procedure initially to describe the relationships between them and the 3D molecular properties of the ligands. The resulting model (or models) can then be used to the design of new compounds with optimized properties. To this end, the structure activity relationships (SAR) coined at the end of XIX century is thoroughly used. When the quantitative biological descriptor is available, QSAR can be developed by using either 2D or 3D information from chemical structures. In these cases, specific biological

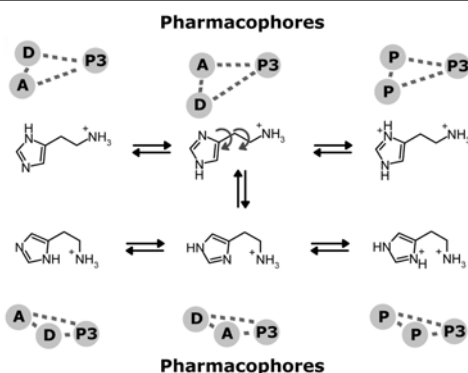


Figure 1. Pharmacophore hypotheses generated by different conformational and protonation states of the histamine.

functions have to be established, since the molecular modification has to respond to the same target according to the complementary intermolecular interaction. Pharmacodynamics (PD) of ligand-receptor complex, PK properties, log P, all respond well to this quantitative relationships. However, only advanced statistical models are to be found of any value when the 3D structure of the small molecule is to be described as being of any importance to the ligand-receptor complex. Validation methods have to be applied along with multiple regressions that allow the inclusion of a myriad of independent variables. The choices of the suitable descriptors, however, are not a simple task. The Buffon's needle problem [28] emphasizes that to find out the probability that a needle of length l will land on a line, given a floor with equally spaced parallel lines at distance d apart, depends on its length, color, composition, texture and orientation. This means that molecules have to be described in terms of their constitutional (molar mass, number of atoms), topological (connectivity, 2D fingerprint, 3D topographical indices – pharmacophores), electrostatics (polarity, polarizability, partial charges), geometry (length, width, volume), and chemistry (lipophilicity, HOMO/LUMO energies, vibrations, bond orders, total energy, free Gibbs energy, enthalpy and entropy) through 1D, 2D and 3D representations. To the purposes of 3D QSAR multiple regressions, principal component analysis (PCA), partial least square (PLS), neural networks (NN), etc, have the goal of connecting descriptors to structural elements to allow for interpolation and with care for data extrapolation. Nonetheless, it has to be realized that structural elements can yield information about the bioactivity as well as properties can do so but only to a certain extent. This means that there exists a certain chemical-biological space that is overlaid and no informative distinction can rise from such regions. For instance, when we say that there is an H-bond donor group interplaying a relationship we cannot state if it comes from an –OH or –NH group. This is shown in the Venn diagram below (Figure 2).

Thus, from activity itself we can only infer about structure and property.

There are some pre-qualifications that need to be met before any QSAR can be envisaged. The relationship of correlations vs dependency has to be settled. For correlation we understand that two or more variables can be correlated to the same

property of a certain system without interfering with each other. When this correlation is established by the fact that one change is due to the other, no correlation actually can be set forth. On the contrary, in this case there is a clear dependency between the variables or descriptors. For example, the legs and head of an elephant are correlated regarding their sizes, but they do not depend on each other.

Structure property correlations (SPC) rely on that the correct interpretation of the whole molecular structures is dealt with their own described properties: intrinsic (molar mass, volume, connectivity, charge), chemical (pKa, log P, solubility, and stability), and biological (activity, toxicity, biotransformation, PK). So, to circumvent the problem of intercorrelated variables, the Craig type diagram has to be used to correctly choose the independent variables. For example, in the graph below (Figure 3) that accounts for the relationships between the Hansch lipophilicity constant, π , and the Hammett σ constant, the selected substituents in a congener set of compounds appropriate for a 2D QSAR analysis are to be chosen to grant their lack of intercorrelation. In other words, if we find the biological response to be correlated with π there is no role played by σ .

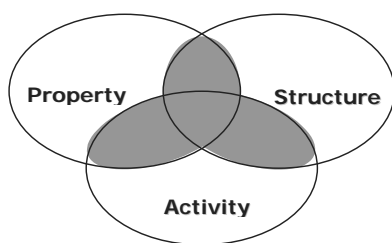


Figure 2. The relationships between property, structure and activity.

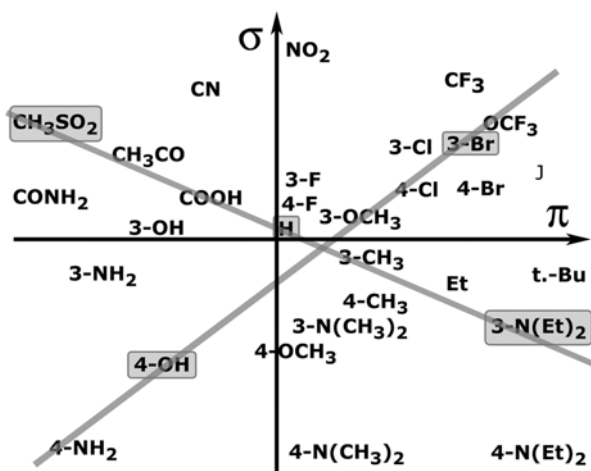


Figure 3. Craig plot from a congener set design.

Variance among the range of biological response is an early verification that has to be done to avoid overfitting and unmet correlations that modifies according to the way the fitted model has treated the raw data. In the following graph, the second normal distribution is the one most appropriated to be used in the analysis (Figure 4).

One good way of doing this is by carrying out a cluster analysis. For instance, the graph below shows a proper similarity analysis for the choice of the training and test sets (see below) well distributed within the whole range of bioactivities (fictitious data), Figure 5.

Instead of calculating or obtaining experimentally all properties, the chemical structural space has to be built according to the correct biological space. For instance, if we intend to synthesize benzodiazepine derivatives (Figure 6) whose structures are proposed from a 3D QSAR method using the following template, we do not need to do it for 759.10^6 molecules! Instead, it might be feasible only for 15 molecules!

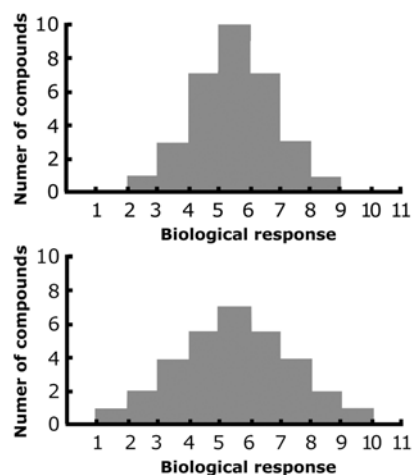


Figure 4. Two kinds of normal distribution in a data set.

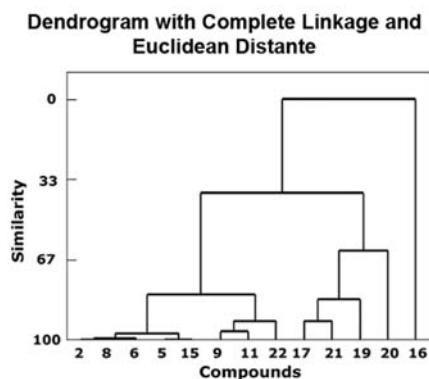


Figure 5. Dendrogram analysis of a compound series.

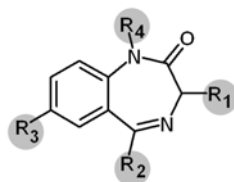


Figure 6. Benzodiazepine template and four putative positions for chemical modifications.

This, however, is only possible if the chemical-biological space is appropriately defined. To accomplish with that, let us consider the partition coefficients (P) of drugs that act in the central nervous system (CNS), Figure 7.

Y axis represents the number of drug occurrences with a desired $\log P$ value. If that is so, we should look for drug molecules that bear $\log P$ of value ca. 2-3. This is also because the chance of finding out molecules with drug-like characteristics would be easier and well widespread along with the biological island of interest.

Variable selection is a statistical procedure very useful in regression analysis. If we have n structures represented by p descriptor variables, some of these can be removed and prediction rules are calculated for the left set of variables. The few selected variables diminish the prediction error and also leads to easier interpretation of the final model. In this parsimony way of obtaining the model, only the most representative descriptor variables are used to describe the structure activity relationships. Many suitable search algorithms can be applied for carrying out variable selection: forward selection, backward elimination, simulated annealing, genetic algorithms, fractional factorial design, etc.

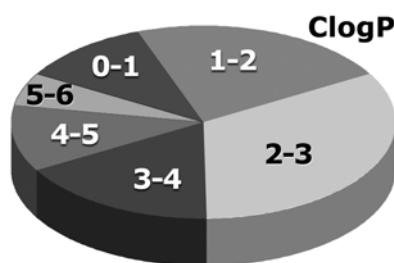


Figure 7. Partition coefficients observed for CNS drugs.

Drug-receptor interactions

QSAR include all statistical methods from which the biological response is correlated to structural elements (Free-Wilson), physicochemical parameters (Hansch-Fujita) or molecular interaction fields (3D QSAR). In all 3D QSAR analyses, the 3D structures of molecules are correlated to molecular fields. There is no complete need for congener set of compounds, but the same mode of action has to be warranted. This means the co-complex ligand-receptor must also have very similar mode of binding (MOB). Prediction of isosteric effect might be envisaged and affinity is to be correlated

with energy of interaction. Although the biological response can be potency, affinity is the desired property to be correlated to it.

Some of the intermolecular forces used to hold together ligands and their receptors are depicted below (Figure 8).

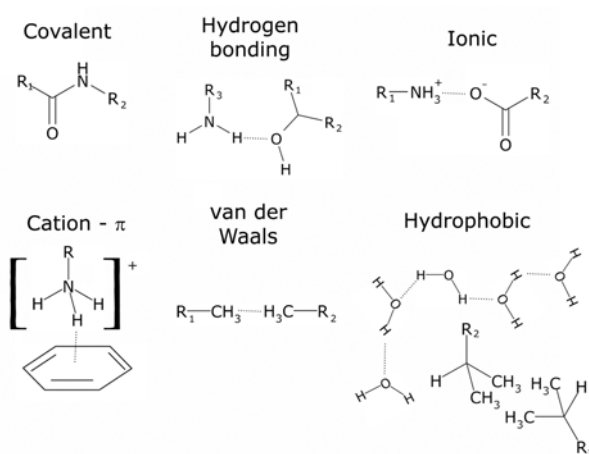


Figure 8. Some intermolecular forces that mediate the drug-receptor interactions.

The drug-receptor interactions have to be described in terms of such forces. In order to accomplish with this, the estimation of binding affinity of the ligand through their free energy of binding has to be carried out.

The major drug-receptor interaction forces are of utmost importance in describing any 3D QSAR model. They comprise the long-range attractions between the drug and receptor that are of electrostatic character. The resulting noncovalent bonds charge-charge, charge-dipole and dipole-dipole are all important to hold the two molecules together. Although the charge-charge interaction is stronger because many drugs contain charged functional groups, the others are more widely found along the whole structures due to differences in electronegativities of all interacting atoms. Inductive interactions like charge redistribution (polarization) occur when temporary dipoles are induced. Although of widespread occurrence in drug-receptor interactions, the total energy accounted for the ion- and dipole-induced interactions is not always taken into consideration when energy calculations are performed. Charge transfer occurs when electron donors make close contact with electrons acceptors and thus allow for an electron transfer to the low-energy molecular orbital of the latter. Nonpolar interactions are attractive forces taking place within nonpolar substituents together. This is so because of the fluctuating dipoles induce dipoles in each other. The resulting forces are of attractive energy and can be very significant when there is a close contact between the drug and the receptor. Hydrogen bonds play a pivotal role in biological systems. Those of oxygen and nitrogen atoms are responsible for maintaining tertiary structure of proteins and nucleic acids. Hydrogen bonds also play their role when binding of small molecules in the receptor site. From the hydrophobic interactions there are highly ordered water molecules that form cages around solute molecules with nonpolar alkyl

chains. Ordered water molecules interact around the drug and receptor. When the drug-receptor interaction takes place, disordered water molecules are displaced and the co-complex is stabilized by hydrogen bonding, ionic and hydrophobic interactions. There is a penalty due to the loss of translational and rotational movements along with the restricted conformational freedom of flexible molecules when it happens. Nonetheless, the decrease in such entropy terms is somewhat accompanied by vibrational degrees of freedom that rises when the complex is formed. The representation below depicts the three ways molecules storage energy (Figure 9).

The whole process of drug-receptor interactions results in the total free energy change, ΔG , which is associated with all of the electrostatic, inductive, nonpolar and hydrophobic interactions less any conformational energy and rotational, translational or conformational entropy costs allied with the interactions.

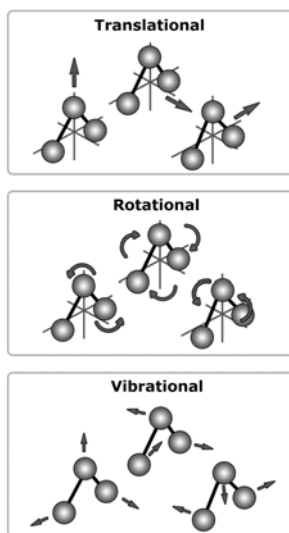


Figure 9. Molecular movements.

A thorough theoretical description of all of the energy contributions for solving the drug-receptor complex surrounded by solvent and any other pertinent solute would be a keystone in 3D QSAR. Instead, the role a functional group plays in the drug molecule can be assessed to describe any increase in binding to the receptor.

One well guessed attempt to do that comes from the Andrews analysis [29] (Figure 10 and Equation 1):

$$\begin{aligned} \Delta G_{\text{avg}} = T \Delta S_{\text{rt}} + n_{\text{DOF}} \cdot E_{\text{DOF}} + \sum n_x \cdot E_x = \\ -59 - 3.0n_{\text{DOF}} + 3.0n_{\text{Csp}^2} + 3.4n_{\text{Csp}^3} + 48n_{\text{N}^+} + 5.0n_{\text{N}} + 34n_{\text{COO}^-} \\ + 42n_{\text{PO4}^{2-}} + 10.5n_{\text{OH}} + 14.2n_{\text{C=O}} + 4.6n_{\text{O,S}} + 5.4n_{\text{Hal}} \end{aligned}$$

Equation 1. Calculation of the free energy according to Andrews.

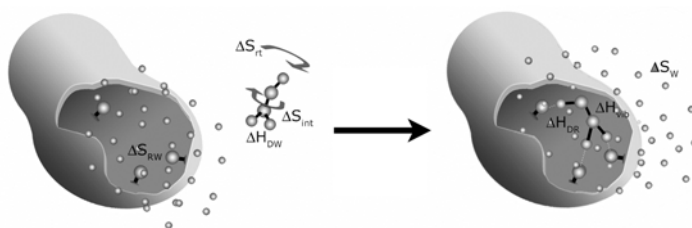


Figure 10. Free energy components of a drug-receptor interaction.

The free energy average of binding interactions is related to the degrees of freedom (DOF), all the intermolecular forces (carbon atoms sp^2 and sp^3 hybridized, quaternary and neutral nitrogens, carboxylates, phosphates, hydroxyl, carbonyl, oxygen, sulphur, halogens) and the entropy. Although entropy is a term of questionable utility in this equation, from such analysis it is possible to realize the whole different types of molecular constituents might have when the intermolecular complementary interactions take place, Equation 2.

Estimates of the free energy of binding for a given protein-ligand complex of known 3D structure has also been developed from empirical function [30].

$$\begin{aligned} \text{Log } 1/K_i &= 1.4(\pm 0.4) \text{ ionic hydrogen bonds} \\ &+ 0.83(\pm 0.3) \text{ neutral hydrogen bonds} \\ &+ 0.03(\pm 0.01) \text{ lipophilic contact surface area} \\ &- 0.25(\pm 0.1) \text{ number of rotatable bonds} - 0.91(\pm 1.4) \\ &(n = 45, r^2 = 0.766, s = 1.40, F = 32.8) \end{aligned}$$

Equation 2. Major intermolecular interaction terms that contribute to the potency.

The logarithmic relationship between free energy of binding and dissociation constant potency means that every ΔG change of $-1.4 \text{ kcal mol}^{-1}$ results in a 10-fold change in potency. Kuntz et al. [31] surveyed the dissociation or IC_{50} values of ~ 150 ligand complexes and concluded that the maximum affinity per atom for organic compounds is $-1.5 \text{ kcal mol}^{-1}$ per non-hydrogen atom. Ligand efficiency [32,33] is a way of normalizing the potency and molecular weight (MW) of a compound to provide a useful comparison between compounds with a range of MWs and activities: $\Delta g = \Delta G/N_{(\text{non-hydrogen atoms})}$.

Cheminformatic tools

Computer aided molecular design (CAMD) or computer assisted molecular design relies on the fact the computers play a fundamental role in the process of discovery and development of new chemical entities. It is considered that the beginning of the 1990's corresponds to a change in the way computers impacted the drug design efforts. Up till then, the candidate drug molecule would arise after a "linear" discovering fashion process. From diverse libraries, screening would identify hits whose optimized potencies

would serve in the traditional QSAR modeling. The optimization of the pharmacokinetics properties (ADME) would follow next and then the toxicity profile of molecules would be incorporated to the arsenal of information needed to label the compound as a good candidate drug molecule or not. Latter, parallel optimization of potency, selectivity and ADME/Tox properties via predictive computational models takes place mainly due to the role cheminformatics play in the drug discovery and development process.

Cheminformatics [34] represent the computational approach capable of dealing with a compound chemical library with great diversity and with varied structural processing technologies for diversity analysis.

Cheminformatics (chemoinformatics or chemo-informatics) is a discipline emerged from computational chemistry, chemometrics, QSAR, chemical information, etc., that uses computer technologies to process chemical data. Since we are dealing with chemical structures, many requirements to work with chemical data entail representation, storage, retrieving structures (libraries), design and data organization, prediction of structures and properties, drug-like and lead-like properties, molecular similarity [35] and dissimilarity, etc. Accordingly, cheminformatics stands for management, maneuvering and optimization of properties that maybe useful in the process of discovery and development of new drug molecules.

Methods in cheminformatics will be used for the description of the molecular structure from its one-dimensional characteristic till the stereoelectronics, going through 2D, 3D, colligative properties and stereodynamics. A variety of computer programs will be necessitated for solving this multitude of activities when designing new candidate drug molecules: CoMFA, HQSAR, VolSurf, Almond, 3D TSAR, GRID/GOLPE, AMBER, CATALYST, etc., all integrated with experimental procedures.

Quantitative structure activity relationships (QSAR)

The drug-receptor interactions are a function of the three-dimensional structure of the ligands (Figure 11).

From such assumption, the 3D QSAR model can be built by following some simple steps, despite the fact that they are not so easily solved. These steps are done by the comparative molecular interactions field (CoMFA) but they are readily available for every 3D QSAR model generation with the suitable adjustment. To start with, it is necessary to create a 3D database that contains all three-dimensional structures of the study molecules. Of course, appropriate biological responses are granted for such studies. The first generated 3D structures might be obtained by simply applying one method of 2D conversion to 3D structures, like Concord[®] and Corina[®]. Just sketching the 2D structure and saving it as SMILES notation ready to be applied in the conversion software, or, of course, directly from it.

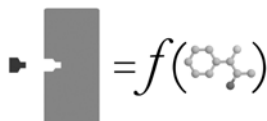


Figure 11. The drug-receptor interactions as function of the 3D structures.

In doing so, however, it is necessary to bear in mind the quality of the 3D structure generated. First of all, it has to be assessed the quality of the compound's structure. Secondly, it has to be decided which conformations of the compounds have to be stored in the database. On the other hand, molecules are flexible and this means that their generation must not take account only for the atom bearing connectivity but also the environment in which they are located at. For instance, the question to be answered here might be: what sort of conformation should one use for? In the cartoon below [36], we can see clearly the need for correctly describing the right conformation it is needed for the 3D QSAR analysis. Except the one that represents the conformation in the receptor site as being the bioactive conformation, all the others even being fully optimized do not resemble the needed one (Figure 12). It is widespread known that the use of *in vacuo* conformation would have no correspondence to the bioactive conformation. The reasoning for this is the fact that the bound conformation is not the most stable one but, on the contrary, it is a conformation with higher free energy of interaction due to the intermolecular interplay around the sites of interaction. This escaping from the fully optimized conformational energy has to be addressed properly to accomplish with the correct description of the receptor-bound conformation.

In the structure based drug design (SBDD), the conformations we can obtain from the protein databank (PDB) [37] - the crystal structures of the target along with the ligand X-ray crystal structure, stand for by the condensed phase in the cartoon above. While this is a good compromise of choosing a ligand conformation taken from the receptor site, not every ligand will have the needed coordinates to allow for correctly representing the 3D structure. In such cases, the rest of the conformations for the purpose of 3D QSAR can be generated by similarity calculations as depicted in the illustration that can be found in the example section of this chapter. Thus far, the issue of flexible conformation must be taken into account when dealing with the 3D QSAR modeling at the time it is being carried out.

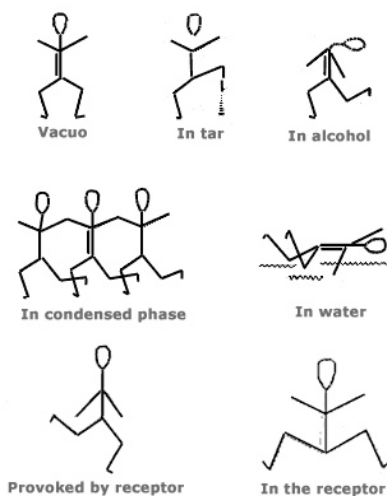


Figure 12. The influence of the environment on the conformation.

Calculating charges for each of the compound - either Gasteiger-Marsilli or Gasteiger-Hückel - maybe appropriate if we are only interested in the prediction of biological activity from 3D structures of molecules. They are quite reliable and very fast to calculate. Nevertheless, when dealing with electrostatic interactions, charges maybe of paramount importance in obtaining a 3D QSAR model. So, it is often useful to assume that the calculated charges may change the overall results. Moreover, by changing conformation it is required to re-calculate charges for the new generated one. Concerning the calculation of charge distribution in cases where insight into the physicochemical properties of the molecule that can be used to describe the drug-receptor interactions by means of their complementarities, a systematic study of the correct charges to be calculated for a molecule has to be effectuated. In such cases the calculation of charge distribution for a molecule may include the ESP keyword option of MOPAC, the standard MOPAC method, the CHELPG method and so forth. In this case, the PLS analyses (see below) have to be compared to take into account the correct choice of the charges to be used. Any evidence of reliability has to be disclosed.

After the charges are correctly calculated, minimizations of the structures within the receptor site if the 3D structure of the target is known and the alignment have to be accomplished. It is required that even at the molecular mechanics level, a molecular dynamics should be performed since the maneuvering of the ligand in the receptor site is not good enough to ascertain for a correct docking within the “desired” mode of binding [38].

Without the information of the 3D structure of the target, the pharmacophore hypothesis, 3D similarities and 3D QSAR can be used to attempt to interpret the structural complementarities compounds might have when they bind to receptors.

When the 3D structure of all molecules in the series are generated and optimized accordingly, the calculation of the electrostatic and steric field energies with a cut off of $\pm 30 \text{ kcal mol}^{-1}$ using a $+1 \text{ sp}^3$ carbon probe in 2 Å grid can be carried out.

A set of probes and cut offs can be applied in the generation of the molecular interaction fields. When the chosen probe is the $+1 \text{ sp}^3$ carbon the standard force field maybe applied as above. However, this may not be the case for when hydrogen-bonding properties are of interest. In such cases, the force field has to be changed accordingly. For instance, by reducing the role of steric field there is an increase in the electrostatic field. On the other hand, any hydrogen bonding calculation will render the appropriate field if hydrogen with charge of $+1.0$ or an sp^3 oxygen with charge of -1.0 are used without field smoothing around the cut off value.

Doing the regression analyses via partial least square (PLS) and performing full cross-validation with leave-one-out method are the next steps to be followed. First, performing a test with SAMPLS[®] to glimpse the results is a good practice at the beginning of the analysis. When the model is found to be appropriate, the leave-many-out method should be applied. By evaluating the fitting term (R^2) and the prediction power of the model (Q^2), the 3D QSAR can be used to disclose the contour maps, which represents the positions where substitutions are to be placed in order to improve the biological response.

Ehrlich at the turn of the XIX to XX century conceptualized that binding conformations of different drugs with similar steric and electronic patterns rely on the receptor site interaction with the *same* mode of binding. This means that a set of

functional groups in the receptor site (the binding-site hypothesis) recognizes the functional groups (the pharmacophores) in the drug molecule. In other words, in the pharmacophore hypothesis there is a correspondence of functional groups in different congeneric series of drugs while in the binding-site hypothesis the same binding sites overlap the corresponding functional groups of drug molecules. The pharmacophore is the 3D arrangement of molecular fragments or even the distribution of electron density that is recognized by the receptor. That is, some parts of the molecule are of essential importance in the recognition process and eliciting the biological response. Other parts of the molecule can be changed since the scaffold holding the pharmacophoric groups does not alter the activity. Nevertheless, the existence of multiple binding modes can be of no utility when deriving any structure-activity relationships (SAR) for the whole set of molecules. In such cases, new and separate SAR have to be developed.

Two drawbacks are quite important when doing such analyses. The first one is related to the active conformation (the so called “pharmacophore conformation”). Rules of superposition using the active analogue can be defined. Sometimes it is also useful to have a look in the inactive analogues to find out what might be falling apart in getting the biological response. The definition of the alignment rules is crucial. Pharmacophoric and surface properties can be used to help in defining them (Figure 13).

Distances are shown in angstroms. H-don: hydrogen donor. H-acc: hydrogen acceptor. N⁺: nitrogen positively charged.

In this fictitious example, all molecules in the database have to be aligned over the hypothesized pharmacophoric definitions established so far.

Surfaces themselves can be used as a rule for alignment. The problem it is facing here refers to which surface we should rely on to carry on with the alignment. For instance, for simple aromatic compounds some different surfaces maybe drawn (Figure 14).

In this example case, van der Waals (a), dipole (b), Grid NH donor probe (c) and Grid C=O acceptor probe (d) are used to depict some of the suitable surfaces for proper alignment.

The correct alignment relies on superposition of such elements that encompasses the most probable that all molecules in the set bind in a similar way in the receptor site. So by choosing it right is a matter of hard work and represents the greatest challenge in 3D QSAR studies. For instance, if we were to superimpose WIN molecules according to their structures we would certainly miss the right interactions taking place in the human

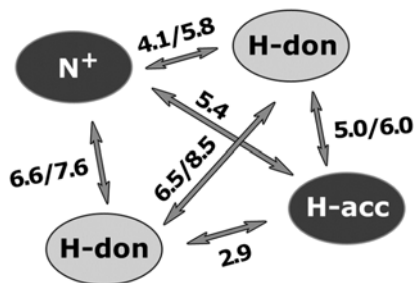


Figure 13. Pharmacophore definition for a set of compounds.

14-rinovirus. We would certainly be tempted to align the molecules as shows in the example below (Figure 15).

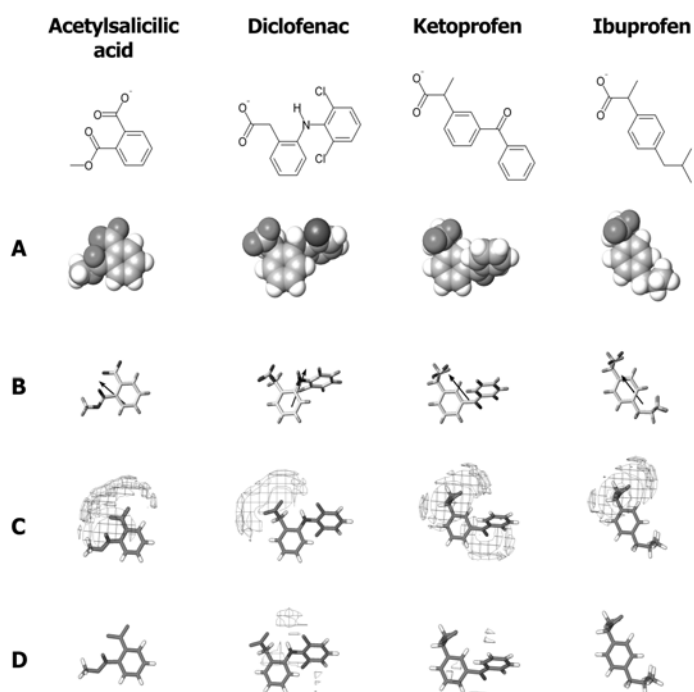


Figure 14. Surface representations of some aromatic molecules.

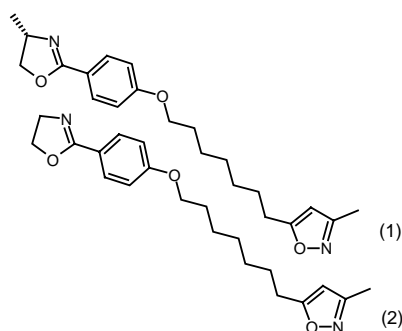


Figure 15. Putative alignment of two WIN molecules.

Provided we know the inhibitory concentrations of the two molecules, it would be resilient to explain that their values are 30 and 600 nM, respectively. However, when looking at their van der Waals surfaces we promptly realize why these two very similar molecules have 20-fold difference in the inhibitory concentrations (Figure 16).

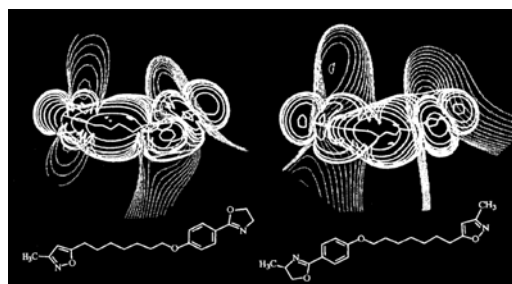


Figure 16. Isopotential surfaces of the two WIN molecules. Reprinted with permission from Prof. Dr. Gerhard Klebe. "Structural Alignment of Molecules. In 3D QSAR in Drug Design. Theory, Methods, and Applications" Escom: Leiden, 1993, pp. 173-199. Copyright Springer. Permission is also granted from Springer Science and Business.

The two molecules enter the receptor site in their reverse sideways. The superposition of their van der Waals surfaces shows clearly how the recognition patterns are seen in the intermolecular complementary interactions. This is due to the high level of similarity the two surfaces share when occupying the receptor site.

When the 3D structure of the receptor is known all molecules can be docked into the receptor site. By doing this it is possible not only to address the alignment task but it is also possible to get information on the bioactive conformations. Nevertheless, most of the available procedures accounts for flexibility of the ligands only. This means that the receptor site does not change during the docking procedure, thus avoiding any flexible side chain to be adjusted to the conformations these ligands can get when binding to the site. In such cases, there is also the need for carrying out some molecular dynamics to better describe the flexibility of ligands and targets. For most of the purposes in 3D QSAR this is not an easy task because it is very time consuming. Nonetheless, it has to be assumed a compromise in order to allow for the most acceptable 3D structures to be used in the next steps of the 3D QSAR analysis.

When appropriate charges, conformations and alignment are settled, all molecules in the set have to be put inside a grid box (Figure 17). The grid box can have different coordinate lattice points, but to start with 2 Å is the default and most suitable one. It has to be pointed out the need for adjusting all molecules inside the box. Carefully moving the lattice along the X, Y and Z axes make any artifacts of the grid box location to arise and thus take into consideration the positions where no distance dependence is found. So, the positioning of the molecules in the box has also to be tested by focusing the regions when necessary. Some methods like GRIND (GRid-INdependent Descriptors) [39], a relatively new generation of alignment independent 3D-molecular descriptors, do not need this because the fields are aligned independently of the original grid positions.

A virtual chemical probe has to be selected. This probe will have to have as a feature three different kinds of intermolecular interaction forces. Since this is not always feasible, different probes with certain features will be used instead. It can be a single or a multiple atom probe, depending on the feature it will pertain. Nonetheless, the probes will have to be able to describe interactions within three major forces: electrostatic, steric and hydrophobic. The resulting moving around the aligned molecules in the grid is related to favorable and unfavorable energy interactions, which are then set into the

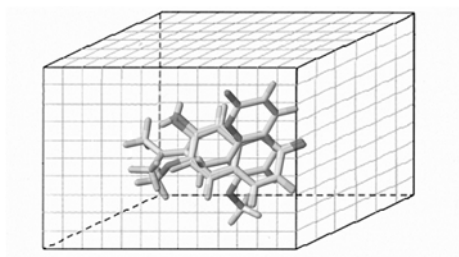


Figure 17. Molecules inside the grid box.

QSAR table. The QSAR table is now read to be analyzed. Since it has thousands of variables and the X-matrix contains much less molecules than variables, the use of multiple linear regressions (MLR) is not appropriate. Instead, a PLS analysis [40] has to be carried out on them.

In PLS regression [41], the X block of independent variables (descriptors) is correlated with the Y vector (activity) in such a way that the projected coordinates T are good predictors of Y. In this case, the biological activities are included in the decomposition procedure. The goal of PLS regression is to build a model that will be able to predict Y from X and to describe their common structure. The best model obtained should be useful to calculate reliable predictions of Y values for new molecules, not included in the model. PLS decomposes the X-matrix as the product of two smaller matrices, P and T. The main difference is that PCA obtains PCs that represent at best the structure of the X-matrix, while PLS obtains the latent variables (LVs) that represent the structure of the X and Y matrices, maximizing the fitting between the Xs and Ys. Selecting the correct number of LVs (dimensionality) and testing the predictive power of the model is of critical importance in PLS. The predictive ability of a regression model is evaluated usually through an internal cross-validation (CV). This procedure creates reduced models (where some of the objects are removed), and are used to predict the Y variables of the objects held out. The regression model is usually validated through an internal leave-one-out (ideally, leave-two-out or leave-many-out) cross-validation procedure, a useful tool to verify its ability for future predictions.

In general, the parameters used to assess the statistical quality of the model are the correlation coefficient (R^2) and the cross-validated correlation coefficient Q^2 (r_{cv}^2). The CV is a valuable technique because it performs an internal validation of the model and obtains an estimation of the predictive ability without the help of external data sets. The best way to examine the information from PLS is to plot the matrices obtained. The 2D T-U score plots represent objects in the space of X-scores (T) against the Y-scores (U). It gives an idea of the correlation between the Xs and the Ys obtained in the model for each one of the LVs.

The final model should have a correlation coefficient ($R^2 > 0.6$) and cross-validated correlation coefficient ($Q^2 > 0.5$) 'to ring the bell'. As a matter of fact, the higher the correlations are the better the model is. On the other hand, there are some other ways of checking out the statistical validity of the model. The values of Q^2 have to be evaluated with care since they may be overestimated depending, for instance, on the way training and test sets are defined.

At this stage “scrambling” of data analyses has to be envisaged to bring about the strength of the model [42]. To predict the biological response (BR) of a new ever synthesized compound, the created model must have mutually independent variables and BR be represented by a single normal distribution spread evenly across the range of responses. The results are reliable when the covariances between and among descriptor variables are negligible. If these are not met, as is most of the time, principal component regression (PCR), partial least square (PLS) or artificial neural networks (ANN) are used as alternative approaches.

In the progressive scrambling procedure, biological response can be sorted out into a new matrix table let us say by starting from the most potent molecules towards the least one. Groups of potencies are then created and scrambled within each one. New fit (R^{*2}) and prediction (Q^{*2}) are obtained and evaluated against the original ones. If good Q^{*2} is still obtained, the QSAR model has no reliable robustness and it is simply fitting noise. But, be aware that the inverse result is not solely a guarantee of the robustness of the Q^{*2} !

The number of components (NoC) to be in the final model has to be assessed by graphing R^2 and Q^2 against NoC, and thus helping to avoid data overfitting (Figure 18). The chosen NoC will be the one with parallels R^2 and Q^2 in the ‘plateau’ that represents the stability of these two fit and prediction tools. In the example, a maximum number of components should be equal to three, because Q^2 value decreases after the third LV.

The 3D QSAR equation is not actually the one we look for in the CoMFA method. Instead we see it through the molecular interaction fields allocated along with the molecule as the information from regression model displayed in the 3D structure space. They maybe depicted as steric CoMFA contour map or electrostatic CoMFA contour map. For instance, the maps for steroid binding to testosterone-binding globulin (TBG) [43] can be depicted as follows. The codes indicate regions where (a) electronegative substituents enhance (1) or reduce (2) the binding affinity and (b) regions where substitution enhances (3) or reduces (4) the binding affinity (Figure 19).

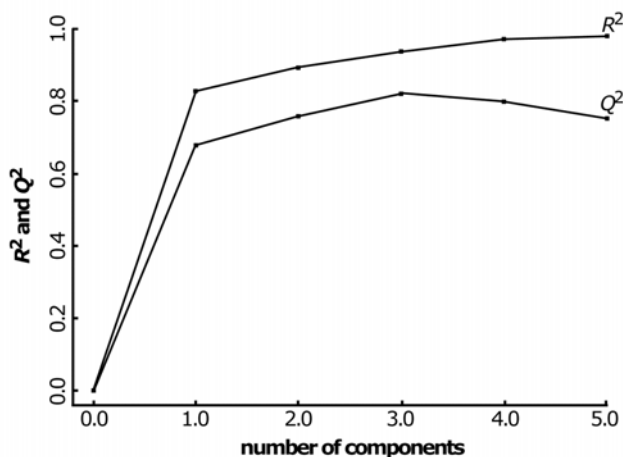


Figure 18. Selection of the number of components based on the R^2 and Q^2 plots.

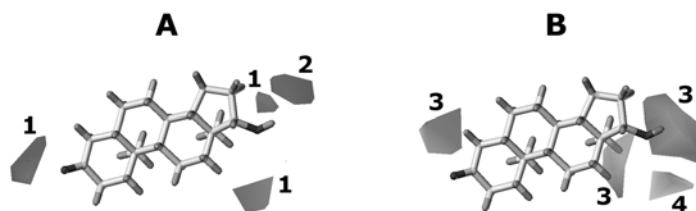


Figure 19. Molecular fields for a steroid compound (a) electrostatic and (b) steric.

CoMSIA

The van der Waals potential as applied through the Lennard-Jones function and the Coulomb potential are used in the CoMFA method according to the Morse diagram, whereas the comparative molecular similarity indices analysis (CoMSIA) [44,45], the CoMFA version for similarity fields, uses a Gaussian approximation for them. The important advantage that rises from this is that the cut off values can be avoided by CoMSIA method. Here, probes are used to calculate their 'similarity indices' to the investigated molecules in the different grid points.

When applying CoMFA, the resulting contour maps represent regions apart from the molecules where interactions with probes take place. However, in the CoMSIA maps the regions are within the area occupied by the ligand skeletons. This is so because the fields are based on similarity indices of drug molecules that bear a common alignment. As a result the spatial similarity indices unveil more significant power when designing novel compounds. For instance, we have recently shown that the CoMSIA fields for non-steroidal ligands that interact with the estrogen receptor (ER) have much higher predictive power ($Q^2 = 0.835$) as compared to CoMFA results ($Q^2 = 0.584$) [46].

Molecular interaction fields (MIF)

Molecular interaction fields can be calculated for molecules with known 3D structures. MIF are used to describe the interaction energy between the molecules and virtual chemical probes in the surrounding volumes. The regions depicted in the MIF can be used to indicate the energy of interaction: for large positive regions the probe is repelled, while for negative ones the probe depicts a favorable region to where ligands or fragments maybe placed to exploit the design of new ligands.

Molecular interaction fields can be calculated either for 3D structures of ligands and targets. In the latter, the *de novo* design or docking in the binding site are used in order to design ligands.

In 3D QSAR, MIF for different compounds are generated and then analyzed statically. If the resulting statistical model is sound, the MIF serve as a hint to the common shared field molecules when interacting with the target. In such cases, the MIF can be used in the optimization process of the candidate drug molecule.

This section is dedicated to the GRID method [47,48] of calculating MIF. By applying proper chemometric tools to the resulting MIF calculations, the 3D QSAR is analyzed in a similar way to that of CoMFA, for instance. Actually, the GRID MIF can easily be incorporated into the CoMFA tables in order to include the MIF displayed by many different kinds of single or multi-atom probes. Moreover, MIF as calculated by

GRID method are also appropriate to be used when the 3D structure of the target is known. When the MIF are calculated for the target macromolecule itself, the grid box may be chosen to cover the whole receptor or just the receptor site. The specific binding sites are then located within the receptor site, which are used to identify regions of influence to the potency.

Forces that we use to describe MIF were shown above. Out of all of them the hydrogen bonding calculations in the GRID method are quite accurate to the level of finding water molecules that may 'co-crystallize' with the target when drug-receptor interactions are to be studied by X-ray crystallography. This is also possible to find out water molecules that may bridge the interactions between the probe and the target. On the other hand, the GRID method can be used by explicitly considering such water molecules without having to be representative of the continuum of the solvent as in the electrostatic interactions normally calculated in many other 3D QSAR studies.

The GRID force field introduced by Peter Goodford more than 15 years ago [49] contains a term to model hydrogen bonds with three components: the energy dependent on the separation between target and probe atoms and the energies on the hydrogen bond angles between the target and probe. The chemical nature of the hydrogen-bonding atoms is also considered as for the Morse function. Different functional forms are used in the GRID method to allow the rotation of atoms to form optimally oriented hydrogen bonds since these are directional noncovalent interactions. For each calculation with GRID not only hydrogen atoms and lone-pairs of electrons are allowed to move around, but amino acid side chains are also allowed to move in order to identify better probe positioning. This is great because it includes the flexibility of the target when MIF are calculated. So far, the regions depicted by MIF are representative of the whole space the site may allocate for ligand to be placed in.

Molecular interaction fields maybe used for the design of ligands binding to a target macromolecule whose 3D structure is known. *De novo* design is useful when there is no prior information of what compounds bind to the target. By allocating the energy minima for a certain probe, the position maybe used to place a ligand's fragment. Modification of the lead compound is also possible. Placing substituents in the scaffold of a ligand may position them near the energy minima found in the target. Docking of a particular fragment that may represent any part of a ligand or a set of ligands when both 3D structures of ligand and target are known is another way of using GRID. GRID can also be used for comparison purposes among a set of ligands and targets when selectivity is a parameter to be considered in the search for selective ligands. Pharmacophore detection and molecular superposition of ligands in deriving SAR find application from MIF calculations using the GRID method. Since MIF are dependent on the conformation of the target molecule (ligand or macromolecule), the interactions maybe used to compare binding energies. However, MIF can be less sensitive to conformations when appropriate transformations can be carried out to yield descriptor variables of pharmacokinetic interest, like, for instance, in the VolSurf method [50,51].

Examples of 3D QSAR models

Molecular modeling and chemometrics, all included in the field of cheminformatics, represent the basis of 3D QSAR. By joining together such tools the robustness of the desired models maybe accomplished.

The example that will be shown step-by-step is related to the search of rigid analogues of the drug pentamidine, which is used for the treatment of different kinds of leishmaniasis.

The antimicrobial activity of aromatic bisamidines against a widespread range of microorganisms, like *Acanthamoeba*, *Babesia canis*, *Crithidia fasciculata*, *Cryptosporium parvum*, *Giardia lamblia*, *Leishmania spp.*, *Plasmodium spp.*, *Toxoplasma gondii* and *Trypanosoma spp.* is well known [52]. In the case of visceral leishmaniasis, pentamidine isethionate (Pentacarinat®-Rhodia), which belongs to this class of compounds, has been used clinically to treat patients affected by *Pneumocystis carinii* pneumonia (PCP), occurring in ca. 80% of the HIV-infected individuals, as well as antimony resistant leishmaniasis patients. Side effects are expected along the treatment, like abscesses at the site of intramuscular injection, severe hypotension, leukopenia, thrombocytopenia, hypocalcaemia, hepatic, renal and pancreatic complications. In case of leishmaniasis, if the treatment is kept for too long, more serious adverse effects may occur, such as hypoglycemia and acute pancreatitis, leading to diabetes. Hence, pentamidine is second-line in the treatment of PCP, although it is as effective as the combination of trimethoprim and sulfamethoxazole (TMP-SMZ). Leishmaniasis and HIV virus co-infection has drawn a new profile for this disease in the last years.

It has been suggested that aromatic bisamidines bind specifically through a noncovalent and nonintercalative mode to AT-rich regions of B-DNA. The ligand adopts an 'isohelical' conformation to fit the minor-groove [53], in which both the aromatic rings stand in a 'twisted' position, following the helix curvature, as illustrated below for pentamidine, Figure 20.



Figure 20. Mode of binding (MOB) of pentamidine in the DNA A2T2 region sequence.

The structure-based ligand design emphasizes the discovery of new lead compounds using the 3D structure of the receptor, towards a continuous increase in the ligand-receptor complementarities. A comparison based on molecular mechanics (MM) energy calculations is made among classical minor-groove binders (MGBs) and some rigid analogues, in order to rank them in their self-complementarity and binding to d(CGCGAATTCGCG)₂ DNA sequence also known as Dickerson Drew Dodecamer, (DDD). By integrating both structure based and ligand based drug designs some rational bases for the proposal of new chemical substances that bind to DNA can be fulfilled.

We start with the classical QSAR equation of the antileishmaniasis activity (IC_{50}) for a set of 37 pentamidine analogues acting against *Leishmania mexicana amazonensis* whose general structures are depicted as follows [54], Figure 21.

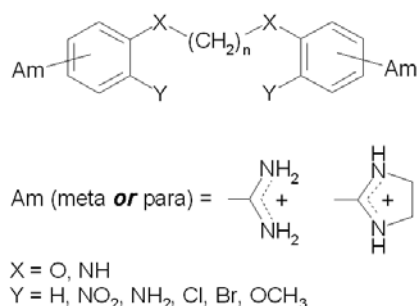


Figure 21. Chemical structures of a series of aromatic amidines.

In order to derive the QSAR equation below, we applied the combined Hansch-Fujita and Free-Wilson method for a set of descriptors calculated for the 3D structures of all pentamidine analogues. Out of some QSAR, the best model fitting the biological response includes indicative variables (I) and the partitioning (P), along with the minimum volume molecules disclose in terms of their shape.

$$\begin{aligned}
 \text{Log } 1/IC_{50} = & 0.528(\pm 0.097)I_{p-Am} - 0.182(\pm 0.059)I_{o-subst} \\
 & + 0.330(\pm 0.091)I_{NH} - 1.026(\pm 0.275)I_{n=2} - 0.314(\pm 0.057)I_{n=4} \\
 & + 0.230(\pm 0.117)I_{n=6} + 0.188(\pm 0.056)\log P \\
 & + 0.004(\pm 0.002)V_{min} - 1.467(\pm 0.380) \\
 (n = 37, R^2 = 0.921, s = 0.170, Q^2 = 0.623)
 \end{aligned}$$

Equation 3. Classical QSAR for a set of pentamidine analogues.

From this QSAR equation we can clearly see that the amidine group is to be allocated at the *para*-position of the phenyl ring to be the one to increase the IC_{50} values. When the group was placed at the *meta*-position a detrimental contribution to IC_{50} resulted in the diminution of the potency. A bioisosteric replacement of -O- to -NH- increases potency. The homologation up till four methylene groups is also detrimental to activity but, when the number of (CH_2) groups is six, there is a positive effect and the potency increases again. Partitioning seems to be of valuable importance to describe the potencies. We have reasoned it as to the fact that since the inhibitory concentration is being measured in cell systems we would think of it as being due to the crossing of membranes such molecules have to go through in order to reach the target within the cell.

The statistics of the above QSAR equation seem quite good for the kind of measurement that was carried out. However, despite the fact that the fit curve resembles linearity, the predictive power of the model just ring the bell (ca. 62%). Nevertheless, it seems to be a bit poorer than the one we would expect. On the other hand, as thoroughly

described in this section, the variable descriptors selected in the QSAR suggest there is good chance that we could use them to rather gather information on the drug-receptor interactions. The amidine group at the *para*-position, not *meta*, the length of the bridge linking the two amidine groups and the bioisosteric replacement, all might be associated to the binding of pentamidine analogues to the B-DNA. In order to verify on that, we carried out a conformational analysis in all study molecules. To accomplish with this, we retrieved the X-ray crystallographic structure of the pentamidine from the Cambridge Crystallographic Data Centre (CCDC) and the one from the PDB as co-complexed to B-DNA.

Pentamidine itself bears an extended conformation whereas when bound it holds the B-DNA isohelical conformation.

All penamidine analogues including it are very flexible molecules. The combinatorial complexity of such a problem can easily be felt by calculating the number of conformations needed to describe the whole conformational space of these molecules. The number of van der Waals comparisons, V , for doing that is $V = (360/A)^T \times N(N-1)/2$, where N is the number of atoms of a molecule with T torsional degrees of freedom (rotatable bonds). For each torsional degree of freedom, T , explored at a given angular increment in degrees, A , there are $360/A$ values to be examined for each T [55], that is for checking steric clashes. If we take the drug pentamidine ($N = 51$, $T = 8$, $N(N-1)/2 = 1,275$) as an example in the set, we would have to check $V = 921.10^{15}$ different conformations to sample the desired conformational space if we were to carry out a proper analysis by using increments of 5° ! Of course, this is an assumption that is out of question for many QSAR studies.

An easy solution for that is to hypothesize that the binding of pentamidine to B-DNA is the mode of action for the antileishmaniasis activity of these amidines. If that is so, we have just to generate 37×2 conformations; one due to the extended and other to the isohelical conformations. In doing that we have the opportunity to jointly derive a QSAR model based on the ligands and target structures (putting together ligand based drug design (LBDD) and structure based drug design (SBDD)).

Pruning the combinatorial tree to work for the slightest conformation number we face another problem related to the existence of only two out of the 74 conformations needed to carry out the analysis. One way of dealing with this problem is to generate the rest of the conformations by similarity calculations. When two molecules A and B are superposed in some defined orientation their similarities can be obtained for a set of n points. So, we have done that for all 37 molecules in the two different conformations and generated two sets of 37 conformations for each member of the series in the extended and isohelical conformations. We carried out the quantum-chemical similarity calculation suggested by Carbo [56,57], which is based on the electron densities ρ_A and ρ_B , derived from the wave functions of molecules A and B, as fully implemented in the 3D TSAR software (Accelrys, Inc.), whose solutions are presented below (Figure 22).

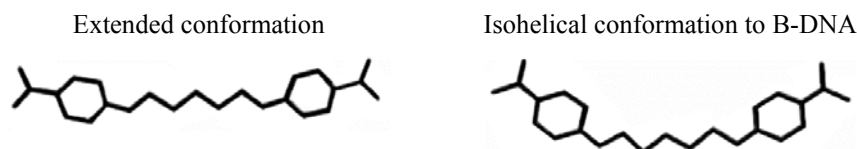


Figure 22. Pentamidine conformations used to calculate the descriptors.

The best QSAR model arose with the following variable descriptors (Equation 4).

$$\begin{aligned} \text{Log } 1/\text{IC}_{50} = & 9.543(\pm 2.722)\text{Carbo}_{\text{isohelical}} + 0.165(\pm 0.031)\text{Clog } P \\ & - 0.145(\pm 0.070)S_{(i)} - 9.220(\pm 2.310) \\ (n = 37, R^2 = 0.685, s = 0.317, Q^2 = 0.513) \end{aligned}$$

Equation 4. Relationship between the isohelicity and the potency.

The inhibitory concentration is now described only by three variables where the Carbo similarity index for isohelical conformation to B-DNA is the one to be of most favorable increase to the IC_{50} values. No similarity index for extended conformations appeared to be of importance in the equation above. Based on this, we have postulated that pentamidine analogues ‘travel’ through cell membranes towards the inn-cells to reach the B-DNA receptor. By interacting in the minor grooves of the B-DNA the drug molecules preclude it of being of any further utility for the cell surviving and the parasite is killed.

In the above equation there is another descriptor variable that is linked to the binding of pentamidine analogues do B-DNA. This is the electrotopological index ($S_{(i)}$) [58,59]. It accounts for the electronegativity of the bioisosteric replacement of -O- to -NH- by considering the topology of the group atoms in the desired conformation. The values of $S_{(i)}$ for pentamidine is 5.657 and 3.372 for its -NH- bioisoster, which is the one most potent in the whole series. This means that bioisosteric change increases the potency by allowing further hydrogen-bonding to take place when the molecule is seated in the A2T2 grooves of the B-DNA.

A closer inspection on the statistics of the above equation reveals, however, that albeit the prediction power and the linear fit are appropriate for this relationship, their values would also just ring the bell besides the whole evidences for binding to the receptor target.

The equation was raised from molecular modeling and chemometrics that constitute the basis of the 3D QSAR model depicted. In this case, theoretical descriptors of a series of molecules were correlated to inhibitory concentrations of *L. mexicana* yielding the appropriate probes for the biological responses measured so far. Because of our expertise in problem solving by chemometric methods, we have reasoned the 3D QSAR equation above would have only to be re-written by using another chemometric tool instead of the multiple regression analysis carried out in the data. Out of the three variables descriptors used to derive the QSAR model, partitioning is the one regarded to behave non-linearly. If this would suggest that the relationship between structure and activity is nonlinear, the use of artificial neural networks (ANN) would be sought as the multivariate nonlinear chemometric tool for addressing the problem of the depicted model. Moreover, from the application of such method we would prevail to the QSAR model of being obtained either with a poor statistics or the variable descriptors would be selected just by chance. We then applied the ANN as proof of consistency regarding the correct use of the three variable descriptors as being the ones needed to describe the antileishmaniasis activity of all pentamidine analogues [60]. The configuration used is shown below (Figure 23).

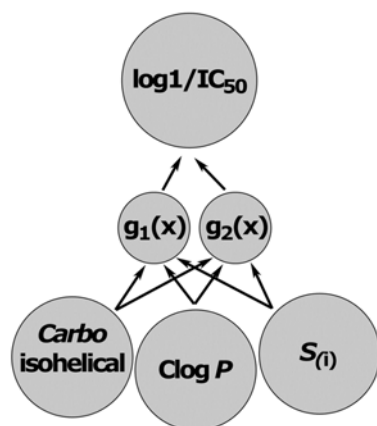


Figure 23. Organization of the artificial neural network.

The input variables were the three selected for obtaining the QSAR model discussed above, and the output variables were the IC_{50} values. Then we raised the relationship between calculated and experimental values to yield a very good correlation between the two to ascertain the high predictive power of the variable descriptors (X) in disclosing the appropriate IC_{50} values (Y), Equation 5.

$$\log 1/IC_{50(\text{predicted})} = 1.018(\pm 0.068) \log 1/IC_{50(\text{experimental})} + 0.029(\pm 0.049)$$

$$(n = 37, R^2 = 0.849, s = 0.210, Q^2 = 0.818)$$

Equation 5. QSAR obtained by applying the ANN.

The resulting ANN analysis of the 37 bisamidines gives rise to model explaining 85% of the variation of Y. The significance of the model was tested by cross-validation, which gave a Q^2 of 0.82 (the predictive R^2). In the context of 3D QSAR, the use of ANN for modeling and data analysis is clearly showing that the selected descriptor variables can be considered as good ones to describe the biological responses. Moreover, it is feasible to make the relationships with the pentamidine analogues binding to the B-DNA by using its isohelical conformation.

If the above assumptions are correctly made we would expect that by synthesizing pentamidine analogues with isohelical conformations constrained, as rigid counterparts, the binding to B-DNA would be tighter. So, in order to validate this hypothesis we docked some pentamidine rigid analogues into the A2T2 B-DNA to verify that this is just the case. That is, rigid analogues that bear the isohelical conformations to B-DNA bind tighter to it [61,62].

Next, we carried out an external validation of the model by deriving the 3D QSAR using the molecular interaction fields (MIF) of a set of classical minor-groove binders (MGB) in order to rank them through their complementarity to the Dickerson Drew

Dodecamer (DDD) according to their interaction energies with B-DNA. The comparative molecular field analysis (CoMFA +1 sp^3 carbon probe) contour map to the MGB HD33258 molecule [63] illustrates the importance of relatively bulky positively charged groups attached to the aromatic amidine moieties (1), and the negatively charged ones (2) into the middle chain (Figure 24).

The two charged benzamidine ends portray the needed recognition for the A2T2 minor groove binding whereas in the middle chain we can see the necessity for opposite electrostatic interactions as fully pictured in the grooves of B-DNA, and represented below as such for pentamidine (Figure 25).

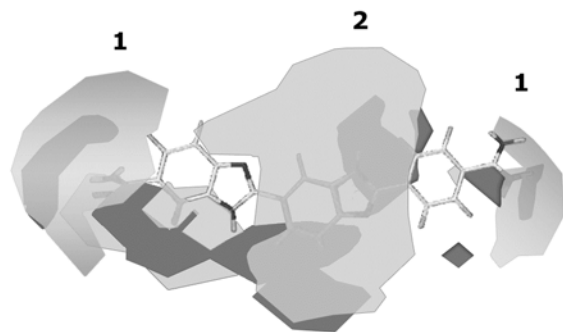


Figure 24. Molecular interaction fields to the molecule MGB HD33258.

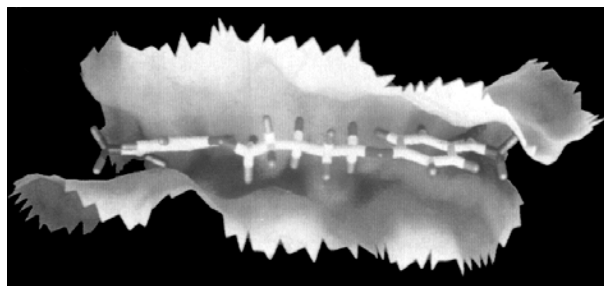


Figure 25. Representation of pentamidine inside the A2T2 B-DNA.

The GRID MIF can be used to identify the selectivity of the set between the minor against the major grooves of the B-DNA. In order to do so the aromatic amidine probe available in the GRID program was screened against the DDD as obtained from the PDB. In the first run (Figure 26A) we deleted the pentamidine drug and the water molecules while in the second one (Figure 26B) only the drug molecule was deleted. The results are pictured below (Figure 26).

The aromatic amidine recognition sites for A2T2 $\text{d}(\text{CGCGAATTCGCG})_2$ with and without water, at $-16 \text{ Kcal mol}^{-1}$, are clearly found in the minor grooves of the B-DNA. These energy minima encompass all the previous findings and assure the favorable interaction site where such groups should be placed in the scaffold molecule

to be of any importance for binding within the AT-rich region of B-DNA minor groove.

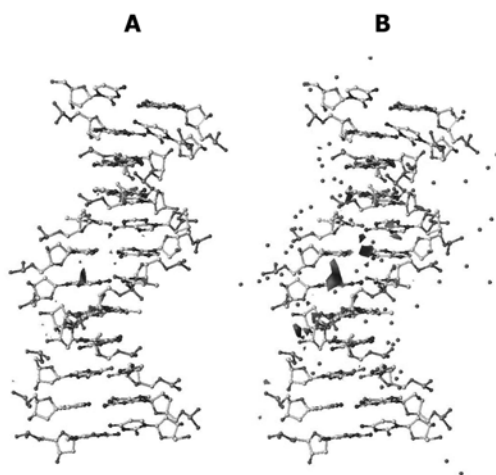


Figure 26. Molecular interaction fields from the aromatic amidine probe in the A2T2 region.

Conclusions

Three-dimensional quantitative structure-activity relationships (3D QSAR) represent the key milestone in medicinal chemistry to define the structure-activity relationships (SAR) properly. The CoMFA models can rationalize SAR by adding information regarding the molecular interaction fields (MIF) obtained from 3D structures of small molecules. It can be applied when no 3D structural information of the target is known. However, if the 3D structure of the receptor target is known, CoMFA cannot only simply indicate where fragments should be placed in the small molecule scaffold but it is also amenable to confirm and rationalize both descriptions (LBDD and SBDD) to be in agreement. In such a case, the robustness of the CoMFA model maybe of invaluable importance in the drug design arena. In this situation, the conclusion we may draw to describe the nature of the receptor structure represents the output we would like to see from quantitative. Nevertheless, it cannot be thoroughly extrapolated! On the contrary, it has to be applied in a finite set of molecules, but with enough detail to entail the agreements we are seeking in the design of new candidate drug molecules. If that is so, the pharmacophoric model can be sorted out. Of course, when the 3D structure of the receptor is known life becomes easier and the models more robust. This is so because it is easier to draw the 'correct' conformation and also to allow for proper superposition of molecules inside the grid box. Actually, it seems quite clear from recent literature search that when it is applied as *in vitro* assay the results are seemingly much more consistent with its foreground theory. Nonetheless, different applications for CoMFA models can be traced back from the search and these include lots of studies where no information about the mode of action or even mode of binding is known! Although such results maybe taken cautiously, it is our thinking that 3D QSAR methods like CoMFA have a place in the field of drug discovery where 3D molecular structures are being taken into

consideration. With the advent of pharmacogenomics, where all of target macromolecules with therapeutically interest might be available, these methods should be of even more relevance in finding robust structure-activity relationships. Overall, if the 3D QSAR model can be used for 3D search in databases we would prefer to test molecules that fit to such model firstly. If we can manage to include GRID molecular interaction fields into the CoMFA study this concluding result can be even more enthusiastic. In the cheminformatics era joining together CAMD tools is of utmost importance!

Acknowledgments

We gratefully acknowledge financial support from CNPq, CAPES and FAPESP, Brazil.

References

1. Keseru, G. M.; Makara, G. M. *Drug Discov. Today*, **2006**, v. 11, p. 741.
2. Kramer, R.; Cohen, D. *Nature Rev. Drug Discov.*, **2004**, v. 3, p. 965.
3. Belfield, G. P.; Delaney, S. J. *Biochem. Soc. Trans.*, **2006**, v. 34, p. 313.
4. Vanbogelen, R. A.; Beushausen, S. J. *Proteome Res.*, **2006**, v. 5, p. 1819.
5. Thompson, R. B. *FASEB J.*, **2001**, v. 15, p. 1671.
6. Trigg, D. J. *Drug Develop. Res.*, **2003**, v. 59, p. 269.
7. Istvan, E. S.; Deisenhofer, J. *Science*, **2001**, v. 292, p. 1160.
8. Balunas, M. J.; Douglas Kinghorn, A. *Life Sci.*, **2005**, v. 78, p. 431.
9. Mang, C.; Jakupovic, S.; Schunk, S.; Ambrosi, H.-D.; Schwarz, O.; Jakupovic, J. *J. Comb. Chem.*, **2006**, v. 8, p. 268.
10. van de Waterbeemd, H.; Gifford, E. *Nature Rev. Drug Discov.*, **2003**, v. 2, p. 192.
11. Huwe, C. M. *Drug Discov. Today*, **2006**, v. 11, p. 763.
12. Haggarty, S. J. *Curr. Opin. Chem. Biol.*, **2005**, v. 9, p. 296.
13. Langer, T.; Wolber, G. *Drug Discov. Today Technol.*, **2004**, v. 1, p. 203.
14. Kuntz, I. D. *Science*, **1992**, v. 257, p. 1078.
15. Henry, C. M. *CENEAR*, **2001**, v. 79, p. 69.
16. Klebe, G. *Drug Discov. Today*, **2006**, v. 11, p. 580.
17. Limbird, L. E.; *Cell Surface Receptors: A Short Course on Theory and Methods*, Springer-Verlag: New York, 2004.
18. Lipinski, C. A. *Drug Discov. Today Technol.*, **2004**, v. 1, p. 337.
19. Vieth, M.; Siegel, M. G.; Higgs, R. E.; Watson, I. A.; Robertson, D. H.; Savin, K. A.; Durst, G. L.; Hippskind, P. A. *J. Med. Chem.*, **2004**, v. 47, p. 224.
20. Vieth, M.; Sutherland, J. J. *J. Med. Chem.*, **2006**, v. 49, p. 3451.
21. Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. *Drug Discov. Today*, **2003**, v. 8, p. 876.
22. Rees, D. C.; Congreve, M.; Murray, C. W.; Carr, R. *Nature Rev. Drug Discov.*, **2004**, v. 3, p. 660.
23. Carr, R. A. E.; Congreve, M.; Murray, C. W.; Rees, D. C.; *Drug Discov. Today*, **2005**, v. 10, p. 987.
24. Ringe, D.; Mattos, C. *Med. Res. Rev.*, **1999**, v. 19, p. 321.
25. Gohlke, H.; Klebe, G. *Angew. Chem. Int. Ed.*, **2002**, v. 41, p. 2645.
26. Müller, G. *Drug Discov. Today*, **2003**, v. 8, p. 681.
27. Hugo Kubinyi's drug design lectures: <http://www.kubinyi.de/>
28. Buffon, G. Editor's note concerning a lecture given 1733 by Mr. Le Clerc de Buffon to the Royal Academy of Sciences in Paris. *Histoire de l'Acad. Roy. des Sci.*, **1733**, p. 43.
29. Andrews, P. R.; Craik, D. J.; Martin, J. L. *J. Med. Chem.*, **1984**, v. 27, p. 1648.

30. Böhm, H.-J. *J. Comput.-Aided Mol. Des.*, **1994**, v. 8, p. 623.
31. Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. *Proc. Natl. Acad. Sci. USA*, **1999**, v. 96, p. 9997.
32. Hopkins, A. L.; Groom, C. R.; Alex, A. *Drug Discov. Today*, **2004**, v. 9, p. 430.
33. Abad-Zapatero, C.; Metz, J. T. *Drug Discov. Today*, **2005**, v. 10, p. 464.
34. Xu, J.; Hagler, A. *Molecules*, **2002**, v. 7, p. 566.
35. Wermuth, C. G. *Drug Discov. Today*, **2006**, v. 11, p. 348.
36. Tute, M. S. Pfizer, Inc. and University of Kent at Canterbury. Personal communication, 1995.
37. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.*, **2000**, v. 28, p. 235.
38. Alonso, H.; Bliznyuk, A. A.; Gready, J. E. *Med. Res. Rev.*, **2006**, v. 26, p. 531.
39. Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. *J. Med. Chem.*, **2000**, v. 43, p. 3233.
40. Baroni, M.; Clementi, S.; Cruciani, G.; Costantino, G.; Riganelli, D.; Oberrauch, E. *J. Chemometrics*, **1992**, v. 6, p. 347.
41. Montanari, M. L. C.; Andricopulo, A. D.; Montanari, C. A.; *Thermochim. Acta* **2004**, v. 417, p. 283.
42. Clark, R. D.; Fox, P. C. *J. Comput.-Aided Mol. Des.*, **2004**, v. 18, p. 563.
43. Cramer III, R. D.; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.*, **1988**, v. 110, p. 5959.
44. Klebe, G. Comparative Molecular Similarity Indices Analysis: CoMSIA. In *3D QSAR in Drug Design*, Vol. 3, Kluwer Academic Publishers: Dordrecht, 1998, pp. 87-104.
45. Klebe, G.; Abraham, U.; Mietzner, T. *J. Med. Chem.*, **1994**, v. 37, p. 4130.
46. Menezes, I. R. A.; Leitão, A.; Montanari, C. A. *Steroids* **2006**, v. 71, p. 417.
47. Wade, R. C.; Goodford, P. J. *J. Med. Chem.*, **1993**, v. 36, p. 148.
48. Wade, R. C.; Clark, K. J.; Goodford, P. J. *J. Med. Chem.*, **1993**, v. 36, p. 140.
49. Reynolds, C. A.; Wade, R. C.; Goodford, P. J. *J. Mol. Graph.*, **1989**, v. 7, p. 103.
50. Cruciani, G.; Pastor, M.; Guba, W. *Eur. J. Pharm. Sci.*, **2000**, v. 11 Suppl. 2, p. S29.
51. Cruciani, G.; Crivori, P.; Carrupt, P. A.; Testa, B. *J. Mol. Struct. (THEOCHEM)*, **2000**, v. 503, p. 17.
52. Oliveira, A. M.; Custódio, F. B.; Donnici, C. L.; Montanari, C. A. *Eur. J. Med. Chem.*, **2003**, v. 38, p. 141.
53. (a) Edwards, K. J.; Jenkins, T. C.; Neidle, S. *Biochemistry*, **1992**, v. 31, p. 7104. (b) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E.; *Nucleic Acids Res.*, **2000**, v. 28, p. 235. (c) PDB structure number: 1D64.
54. Montanari, C. A.; Tute, M. S.; Beezer, A. E.; Mitchell, J. J. *Comput.-Aided Mol. Des.*, **1996**, v. 10, p. 67.
55. Marshall, G. R. Binding-Site Modeling of Unknown Receptors. In *3D QSAR in Drug Design. Theory, Methodos, and Applications*, Escom: Leiden, 1993, pp. 80-116.
56. Carbo, R.; Leyda, L.; Arnau, M. *Int. J. Quantum Chem.*, **1980**, v. 17, p. 1185.
57. Carbo, R.; Domingo, L. *Int. J. Quantum Chem.*, **1987**, v. 32, p. 517.
58. Kier, L. B.; Hall, L. H. *Pharm. Res.*, **1990**, v. 7, p. 801.
59. Hall, L. H.; Mohnney, B.; Kier, L. B. *J. Chem. Inf. Comput. Sci.*, **1991**, v. 31, p. 76.
60. Montanari, C. A.; Tute, M. S. *Quant. Struct.-Act. Relat.*, **1997**, v. 16, p. 480.
61. Montanari, C. A.; Trent, J. O.; Jenkins, T. C. *J. Braz. Chem. Soc.*, **1998**, v. 9, p. 175.
62. Menezes, F. A. S.; Montanari, C. A.; Bruns, R. E. *J. Braz. Chem. Soc.*, **2000**, v. 11, p. 393.
63. Willis, B.; Arya, D. P. *Biochemistry*, **2006**, v. 45, p. 10217.