Fabien Fontaine
Computer-assisted drug design laboratory
Research Group on Biomedical Informatics
Pompeu Fabra University

DEVELOPMENT AND APPLICATIONS OF NEW 3D MOLECULAR DESCRIPTORS

Barcelona, 2004

Ph.D. thesis directed by Manuel Pastor
Pompeu Fabra University

Dipòsit legal: B.13422-2005
ISBN: 84-689-1255-7

Acknowledgments

During this last year of Ph.D. I have faced the most difficult adversity of my life. I really thought that I would never be able to write this thesis. That is why I would like to thank all the people who have made me feel better during this year.

I would like to thank Manuel Pastor and Ferran Sanz for giving me the opportunity to carry out my Ph.D. in the GRIB. Manolo, thank you for giving me the freedom to develop my own work. Thank you also for understanding so well my health problem and for giving me all the support I needed in these difficult moments.

Thank you to all the lab. Ramon, Cristina, Lulla, Josep, Bet and Esteban, we had so much fun in Zaragoza. Maybe we will be able to repeat that again. Robert, you are really the king of the Spanish omelet and my favorite Barceloneta neighbor. Sergi and Genis, finally you don't go to San Louis together, anyway I hope to see you next year in the US. Pep, It was good to have lunch with you, even if I did not understand all your jokes. Nicholas, thank you for the fishes even if I have no idea of what to do with them now. Montse, all the best for your pregnancy, I am sure that you will be a wonderful mother as you have been a wonderful friend. Alfons and Oscar, thank you for giving me a bit of your precious time to solve the problems of my crazy computer. Jordi, you are the cleverest men I know, is there a question of which you don't know the answer ? Hugo, you are so far from us now, we had so good discussions together, I wish you the best for your life with Raquel and Lola. Jorge and Cristina, I hope that you don't have bad dreams about the link3D, it is really a wonderful application. Xavi, I remember how great it was to climb and to ski with you, I wish you all the best for your life with Renata.

I would like to thank all the members of the Uruguayan family. Hugo and Dami, I will miss you so much in the US. Hopefully, we will have a

tremendous party all night for my wedding. Vero and Mariana, I had so much fun playing hockey with you, the every Wednesday games were one of the great moment of my life in Barcelona.

Citlali you are the nicest Mexican girl I have ever meet, we had great times skating together, I wish I could have been rich to buy you new roller-skates, I really hope that your life with Andres will be a success.

Thank you also to the Chilean family. Ray, you are the king of the prbb and of the barbecues. Eduardo, I swear if I had the possibility I would have bough a Ixus, and I know that the Frenchs played poorly for the European football championship but a bit better than the Spanishs.

Thank you to the lovely Argentinean girl Vicky. The "alfajoles" were so good, when are you going to Argentina again ? Anyway thank you so much for your advices during our walks on the beach.

I would like to thank Hagar and Miguel for visiting us in Washington D.C., I promise I will try to find a house with a hot tub for your next visit.

A great thank you to the Maria-Isabel's laboratory. Especially to Mireia, I remember the great moments we had with Pau and the yellow submarine in Menorca.

A very special thanks to grandpa, grandma, mum, dad, Eric and Carole. Living far from you is not easy, fortunately there are the holidays to spend a bit of time with you. Even if I live farther and farther, I give you all my love and affection.

Finally, I would like to thank the most important person for me. Maria-Isabel, by chance about four years ago we discovered that we lived at just 100 meters one from the other. Our lives have converged rapidly, first we

became friends, then lovers, fiancée and soon we will become wife and husband. In the Lord of the ring, Sam could not carry the ring of Frodo so he carried Frodo and his ring up to the volcano. You did the same for me, you could not do this thesis for me but you gave me the force to start and finish it. Without you this work would not have been possible, that's why I dedicate it to you. I love you so much.

Preface

The field of computer-assisted drug design started to attract my attention during my third year at the biotechnology department of the Luminy School of Engineering. A that time a senior scientist working for the pharmaceutical industry who was visiting our school told me that the people working in this field were mostly all chemists. I felt like I had made a big mistake. The studies I was doing would lead me away from this field that for some reasons had risen my interest. Fortunately, I had the possibility to spend a training period with Anne Imberty at the laboratory of structural glycobiology. Now, I must recognize that at that time I used the molecular modeling packages as black boxes, since my knowledge in this field was very limited. After this first contact with the world of molecular modeling I was very enthusiastic and applied for a six month training period in this field at the end of my last year of school. I realized that I had to learn programming if I wanted to make my CV more appealing. That's why my project was related with the programming of pharmacophore fingerprints at the molecular design group of AstraZeneca, UK.

Now that I am about to finish my Ph.D. studies, I do not use programs like black boxes anymore and my knowledge of programming languages is much more advanced than it was at the beginning of my work. This thesis is one of the main achievement of my work, as is the last version of the program Almond. I had the luck to contribute to this fantastic software for drug design, and I thank doctor Manuel Pastor for that. After seven years, my interest for the field of drug design is still growing, I hope that you will enjoy this contribution to this field as much as I enjoyed working on it.

List of Papers

Paper I. Fontaine F, Pastor M, Gutierrez-de-Teran H, Lozano JJ, Sanz F. Use of alignment-free molecular descriptors in diversity analysis and optimal sampling of molecular libraries. *Mol Divers* 2003;**6**:135-47.

Paper II. Fontaine F, Pastor M, Sanz F. Incorporating molecular shape into the alignment-free Grid-INdependent Descriptors. *J Med Chem* 2004;**47**: 2805-15.

Paper III. Brea J, Masaguer CF, Villazon M, Cadavid MI, Raviña E, Fontaine F, Dezi C, Pastor M, Sanz F, Loza MI. Conformationally constrained butyrophenones as new pharmacological tools to study 5-HT_{2A} and 5-HT_{2C} receptor behaviours. *Eur J Med Chem* 2003;**38**:433-40.

Paper IV. Fontaine F, Pastor M, Zamora I, Sanz F. Anchor-GRIND: Filling the gap between standard 3D QSAR and the GRid-INdependent Descriptors. To be submitted

Abbreviations

1D: One-Dimensional
2D: Two-Dimensional
3D: Three-Dimensional
CoMFA: Comparative Molecular Fields Analysis
DNA: Desoxyribo Nucleic Acid
EM: Extrathermodynamic Methodology
FFD: Fractional Factorial Design
GOLPE: Generating Optimal Linear PLS Estimation
GRIND: Grid-INdependent Descriptors
HOMO: Highest Occupied Molecular Orbital
LOO: Leave-One-Out
LUMO: Lowest Unoccupied Molecular Orbital
LV: Latent Variable
MACC-2: Maximum Auto- and Cross- Correlation
MEP: Molecular Electrostatic Potential
MIF: Molecular Interaction Field
MIP: Molecular Interaction Potential
MLR: Multiple Linear Regression
PC: Principal Component
PCA: Principal Component Analysis
PCR: Principal Component Regression
PLS: Partial Least Square or Projection on Latent Structures
QSAR: Quantitative Structure-Activity Relationships
RMS: Root Mean Square
SDEP: Standard Deviation of Error of Prediction

Table of contents

INTRODUCTION

Molecular descriptors and drug discovery

Evolution of the drug discovery process

New drugs have not always been discovered as they are nowadays. Medicinal plants have been used since antiquity for the treatment of health disorders. Numerous plant extracts have been the source of new drugs, for example the bark of some trees is rich in methyl-salicylate. As an example of chemical extraction, Sertüner isolated morphine from opium in 1806.

At the end of the 19th century, Paul Ehrlich, one of the fathers of the medicinal chemistry, introduced the concept of "chemoreceptors".(1) Ehrlich was the first to argue that differences in chemoreceptors between species may be exploited therapeutically. A more functional concept was introduced into pharmacology by J. N. Langley in 1905 in which the receptor serves as a "switch" that receives and generates specific signals and can be either blocked by antagonist or turned on by agonists.

Later came the secondary metabolites of mammals as another source of drugs. One of the most famous example is insulin which was purified in 1922 by Banting and Best. Vitamins were also identified during the middle of the 20th century. Another major breakthrough of the drug discovery history is the discovery of antibiotics, with the preparation of penicillin by Chain and Florey in 1940.

With the development of the organic synthesis, numerous reaction intermediates were synthesized, which led to the discovery of some new drugs. For example, the structure of the benzodiazepine Librium was discovered as an unexpected product of reaction.

Up to 1960s, the biological activity of a compound was essentially determined on entire animals, and the testing on cells was an exception.(2)

Later, progresses made in molecular biology and biochemistry allowed the development of more sophisticated biological assays, introducing the possibility to test receptor-ligand interaction *in-vitro*. Further progresses in molecular biology also allowed the production of recombinant proteins. Tissue plasminogen activator and erythropoietin are examples of successful recombinant proteins. In current drug discovery projects, molecular biology is now a key tool to understand the disease process at molecular level and to find out the suitable molecular targets for novel drugs.

In the seventies, the development of X-ray crystallography and nuclear magnetic resonance provide the first 3D structures of the biological targets, sometimes as complexes with a ligand bound. This new source of structural information opened the door to the structure-based drug design. Later in the nineties, the advances in combinatorial chemistry allowed the creation of extensive collections of compounds for testing. Robotics and the miniaturization favored the development of high throughput screening platforms able to perform biological tests on thousands of compounds a week. Natural product are still used but now the trend shifted towards the screening of libraries of compounds focused towards one or a family of targets.

Nowadays, the drug discovery process is complex. The pharmaceutical Industry has successfully integrated all the techniques described to obtain an optimum drug discovery methodology and is also an important contributor for the development of new technologies.

Molecular modeling

Molecular modeling is now one of the key techniques used during the drug discovery process. In Molecular modeling, a molecule is generally represented by a set of atoms and its coordinates. This model is the starting point for molecular simulations in different conditions and for the computing

of molecular properties. Some properties, e.g. molecular interactions, are particularly relevant for understanding drugs action mechanism.

Molecular modeling is a young but very diverse domain of research. The size and the number of the molecules under study can vary greatly and consequently the type of computational resources required to perform the calculations can be very different (Figure 1). The level of simplification used to approximate the reality also affect the type of computational resources required. We will give more details about three types of approximation: quantum mechanics, molecular mechanics, and rule based system.

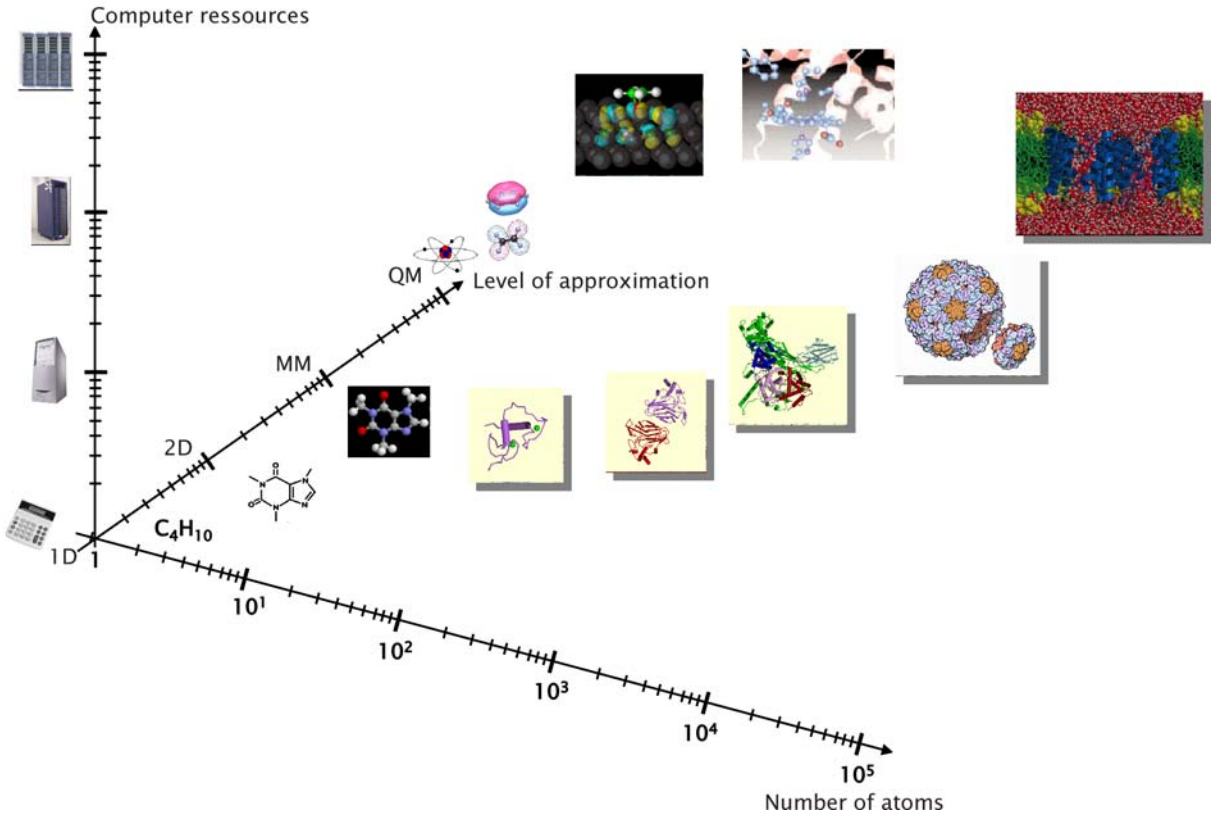


Figure 1. Exploring molecular modeling space

Quantum mechanics

In quantum mechanics the electrons are explicitly considered in the calculations, and so it is possible to derive properties that depend upon the electronic distribution and to investigate chemical reactions in which bonds are broken and formed. In opposition to classical mechanics the motion of the electrons is not along a trajectory, instead the electrons are spread through space like a wave.(3) For each specific location there is a probability to find the electrons at this position. The probability of finding the electrons depends on the value of the wavefunction, the higher the square of the wavefunction in a region of space the higher is the probability to find the electrons in that region. The Schrödinger equation allows to find the wavefunction of a collection of electrons. Under the Born-Oppenheimer approximation the nuclei move relatively slowly and may be treated as stationary while the electrons move around them. We can therefore think of the nuclei as being fixed at arbitrary locations, and then solve the Schrödinger equation for the wavefunction of the electrons alone. According to the molecular orbital theory the electrons spread throughout the whole molecule, and it is possible to define its wavefunction by a linear combination of the atomic orbitals. There are two families of calculation, the *ab initio* method where an attempt is made to calculate structures from first principles and the atomic numbers of the atoms present, and the semi-empirical method where certain integrals are set equal to parameters that have been chosen to lead to the best fit to experimental quantities. Thanks to modern computer applications it is possible to compute thermodynamic an structural properties such as enthalpies of formation.

In drug design, quantum mechanic calculations have a full range of application, for example precise energy minimization of molecular structures, location of transition structures, computation of molecular descriptors such as the dipolar moment, partial charge distribution for molecular mechanics simulations and molecular electrostatic potential computation.

Molecular mechanics

Molecular mechanics describes the atomic interactions using Newtonian mechanics. The system is described with the position of the nuclei while the charge distribution is considered to remain constant. Basically the interactions can be divided into two parts : the bonded and non-bonded interactions.

The balance between such interactions is defined by a force field, which is a set of equations representing the potential energy surface with respect to changes in the geometry of the molecule. In its simplest form the interaction potential *Vb* can be formulated as:

$$Vb = \sum E_{bonds} + \sum E_{angles} + \sum E_{dihedrals} + \sum E_{vdW} + \sum E_{electrostatic}$$

$$Vb = \sum \frac{1}{2} K_b (b - b_0)^2 + \sum \frac{1}{2} K_\theta (\theta - \theta_0)^2 + \sum K_\phi [1 + \cos(\phi - \delta)]$$

$$+ \sum \left[C_{12}(i,j) / r_{ij}^{12} - C_6(i,j) / r_{ij}^{6} \right] + \sum q_i q_j / (4\pi\varepsilon_0 \varepsilon_r r_{ij}) \qquad \text{eq. 1}$$

In the GRID(4) force field, which is extensively used in this work, only non-bonded interactions are considered. The force field contains a van der Waals term, an electrostatic term, and a hydrogen bond term which has been added to better represent the directionality and the strength of the hydrogen bonds. Another GRID force field peculiarity is that the value of the dielectric constant changes with the local environment of the probe.

One of the most critical aspects of any force field is its parameterization. The quality of the simulation depends on the quality of the parameters. Good force fields have transferable parameters, which means that the parameter can be transferred from one molecule to another without the need to derive new parameters for each new molecule studied.

With force fields for the bonded and non-bonded interactions it is possible to run advanced calculations such as structural optimization,

conformational analysis and molecular dynamics. If the aim is to locate the global energy minimum, structural optimization is a difficult problem due to the complexity of the potential energy surface. Minimization algorithms generally find only the local minima, while other methods more related with conformational analysis allow a better sampling of the energy surface. The main methods used for conformational sampling are distance geometry, genetic algorithm, systematic search, random search and molecular dynamics.(5) In molecular dynamics the molecular motions are simulated by integrating the Newton's equations of motion for each atom and incrementing the position and velocity of each atom by use of a small time increment.

Apart form energy minimization and conformational analysis, force fields have a wide range of application in drug design. Force field are well suited for the simulation of biopolymers such as proteins and DNA. Molecular simulations such as Monte Carlo and molecular dynamics provide structural information about the conformational changes in molecules and the distributions of molecules in a system.(5) Besides, simulations allow to compute properties of a system such as its internal energy. An interesting tool for *de-novo* design is MCSS.(6) This program places multiple copies of small fragments in a defined search area with a receptor. These chemical fragments are then subjected to an energy minimization protocol, which considers interactions between the receptor and fragments but ignores interactions between the fragments. The result is a series of low energy fragment poses encompassing several local minima.

The GRID force fields(4) is particularly useful to compute molecular interaction fields (MIF). In drug design MIF have two types of applications: In structure based design MIF are essentially used to find sites of favorable interaction for a chemical group in a protein binding site. In ligand based design MIF provide a virtual receptor sites which represents the type of interactions that a compound can make. The use of GRID and MCSS for functional group placement in protein binding sites have been compared recently.(7)

Rule based systems

It is also possible to obtain a reasonable structure of almost any compounds without solving any equation thanks to expert rule-based programs. For example, the program CONCORD(8) and CORINA(9) allow the automatic generation of three-dimensional atomic coordinates from the topological description of a molecule as expressed by a connection table. The program uses a database of bond lengths, angles and of ring conformations while the acyclic fragments are built in extended conformation.

Conformational search can also be performed following a similar approach, for instance the docking program Dock(10) use precomputed table of torsional angles for the fast generation of conformers.

The philosophy of such approaches is that in drug design it is often not necessary to provide a very accurate representation of the structure of the compounds studied. Instead, a good approximation allows to speed up the computation and to process massive amounts of data.

Quantitative structure-activity relationships

Very often, quantitative structure-activity relationships (QSAR) models are used in drug design. The idea behind the QSAR formalism is that the activity of a compound depends on its structure as described by electronic, hydrophobic and steric properties.(11) In medicinal chemistry the most interesting pharmacological properties depend on the interaction with a certain receptor. The strength of the interactions is correlated with the affinity for the receptor, and therefore only molecular descriptors that allow to quantify in some way such interaction can be used for QSAR purposes.

In QSAR there is no defined theory about the intrinsic nature of the relationship between structure and activity. The phenomena studied are not

well known so it is not possible to describe them with an equation based on a underlying theory. For this reason, QSAR models belong to the family of empirical models which provide only an approximate solution. Some assumptions must be made regarding the form, the continuity and the range of application of the relation.

Probably the most simple QSAR equation is the one which uses discrete parameters as the Free-Wilson and Fujita-Ban analyses.(12) In such analyses, the activities of a series of derivative of a reference structure is described by means of equation 2:

$$BA = \sum a_i I_i + \mu \qquad \text{eq. 2}$$

where BA is the biological activity of each product, a_i is the contribution to the activity of each substituent i, I_i is a binary variable which takes the value 1 when the substituent i is present and 0 when the substituent i is absent. The μ constant corresponds to the mean activity of the series in the Free-Wilson method and to the activity of the product without substitution in the Fujita-Ban model. Models of this type are only valid for congeneric series and are only useful to determine the optimal combination of substituents.

Instead of using discrete values, models which use continuous parameters are at a higher level of complexity. Parametric models are expressed by a mathematical function such as equation 3:

$$\log A = a_1 x_1 + a_2 x_2 + ... + a_n x_n + cte \qquad \text{eq. 3}$$

The first QSAR equation of this type, as it is still used nowadays, was published 40 years ago by Hansch et al.(13) to explain the activity of plant growth regulators using Hammet(14) constants and hydrophobicity parameters. The Linear Free Enthalpy Relationship method proposed by Hansch is also called "Extrathermodynamic Methodology" (EM) since the structure-activity relationships are described with thermodynamic terms (such as ΔG) without deducing them from the thermodynamic laws. The basic

postulate of the EM is that ΔG can be discomposed in several terms, each one taking into account different types of interaction between the ligand and its receptor. The fundaments of the EM can be summarized in a series of statements:

1. The biological activity is a function of the structure of the compound.
2. The structure of the compound imply global and local properties.
3. These global and local properties can be quantified by means of some parameters.
4. There is always a function which correlates the changes of biological activity with the changes of the global and local properties, although this relation may be not simple to determine.

The global parameters can be of two types: experimental or calculated. Probably the most used experimental parameter is the octanol/water coefficient of partition (log P) which is used to measure the hydrophobicity of the compound. It is a parameter difficult to obtain since it requires an experiment although it is widely accepted. As an alternative solution, computational methods have been developed to calculate it from the topology of the molecule.

The hydrophobicity can also be expressed by means of local parameters, i.e. parameters for each one of the substituents of the series considered. The parameter π(15) is equal to the difference between the octanol/water partition coefficient of a standard compound (e.g. benzene) with an hydrogen atom and the same compound with the substituent considered.

In order to avoid costly experiments plenty of computed molecular descriptors have been developed, for instance the program Dragon(16) computes up to 1800 descriptors. Therefore, a simple enumeration of the existing molecular descriptors is out of the scope of this introduction (for a review see (17, 18)). Instead a common classification based on the dimensionality of the structure used to compute the descriptor will be given.

1D descriptors are properties that do not require the knowledge of the topology or the tri-dimensional structure of the compounds. Therefore 1D descriptors are related to global properties of the molecule. Molecular weight is probably the most used descriptor of this category although many others exist, for example the number of atoms of a given type.

2D descriptors are based on the molecular connectivity of the compounds. Most of the calculated log P are based on fragmental approach and are therefore 2D descriptors. Molecular connectivity indices first described by Randic(19) and then extensively investigated by Hall, Kier and co-workers(20, 21) are well known 2D descriptors. Another type of 2D descriptors are the so-called "fingerprints" where the presence of a given fragment is encoded into a bit string.

3D descriptors are computed from a three-dimensional structure of the compounds. The properties can be global, for example the HOMO, the LUMO energy, and the dipolar moment which are computed by quantum mechanics softwares. Comparative Molecular Field Analysis (CoMFA)(22) and related approach are the most applied 3D methods generating local parameters. The 3D descriptors consist of the energy of interaction computed at thousands of grid points, consequently CoMFA models are at a higher level of complexity compared to the parametric approach of Hansch. The alignment-free descriptors GRIND(23) and VolSurf(24) used in this work represent a second generation of 3D descriptors developed to overcome the problems inherently associated to typical 3D descriptors. VolSurf descriptors are essentially global properties of the molecule while GRIND are local properties.

Once the biological values have been obtained and the suitable descriptors have been generated either from the experiment or from the computer, the next step is to obtain a mathematical relationship between structural and activity data. The most used multivariate techniques are Multiple Linear Regression (MLR), Principal Component Regression (PCR) and Partial Least Square (PLS). Often, the choice of the type of multivariate analysis employed is influenced by the type of descriptor used. For example,

MLR can be applied with the EM since the number of variables is low. In CoMFA, the method of choice is PLS due to the presence of a large number of correlated variables.

Molecular similarity and diversity

Molecular diversity analysis relies on the concepts of molecular similarity and molecular dissimilarity. In the context of drug discovery, it is assumed that similar compounds have more chance to interact in the same way with a given receptor than dissimilar compounds.(25) When an active compound is already known, similar compounds can be searched in a database hoping that such compounds will have similar biological properties. If no bioactive compound is known, then a sample of dissimilar molecules are often searched in order to maximize the probability to obtain an active molecule from the biological screening.

Computational approaches have been developed to provide practical solutions to the problem of diversity sampling. Since it is impossible to synthesize all the possible compounds and to test all the available ones, it would be more reasonable to pick a subset of dissimilar molecules of realistic size, hoping that at least one of molecule will have a significant response during the biological assay.(26) When such "hit" molecule is detected, the molecules excluded from the initial assay but with similar properties can be tested for rapid optimization of the desired pharmacological profile. This means that biological assays performed on diverse subset of compounds are not intended to obtained a lot of hits, but to obtain at least one hit in all the biological assays that can be performed on a screening platform.

Molecular diversity approaches have some limitations, even if it is more likely that similar compounds produce the same biological response during the assay, it is never guaranteed that two compounds even differing just for a single atom share the same biological profile, particularly in screening involving protein-ligand interactions.(27) Some measures of similarity are more

relevant than others. No ideal measure of dissimilarity is available and, each similarity/diversity measure has its own weaknesses.

Alignment-dependent descriptors

Molecular interaction potentials

It is admitted that the activity of a compound depends on the interactions that it can make with the targeted receptor. Therefore a powerful methods for QSAR analysis is to compute the interactions that the compounds can make with its receptor and to compare them. Molecular Interaction Potentials (MIP) contain the energies of interaction between molecular probes and the compound studied in all points of space (Figure 2). The simplest probe is a proton and in such cases the potential is called Molecular Electrostatic Potential (MEP). In more complex cases the probe can be a small molecule (e.g. water) or a chemical group such as the amide nitrogen of the peptide bond. For example, the program GRID(4) used in this study contains more than 50 probes of different types. MIP are used in two ways: on proteins they identify the regions where a ligand would bind favorably whereas on ligands they allow to determine the kind of interaction that the ligands can make with the receptor binding site.

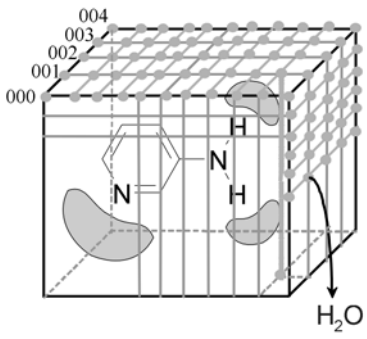


Figure 2. MIP calculation

MIP can be computed analytically by means of QM methods or approximated. Often, the target-probe energy is computed at regular intervals, inside a box that surround the molecule or the regions to be studied.

CoMFA and related approaches

Alignment of molecular structures

A MIP contains a full set of information related to the potentiality of interaction of a molecule. If the MIP of two molecules differs at a particular region of space, it indicates that they would interact differently at this region. This difference of interaction may be responsible for the difference of activity of the compounds and the objective of the CoMFA model is to elucidate if this relation is true or not, finding correlations between the differences in MIP and differences in biological activities. Unfortunately, the values of each MIP are sensitive to the orientation of the structure used to generate the MIP. When series of compounds need to be compared on the basis of their MIP, the comparison cannot be performed directly. A previous step of structural alignment of the compounds is required so that the same grid box can be used for all the compounds. Structural alignment is a complex task, especially if the compounds to align are structurally diverse. Actually, the amount of similarity between the compounds may influence the choice of the method used to overlay the compounds. Lemmen and Lengauer(28) provide a good review for the method developed to align molecules. Basically the main methodologies to align the molecule are :

1. RMS-fitting(29) of rigid-body objects which is possible when a common structural core is shared by the compounds of the series. If the structures share a flexible fragment the directed tweak(30) is preferable to the RMS-fitting.

2. Volume overlap optimization. In such methodology the molecules are represented by a set of spheres(31) or Gaussians(32-39) or MIP(40), and the overlap between them is quantified by means of a similarity measure. The fit is optimized by maximizing the similarity between the moving and the fixed compound.

3. Geometric hashing(41). The technique comes from the field of computer vision. It is based on the encoding of a set of geometric information in a hash-table which is invariant under rotation and translation. During the structural matching the hash table is queried with structural features from the molecule to align. The position in the hash-table that receives most queries corresponds to a transformation which is more likely to superimpose essential structural features of the two molecules.

4. Pharmacophore mapping by use of a clique detection algorithm. It is implemented in the program DISCO(42) in order to find the maximum common subgraph between all members of a set of molecules. The nodes of the graph are a set of pharmacophore points while the edges are the distance between them. The matching procedure utilizes clique detection to determine overall valid distance constraints.

5. Distance geometry.(43-46) This method is normally used in conformational analysis, but it can also be used to align structures if suitable constraints are provided.

Multivariate analysis

Once the compounds are aligned, a GRID calculation can be performed using a grid box, of at least, the size of the largest compound. The resolution of the grid is generally between 0.5 and 1 Å which means that thousands of variables are generated by the calculation (Figure 3).

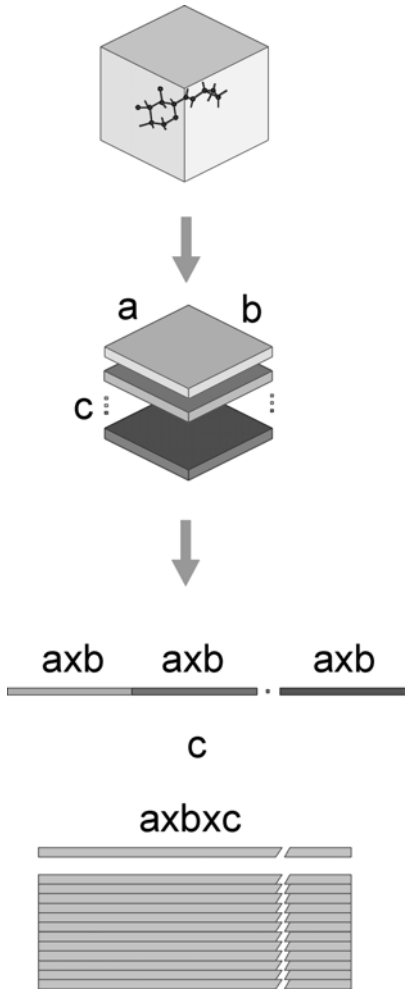


Figure 3. Decomposition of the MIP variables

Multivariate techniques are required to handle such amount of data, in this work Principal Component Analysis(47) (PCA) is used as a descriptive method whereas the Projection on Latent Structure or Partial Least Square(48) (PLS) is used as a regression method.

PCA is a useful technique to discover patterns and trends in the objects. In some way it summarizes the information contained in the **X** matrix and puts it in a form understandable by human beings. In PCA the original X-matrix is decomposed in two smaller matrix **P** and **T** (Figure 4) so that:

$$\mathbf{X} = \mathbf{1} \cdot \overline{\mathbf{x}'} + \mathbf{T} \cdot \mathbf{P} + \mathbf{E}$$

$\mathbf{1} \cdot \overline{\mathbf{x}'}$ represents the variable averages. If we subtract the variable averages to the data, the equation is simplified to:

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P} + \mathbf{E}$$

P is the loading matrix, it contains information about the variables. It describes a few vectors (the so called Principal Components, PC) which are obtained as lineal combinations of the original X-variables. **T** is the score matrix. It contains information about the objects. Each object is described in terms of their projections onto the PC, instead of the original variables.

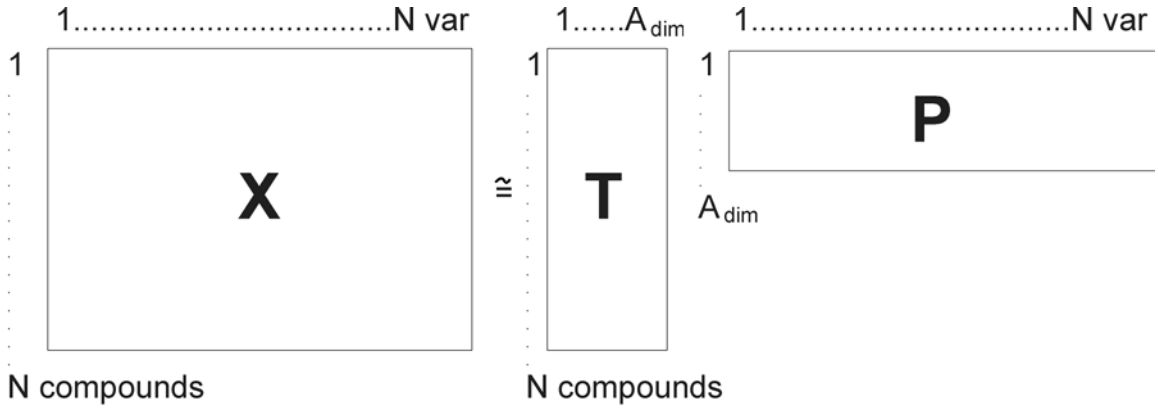


Figure 4. Decomposition of the X-matrix

The residual **E** matrix has the same dimensionality as the X-matrix. It contains all the information that is not explained by the product of the score and the loading matrix. One interesting property of the PC is that each PC is

orthogonal to each other. There is absolutely no correlation between the information contained in different PC.

In 3D-QSAR the number of variables is often much higher than the number of objects so that multiple linear regression cannot be used to correlate the variables with the activity. Instead, projection methods such as PLS should be used. As PCA, PLS deals with the matrix of variables X but also with one or more dependent variable Y.

$$\mathbf{Y} = f(\mathbf{X}) + \mathbf{E}$$

As for PCA the $\mathbf{X}$ matrix is decomposed as the product of the weight matrix $\mathbf{W}$ and the score matrix $\mathbf{T}$. The loading matrix contains few vectors (the so called latent variables (LV), which are obtained as linear combinations of the original X-variables. The concept of LV is quite equivalent to the PC in PCA which means that each LV is orthogonal to each other. The scores matrix contains information about the objects. Each object is described in terms of the LV, instead of the original variables. The PLS algorithm optimizes the values of the LV under two constrains: The LV have to represent the structure of the $\mathbf{X}$ matrix and the $\mathbf{Y}$ matrix and the LV have to maximize the fitting between the X's and the Y's.

The ideal number of LV cannot be determined from the quality of the fitting because the algorithm may overfit the data if too many LV are added. The predictive ability, measured in objects not included in the regression is the best way to really evaluate the quality of the regression model. The most used way to estimate the quality of a model is the cross-validation. In practice, models are built with one or several objects removed from the original dataset and then the models are used to predict the Y of the objects held out. The procedure is repeated until at least every object has been held out once. Then the experimental Y are compared with the predicted Y and, the SDEP (Standard Deviation of Error of Prediction) and the q^2 (predictive correlation coefficient) are calculated.

$$SDEP = \sqrt{\frac{\sum (Y - Y')^2}{N}}$$

$$q^2 = 1 - \left[\frac{\sum (Y - Y')^2}{\sum (Y - \overline{Y})^2} \right]$$

where Y is the experimental value, Y' is the predicted value, $\overline{Y}$ is the average Y value, and N is the number of objects. The Leave-One-Out (LOO) predictive correlation coefficient (q^2_{LOO}) is the most used. In LOO cross validation procedure, models are built keeping one object at a time out of the analysis and repeating the procedure until all the objects are kept out once. As a rule of thumb a q^2_{LOO} superior to 0.5 is necessary to obtain an acceptable model.

Variable selection

Usually not all the variables contribute in the same way to explain the **Y** matrix, and some of the variable only add noise to the model. However, every X variable, even if it does not contribute to explain the Y variables, certainly contributes to the structure of the **X** matrix. As the solution provided by PLS has the constrain of explaining the structure of the **X** matrix, this structure only makes more difficult to find a solution satisfying both constraints. Therefore the quality of the models may be increased by the appropriate variable selection. In this work we use the Fractional Factorial Design (FFD) procedure as implemented in GOLPE (Figure 5).(49) The idea is to remove some variables from the model, compute the SDEP and see if the model is improved or not. Since it would be too time-consuming to test every combination of variables to know its impact on the model, a design matrix is used instead. The effect of a variable in the model is equal to the average SDEP for all models that include the variable minus the average SDEP for the models that do not include it. The statistical significance of the effect of a variable is validated by

comparing the effect of this variable with the average effect of dummy variables by mean of a Student t test.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	…	x_p
Model 1	-	-	-	-	-	+	+	+	+	...	+
Model 2	-	-	+	+	+	+	-	-	+	...	-
Model 3	-	+	+	-	+	-	-	+	-	...	+
.....	...	...	...	...	...	...	...	...	...	...	...
Model N-1	-	-	-	-	+	+	+	+	-	...	+
Model N	+	-	+	-	+	-	+	-	+	...	-

Figure 5. GOLPE Fractional Factorial Design

Strengths and limitations of alignment-dependent approaches

Alignment-based strategies are powerful but limited by the intrinsic problems of the method. The description is highly specific since each variable corresponds to a tiny region of three-dimensional space, therefore, relevant regions of the field can be easily identified, which makes easier the interpretation of the models and the design of new compounds. However, the quality of the model depends strongly on the quality of the alignment. This means that little inconsistencies in the alignment can affect largely the quality of the models. The alignment is always biased towards a given solution since multiple solutions are generally available. In addition, for large series of compounds the number of variables may be so huge that the time required for variable selection can make this operation prohibitive.

For these reasons, alignment independent methods were developed. The idea of these methods is to retain the MIP information that is relevant to explain the desired properties. The information is compacted in a reduced set of descriptors which makes easier its analysis.

Alignment-free descriptors

We call alignment-free descriptor any type of 3D molecular descriptor that is translational and rotational invariant, i.e. that does not require the structural superimposition of the compounds studied. The most simple alignment-free descriptor is the dipolar moment, but there are many more:

1. In the pharmacophore fingerprint approaches, a set of feature points are computed such as hydrophobic centers, hydrogen bond donor and hydrogen bond acceptor atoms. Then all the combinations of 2-points,(50) 3-points,(51, 52) or 4-points(53) pharmacophores are computed. For each combination of feature points at a given distance range a bit in the fingerprint is set to one. The fingerprint may be hashed to save space, which means that the same bit may be set to one by two different pharmacophore triplets or quadruplets.
2. Distance Profiles(54) (DiP) method uses count of atom-pairs to obtain an histogram instead of a fingerprint. For each type of atom pair, e.g. oxygen-nitrogen, a set of distance bins are defined, and for each atom-pair, the value of the corresponding distance bin is incremented by 1.
3. The geometric Start-End Shortest Path(55) (SESP) vectors are a mix of 2D and 3D descriptors. The method weights the shortest paths between atoms by the Euclidean distance between them. For each pair of atom type at a given shortest path distance, the sum of the weighted shortest paths is computed. This allow to differentiate molecule with similar topology but different geometry like the boat and chair conformation of hexane.
4. The EVA(56) and EEVA(57) belong to the family of spectroscopic QSAR descriptors. In the EVA approach normal coordinate frequencies are calculated from the 3D coordinates of the molecule, then the normal coordinate eigenvalues are projected onto a bounded frequency scale

which is smoothed by a Gaussian function. The resulting spectrum is sampled at fixed intervals to provide a set of values which describes each molecule.

5. The Comparative Spectra Analysis(58) (CoSA) uses molecular spectra as molecular descriptors for 3D-QSAR. Experimentally determined ^{1}H NMR, mass and IR spectra as well as simulated IR and ^{13}C NMR spectra are converted into matrices of descriptors for PLS analysis.

6. The Weighted Holistic Invariant Molecular(59) WHIM descriptors are based on a principal components treatment of atomic coordinates that is claimed to be invariant to rotation and translation.

7. The Comparative Molecular Moment Analysis(60) (CoMMA) makes use of molecular moments such as the principal moments of inertia and properties derived from the dipole and quadrupole moments to characterize compounds in a CoMFA-like study.

8. The 3D-MoRSE(61) (Molecule Representation of Structures based on Electron diffraction) converts the three-dimensional structure into a fixed number of variables. The method uses a modified version of the molecular transform used in electron diffraction studies. Various atomic properties such as atomic mass, partial atomic charge, residual atomic electronegativities, and atomic polarizabilities can be encoded into a set of 3D-MoRSE value that are independent of the orientation of the molecule.

9. In the Internal Distance Analysis(62) (IDA) a common reference frame is defined from the center of mass and two points inside the molecules. The reference frame is used to calculate the polar coordinates of the points on the solvent accessible surface of the molecules. For each increment of θ and φ, a steric and electrostatic descriptors is computed, which leads to a translationally and rotationally invariant matrix of descriptors.

10. Autocorrelation methods(63-66) transform spatial alignment dependent properties into vectors independent of the orientation of the molecules which can be correlated with biological activity. For example, Wagener et al.(64) used autocorrelation of molecular surface properties to compress the information contained in the electrostatic or hydrophobic potential at the van der Waals surface.

11. The MaP(67) (mapping property distributions of molecular surface) approach encodes the distribution of molecular properties that are mapped on the molecular surface. The methodology uses a grid-based surface with equally distributed surface points and assumes that the count of the occurrence of a given pair of properties at a given distance range is meaningful. The properties mapped are hydrophobicity, hydrophilicity, H-bond acceptor and H-bond donor.

GRID-based alignment-independent descriptors

VolSurf

Volsurf(24) descriptors were designed to represent properties which are relevant to describe the pharmacokinetic properties of drugs. The descriptors contain information about the size and shape of polar and hydrophobic patches, as well as the balance between them. The computation of the descriptors from the 3D structure of the compounds is a two-steps procedure (Figure 6). In the first step, a MIP is calculated at least using the water and the hydrophobic probe of the program GRID. In the second step, the MIP are analyzed and the information required to compute the descriptors is extracted. The VolSurf descriptors are summarized in table 1.

The basic concept of VolSurf is to extract the information present in the MIP into few quantitative numerical descriptors which are easy to understand and interpret. Some of these descriptors are specifically designed to characterize global properties of the molecule (e.g. hydrophilic descriptors), while others are more focused on the local properties of the molecule (e.g. integy moment) and are therefore more sensitive to changes of conformation. VolSurf parameters can be computed for a wide set of GRID probes, however the most important are the one for the water and hydrophobic probes since there are more relevant to explain pharmacokinetic properties.

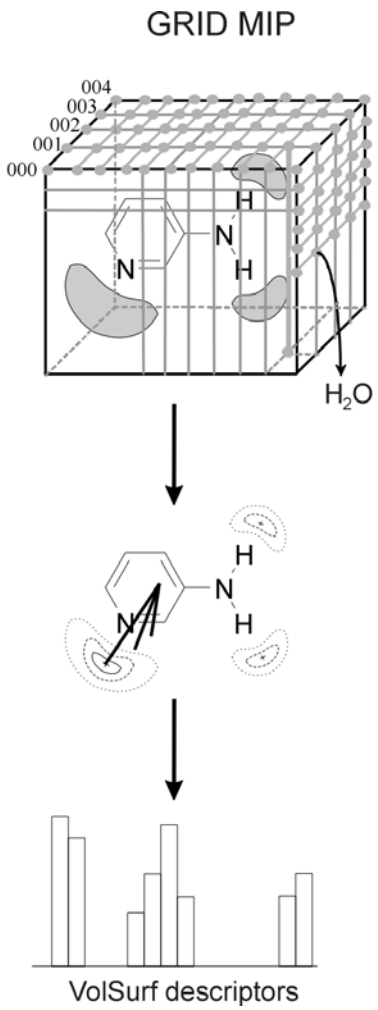


Figure 6. Calculation of the VolSurf descriptors

The most important descriptors to explain if a compound can cross a cellular membrane are the hydrophilic regions and the capacity factors. These two types of descriptors correspond to the capacity of a molecule to interact with water molecules. Hydrophilic regions are defined as the molecular envelope which is accessible to and attracts water molecules while capacity factors represent the ratio of the hydrophilic surface over the total molecular surface. In other words hydrophilic regions measure the total hydrophilicity while capacity factors measure the relative hydrophilicity.

Table 1. VolSurf descriptors

Type	Name	Meaning
Size	MW	Molecular Weight
	Molecular volume	Volume and surface are computed from the
	Molecular surface	positive part of the MIP at an energy cutoff of 0.2 kcal/mol
Shape	Volume/Surface ratio	The ratio is a measure of rugosity of the molecule
	Molecular globularity	Defined as S/S_{eq} with S_{eq} being the surface area of a sphere of molecular volume V
Hydrophilic	Hydrophilic descriptors	Define the volume of the hydrophilic envelope at various levels of interaction (-0.2 to -6 Kcal/mol)
	Capacity factors	Ratios of the hydrophilic surface over the total molecular surface
	Integy moment	Vector from the center of mass to the center of the hydrophilic regions at a given energy level
Hydrophobic	Hydrophobic descriptors	Define the volume of the hydrophobic envelope at various levels of interaction (0.0 to –2.0 Kcal/mol)
	Integy moment	Vector from the center of mass to the center of the hydrophobic regions at a given energy level
Mixed	Local interaction energy minima	Energy of the best three local energy minima
	Energy minima distances	Distances between the energy minima
	Hydrophilc-lipophilic balance	Describes which effect hydrophilic or lipophilic dominates in the molecule

Table 1. continued

Type	Name	Meaning
Mixed	Amphiphilic moment	Vector pointing from the center of the hydrophobic domain to the center of the hydrophilic domain
	Critical packing	Parameter which predicts molecular packing such as in micelle formation
	Hydrogen bonding	Capacity to bind to a polar probe other than water
	Polarisability	Estimate of the average molecular polarisabilty

The distribution of the hydrophilic and hydrophobic interactions are encoded into the integy moments. For hydrophilic regions an integy moment is a vector pointing from the center of mass to the center of the hydrophilic regions at a given energy level. Molecule with a big integy moment have a clear concentration of hydrophilic regions in only one of their extremities. Molecules with a small integy moments have polar moieties either close to the center of mass or at opposite ends of the molecule. A high integy moment helps the molecule to cross biological membranes.

All these descriptors together constitute the VolSurf descriptors. Such descriptors have been applied to a wide range of prediction of biological properties which are summarized in table 2.

Table 2. Applications of the VolSurf descriptors

Applications	**References**
Blood-brain barrier permeation	(68, 69)
Membrane partitioning of N-methylated oligopeptides	(70)
Absorption for Caco-2 and MDCK cell monolayers	(24) (71)
Parallel artificial membrane permeation	(72)
Skin permeation	(73)
ΔG of hydration	(73)
Plasticizing efficiency of starch acetate	(74)
Lymphatic transfer of lipophilic compounds	(75)
Surface properties of amino acids	(76)
Binding affinity	(77)
Database mining	(78)
Chemical space navigation	(79)
QSAR	(80, 81,82, 83)
Chemical diversity	(84)

Taken together, all the applications listed in table 2 show the versatility of the VolSurf descriptors. Although they are more suited to problems related with physicochemical and pharmacokinetic properties of drugs, some successful examples related with pharmacodynamic properties have also been published.

A particularly interesting application is the prediction of blood-brain barrier (BBB) permeation by Crivori et al.(68) A PLS discriminant model that correctly predicted more than 90% of the BBB data was made. The PLS coefficients showed that hydrophilic regions, capacity factors, and H-bonding inversely correlated with BBB permeability. Integy moments, the hydrophobic regions and the critical packing are directly correlated with BBB permeation, but their role appear less important than that of polar descriptors. For skin permeation, a model was built by Cruciani et al. (73) for a

diverse set of drugs. The authors conclude that hydrophilic regions should be increased for improved skin permeation, but only with polar regions well distributed over the molecular surface.

Original applications of VolSurf include database mining,(78) which is the process of handling databases by means of informatic technologies. Results show that VolSurf is not suitable for clustering compounds according to their pharmacodynamic properties whereas it is highly recommended for pharmacokinetic profiling of drugs.

Another interesting application of VolSurf is the so-called chemical space navigation presented by Oprea et al.(79) In their approach, the authors use a set of 423 diverse molecules to define a chemical space based on the VolSurf descriptors. The PCA of the 423 objects shows that the first principal component is well correlated with passive permeability while the second principal component correlates with solubility, thus providing a convenient map of pharmacokinetic properties.

GRid-INdependent Descriptors

The GRid-INdependent Descriptors(23) (GRIND) are another type of alignment-free GRID based molecular descriptors specifically designed to characterize ligand-receptor interactions. The aim of the GRIND is to identify relevant regions of interaction and to describe the relative position of these regions.

The GRIND calculation starts with the computation of one or several GRID MIF (Figure 7). Generally, the GRID probes used for the calculations are the hydrophobic probe (DRY), the hydrogen bond acceptor carbonyl oxygen probe (O), and the hydrogen bond donor amide nitrogen probe (N1). These three probes were chosen because they represent the most characteristic non-bonding interactions found in biological receptors. In the present work, a fourth probe describing the molecular shape will be added.

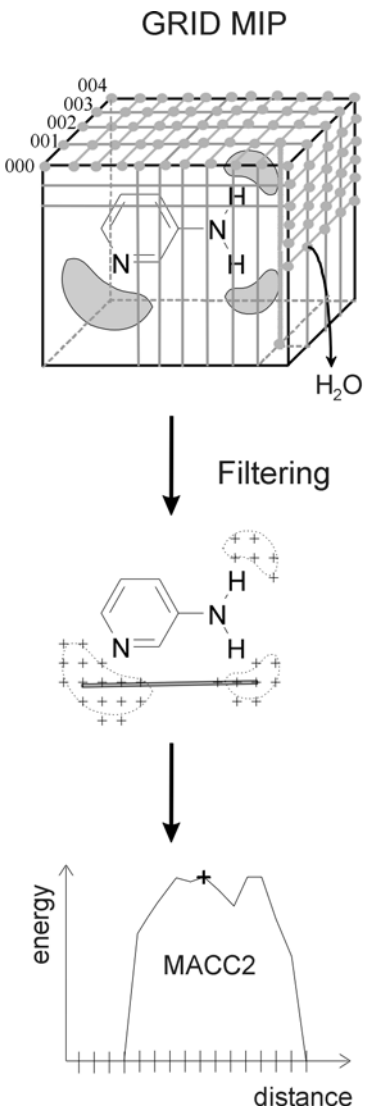


Figure 7. Calculation of the GRIND

The MIF are filtered by a function based on the intensity and the distance of the MIF nodes. The role of this function is to identify well defined regions of relevant interactions. The next step is to multiply each node value by a precomputed scaling factor so that most node values will take values in the range 0 and 1. From each possible pair of scaled filtered MIF a Maximum Auto- and Cross- Correlation (MACC-2) transform is applied. The purpose of the transformation is to obtain descriptors independent of the orientation of the molecule. In practice the scaled energy product of each pair of node are calculated, and the values are put into a distance bin according to the node to node separation. For each distance bin, only the highest energy

product is kept, consequently the transformation allows to trace back what is the pair of nodes responsible for the highest energy product located in a distance bin, which is useful for the chemical interpretation of the models. Each ensemble of distance bins for a given pair of MIF is called a correlogram. The default set of probe DRY, O, N1 generates three auto-correlograms (DRY-DRY, O-O, N1-N1) and three cross-correlograms (DRY-O, DRY-N1, O-N1). Taken all together, the set of 6 correlograms define a pattern representative of the interactions that can make a compound. To summarize, the GRIND encode the geometrical relationships between the relevant sites of interaction of the MIF into an orientation-independent set of variables. Some published applications of the GRIND are shown in table 3.

Note that the GRIND are alignment independent but not conformation independent. Nevertheless GRIND provide a rough description which is much less affected by small conformational inconsistencies than CoMFA.(85) Conformations obtained from automatic structure generators such as CORINA are normally in an extended form. Consequently, the models are biased by the use of extended conformations of the compounds which are not necessarily the bioactive ones. However, such conformations have the advantage of being generally consistent one with the others, which is an important prerequisite for any QSAR model building. Using extended conformations is not a problem if the aim of the model is to obtain some information about the structural factors essential for the activity of the compounds. Whatever the conformation, these structural factors should be the same. Nonetheless, the distances between the site of interaction are completely dependent of the conformation chosen and care must be taken if the objective is to derive a pharmacophore from the model.

The first published application of the GRIND(23) demonstrates the alignment independency of the GRIND with a dataset of glucose analogue inhibitors of the glycogen phosphorylase.(23). Three copies of the same molecule were taken at different orientation, and it was shown that the same molecules clustered nicely in the PCA and PLS plots.

Table 3. Application of the Grid-Independent Descriptors

Applications	References
QSAR	(23) (86, 87) (88, 89)
CYP2C9 inhibition	(90, 91)
CYP3A4 enzyme stability	(92)
Database mining	(78)
Identification of structural patterns among databases	(93)
Protein binding sites comparison	(94)

An interesting application is the combination of the GRIND with the flexible GRID fields approach.(91) The flexible regions of the ligands are able to move in response to the interaction with the probe. The result is a map of interactions for each probe that describes the most energetically favorable possibilities that a ligand has when it is allowed to adjust to the surrounding. The method was validated by comparing the flexible GRID fields with the fields obtained from the merging of 100 GRID fields generated by a random conformational search. The resulting GRIND model using the flexible GRID filelds had the same predictivity as the model with the docked conformations generated with the program GOLD(95).

As for VolSurf, Cruciani et al.(78) studied the potential use of the GRIND descriptors for database mining. For pharmacodynamic properties, the GRIND performed well but not as good as the Unity fingerprint. The results of consensus PCA indicated that the information contained in the GRIND is different from all the other tested methodologies. Interestingly, the PCA on the blood/brain barrier permeation dataset showed a nice clustering of the compounds according to its pharmacokinetic properties.

The GRIND can also be used to compare protein binding sites. For instance, Gutierrez-de-Terán et al.(94) computed the GRIND for a set of experimental ribose binding sites and a hypothetical one, then they calculated the similarity between them using the Hodgkin index. The similarity

between the putative binding site and some of the experimental ones was higher than the similarity observed between the experimental binding sites. These results further confirm that the putative ribose binding site identified by the docking program is similar to other binding sites.

RESULTS AND DISCUSSION

In the previous section we described the principle, applications, and developments of the alignment-free molecular descriptors GRIND and VolSurf. In this section, we will introduce the new applications and developments contributed by the author and published in 4 articles.

In **publication 1** we explain the selection of a diverse sample of primary amine using such descriptors. A great part of the paper is dedicated to the type of description provided by the descriptors and to compare them. Although VolSurf and GRIND use the same type of information, i.e. the GRID MIF, the information contained in each set of descriptors is different. Both set of descriptors describe first global features of the compounds, and in this part the description is fairly similar, however for positional features the description is more specific to the method used. From a point of view of the diversity, both method perform equally although the content of the clusters is very different. The examples given in **publication 1** showed some limitations regarding the descriptions of the GRIND.

These drawbacks are in part corrected in **publication 2** by the development and incorporation of shape descriptors into the GRIND. Three 3D-QSAR models have been published with the shape field included to the original GRIND, two in **publication 2** and one in **publication 3**. The objective of **publication 2** is to show a model with favorable shape effects detected and a model with unfavorable shape effects detected. Although the shape field is not properly a GRID MIF, it is perfectly integrated in the GRIND formalism and its interpretation is consistent with the GRIND philosophy.

In **publication 4**, a new methodology, anchor-GRIND, is presented. This methodology makes use of an anchor point to improve the specificity of the description. The idea of using a specific anchor point was introduced in **publication 1**. Good results were obtained for a congeneric, a combinatorial chemistry and a non-congeneric dataset. The model obtained are simpler to analyze because the reference point for comparison is always the same. The

special filtering applied to the scaffold improves the information content of the correlograms.

All together, these publications show the evolution of the GRIND descriptors during the past few years, from its application to diversity selection to the anchor-GRIND methodology.

Alignment-free descriptors and molecular diversity

Paper 1 tests the suitability of the alignment-free molecular descriptors VolSurf and GRIND for diversity sampling of a library of primary amines. The first part of the paper is focused on the type of description provided by each set of descriptors and the difference between them. VolSurf and GRIND descriptor are compared with descriptors obtained from quantum-mechanical calculation with the program AMSOL. VolSurf description is more similar to GRIND description than AMSOL description since the two programs use the same source of information, i.e. the GRID MIF. A PCA analysis is performed in order to remove redundant information of the original descriptors. The first PC of GRIND and VolSurf analyses differentiate the compounds in a similar way which is related to the hydrophilicity of the compounds. The remaining PC indicates that the descriptions provided by VolSurf and GRIND are different, especially for positional information. In the second part of the article, the sampling of the database is performed on the basis of k-means clustering on the space of the selected PC. Diversity sampling is compared with random sampling using a R^2-like diversity coefficient. The coefficient indicates that the sample size has a great impact on the diversity and that random sampling is not suitable for obtaining optimally diverse samples, especially for a small sample size.

VolSurf and GRIND descriptors are easy to obtain and fairly fast to compute, which make them suitable for routine diversity selection. For both method the compounds are separated first according to global features such as hydrophilicity and after by positional features such as the integy moment.

This means that most of the variance of the descriptors is due to global features, and that the complexity of the description increases with the rank of the principal component. As a consequence, the first PC are easier to interpret, and the description has been limited to the number of PC that can be linked to well defined structural features.

From the point of view of database sampling, more advanced selection could be performed. It would be interesting to selected the compounds in a subspace of descriptor to exclude compounds with undesired properties. For example, too hydrophilic reagents should be removed from the selection to avoid poor intestinal membrane permeability. The first PC is ideal for that but the problem remains to know where to cut out, because it is the hydrophilicity of the final product that really matters and not the hydrophilicity of the reagents. However, since hydrophilicity is essentially an additive property, a cut-off limit could be defined.

In the context of focused library design the GRIND descriptors have a great potential. The library could be designed around a given pharmacophore of the reagent. There is a problem if the pharmacophore is between two different reagents since the pharmacophore is not an additive property. In such cases, the selection should be performed preferably on the product space instead of the reagent space.

More than a mere diversity selection the design of a library is often a multicriteria process.(96, 97) For instance cost and drug-likeness are two parameters that may be considered at the time of performing the library sampling. Among the methods used for multiobjective selection of compounds, the most described are simulated annealing and genetic algorithms. Particularly interesting is the method described by Jamois et al(98) where the descriptors are calculated on the fly for the selection. This avoid the computation of the descriptors for all the virtual library and therefore would be suitable for descriptors such as VolSurf and GRIND which are fairly slow to compute compared to common descriptors used for this kind of study (e.g. rule-of-five parameters).

In the conclusion of **publication 1** we state that the right choice of descriptors depends on the kind of problem that the selected series is intended for. In the data mining publication of Cruciani et al,(78) the author claims that GRIND are better than VolSurf descriptors to deal with pharmacodynamic property. According to this conclusion, the GRIND descriptors should be preferred for a diversity sampling designed to improve the probability of obtaining a compound that interacts with a set of receptors. In addition, the GRIND descriptors have been especially designed for QSAR which means that there are more promising for obtaining biologically relevant discrimination. However in our case, we extend VolSurf description to the fourth PC while Cruciani et al. just considered the two first components. The additional information contained in the two last PC is more positional and may be relevant regarding the binding of the compounds to a hypothetic receptor. Perhaps the best description would be a consensus between VolSurf and GRIND with the drawbacks of one set of descriptors compensated by the other. VolSurf would be used to obtain drug-like molecules with a good pharmacokinetic profile and GRIND would be combined with VolSurf to focus on a given motif of interaction

In any case, the ability of the descriptors to represent well the molecular interactions of each compounds are limited by the GRID force field limitations. For example, the hydrophilic interactions of the AMSOL neighbor for the seed 2 in table 5 are much higher for VolSurf than for AMSOL, indicating that the interactions with water are treated differently at the molecular mechanic level than at the AM1 level. It should be noticed that the possibility to calculate atomic charges using AM1 semi-empirical calculation is now available in VolSurf, and that the authors of the program claim that better results are obtained with such charges.

Regarding the diversity sampling, the coverage of the database as expressed by the R^2-like coefficient follows a hyperbolic relationship with respect to the number of compounds. It is necessary to sample around twice as many reagent with random selection than with diversity selection to obtain the same coverage. Interestingly, a similar relationship is found by Potter and

Matter with the coverage of the database expressed as the percent of biological class covered.(25) The efficiency enhancement is of the same order of magnitude: between 1,5 and 3,4 more compound need to be tested by random selection to obtain the same coverage as the diversity one.

One of the drawbacks of the GRIND is that atoms that do not have a strong interaction with the probe are not well described. In table 5 of **publication 1** some compounds are ranked according to their distance to a seed compound. For example, for the first seed, the compound that is ranked third with the GRIND descriptors is highly substituted with chlorine atoms, while the seed is not. The shape probe described in the next publication is an attempt to solve this problem.

GRIND. Improving the description by considering shape properties

A major drawback of the original GRIND methodology is that it doesn't consider the shape explicitly. **Publication 2** describes the development and the integration of shape descriptors into the GRIND. Such descriptors are based on the local curvature of the molecular surface. The spatial extents of the molecules are recognized by selecting the regions of maximum convexity. Such regions allow to compare and visualize the protrusions of the substituents from the molecular scaffold. The descriptors are stored in a similar way as a GRID MIF, which facilitates their integration into the GRIND. Two 3D-QSAR studies were performed, in both of them the quality of the models was improved with the inclusion of the shape descriptors. The first example is about xanthine-like antagonists of the A_1 adenosine receptor, and illustrates how the shape field can be used to identify favorable shape matching between the ligands and its receptor. The second example is about inhibitors of the plasmepsin II aspartyl protease of *Plasmodium falciparum*, and shows how unfavorable steric interactions can be identified.

In **publication 3**, another 3D-QSAR model is performed on a set of 52 antagonists of the 5-HT_{2A} receptor. The model includes the shape descriptors,

which are important to explain the activity of the compounds of the series. Particularly important is an optimal distance between a hydrogen bond acceptor region generated by the protonated amine and the farthest extreme of the molecule represented by the molecular shape field.

In these studies, the description of the shape based on curvature is rough compared to other methods which allow to differentiate between planes and saddle points.(99, 100) In the current context, the description is precise enough since only the most convex regions are conserved for the analysis. However for other applications such as binding site description, an algorithm giving a more detailed description could be useful.

The shape field aims at the identification of potential pocket of favorable or unfavorable interactions as it is exemplified in figure 5 of **article 2**. Recent studies of the binding of butyrofenone by Dezi(101) indicate that the shape pocket identified in the figure 3b of **publication 3** corresponds to a pocket of interaction in the receptor. However each shape patch does not necessarily correspond to a binding pocket, for instance the shape patch on the top of the glutamine residue in figure 9a of **publication 2** does not correspond to a pocket but instead to a region of possible steric interactions (figure 10 **publication 2**). This example illustrates well that it is not possible to produce an accurate description of the shape of the receptor in an indirect way.

Further analyses of the description provided by the shape field indicate that not only shape differences are considered but also size differences. Indeed curvature is a 'pure' shape property but the distance between two nodes with a given curvature is not, and it depends also on the size of the substituents situated between these two nodes. For example, in figure 3b of **publication 3**, the optimal distance between the hydrogen bond donor site and the farthest extreme of the molecule depends on the size of the substituent. In practice, both the size and shape are important for describing steric interactions and this is why it is important to consider them together.

One crucial point is the problem of conformation and how the shape of the compounds are affected by the conformation generated by CORINA.

Since the compounds are generated in extended conformations there is a maximum displacement of the shape patches upon size change. However, during the binding some groups may fold and therefore have the same shape as smaller groups. This cannot be detected with the approach described in **paper 2** since only one extended conformation is used.

Another problem is that part of the shape information contained in the shape field is lost after the MACC-2 transform. A molecule with the shape of a equilateral triangle such as clozapine will have 3 convex patches, one at each extremities of the molecule, but it will have only 2 shape peaks in the TIP-TIP correlogram as a linear molecule of the same size. One potential solution to this problem would be to implement a MACC-3 transform were the correlogram is the product of 3 grid node values. This solution has the inconvenient of increasing the number of variable drastically. For example for 3 probes, the number of correlograms is increased from 6 to 10 and the number of variables per correlogram increases with the size of the molecule to the power of 3. Consequently, an improvement in the specificity of the description must be followed by an increase of the number of variable, which complicates the interpretation of the models.

GRIND. Improving the geometrical description from chemical knowledge

In **publication 1** we saw that the comparison of the distribution of the MIF requires a reference point for comparison. In the case of VolSurf this reference point is the center of mass of the molecule, in the case of GRIND, for the series analyzed there, the reference points are the sites of interaction with the amine nitrogen. In order to improve the specificity of the description, we propose to define explicitly a specific atom as the reference point of each compound. This "anchor point" is defined using the topology of the compounds.

The GRIND are less specific than CoMFA or related approaches. Consequently they are less sensitive to little inconsistencies in the conformation of the ligands, however it is sometimes impossible to differentiate some structural features from others. The objective of **publication 4** is to provide a description not as detailed as CoMFA but more specific that the standard GRIND methodology. The anchor point allows a more detailed comparison of the spatial distribution of the MIF of each substituent. Furthermore, it may be not the distances between the MIF of the substituent and the anchor point that are important for the activity but the distance between features of the substituent themselves. For this reason The MIF-MIF block of descriptors has been introduced, providing extra information and helping to describe better symmetric regions. In the first dataset of **publication 4**, i.e. hepatitis C virus NS3 protease inhibitors, the MIF-MIF block improves the quality of the QSAR model, which indicates that in some cases the anchor point variable are not enough to describe the relative position of all the interactions made by the compounds. The second dataset, the acetylcholinesterase inhibitors, shows the application of the anchor point methodology in combinatorial chemistry while the third dataset, the benzamidine factor Xa inhibitors, is a case where the anchor point is used for the analysis of a large set of ligands.

The anchor point methodology is limited to series of compounds with a region of common topology which is supposed to bind in the same way for all the compounds. Nevertheless the panel of datasets that can be studied with this methodology is huge, particularly with the increasing number of combinatorial chemistry datasets.

For congeneric series, often found in medicinal chemistry, (e.g. in the first dataset of **publication 4**) it is fairly easy to define an anchor-point from the scaffold. Even non-congeneric series can be treated by the method (e.g. the factor Xa dataset of **publication 4**). Provided there is a common chemical group that can be identified, any non-congeneric series can be studied with the anchor point approach. Among the promising candidate series we can

find the 5-HT receptor antagonists with its charged nitrogen or the PPAR inhibitors with its charged carboxyl moiety.

Another particularity of the anchor point approach is that the interactions of the scaffold are filtered out. Since the scaffold is the same for all the compounds it is not necessary to describe the interactions with this part of the molecule. This is particularly interesting for series of compounds with a big scaffold and small R substituents. In such cases the correlograms are 'saturated' by energy products that are the results of two sites of interactions of the scaffold. The filtering cleans the correlograms of useless information and thus allow a more detailed description of the relevant parts of the molecule, i.e. the R groups.

Apart from the requirement of a common structural feature, the anchor point methodology may suffer from other limitations. The position of the anchor point may be not trivial, for example if several possibilities are available to the user, the quality of the QSAR model may change depending on the position of the anchor point chosen. Some structural differences between the compounds may be well described with a given anchor point and not so well with another, and it is not possible to know in advance which one will give the best results. In addition, if the anchor point is far from the R group intended to describe there is more risk of inconsistencies in the conformation of the scaffold, which affects the orientation of the R group.

GRIND. Future works

In this section we will discuss the potential improvements that could be incorporated to the GRIND. We will see that both the filtering and the encoding are concerned.

Regarding the filtering, the algorithm can be modified. The current method forces the user to define a specific number of nodes to filter in. The value that must be set by the user depends on the size and the type of interaction of the compounds of the series. If this number is too low some

regions of interaction may be missed, if it is too high the filtering procedure becomes very slow and irrelevant nodes are selected. To ensure a correct filtering, the user must control the results by performing a visual inspection. A better algorithm would find the ideal number of nodes automatically. In order to achieve that, ad-hoc clustering algorithms, based on the energy of the nodes, should be developed.

Regarding the encoding, one of the potential improvement of the method is to use a more complex geometrical description than pairwise distances. In their current form the GRIND descriptors cannot provide a full description of all the geometrical relationship between the regions of interest. For example, a molecule with the shape of an equilateral triangle such as clozapine will have 3 convex patches, one at each extremity of the molecule, but it will have only 2 shape peaks in the TIP-TIP correlogram as linear molecule of the same size. There are a wide range of possibility to increase the specificity of the geometrical description but with the drawback of increasing the number of variable drastically. Apart from the problem of memory overloading, new ways of visualizing the variables must be designed too.

Dealing with the conformational problem is a big challenge for the GRIND. Until now, a lot of models have been obtained with a single conformation generated by the program CORINA, which in general produces consistent conformation although this is not always the case. The quality of the results obtained showed that using an extended conformation is a good first approximation. Would it be preferable to use multiple conformations instead of a single one? One possible option could be to select a hypothetic bioactive one, for example using a structure as a template and selecting the conformation of the other molecules on the basis of its similarity to the template. These questions remain open until further work have been done.

Diversity sampling and 3D-QSAR are not the only possible applications of the GRIND descriptors. One of the most straightforward applications of the 3D-QSAR models is inverse QSAR. In inverse QSAR the objective is to find new compounds with a desired activity using the initial QSAR model. To the date,

no example of such study with the GRIND descriptors has been published. The main difficulty is to obtain good quality 3D-QSAR model that are robust enough to predict the activity of compounds from structural class other than the ones included in the training set. It should be noticed that there is no restriction concerning the method used to build the GRIND model. In this work only PLS has been applied but neural networks(102) or recursive partitioning(103) are other multivariate analysis method that can be applied, to name but a few.

Molecular similarity is another field which is related to inverse-QSAR and virtual screening, where the GRIND have a great potential. By means of the suitable similarity coefficient, e.g. the Hodgkin index, it is possible to compare the correlograms of two compounds, thus allowing to find new chemical entities similar to a given compound having the desired activity. This concept has already been applied with 3D-fingerprints,(104) where the presence of a given pharmacophore is encoded in a bit string. The GRIND consider only pairwise distances, so the geometrical description they provide is less specific than the 3D-fingerprints which normally encodes 3 to 4 points pharmacophores. However the GRIND present some advantages over the 3D-fingerprints, since the GRIND paradigm is not based on pharmacophoric points but rather on pharmacophoric regions which change smoothly of intensity. Using GRIND, it is easier to encode the relative position of a surface feature such as hydrophobic regions than with a single point based method. Consequently the GRIND should be less sensitive to the size of the distance ranges and to small changes of conformations. Moreover, the GRIND descriptors encode numerical values so that the intensity of the interactions are taken into account, whereas only categorical values are encoded into 3D-fingerprints. Preliminary results on a set of cannabinoids ligands showed that if two conformations of two molecules have a high similarity score there is more probability to align their structures successfully. Consequently the GRIND can be applied as a first filter for database query before using a more time-consuming but more specific method such as structural superimposition.

One step further than similarity search but also possible with the GRIND is pharmacophore query. In classical pharmacophore search, a set of special features (e.g. hydrogen bond donor atom) separated by a given distance ranges are defined, and if a compound matches the pharmacophore it is included in the hit list. In a GRIND pharmacophore search the operating mode would be different since the method deals with numerical values. The user would define the distance bins containing the pharmacophore, then, for each compound, the value of the energy products of each pharmacophore bin would be multiplied or summed together in order to obtain a matching score. The result of the query would be a hit list sorted according to the score obtained by each compound.

Another potential application of the GRIND is protein binding site classification. Each binding site contains a set of specific interactions that can be identified by means of MIF computations. As it is performed for ligand based studies, the geometrical relationships between the MIF can be encoded into the GRIND variables, with the difference that the coordinates of the grid box are set to contain the whole binding site. The design of an automatic procedure to define the grid box would, in theory, allow the comparison of any type of binding sites. The first example of such binding site comparison was performed by Gutierrez de Teran et al.(94) In their publication, the authors compared an hypothetic binding site of ribose identified by docking with known binding sites of ribose extracted from the Protein Data Bank. They concluded that the binding site identified by docking was a valid candidate since its similarity with one of the crystallographic ribose binding sites was higher than the similarity between most of the ribose binding sites. At a larger scale, a classification of the binding sites of the Protein Data Bank would be successful if similar binding sites identified by the method would bind similar ligands.

CONCLUSIONS

This work describes recent developments and applications of the alignment-free descriptors GRIND. We have reached the following conclusions:

1. We demonstrated that GRIND can be used for diversity sampling of molecular libraries by means of principal component analysis and k-means clustering. The information given by the GRIND was compared with the information provided by VolSurf and quantum-mechanical descriptors. GRIND and VolSurf provide a similar global description but are very different for describing topological and positional features of the molecules, consequently the content of the clusters can be very different. Random sampling required two times more compounds to obtain the same diversity as the k-means sampling described.

2. We developed new shape descriptors based on the surface curvature of the compounds. The descriptors were able to identify highly relevant shape characteristic, such as the spatial extents of the molecules. In addition, they were perfectly integrated into the GRIND, which made the interpretation of their weight on the models easier. The descriptors were useful for identifying favorable as well as unfavorable shape effects. The unfavorable shape effects were in good agreement with the structural information available.

3. We developed a new methodology called anchor-GRIND, which is restricted to cases where there is at least a single point in the structure which can be recognized as common for either chemical or biological reasons. This common point is used for MIF comparison by means of alignment-independent descriptors. The anchor-GRIND methodology provides a simple but efficient way of obtaining 3D-QSAR models, both for congeneric and for

non-congeneric series. Again we found good agreement between the descriptors important for the models and the structural information available.

This work continues the development of the alignment-independent descriptors. We wish further development and application will come in the next future.

BIBLIOGRAPHY

1. Drews J. Drug discovery: a historical perspective. Science 2000;287(5460):1960-4.
2. Monge A. Evolución de los metodos de busqueda y descubrimiento de farmacos. In: Avendaño C, editor. Introducción a la química farmacéutica. Madrid: McGraw-Hill; 2001. p. 25-40.
3. Atkins P. The elements of physical chemistry. Third ed. Oxford: Oxford university press; 2001.
4. Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem 1985;28(7):849-57.
5. Leach AR. Molecular Modelling. Principle and applications. Second ed. Harlow: Pearson Education Limited; 2001.
6. Miranker A, Karplus M. Functionality maps of binding sites: a multiple copy simultaneous search method. Proteins 1991;11(1):29-34.
7. Bitetti-Putzer R, Joseph-McCarthy D, Hogle JM, Karplus M. Functional group placement in protein binding sites: a comparison of GRID and MCSS. J Comput Aided Mol Des 2001;15(10):935-60.
8. Pearlman RS. "CONCORD User's Manual," distributed by Tripos Inc., St. Louis, MO.
9. Gasteiger J, Rudolph C, Sadowski J. Automatic generation of 3D atomic coordinates for organic molecules. Tetrahedron Comp Method 1990;3:537-547.
10. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule-ligand interactions. J Mol Biol 1982;161(2):269-88.
11. Selassie CD. History of quantitative structure-activity relationships. In: Abraham DJ, editor. Burger's medicinal chemistry and drug discovery. Hoboken: John Wiley & Sons, Inc.; 2003. p. 1-48.

12. Kubinyi H. Free Wilson Analysis - Theory, Applications and Its Relationship to Hansch Analysis. Quantitative Structure-Activity Relationships 1988;7(3):121-133.
13. Hansch C, P.P. M, T. F, R.M. M. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. Nature 1962;194:178-180.
14. Hammet LP. Physical Organic chemistry. Reaction Rates, Equilibria and Mechanism. 2nd ed. New York: McGraw-Hill; 1970.
15. Fujita T, Hansch C, Iwasa J. New Substituent Constant Pi Derived from Partition Coefficients. Journal of the American Chemical Society 1964;86(23):5175-&.
16. Todeschini R, Consonni V, Pavan M. Dragon. Software for the calculation of molecular descriptors, version 1.11; http://www.disat.unimib.it/chm/Dragon.htm.
17. Todeschini R, Consonni V. Handbook of molecular descriptors. Weinhem: Wiley-VCH; 2000.
18. Livingstone DJ. The characterization of chemical structures using molecular properties. A survey. J Chem Inf Comput Sci 2000;40(2):195-209.
19. Randic M. Characterization of molecular branching. J Am Chem Soc 1975;97(23):6609-6615.
20. Kier LB, Hall LH, Murray WJ, Randic M. Molecular connectivity. I: Relationship to nonspecific local anesthesia. J Pharm Sci 1975;64(12):1971-4.
21. Kier LB, Murray WJ, Hall LH. Molecular connectivity. 4. Relationships to biological activities. J Med Chem 1975;18(12):1272-4.
22. Cramer RD, Patterson DE, Bunce JD. Comparative Molecular Field Analysis (CoMFA): 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 1988;110(3):5959-5967.
23. Pastor M, Cruciani G, McLay I, Pickett S, Clementi S. GRid-INdependent descriptors (GRIND): a novel class of alignment- independent three-dimensional molecular descriptors. J Med Chem 2000;43(17):3233-43.

24. Cruciani G, Pastor M, Guba W. VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. Eur J Pharm Sci 2000;11 Suppl 2:S29-39.
25. Potter T, Matter H. Random or rational design? Evaluation of diverse compound subsets from chemical structure databases. J Med Chem 1998;41(4):478-88.
26. Makara GM, Nash H, Zheng Z, Orminati JP, Wintner EA. A reagent-based strategy for the design of large combinatorial libraries: a preliminary experimental validation. Mol Divers 2003;7(1):3-14.
27. Martin YC, Kofron JL, Traphagen LM. Do structurally similar molecules have similar biological activity? J Med Chem 2002;45(19):4350-8.
28. Lemmen C, Lengauer T. Computational methods for the structural alignment of molecules. J Comput Aided Mol Des 2000;14(3):215-32.
29. Kabsch W. Solution for Best Rotation to Relate 2 Sets of Vectors. Acta Crystallographica Section A 1976;32(SEP1):922-923.
30. Hurst T. Flexible 3d Searching - the Directed Tweak Technique. Journal of Chemical Information and Computer Sciences 1994;34(1):190-196.
31. Masek BB, Merchant A, Matthew JB. Molecular Shape Comparison of Angiotensin-Ii Receptor Antagonists. Journal of Medicinal Chemistry 1993;36(9):1230-1238.
32. Lemmen C, Hiller C, Lengauer T. RigFit: A new approach to superimposing ligand molecules. Journal of Computer-Aided Molecular Design 1998;12(5):491-502.
33. Parretti MF, Kroemer RT, Rothman JH, Richards WG. Alignment of molecules by the Monte Carlo optimization of molecular similarity indices. Journal of Computational Chemistry 1997;18(11):1344-1353.
34. Nissink JWM, Verdonk ML, Kroon J, Mietzner T, Klebe G. Superposition of molecules: Electron density fitting by application of Fourier transforms. Journal of Computational Chemistry 1997;18(5):638-645.
35. McMahon AJ, King PM. Optimization of Carbo molecular similarity index using gradient methods. Journal of Computational Chemistry 1997;18(2):151-158.

36. Grant JA, Gallardo MA, Pickup BT. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. Journal of Computational Chemistry 1996;17(14):1653-1666.
37. Petitjean M. Geometric Molecular Similarity from Volume-Based Distance Minimization - Application to Saxitoxin and Tetrodotoxin. Journal of Computational Chemistry 1995;16(1):80-90.
38. Good AC, Hodgkin EE, Richards WG. Utilization of Gaussian Functions for the Rapid Evaluation of Molecular Similarity. Journal of Chemical Information and Computer Sciences 1992;32(3):188-191.
39. Mestres J, Rohrer DC, Maggiora GM. MIMIC: A molecular-field matching program. Exploiting applicability of molecular similarity approaches. J Comput Chem 1997;18(7):934-954.
40. Sanz F, Manaut F, Rodriguez J, Lozoya E, Lopezdebrinas E. Mepsim - a Computational Package for Analysis and Comparison of Molecular Electrostatic Potentials. Journal of Computer-Aided Molecular Design 1993;7(3):337-347.
41. Nussinov R, Wolfson HJ. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. Proc Natl Acad Sci U S A 1991;88(23):10495-9.
42. Martin YC, Bures MG, Danaher EA, Delazzer J, Lico I, Pavlik PA. A Fast New Approach to Pharmacophore Mapping and Its Application to Dopaminergic and Benzodiazepine Agonists. Journal of Computer-Aided Molecular Design 1993;7(1):83-102.
43. Sheridan RP, Nilakantan R, Dixon JS, Venkataraghavan R. The ensemble approach to distance geometry: application to the nicotinic pharmacophore. J Med Chem 1986;29(6):899-906.
44. Crippen GM. Distance geometry approach to rationalizing binding data. J Med Chem 1979;22(8):988-97.
45. Crippen GM. Quantitative structure-activity relationships by distance geometry: thyroxine binding site. J Med Chem 1981;24(2):198-203.

46. Crippen GM. Quantitative structure-activity relationships by distance geometry: systematic analysis of dihydrofolate reductase inhibitors. J Med Chem 1980;23(6):599-606.
47. Carey RN, Wold S, Westgard JO. Principal component analysis: an alternative to "referee" methods in method comparison studies. Anal Chem 1975;47(11):1824-9.
48. Hoskuldsson A. PLS regression methods. J Chemometrics 1988;2:211-228.
49. Baroni M, Costantino G, Cruciani G, Riganelli D, Valigi R, Clementi S. Generating Optimal Linear Pls Estimations (Golpe) - an Advanced Chemometric Tool for Handling 3d-Qsar Problems. Quantitative Structure-Activity Relationships 1993;12(1):9-20.
50. Makara GM. Measuring molecular similarity and diversity: total pharmacophore diversity. J Med Chem 2001;44(22):3563-71.
51. McGregor MJ, Muskal SM. Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. J Chem Inf Comput Sci 1999;39(3):569-74.
52. Matter H, Potter T. Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets. J Chem Inf Comput Sci 1999;39:1211-1225.
53. Mason JS, Morize I, Menard PR, Cheney DL, Hulme C, Labaudiniere RF. New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. J Med Chem 1999;42(17):3251-64.
54. Baumann K. Distance Profiles (DiP): A translationally and rotationally invariant 3D structure descriptor capturing steric properties of molecules. Quantitative Structure-Activity Relationships 2002;21(5):507-519.
55. Baumann K. An alignment-independent versatile structure descriptor for QSAR and QSPR based on the distribution of molecular features. Journal of Chemical Information and Computer Sciences 2002;42(1):26-35.
56. Ferguson AM, Heritage T, Jonathon P, Pack SE, Phillips L, Rogan J, et al. EVA: a new theoretically based molecular descriptor for use in QSAR/QSPR analysis. J Comput Aided Mol Des 1997;11(2):143-52.

57. Tuppurainen K, Viisas M, Laatikainen R, Perakyla M. Evaluation of a novel electronic eigenvalue (EEVA) molecular descriptor for QSAR/QSPR studies: validation using a benchmark steroid data set. J Chem Inf Comput Sci 2002;42(3):607-13.
58. Bursi R, Dao T, van Wijk T, de Gooyer M, Kellenbach E, Verwer P. Comparative spectra analysis (CoSA): spectra as three-dimensional molecular descriptors for the prediction of biological activities. J Chem Inf Comput Sci 1999;39(5):861-7.
59. Todeschini R, Lasagni M. New Molecular Descriptors for 2d and 3d Structures - Theory. Journal of Chemometrics 1994;8(4):263-272.
60. Silverman BD, Platt DE, Pitman M, Rigoutsos I. Comparative molecular moment analysis (CoMMA). Perspectives in Drug Discovery and Design 1998;12:183-196.
61. Schuur JH, Selzer P, Gasteiger J. The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. Journal of Chemical Information and Computer Sciences 1996;36(2):334-344.
62. Klein CT, Kaiblinger N, Wolschann P. Internally defined distances in 3D-quantitative structure-activity relationships. J Comput Aided Mol Des 2002;16(2):79-93.
63. Klein CT, Kaiser D, Ecker G. Topological distance based 3D descriptors for use in QSAR and diversity analysis. J Chem Inf Comput Sci 2004;44(1):200-9.
64. Wagener M, Sadowski J, Gasteiger J. Autocorrelation of Molecular-Surface Properties for Modeling Corticosteroid-Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. Journal of the American Chemical Society 1995;117(29):7769-7775.
65. Broto P, Moreau G, Vandycke C. Molecular-Structures - Perception, Auto-Correlation Descriptor and Sar Studies - Auto-Correlation Descriptor. European Journal of Medicinal Chemistry 1984;19(1):66-70.
66. Fechner U, Franke L, Renner S, Schneider P, Schneider G. Comparison of correlation vector methods for ligand-based similarity searching. J Comput Aided Mol Des 2003;17(10):687-98.

67. Stiefl N, Baumann K. Mapping property distributions of molecular surfaces: algorithm and evaluation of a novel 3D quantitative structure-activity relationship technique. J Med Chem 2003;46(8):1390-407.
68. Crivori P, Cruciani G, Carrupt PA, Testa B. Predicting blood-brain barrier permeation from three-dimensional molecular structure. J Med Chem 2000;43(11):2204-16.
69. Ooms F, Weber P, Carrupt PA, Testa B. A simple model to predict blood-brain barrier permeation from 3D molecular fields. Biochim Biophys Acta 2002;1587(2-3):118-25.
70. Alifrangis LH, Christensen IT, Berglund A, Sandberg M, Hovgaard L, Frokjaer S. Structure-property model for membrane partitioning of oligopeptides. J Med Chem 2000;43(1):103-13.
71. Ekins S, Durst GL, Stratford RE, Thorner DA, Lewis R, Loncharich RJ, et al. Three-dimensional quantitative structure-permeability relationship analysis for a series of inhibitors of rhinovirus replication. J Chem Inf Comput Sci 2001;41(6):1578-86.
72. Ano R, Kimura Y, Shima M, Matsuno R, Ueno T, Akamatsu M. Relationships between structure and high-throughput screening permeability of peptide derivatives and related compounds with artificial membranes: application to prediction of Caco-2 cell permeability. Bioorg Med Chem 2004;12(1):257-64.
73. Cruciani C, Crivori P, Carrupt PA, Testa B. Molecular fields in quantitative structure-permeation relationships: the VolSurf approach. Journal of Molecular Structure-Theochem 2000;503(1-2):17-30.
74. Tarvainen M, Sutinen R, Somppi M, Paronen P, Poso A. Predicting plasticization efficiency from three-dimensional molecular structure of a polymer plasticizer. Pharm Res 2001;18(12):1760-6.
75. Holm R, Hoest J. Successful in silico predicting of intestinal lymphatic transfer. Int J Pharm 2004;272(1-2):189-93.
76. Testa B, Bojarski AJ. Molecules as complex adaptative systems: constrained molecular properties and their biochemical significance. Eur J Pharm Sci 2000;11 Suppl 2:S3-14.

77. Zamora I, Oprea T, Cruciani G, Pastor M, Ungell AL. Surface descriptors for protein-ligand affinity prediction. J Med Chem 2003;46(1):25-33.
78. Cruciani G, Pastor M, Mannhold R. Suitability of molecular descriptors for database mining. A comparative analysis. J Med Chem 2002;45(13):2685-94.
79. Oprea TI, Zamora I, Ungell AL. Pharmacokinetically based mapping device for chemical space navigation. J Comb Chem 2002;4(4):258-66.
80. Filipponi E, Cruciani G, Tabarrini O, Cecchetti V, Fravolini A. QSAR study and VolSurf characterization of anti-HIV quinolone library. J Comput Aided Mol Des 2001;15(3):203-17.
81. Menezes IR, Lopes JC, Montanari CA, Oliva G, Pavao F, Castilho MS, et al. 3D QSAR studies on binding affinities of coumarin natural products for glycosomal GAPDH of Trypanosoma cruzi. J Comput Aided Mol Des 2003;17(5-6):277-90.
82. Leitao A, Andricopulo AD, Oliva G, Pupo MT, de Marchi AA, Vieira PC, et al. Structure-activity relationships of novel inhibitors of glyceraldehyde-3-phosphate dehydrogenase. Bioorg Med Chem Lett 2004;14(9):2199-204.
83. Cianchetta G, Mannhold R, Cruciani G, Baroni M, Cecchetti V. Chemometric studies on the bactericidal activity of quinolones via an extended VolSurf approach. J Med Chem 2004;47(12):3193-201.
84. Staerk D, Skole B, Jorgensen FS, Budnik BA, Ekpe P, Jaroszewski JW. Isolation of a library of aromadendranes from Landolphia dulcis and its characterization using the VolSurf approach. J Nat Prod 2004;67(5):799-805.
85. Fontaine F, Pastor M, Sanz F. Potential usefulness of the GRIND descriptors for obtaining 3D-QSAR models without supervision. In: Poster presented at the XIIth nacional congress of the spanish society of medicinal chemistry; 2001; Sevilla; 2001.
86. Benedetti P, Mannhold R, Cruciani G, Pastor M. GBR compounds and mepyramines as cocaine abuse therapeutics: chemometric studies on selectivity using grid independent descriptors (GRIND). J Med Chem 2002;45(8):1577-84.

87. Benedetti P, Mannhold R, Cruciani G, Ottaviani G. GRIND/ALMOND investigations on CysLT(1) receptor antagonists of the quinolinyl(bridged)aryl type. Bioorg Med Chem 2004;12(13):3607-17.
88. Prusis P, Dambrova M, Andrianov V, Rozhkov E, Semenikhina V, Piskunova I, et al. Synthesis and quantitative structure-activity relationship of hydrazones of N-amino-N'-hydroxyguanidine as electron acceptors for xanthine oxidase. J Med Chem 2004;47(12):3105-10.
89. Ballistreri FP, Barresi V, Benedetti P, Caltabiano G, Fortuna CG, Longo ML, et al. Design, synthesis and in vitro antitumor activity of new trans 2-[2-(heteroaryl)vinyl]-1,3-dimethylimidazolium iodides. Bioorg Med Chem 2004;12(7):1689-95.
90. Afzelius L, Masimirembwa CM, Karlen A, Andersson TB, Zamora I. Discriminant and quantitative PLS analysis of competitive CYP2C9 inhibitors versus non-inhibitors using alignment independent GRIND descriptors. J Comput Aided Mol Des 2002;16(7):443-58.
91. Afzelius L, Zamora I, Masimirembwa CM, Karlen A, Andersson TB, Mecucci S, et al. Conformer- and alignment-independent model for predicting structurally diverse competitive CYP2C9 inhibitors. J Med Chem 2004;47(4):907-14.
92. Crivori P, Zamora I, Speed B, Orrenius C, Poggesi I. Model based on GRID-derived descriptors for estimating CYP3A4 enzyme stability of potential drug candidates. J Comput Aided Mol Des 2004;18(3):155-66.
93. Cruciani G, Benedetti P, Caltabiano G, Condorelli DF, Fortuna CG, Musumarra G. Structure-based rationalization of antitumor drugs mechanism of action by a MIF approach. Eur J Med Chem 2004;39(3):281-9.
94. Gutierrez-de-Teran H, Centeno NB, Pastor M, Sanz F. Novel approaches for modeling of the A1 adenosine receptor and its agonist binding site. Proteins 2004;54(4):705-15.
95. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein-ligand docking using GOLD. Proteins 2003;52(4):609-23.
96. Agrafiotis DK. Multiobjective optimization of combinatorial libraries. Mol Divers 2002;5(4):209-30.

97. Gillet VJ, Khatib W, Willett P, Fleming PJ, Green DV. Combinatorial library design using a multiobjective genetic algorithm. J Chem Inf Comput Sci 2002;42(2):375-85.

98. Jamois EA, Lin CT, Waldman M. Design of focused and restrained subsets from extremely large virtual libraries. J Mol Graph Model 2003;22(2):141-9.

99. Cosgrove DA, Bayada DM, Johnson AP. A novel method of aligning molecules by local surface shape similarity. J Comput Aided Mol Des 2000;14(6):573-91.

100. Goldman BB, Wipke WT. Quadratic shape descriptors. 1. Rapid superposition of dissimilar molecules using geometrically invariant surface descriptors. J Chem Inf Comput Sci 2000;40(3):644-58.

101. Dezi C. Personnal communication.

102. Zupan J, Gasteiger J. Neural Networks in Chemistry and Drug Design: An Introduction, 2nd Edition. Weinheim: Wiley-VCH; 1999.

103. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Belmont, CA; 1984.

104. Mason JS, Good AC, Martin EJ. 3-D pharmacophores in drug discovery. Curr Pharm Des 2001;7(7):567-97.

PAPER I

Molecular Diversity, **6:** 135–147.
KLUWER/ESCOM

Full Paper

Use of alignment-free molecular descriptors in diversity analysis and optimal sampling of molecular libraries

Fabien Fontaine, Manuel Pastor, Hugo Gutiérrez-de-Terán, Juan J. Lozano & Ferran Sanz*
Research Group on Biomedical Informatics (GRIB), IMIM, Universitat Pompeu Fabra, C/ Dr. Aiguader 80, Barcelona, Spain
(Author for correspondence, E-mail: fsanz@imim.es, Fax: +34 932 240 875)*

Received 5 May 2003; Accepted 11 July 2003

Key words: Almond, diversity, GRIND, k-means clustering, molecular descriptors, molecular library sampling, principal component analysis, quantum-mechanical descriptors, VolSurf

Summary

The selection of a sample of diverse compounds is a common strategy for exploring large molecular libraries. However, the success of such approach depends on the selection of relevant molecular descriptors and the use of appropriate sampling methods. In the context of pharmaceutical research, the molecular descriptors should be based on physicochemical properties related with the pharmacological behaviour of the compounds. In this sense, the alignment-free GRIND and VolSurf molecular descriptors are promising candidates since they have been successfully used in the modelling of both pharmacodynamic and pharmacokinetic properties of drugs. This work describes the use of such descriptors in the diversity sampling of a library of primary amines and compares the results with those obtained in a previous study that used quantum-mechanical descriptors. As in the previous work, principal component (PC) analysis was applied to reduce the dimensionality and remove redundant information of the original descriptors, and the compounds were sampled on the basis of k-means clustering on the space of the selected PCs. The results of the present study show that VolSurf and GRIND provide similar quality sampling regarding global features of the molecules such as hydrophilicity, however the topology of the compounds is considered differently. The similarity between particular compounds strongly depends on the original descriptors used. However all the sample selections done in the PC space after k-means clustering provide the same apparent diversity in comparison to the whole dataset. The results indicate that there is no best set of descriptors on a diversity basis. The selection of descriptors must be based on the drug features to be investigated.

Abbreviations: 2-D: 2-Dimensional; 3-D: 3-Dimensional; GRIND: GRid-INdependent Descriptors; MACC2: Maximum Auto- and Cross-Correlation; MIF: Molecular Interaction Field; PC: Principal Component; PCA: Principal Component Analysis; QM: Quantum-Mechanical; QSAR: Quantitative Structure-Activity Relationships.

Introduction

Nowadays, combinatorial chemistry technologies are able to generate extremely large series of compounds. However, due to practical and economical reasons, it is more convenient to design and synthesize smaller but informative subsets. Such a design requires the selection of appropriate samples of synthesis building blocks from extensive reagents databases. Unfortunately, there is not a single, well-established method for this sampling exercise. In many cases, the evaluated samples have to be informative in relation to the variation of relevant chemical features. Then, random selection is not a relevant choice since reagent databases are often crowded by clusters of compounds that tend to be reproduced by the random-selected samples, and each compound of a cluster is not very informative in comparison to the other cluster members. For this

reason, alternative sampling methods have been postulated to increase diversity, to cover the database better, and to generate a greater amount of hit rates [1, 2]. The idea of looking for diversity in library design is not new and has been extensively reviewed [3–9].

The methods for selecting a subset of diverse compounds differ in two key aspects: the descriptors used to measure the similarity between the compounds, and the mathematical algorithms used for the data treatment and effective selection of the evaluated samples. In a previous paper [10], we postulated the use of principal component analysis (PCA) in order to remove the redundant information of the original descriptors, as well as the sampling of the compounds on the basis of their k-means clustering on the space of the selected PCs. In this previous publication, we used quantum-mechanical descriptors without considering other alternatives.

Many molecular descriptors have been proposed and thoroughly reviewed by other authors [11, 12]. Descriptors are often classified according to their 'dimensionality', which refers to the representation of the molecule used to calculate the descriptor. 1-D descriptors do not make use of the connectivity or the tridimensional geometry of the molecules, (e.g., experimental values or molecular weight). 2-D descriptors are computed from the formulae (molecular graph) of the compound (e.g., count of hydrogen bond acceptor atoms, fingerprint of structural fragments, or connectivity indices). The computation of 3-D descriptors requires information about a feasible geometry of the molecule. They can be as simple as a scalar, such as dipolar moment, or more sophisticated, such as pharmacophoric fingerprints. In addition, diverse types of descriptors can be combined to obtain a more comprehensive characterisation of the studied compounds. For instance, Mason and Beno optimised the diversity of a virtual library of compounds simultaneously using four-point pharmacophoric fingerprints and BCUT chemistry space descriptors [13]. There is still an active research work on molecular descriptors with potential use in the diversity analysis of molecular libraries, and all the three classes of descriptors: 1-D [14], 2-D [15–18] and 3-D [19, 20] have received recent contributions from the scientific community.

In our opinion, the most important characteristic of the molecular descriptors in the framework of drug design projects is their ability to describe structural and physicochemical features that affect the biological behaviour of the compounds. The same requirement of descriptors is shared by Quantitative Structure-Activity Relationship (QSAR) studies, where the choice of relevant molecular descriptors is also critical. This fact led us to focus our attention on some new descriptors published in the QSAR field, in order to check their applicability in molecular diversity analyses.

State-of-the-art molecular descriptors used in QSAR often involve the computation of molecular interaction fields (MIF). MIF can be obtained using several programmes for example GRID [21], which calculates energies of interaction between the studied molecule (e.g., the building block), and relevant chemical probes (e.g., water, amine nitrogen, etc.). The probe is sequentially placed at the nodes of a grid surrounding the molecule, and the energy of interaction is calculated for each position of the probe (Figure 1a). MIF have been used with success for more than two decades in 3-D-QSAR methods such as CoMFA [22] or GRID/GOLPE [23]. However, the use of MIF for the analysis of series of molecules has the drawback of requiring the alignment of the structures as a preliminary step. Such an alignment is always subjective and often time-consuming, introducing important drawbacks from a practical point of view. For this reason, a new generation of MIF-based, alignment-independent descriptors has been developed recently. The first representative of such new kind of descriptors is the VolSurf method [24], created with the aim of predicting pharmacokinetic properties such as blood-brain barrier permeation [25]. Interestingly, VolSurf descriptors were also applied to the prediction of ligand binding affinity although they were not designed for this purpose [26]. Grid-INdependent Descriptors (GRIND) [27], here called Almond descriptors after the name of the software used to generate them, also belong to the new kind of MIF-based, alignment-free generation of descriptors. GRIND were specifically designed to characterise pharmacodynamic properties such as protein-ligand interactions. Published applications of Almond descriptors include traditional 3-D-QSAR studies [27, 28] and cytochrome metabolism analyses [29]. Recently, Cruciani et al. published a comparative study about the use of Unity 2-D fingerprints, Log P, VolSurf and Almond descriptors for database mining [30], and Oprea et al. compared VolSurf vs. common 2-D descriptors for chemical space definition with respect to permeability and solubility [31].

Although Almond and VolSurf descriptors make use of the same source of information (i.e., the GRID

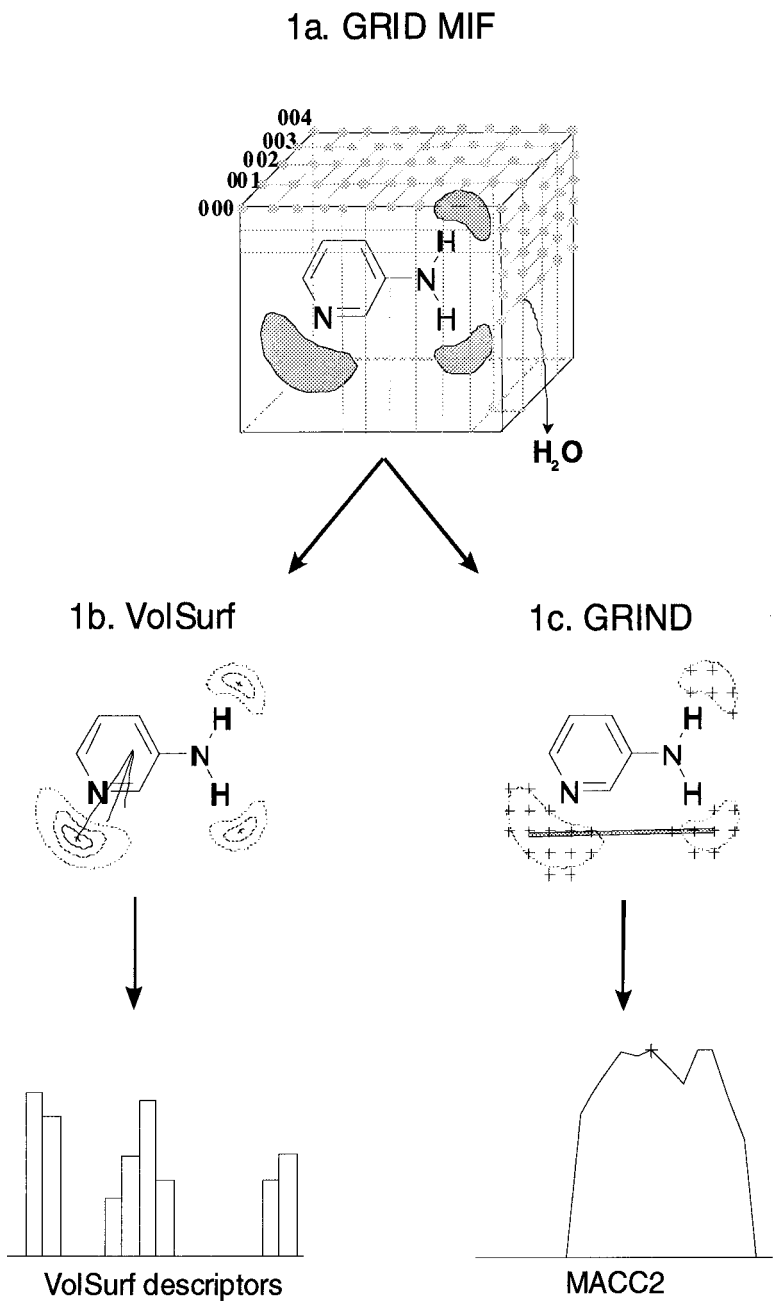


Figure 1. Computation of VolSurf and Almond descriptors. (1a) Calculation of the GRID molecular interaction field (MIF) using the water probe. (1b) Calculation of VolSurf descriptors, e.g., volume of favourable interaction with water, water integy moment, etc. (1c) GRIND calculation; distances between pharmacophoric regions are encoded into Maximum Auto- and Cross-Correlation (MACC2) correlograms.

MIF), they differ significantly on their underlying approaches, which are summarised in Figure 1. The VolSurf descriptors summarize the MIF information in a few quantitative variables easy to understand and to interpret. Basically, the Volsurf descriptors refer to the size and shape of the molecule, to the size and shape of the hydrophilic and hydrophobic regions, and to the balance between them. Conversely, Almond descriptors are focused on the identification of optimal interaction sites, and on the description of the geometrical relationship between such sites.

In order to test the suitability of the aforementioned descriptors (VolSurf and Almond) for the characterization and selection of dissimilar reagents, we present a comparative study aiming to the selection of an optimal subset of primary amines from a database of 746 candidates extracted from the Aldrich catalogue. An analogous study on the same dataset was previously carried out by our group [10]. The only difference consisted in the descriptors used, which were quantum-mechanical (QM) derived parameters, molecular weight and solvent-accessible surface area. The selection of the compounds was carried out after k-means clustering in a space of principal components.

Materials and methods

Molecular structures

Except for the molecular weight, the calculation of all the descriptors requires the conversion of the 2-D structures of the compounds into feasible 3-D structures. In the present study, we have used the CORINA software for such a task [32].

The calculation of the QM descriptors requires carrying out further geometrical optimisation before calculating the descriptors. The optimisation was performed using semi-empirical methods and the AM1 Hamiltonian, simulating the aqueous environment by means of the SM5.4A solvation model [33] as implemented in AMSOL 6.3 [34].

Molecular descriptors

AMSOL

AMSOL 6.3 [34] was used to compute the six descriptors reported in our previous study [10]: Molecular weight, dipolar moment, solvent-accessible surface area and free energies of solvation of the considered compound in water, benzene and octanol. Details about the choice of the descriptors and their computation can be found in the previous publication [10].

VolSurf

VolSurf descriptors were generated with VolSurf 3.0.7c [35]. The chemical probes used were OH2 (water), DRY (hydrophobic) and N1 (amidic nitrogen). N1 is a hydrogen bond donor probe that exclusively highlights the hydrogen bond acceptor centres. Consequently, it offers complementary information in comparison to the water probe, which informs on all the possible hydrogen bonding centres without paying attention to their donor or acceptor character. The VolSurf computation produced a set of 72 descriptors for each compound. A detailed explanation of the VolSurf methodology is given elsewhere [24].

Almond

GRIND were generated with Almond 3.2.0 [36], using the default set of GRID probes: DRY (hydrophobic), O (carbonyl oxygen), N1 (amide nitrogen). For each compound, GRIND variables are grouped into six blocks called Maximum Auto- and Cross-Correlation (MACC2) correlograms. Three of them inform on the presence of pairs of optimal interaction sites with the same probe at particular distances (i.e., DRY-DRY, N1-N1, O-O), and the remaining correlograms inform on pairs of optimal interactions but with different probes (i.e., DRY-O, DRY-N1 and O-N1). Each correlogram is a compact representation of the geometrical relationship between energetically favourable regions of the MIFs. An example of MACC2 calculation is shown in Figure 1.

The Almond filtering parameters were modified to provide a better coverage of the relevant interactions: the number of optimal interaction nodes to extract was set to 150 nodes and the field weight was decreased to 40%. A total of six blocks of 78 variables was obtained for each compound. More details on the methodology can be found in the original GRIND article [27].

Principal Component Analysis (PCA)

In few words, the basic principle of PCA [37] is to project the objects (here the compounds) from the original n-dimensional space of descriptors into another space of lower dimensionality, preserving the major part of the similarities and differences between the objects, but removing the redundancy between the variables. The variables (or axes) of the new lower-dimension space, called principal components or PCs, are obtained by linear combination of the original variables, and therefore it is possible to quantify the contribution of each original variable to each PC.

PCA was carried out using the SPSS 11.0 software [38]. AMSOL and VolSurf descriptors were autoscaled since they contain variables with different units. Conversely, Almond descriptors were not modified, since the scale is meaningful. In our previous work, a Varimax rotation was applied to the AMSOL descriptors in order to obtain an optimal separation

between contributions of the original descriptors for each PC [10]. In order to be consistent with the previous study a Varimax transform was also applied to the PCA on the VolSurf and Almond descriptors.

Descriptors similarity assessment

The overall similarity between the sets of descriptors produced by the three programmes was assessed by comparing the matrices of Euclidean distances between the compounds in the three spaces of descriptors. If the compounds were described in an equivalent manner by two sets of descriptors, the distances between the compounds in each descriptor space should be correlated. The correlation coefficient used to compare distance matrices was the Pearson correlation coefficient (R):

$$R = \sum_{i=0}^{i=N} \frac{(x_i - x_m)(y_i - y_m)}{\sqrt{(x_i - x_m)^2 (y_i - y_m)^2}} ,$$

where N is the number of pairwise distances in the upper triangle of the distance matrices **X** and **Y** (only matrix elements above the diagonal are considered since the distance matrices are symmetrical), x_i and y_i are pairwise distances in such matrices, and x_m and y_m are the means of each series of N distances.

Diversity assessment

The diversity of each dataset sampling was assessed using the refined version of the R^2-like diversity coefficient proposed by Gutiérrez de Terán et al. [10], which was defined as follows:

$$\text{Total diversity prior clustering} = D_T = \sum_{i=1}^{N} d_{ic}^2$$

$$\text{Diversity lost within the clusters} = D_{WC} =$$

$$= \sum_{j=1}^{n} \sum_{i=1}^{n_j} d_{ic_j}^2$$

$$R^2 = \frac{\text{Remaining diversity after clustering}}{\text{Total diversity}} =$$

$$= \frac{D_T - D_{WC}}{D_T} ,$$

where d_{ic} are Euclidean distances computed in the PC space, N and n are the number of compounds of the whole set and the sample respectively, c is the compound nearest to the centroid for the whole set, and n_j and c_j are the size and the compound nearest to the centroid for the jth-cluster, respectively.

Clustering

The sampling of the dataset was carried out by first performing a k-means clustering of the compounds [39] on the basis of Euclidean distances as implemented in SPSS 11.0 [38]. k-means is a non-hierarchical, iterative clustering method. The clustering process starts with the definition of k arbitrary centroids, then the compounds are assigned to the nearest centroid and k clusters are obtained, then centroids are recomputed and compounds are reassigned if a new nearest centroid exists. The two last steps are repeated iteratively until stable classification. Once the clustering algorithm is completed, the sample is obtained by picking from each cluster the compound nearest to its centroid. Obviously, the number of clusters has to be equal to the planned sample size. In the present study and for the intended comparison purposes, analyses were repeated for different sample sizes: 10, 25, 50, 100 and 200.

Results and discussion

Computation of the descriptors

The starting reagent database consists in a series of 923 primary amines, extracted from the Aldrich catalogue. After filtering out of the molecules unsuitable for quantum-mechanical calculation, 746 compounds remained in the database. Their structures were first converted from 2-D to 3-D with CORINA and the resulting structures were used to compute VolSurf and Almond descriptors. The calculation of the AMSOL descriptors required some further geometrical optimisation by means of AMSOL quantum-mechanical calculations. Although VolSurf and Almond descriptors could have been computed using the AMSOL-optimised structures without any additional computational cost in the framework of the present study, we decided to avoid computationally expensive quantum-mechanical optimisations by directly using the CORINA geometries, with the aim of testing a standard and automatic protocol for the use of VolSurf or Almond descriptors in diversity analyses.

PCA including Varimax rotation was performed for the three sets of descriptors: AMSOL, VolSurf

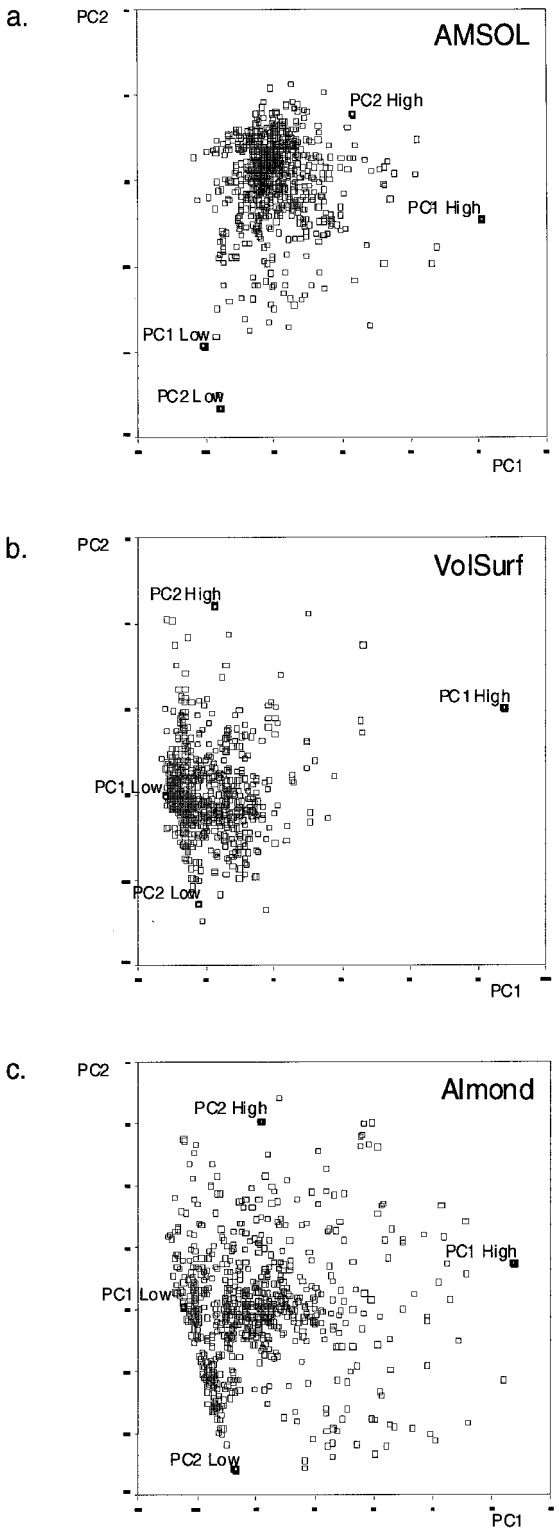


Figure 2. Plot of the scores of the 746 amines of the database in the plane defined by the first two PCs. (a) AMSOL, (b) VolSurf, (c) Almond. Extreme compounds shown in Tables 1, 2 and 3 are marked.

and Almond. The PCA results obtained with AMSOL descriptors were already described in detail by Gutierrez de Teran et al. [10] and are summarized here in Table 1. The results obtained from the PCA on VolSurf and Almond descriptors are summarized in Tables 2 and 3, respectively. The number of PCs chosen was 3, 4 and 4 for AMSOL, VolSurf and Almond, respectively. The position of each compound in the space of the two first PCs for each set of descriptors is plotted in Figure 2.

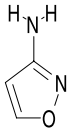
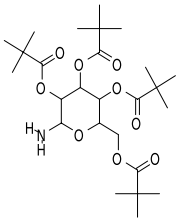
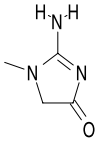
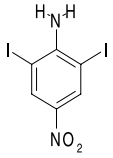
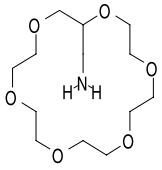
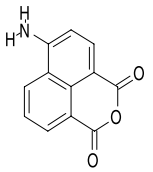
For the three methods, the first two PCs essentially account for global molecular features (i.e., size, solvation, hydrophilicity and hydrophobicity), while the remaining PCs depend more on the spatial distribution of certain substituents of the compounds (i.e., dipolar moment, VolSurf integy moment, specific combination of GRIND variables). The total variance explained is much higher for AMSOL descriptors (around 96%) than for MIF-based methods (around 69% for both methods). This is principally due to the differences in the number of original variables between the different sets of descriptors: AMSOL, VolSurf and Almond descriptors contain 6, 72 and 468 variables, respectively. The higher number of variables of the Almond and VolSurf descriptors is not detrimental for the analysis of the PCA loadings because the variable coefficients are naturally grouped in few blocks along the loading profile.

The first two PCs of VolSurf and Almond descriptors are closely related. PC1 of VolSurf description is mainly influenced by medium and high energy interactions with water, while PC1 of the Almond description is mainly affected by the hydrogen bond acceptor capabilities of the molecule. Consequently, the first PC of both Almond and VolSurf descriptions is highly correlated with the hydrophilicity of the molecules.

Regarding PC2, the interaction with the DRY (hydrophobic) probe is underlying both Almond and VolSurf descriptions. However, we noticed slight differences between the second PCs: the size of the molecule has a higher and more explicit contribution to the VolSurf description than to the Almond one. In addition, the Almond PC2 highlights a particular distance between a hydrogen bond donor site and the amino group (see Table 2 for an example of this feature).

PC3 and PC4 of both Almond and VolSurf descriptions contain information on the spatial distribution of the chemical groups in the structures. In particular, VolSurf PC3 mainly separates the compounds accord-

Table 1. PCA on AMSOL descriptors

Principal component	% variance explained	Lowest PC score representative	Highest PC score representative
PC1 (steric)	41.4		
PC2 (solvation)	36.4		
PC3 (dipolar moment)	18.2		

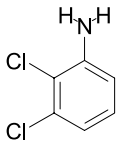
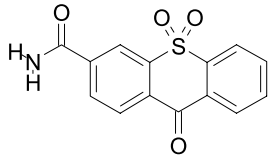
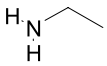
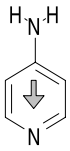
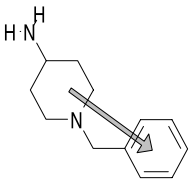
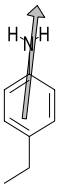
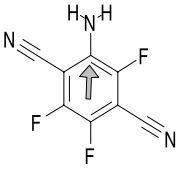
ing to the hybridization state of the amino group: planar amines such as anilines have low PC3 score whilst aliphatic amines have high PC3 score.

It has to be pointed out that VolSurf and Almond descriptors deal with the position of the interacting groups in a different way. In Almond, each GRIND variable corresponds to a particular distance between two pharmacophoric regions. Since the only common hydrogen-bonding centre of all the compounds of the studied series is the amino group, it is involved in most of the variables strongly influencing PC3 and PC4 of the Almond description. In VolSurf, the topology of the interaction sites between the molecule and the DRY and OH2 probes is mainly described with the integy moments. The integy moment of a particular probe is a vector starting from the centre of mass of the molecule and ending at the centre of mass of the interaction regions with the probe at a given energy level (e.g., 3 kcal mol^{-1}). It is therefore an indicator of the global distribution of a specific type of interaction around the molecule. In the present study, the VolSurf and Almond descriptors share a common concept: the comparison of distributions of interaction sites requires a reference point. However, they differ in the implementation of such a concept, in VolSurf the required 'anchor' point is the centre of mass of the molecule, while in Almond the 'anchor' point is the hydrogen bonding centre related with the amino group.

Descriptors similarity assessment

The PCAs provide a good indication of the type of information contained in each set of descriptors but they do not provide a quantification of their similarity. To afford such quantification, matrices of pairwise compound distances were computed for the three PC spaces and the resulting series of distances were compared for every pair of methods using the Pearson correlation coefficient. The correlation coefficients obtained are shown in Table 4. They are fairly consistent with the characteristics of the original variables underlying each PCA. VolSurf and Almond are the most similar (R = 0.58) since they are both based on GRID molecular interaction fields. AMSOL descriptors are little dependant on local features of the considered molecules. Conversely, Almond descriptors mainly rely on the local arrangement of the atoms, which explains the low correlation coefficient (R = 0.37) between

Table 2. PCA on VolSurf descriptors

Principal component	% variance explained	Lowest PC score representative	Highest PC score representative
PC1 (hydrophilicity)	21.6		
PC2 (size and π systems)	18.4		
PC3 (amine geometry and π systems position)[a]	15.6		
PC4 (distribution of hydrophilic interactions)[b]	13.9		

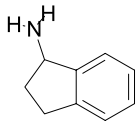
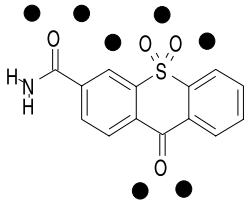
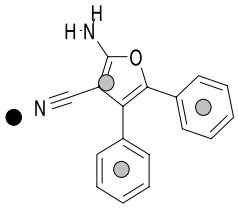
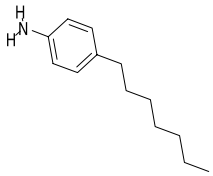
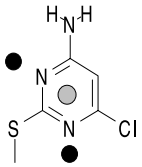
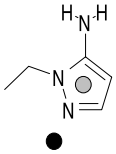
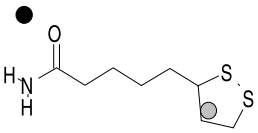
[a] Grey arrow represents DRY integy moment.
[b] Grey arrow represents water integy moment.

Almond and AMSOL descriptors. Lots of VolSurf descriptors have a global nature, which explains the intermediate correlation coefficient between VolSurf and AMSOL (R = 0.50).

In order to further analyse the differences between the three sets of descriptors, we arbitrarily took two compounds from Table 3 and we ranked all the compounds of the database according to their distance to these two seed compounds. The first seed is 5-amino-1-ethylpyrazole, which has a planar amino group, an aromatic heterocycle, a hydrogen bond acceptor nitrogen and a short aliphatic substituent. The second seed is 6-thioctic amide, which has a planar amide nitrogen, a hydrogen bond acceptor atom, and an aliphatic chain terminated with a dithiolane ring. For each seed, a compound that is very close to it in one of the spaces of descriptors, but far in the others was chosen, so that the seed and its neighbour only cluster using a particular set of descriptors. The seeds and their neighbours are shown in Table 5.

For the first seed, all neighbours have approximately the same size and contain a planar amino group and a heterocyclic system. However, they differ in particular features that can be discriminated by some descriptors and not by the others. For instance, when considering the first seed and the first neighbour, it seems that similarity analysis using VolSurf descriptors focus on the existence of two hydrogen-bond acceptor atoms and a planar ring and not on other issues like the different relative disposition of the hydrogen-bond acceptor atoms. When considering the first seed and the second neighbour, it appears that the polysubstitution with chlorines is not detected with Almond descriptors since such atoms are very weak

Table 3. PCA on Almond descriptors

Principal component	% variance explained	Lowest PC score representative	Highest PC score representative
PC1 (Interaction with N1 probe)	21.3		
PC2 (Interaction with DRY probe, N1 interaction far from the NH_2)	20.3		
PC3 (distances between DRY-N1, N1-N1 and NH_2-N1 sites)	19.4		
PC4 (distances between NH_2 and DRY/N1 sites)	7.7		

Grey circles: hydrophobic sites (DRY probe).
Black circles: hydrogen-bond donor sites (N1 probe).

hydrogen-bond acceptors, while such feature is represented by VolSurf and AMSOL descriptors. In the case of the second seed, the VolSurf neighbour share the amide with it, but the shape of the aliphatic moiety is fairly different from that of the 6-thioctic amide. The AMSOL neighbour has similar size but many more hydrogen bond acceptor atoms than 6-thioctic amide, which indicates that the interactions with water are modelled in a very different way at the semi-empirical and at the molecular mechanics levels. The Almond neighbour displays a pyramidal amino group instead of the amide group of the seed. Both groups are somewhat similar because they display both hydrogen-bond acceptor and hydrogen-bond donor capabilities: the lone pair of the pyramidal amino group can interact with the hydrogen bond donor probe like the amide carbonyl of the seed does. However, if the Almond descriptors had been obtained using a protonated amine, the similarity with the seed would have been much lower.

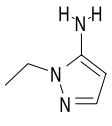
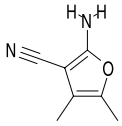
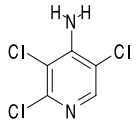
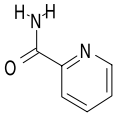
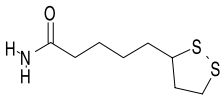
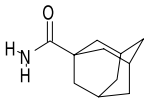
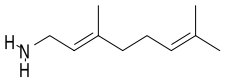
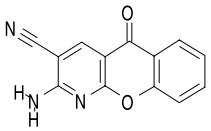
Table 4. Pearson correlation coefficients between similarity matrices

	AMSOL	VolSurf	Almond
AMSOL	1.00		
VolSurf	0.50	1.00	
Almond	0.37	0.58	1.00

Diversity assessment

k-means clustering was carried out on the PC spaces described above (using AMSOL, VolSurf and Al-

Table 5. Comparisons of similarities between compounds

Seed	Neighbour	Ranking		
		VolSurf	Almond	AMSOL
		5	113	48
		598	3	294
		526	276	2
		1	191	263
		228	1	232
		655	182	4

mond descriptors) with the number of clusters ranging between 10 and 200. The sample selection was performed by picking a representative for each cluster (the nearest to the cluster centroid) and therefore the sample size was equal to the number of clusters.

The analyses described in the previous sections shed light on the meaning of the novel descriptors but the central question about their suitability for a diverse-based selection of the reagents is still open. In order to compare the quality of the obtained solutions in terms of remaining diversity in comparison to the whole dataset, we used the R^2-like measure of diversity detailed in the Materials and Methods section. The higher the R^2-like coefficient the better the diversity of the sample compared to the diversity of the full database. When we analyse the evolution of the R^2-like coefficient according to the sample size (Figure 3), a hyperbolic relationship between R^2-like coefficient and the sample size can be appreciated. The optimal number of reagents is between 50 and 100 and any further increase in the number of compounds improves only slightly the diversity of the sample. As shown in Figure 3, k-means sampling generates more diverse samples than random picking: the size

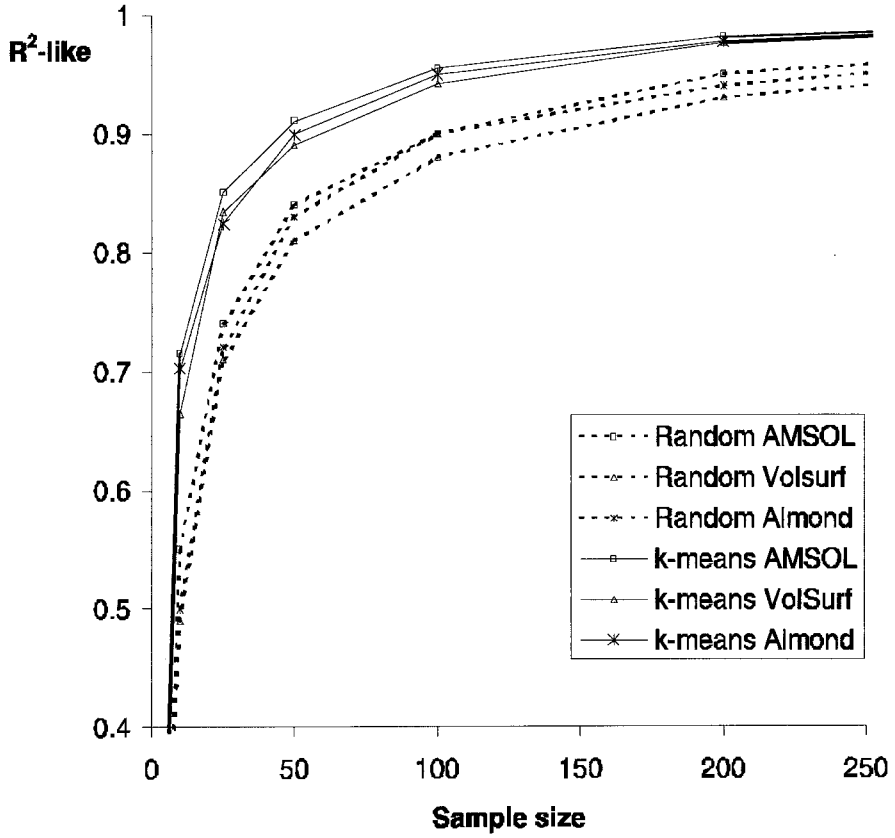


Figure 3. Variation of the R^2-like diversity coefficient according to the sample size, the sampling method (random or k-means) and the descriptors space (AMSOL, VolSurf or Almond).

of a random sample must be the double of a k-means sample to obtain a similar R^2-like coefficient. A problem of the random sampling is that it relies on chance to obtain a good subset. Consequently, the molecular diversity of the sample can fluctuate a lot, particularly when its size is small. For example, a 'lucky' random selection of 50 compounds would generate a R^2-like coefficient as good as a k-means sampling (around 0.90), while an 'unlucky' selection would result in a very bad R^2-like coefficient: the minimum R^2-like coefficient value obtained by random sampling of 50 compounds is indeed around 0.74 for all three methods in our tests. k-means clustering provides a more robust way to obtain a suitable subset. In addition, clustering offers the possibility of having different representatives of the same cluster, as they are sometimes needed for synthetic purposes.

The R^2-like coefficient is roughly the same for all the k-means selections of the same number of compounds, independently of the descriptors used. This could be considered somewhat surprising since we have seen above that the similarity relationships between the compounds strongly depend on the descriptors used. However, we must bear in mind that the R^2-like coefficient is only a measure of conservation of diversity in comparison to the whole dataset, it depends on the clustering/dispersion of the compounds in the descriptors space but not on the particular differences between the compounds. Therefore similar R^2-like values indicate that the compounds are similarly spread in different spaces but the relative position of each particular compound may be totally different. Consequently, the clusters resulting from the subset selection might contain very different compounds, which greatly affects the final series.

It has to be pointed out that the claim that a library of compounds is diverse without indicating the descriptors used for the diversity measurement has no meaning. For example, if we take the 50 compounds selected by k-means clustering in the VolSurf

descriptors space and we calculate the R^2-like measure of this sample in the AMSOL descriptors space, the value decreases from 0.89 to 0.85, which indicates that the diversity-optimisation carried out in a certain molecular descriptors space is always partially effective, since the descriptor space used to perform the optimisation is only a partial representation of the chemical space covered by the compounds. Such results explain why, in some cases, screening of theoretically diverse samples may not perform better than screening of random samples. We must be careful of selecting sets of descriptors as relevant as possible for characterizing properties of the compounds in the framework of drug discovery projects.

Conclusions

The three sets of descriptors (i.e., AMSOL, VolSurf and Almond) provide information about both global and spatially dependent properties of the compounds. AMSOL descriptors are focused essentially towards the global features of the compounds, while Almond descriptors depend more on the position of pharmacophoric regions in the molecules. VolSurf provides an interesting mid-point between the characteristics of Almond and AMSOL descriptors. Despite of the common aim of both VolSurf and Almond to describe the interaction capabilities of the studied compounds, the two programmes consider them in a different fashion. Moreover, the ranking of the molecules according to its distance to particular seed compounds highlighted that the three methods focus on very different aspects of the molecular structure. It remains to be checked which one of the set of descriptors would produce better results in explaining biological screening data.

From a diversity viewpoint, there is no reason to prefer a particular set of descriptors versus the other since the distributions of the compounds are similar. The sample size has a great impact on the remaining diversity up to an optimal number of representatives (between 50 and 100 compounds in the present study). The random sampling is not suitable for obtaining optimally diverse samples, especially for sample sizes below this optimal number since the quality of the sample obtained depends greatly on chance.

In summary, there is no single answer to the question of what are the best descriptors for selecting a good diverse subset. The right choice depends on the kind of problem that the selected series is intended to investigate. In the case of drug-design problems, MIF-based descriptors provide a singular way of characterising compounds that adds extra information with respect to classical descriptors and therefore deserve a place among the tools available to the medicinal chemist.

Acknowledgements

We are grateful to the Fondo de Investigaciones Sanitarias for the partial financial support to this research (Grant No.: FIS 01/1330).

References

1. Potter, T. and Matter, H., *Random or rational design? Evaluation of diverse compound subsets from chemical structure databases*, J. Med. Chem., 41 (1998) 478–488.
2. Zheng, W., Cho, S. J., Waller, C. L. and Tropsha, A., *Rational combinatorial library design. 3. Simulated annealing guided evaluation (SAGE) of molecular diversity: A novel computational tool for universal library design and database mining*, J. Chem. Inf. Comput. Sci., 39 (1999) 738–746.
3. Blaney, J. M. and Martin, E. J., *Computational approaches for combinatorial library design and molecular diversity analysis*, Curr. Opin. Chem. Biol., 1 (1997) 54–59.
4. Bures, M. G. and Martin, Y. C., *Computational methods in molecular diversity and combinatorial chemistry*, Curr. Opin. Chem. Biol., 2 (1998) 376–380.
5. Gorse, D. and Lahana, R., *Functional diversity of compound libraries*, Curr. Opin. Chem. Biol., 4 (2000) 287–294.
6. Willett, P., *Chemoinformatics – Similarity and diversity in chemical libraries*, Curr. Opin. Biotechnol., 11 (2000) 85–88.
7. Tropsha, A. and Zheng, W., *Rational principles of compound selection for combinatorial library design*, Comb. Chem. High Throughput Screen., 5 (2002) 111–123.
8. Beavers, M. P. and Chen, X., *Structure-based combinatorial library design: Methodologies and applications*, J. Mol. Graph. Model., 20 (2002) 463–468.
9. Martin, Y. C., *Diverse viewpoints on computational aspects of molecular diversity*, J. Comb. Chem., 3 (2001) 231–250.
10. Gutiérrez-de-Terán, H., Lozano, J. J., Segarra, V. and Sanz, F., *Molecular diversity sample generation on the basis of quantum-mechanical computations and principal component analysis*, Comb. Chem. High Throughput Screen., 5 (2002) 49–57.
11. Gillet, V. J., *Background theory of molecular diversity*, In Dean, P. M. and Lewis, R. A. (eds.), Molecular Diversity in Drug Design, Kluwer Academic Publishers, Dordrecht, 1999, pp. 43–65.
12. Xue, L. and Bajorath, J., *Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening*, Comb. Chem. High Throughput Screen., 3 (2000) 363–372.
13. Mason, J. S. and Beno, B. R., *Library design using BCUT chemistry-space descriptors and multiple four-point pharmacophore fingerprints: Simultaneous optimization and structure-based diversity*, J. Mol. Graph. Model., 18 (2000) 438–451, 538.

14. Dixon, S. L. and Villar, H. O., *Bioactive diversity and screening library selection via affinity fingerprinting*, J. Chem. Inf. Comput. Sci., 38 (1998) 1192–1203.
15. Lipkus, A. H., *Exploring chemical rings in a simple topological-descriptor space*, J. Chem. Inf. Comput. Sci., 41 (2001) 430–438.
16. Barnard, J. M., Downs, G. M., Von Scholley-Pfab, A. and Brown, R. D., *Use of Markush structure analysis techniques for descriptor generation and clustering of large combinatorial libraries*, J. Mol. Graph. Model., 18 (2000) 452–463.
17. Ivanciuc, O. and Klein, D. J., *Computing wiener-type indices for virtual combinatorial libraries generated from heteroatom-containing building blocks*, J. Chem. Inf. Comput. Sci., 42 (2002) 8–22.
18. Rarey, M. and Stahl, M., *Similarity searching in large combinatorial chemistry spaces*, J. Comput. Aided Mol. Des., 15 (2001) 497–520.
19. Consonni, V., Todeschini, R. and Pavan, M., *Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3-D molecular descriptors*, J. Chem. Inf. Comput. Sci., 42 (2002) 682–692.
20. Makara, G. M., *Measuring molecular similarity and diversity: total pharmacophore diversity*, J. Med. Chem., 44 (2001) 3563–3571.
21. Goodford, P. J., *A computational procedure for determining energetically favorable binding sites on biologically important macromolecules*, J. Med. Chem., 28 (1985) 849–857.
22. Cramer, R. D., Patterson, D. E. and Bunce, J. D., *Comparative Molecular Field Analysis (CoMFA): 1. Effect of shape on binding of steroids to carrier proteins*, J. Am. Chem. Soc., 110 (1988) 5959–5967.
23. Cruciani, G. and Watson, K. A., *Comparative molecular field analysis using GRID force-field and GOLPE variable selection methods in a study of inhibitors of glycogen phosphorylase b*, J. Med. Chem., 37 (1994) 2589–2601.
24. Cruciani, G., Pastor, M. and Guba, W., *VolSurf: A new tool for the pharmacokinetic optimization of lead compounds*, Eur. J. Pharm. Sci., 11 Suppl 2 (2000) S29–39.
25. Crivori, P., Cruciani, G., Carrupt, P. A. and Testa, B., *Predicting blood-brain barrier permeation from three-dimensional molecular structure*, J. Med. Chem., 43 (2000) 2204–2216.
26. Zamora, I., Oprea, T., Cruciani, G., Pastor, M. and Ungell, A. L., *Surface descriptors for protein-ligand affinity prediction*, J. Med. Chem., 46 (2003) 25–33.
27. Pastor, M., Cruciani, G., McLay, I., Pickett, S. and Clementi, S., *GRid-INdependent descriptors (GRIND): A.a novel class of alignment-independent three-dimensional molecular descriptors*, J. Med. Chem., 43 (2000) 3233–3243.
28. Benedetti, P., Mannhold, R., Cruciani, G. and Pastor, M., *GBR compounds and mepyramines as cocaine abuse therapeutics: Chemometric studies on selectivity using grid independent descriptors (GRIND)*, J. Med. Chem., 45 (2002) 1577–1584.
29. Afzelius, L., Masimirembwa, C. M., Karlen, A., Andersson, T. B. and Zamora, I., *Discriminant and quantitative PLS analysis of competitive CYP2C9 inhibitors versus non-inhibitors using alignment independent GRIND descriptors*, J. Comput. Aided Mol. Des., 16 (2002) 443–458.
30. Cruciani, G., Pastor, M. and Mannhold, R., *Suitability of molecular descriptors for database mining. A comparative analysis*, J. Med. Chem., 45 (2002) 2685–2694.
31. Oprea, T. I., Zamora, I. and Ungell, A. L., *Pharmacokinetically based mapping device for chemical space navigation*, J. Comb. Chem., 4 (2002) 258–266.
32. Gasteiger, J., Rudolph, C. and Sadowski, J., *Automatic generation of 3-D atomic coordinates for organic molecules*, Tetrahedron Comp. Method., 3 (1990) 537–547.
33. Giesen, D. J., Gu, M. Z., Cramer, C. J. and Truhlar, D. G., *A Universal Organic Solvation Model*, J. Org. Chem., 61 (1996) 8720–8721.
34. AMSOL 6.5.2, Hawkins, G. D., Giesen, D. J., G. C., L., Chambers, C. C., Rossi, I., Storer, J. W., Rinaldi, D., Liotard, D. A., Cramer, C. J. and Truhlar, D. G., University of Minnesota, Minneapolis, 1997.
35. VolSurf 3.0.7c, Cruciani, G., Pastor, M. and Mecucci, S., Molecular Discovery Ltd., Perugia, 2002.
36. Almond 3.2.0, Cruciani, G., Fontaine, F. and Pastor, M., Molecular Discovery Ltd., Perugia, 2003.
37. Carey, R. N., Wold, S. and Westgard, J. O., *Principal component analysis: an alternative to 'referee' methods in method comparison studies*, Anal. Chem., 47 (1975) 1824–1829.
38. SPSS 11.0.1, SPSS inc. Chicago, 2001.
39. Downs, G. M. and Barnard, J. M., *Clustering methods and their uses in computational chemistry*, In Lipkowitz, K. B. and Boyd, D. B. (eds.), Reviews in Computational Chemistry, Wiley-VCH, John Wiley & Sons, Inc., 2002, pp. 1–40.

PAPER II

Paper published as:

Fabien Fontaine, Manuel Pastor, Ferran Sanz
Incorporating molecular shape into the alignment-free GRid-INdependent Descriptors.
Journal of Medicinal Chemistry 2004 May 20;47 (11):2805-15

PAPER III

Available online at www.sciencedirect.com
SCIENCE DIRECT®
European Journal of Medicinal Chemistry 38 (2003) 433–440
EUROPEAN JOURNAL OF MEDICINAL CHEMISTRY
www.elsevier.com/locate/ejmech

Conformationally constrained butyrophenones as new pharmacological tools to study 5-HT_{2A} and 5-HT_{2C} receptor behaviours

José Brea [a], Christian F. Masaguer [b], María Villazón [a], M. Isabel Cadavid [a], Enrique Raviña [b], Fabien Fontaine [c], Cristina Dezi [c], Manuel Pastor [c], Ferran Sanz [c], M. Isabel Loza [a,*]

[a] *Facultad de Farmacia, Departamentos de Farmacología, Universidad de Santiago de Compostela, 15782 Santiago de Compostela, Spain*
[b] *Facultad de Farmacia, Laboratorio de Química Farmacéutica, Departamento de Química Orgánica, Universidad de Santiago de Compostela, 15782 Santiago de Compostela, Spain*
[c] *Research Group on Biomedical Informatics (GRIB), IMIM, Universitat Pompeu Fabra, C/Doctor Aiguader 80, 08003 Barcelona, Spain*

Received 16 October 2002; received in revised form 2 January 2003; accepted 6 January 2003

Abstract

This study presents new pharmacological and molecular modelling studies on a recently described series of conformationally constrained butyrophenones. Alignment-free three-dimensional quantitative structure-activity relationship models developed on the basis of GRid Independent descriptors and partial least squares regression analysis, allow feasible predictions of activity of new compounds and reveal structural requirements for optimal affinity, particularly in the case of the 5-HT_{2A} receptor. The requirements for the 5-HT_{2A} affinity consist in a precise distance between hydrogen bond donor (protonated amino group) and hydrogen bond acceptor groups, as well as an optimal distance between the protonated amino group and the farthest extreme of the compounds. Another significant result has been the characterisation of two structurally similar compounds as interesting pharmacological tools (1-[(4-Oxo-4,5,6,7-tetrahydrobenzo[b]furan-5-yl)ethyl]-4-(6-fluorobenzisoxazol-3-yl)piperidine and 1-[(4-Oxo-4,5,6,7-tetrahydrobenzo[b]furan-6-yl)methyl]-4-(6-fluorobenzisoxazol-3-yl)piperidine). In spite of their structural similarity, the first compound shows clearly higher affinity for the 5-HT_{2C} receptor (about 100 fold) and higher Meltzer ratio (1.17 vs. 0.99) than the second. Moreover, the first compound inhibits arachidonic acid release in a biphasic concentration-dependent way in functional experiments at the 5-HT_{2A} receptor and it acts as inverse agonist at the 5-HT_{2C} receptor, behaviours that are not shown by the second compound.

Keywords: 3D-QSAR models; 5-HT_{2A}; 5-HT_{2C}; Butyrophenone; GRIND; Inverse agonism; Receptor conformations

1. Introduction

Clozapine and other 'atypical antipsychotics' have been shown to be effective against the negative symptoms of schizophrenia. Blockade of 5-HT_2 receptors has been postulated as the origin of the distinctive pharmacological profile of these drugs, causing an incidence of tardive diskinaesia and other extrapyramidal side effects lower than those found with classical antipsychotics [1].

The involvement of 5-HT_2 receptors in the pharmacological profile of atypical antipsychotics is supported by many biological, pharmacological and clinical studies [2,3]. Initially, many studies pointed to the 5-HT_{2A}

Abbreviations: 3D-QSAR, three-dimensional quantitative structure-activity relationship; 5-HT, serotonin; AA, arachidonic acid; CHO, Chinese hamster ovary; D, dopamine; FFD, fractional factorial design; GPCR, G protein-coupled receptor; GRIND, GRID independent descriptors; HBA, hydrogen bond acceptor; HBD, hydrogen bond donor; IP, inositol phosphates; LOO, leave-one-out; LV, latent variable; PLA_2, phospholipase A_2; PLC, phospholipase C; PLS, partial least squares.

* Corresponding author.
E-mail address: ffmabel@usc.es (M. Isabel Loza).

0223-5234/03/$ - see front matter © 2003 Éditions scientifiques et médicales Elsevier SAS. All rights reserved.
doi:10.1016/S0223-5234(03)00054-0

subtype as the most involved in schizophrenia [4–11]. Meltzer and coworkers [12,13] suggested that the ratio of the pK_is of antipsychotic agents at the 5-HT_{2A} and D_2 receptors, reflects the atypical profile; this ratio appears to be >1.12 for atypical antipsychotics, and <1.09 for classical antipsychotics.

However, it seems now possible that some of the effects of atypical antipsychotics that have been attributed to their blockade of the 5-HT_{2A} receptor may instead be due to the blockade of 5-HT_{2C}. In particular, affinity for 5-HT_{2C} is another important feature for discriminating classical from atypical antipsychotics [14]: clozapine, olanzapine, seroquel and other atypical antipsychotics have indeed greater affinity for 5-HT_{2C} (and in addition for 5-HT_{2A}) than for D_2 receptors [15]. Blockade of 5-HT_{2C} has been indicated as responsible for the relatively mild extrapyramidal side effects observed for atypical antipsychotics (since 5-HT_{2C} rather than 5-HT_{2A} blockade can prevent the extrapyramidal side effects induced by haloperidol [16]). On the other hand, 5-HT_{2C} antagonists, by enhancing dopamine release in the cortex, would efficiently counteract the hypofrontality, which contributes to the negative symptoms of schizophrenia [17,18]. Furthermore, a 5-HT_{2C} receptor polymorphism (Cys23Ser) has been associated to psychotic symptoms in Alzheimer [19].

On the other hand, since G protein-coupled receptors (GPCRs), as any other biomolecule, exist as collections of conformations in equilibrium [20,21], the affinity and efficacy of the drugs will be related with their absolute and relative affinity for each conformation of each receptor, as well as with the changes that each drug is able to induce in such conformations. A current challenge in receptor pharmacology is the development of experimental and computational methods, as well as pharmacological tools, able to detect different conformational states related with different physiological and pathological mechanisms.

One of the consequences of the existence of pools of conformations is that GPCRs may adopt different states/conformations promoted by agonists that could differentially activate diverse biochemical pathways [22]. It has been described that 5-HT_{2A} and 5-HT_{2C} receptors, in a ligand-dependent way, differentially couple to phospholipase C (PLC) –mediating inositol phosphates (IP) accumulation– and phospholipase A_2 (PLA_2) –mediating arachidonic acid (AA) release– pathways.

Another consequence of the conformational variability of the receptors is the existence of constitutive activity, since it is known the ability of some conformations of the receptors to couple with G proteins and to signal a cellular response in the absence of agonists [20,23–25]. This finding led to the reclassification of the antagonists into inverse agonists and neutral antagonists, as a function of their ability to lower or leave unchanged, respectively, the basal activity of the system.

Taking this into account, the present study presents new pharmacological and three-dimensional quantitative structure-activity relationship (3D-QSAR) studies carried out using 5-HT_2 ligands recently published [26], with the aim of better describing particular pharmacological behaviours that make some of such ligands interesting as new pharmacological tools, as well as describing the structural features related with pharmacological properties of the aforementioned compounds.

2. Chemistry

The present study is based in a series of conformationally constrained butyrophenones, the synthesis and standard pharmacological characterisation of which were recently described [26]. A sample of these compounds is shown in Fig. 1 and Table 1.

3. Pharmacology

3.1. Antagonism of serotonin at 5-HT_{2A} receptors from rat aorta

The experiments were performed as previously described, expressing the antagonistic potency as pA_2 [26].

3.2. Human 5-HT_{2C} binding assays

The experiments were performed as previously described [26] by using [^{3}H]mesulergine to label human 5-HT_{2C} receptors and 1 μM mianserin as non-specific masking ligand. Data were fitted by non-linear regression using GRAPH PAD PRISM v2.01 (GRAPH PAD Software). The affinity of compounds were measured as pIC_{50} (−log of concentration that displace the 50% of total binding).

Table 1
Binding affinities of relevant compounds from Ref. [26], measured at rat 5-HT_{2A} and bovine 5-HT_{2C} receptors

Compound	pK_i 5-HT_{2A}	pK_i 5-HT_{2C}
1 (QF0104B)	8.80±0.80	6.63±0.14
2 (QF0108B)	8.57±0.03	6.98±0.10
3 (QF0307B)	8.60±0.80	6.78±0.07
4 (QF0510B)	8.76±0.20	7.06±0.06
5 (QF0603B)	6.84±0.12	7.09±0.17
6 (QF0703B)	8.97±0.09	7.16±0.01
7 (QF0902B)	8.17±0.18	7.16±0.08
8 (QF1004B)	7.97±0.03	⩽5
9 (QF2004B)	8.80±0.88	7.24±0.06

Values represent the mean±S.E.M. of three experiments.

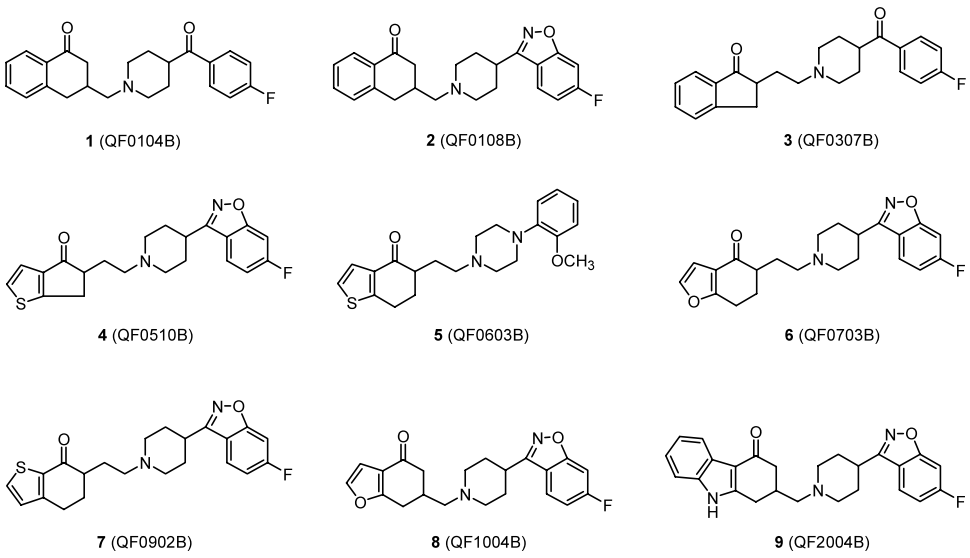


Fig. 1. Structures of relevant compounds.

3.3. IP accumulation and AA release measurement

Both effector pathways were simultaneously evaluated from the same experiment by using the method described by Berg et al. [22]. The experiments were performed at human 5-HT_{2A} receptors transfected in (Chinese Hamster Ovary) CHO cells (200 fmol mg^{-1} protein) and in human 5-HT_{2C} receptors transfected in CHO cells (5–10 pmol mg^{-1} protein). At 5-HT_{2A} receptors, the experiments were performed by inhibiting, with the antagonists/inverse agonists studied, the IP accumulation and AA release induced by 1 μM 5-HT. At 5-HT_{2C} receptors, the experiments were performed by inhibiting, with the antagonists/inverse agonists studied, the basal IP accumulation and AA release. Data were fitted by non-linear regression using GRAPH PAD PRISM v2.01 (GRAPH PAD Software). The potency of the compounds was measured as pIC_{50} (−log of concentration that inhibits the 50% of the maximal stimulation) and their efficacy as the percentage of inhibition of the maximal stimulation.

4. 3D-QSAR analysis

The compounds and pK_i values were extracted from Table 2 of Ref. [26], discarding compounds for which no binding experiment is available or having a pK_i lower than 5. According to these criteria, the 5-HT_{2A} series contained 52 compounds, and the 5-HT_{2C} one 43 molecules. For each compound, 3D structure was obtained starting from its 2D formula and applying the CORINA software [27]. All compounds were considered to have +1 formal charge, which was assigned to the nitrogen atom of the piperidinic ring or to one of the nitrogens of the piperazinic ring. In order to keep the consistency required for the QSAR analyses, in the compounds having a piperazinic ring (which has two protonable positions) the charge was assigned to the nitrogen closest to the cycloalkanone moiety. The series was analysed using a novel 3D-QSAR methodology based in GRid independent descriptors (GRIND) [B], which is implemented in the ALMOND 3.0.3 software [28]. This methodoly has the advantage, in comparison with other 3D-QSAR methods, that the compounds do not require to be superimposed, thus removing one of the most subjective and time-consuming steps of the analysis. Moreover, the molecular description used is rather tolerant with respect to the conformation of the ligands. Details about GRIND can be found elsewhere [29]. Here, the method was applied with the following options: O, N1 and TIP (molecular shape) probes, 100 nodes and 50% of importance given to the field values. The GRID ALM directive was set equal to 1. The method produced 6 correlograms of 67 variables each, thus giving a total of 402 variables.

The regression analysis of the pK_is vs. the GRIND variables was carried out using the partial least squares (PLS) method, also implemented in the ALMOND software. The optimal number of latent variables (LV) was selected on the basis of the cross-validation analysis

Table 2
Potencies of competition for [^{3}H]ketanserin and [^{3}H]mesulergine at human 5-HT_{2A} and 5-HT_{2C} receptors respectively, functional potencies (pA_2) measured at rat 5-HT_{2A} receptors and Meltzer ratios of compounds **6** and **8**

Compound	pK_i 5-HT_{2A}	pK_i 5-HT_{2C}	pA_2	Meltzer ratio
6 (QF0703B)	9.14±0.11	7.46±0.30	9.25±0.05	1.17
8 (QF1004B)	8.26±0.12	5.40±0.52	7.95±0.07	0.99

Data are expressed as mean±S.E.M. of three experiments.

results, but LV producing only small increases on the values of the q^2 were not incorporated into the models. Cross-validation tests were carried out using the standard leave-one-out method. In all the models, a soft variable selection strategy consisting in two sequential fractional factorial design [30] runs was also applied.

5. Results and discussion

In previous works [26,31–35] we studied the antipsychotic profile of different series of compounds at D_2, D_4, 5-HT_{2A} and 5-HT_{2C} receptors, focusing in 5-HT_{2A} and 5-HT_{2C} receptors. Moreover, in order to identify the structural requirements determining the affinity against 5-HT_{2A} and 5-HT_{2C} receptors, 3D-QSAR analyses using the CoMFA and GRID/GOLPE methodologies were carried out. These methodologies required a certain hypothesis on the superposition of the compounds.

The alignment-free 3D-QSAR model for the 5-HT_{2A} affinities presented in the present article has been developed using a series of 52 compounds, which includes the reference compounds haloperidol, ketanserin and risperidone. For this series, the analysis produced a PLS model of rather good quality (LV = 2; $r^2 = 0.85$; $q^2 = 0.74$). However, the observation of the scatterplot of experimental vs. calculated activities permits to identify an important outlier (compound QF1007B). This compound has a structure nearly identical to that of QF1008B but their affinity differs in more than 2.5 log units. The source of this surprising activity difference is being investigated but in the meanwhile the removal of compound QF1007B from the series increased considerably the model quality ($n = 51$; LV = 2; $r^2 = 0.89$; $q^2 = 0.81$). Fig. 2a shows the resulting scatterplot of experimental vs. calculated activities. A detailed interpretation of the model obtained is out of the scope of this article. However, it is noteworthy that the model identifies some structural features important for the affinity for the 5-HT_{2A} receptor, like the presence of groups able to interact at a precise distance with hydrogen bond donor (HBD) (protonated amino group) and hydrogen bond acceptor (HBA) groups (Fig. 3a), and also provides also a precise optimal distance between the favourable interaction region originated by the protonated amino group and the farthest extreme of the compounds (Fig. 3b). Both findings are compatible with the binding mode suggested in a previous work [26] where we postulated the participation of Ser159 in the binding, being such residue located at a distance of Asp155 that is in agreement with the distance now found in the GRIND model and shown in Fig. 3a. On the other hand, the distance shown in Fig. 3b is probably delimiting the depth of the binding pocket or, alternatively, showing the position where interactions with hydrophobic residues could take place.

GPCRs, as any other molecule, exist in many conformations as a consequence of the thermal excitation. Subsets of such conformations ('ensembles') are able to interact with certain G proteins mediating the corresponding physiological functions. Therefore, the manifestation of the receptor functions result from the relative populations of the different conformations. The modification of the relative populations of the functional ensembles by means of formation of drug-receptor complexes is one of the bases of drug effect

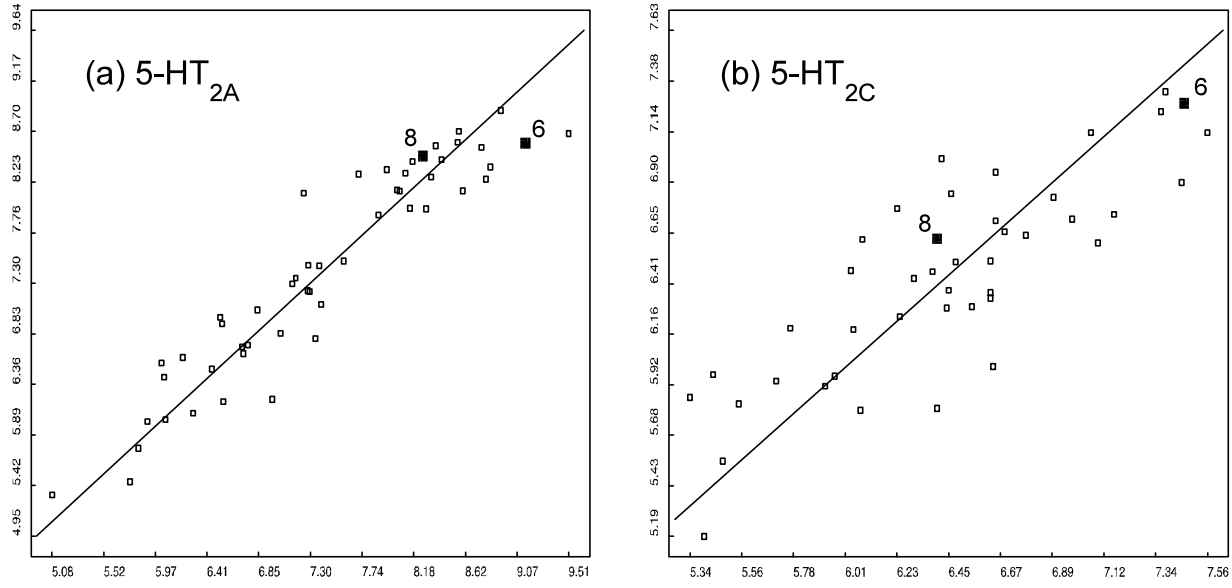


Fig. 2. Scatterplot of experimental vs. calculated binding affinities obtained for the (a) 5-HT_{2A} and (b) 5-HT_{2C} series. Compounds **6** and **8** are highlighted.

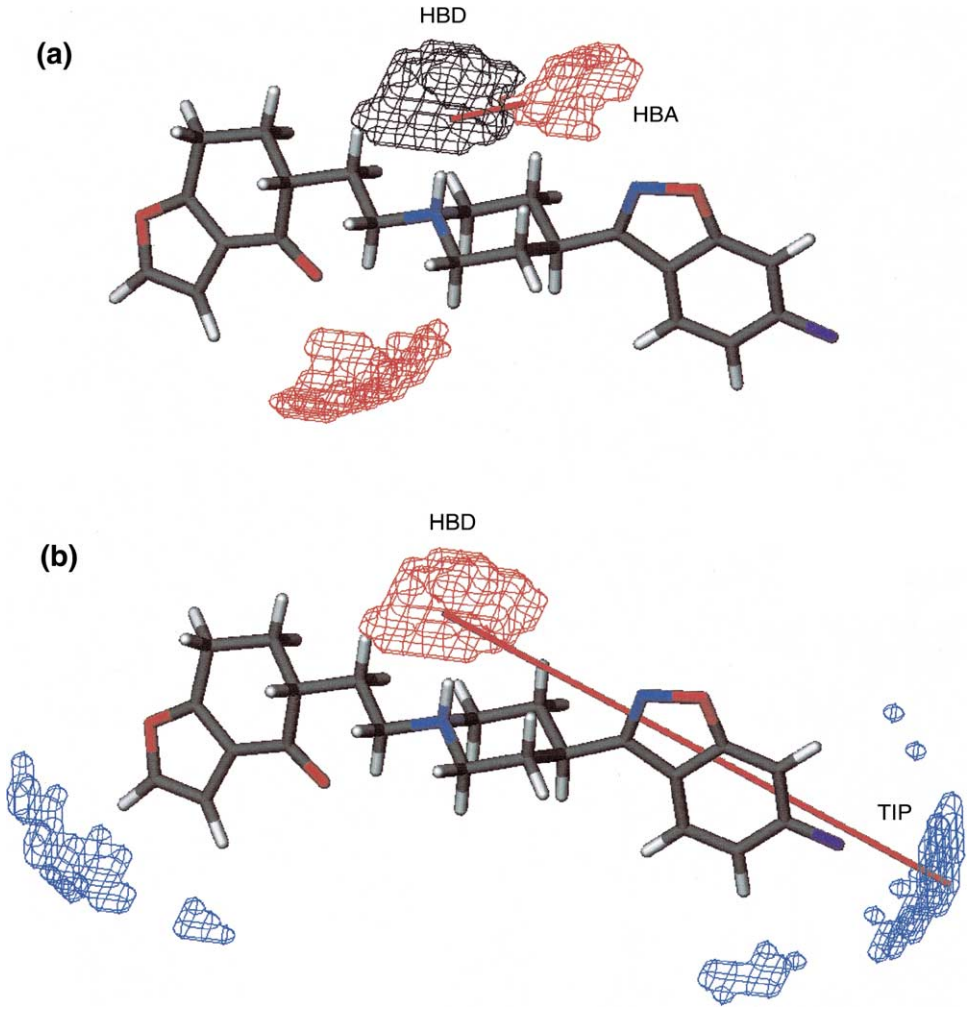


Fig. 3. Regions identified as important for increasing the 5-HT_{2A} binding affinity. Fig. 5a shows the optimal relative location of HBD and HBA regions. Fig. 5b displays the optimal distance between the HBD region generated by the protonated amino group and the farthest extreme of the molecule (TIP) represented by the molecular shape field.

[20]. The present work aims to carry out a sound study of the activation of 5-HT_2 receptors that could unveil conformation-dependent mechanisms.

Since it is difficult to establish structure-activity relationships for series of drugs that have both different pharmacological profiles and strong structural differences, we selected the compounds **6** (QF0703B) (1-[(4-Oxo-4,5,6,7-tetrahydrobenzo[b]furan-5-yl)ethyl]-4-(6-fluorobenzisoxazol-3-yl)piperidine) and **8** (QF1004B) (1-[(4-Oxo-4,5,6,7-tetrahydrobenzo[b]furan-6-yl)methyl]-4-(6-fluorobenzisoxazol-3-yl)piperidine) for the aforementioned pharmacological study, both being closely related from the structural point of view and active at 5-HT_{2A} and 5-HT_{2C} receptors, but showing different pharmacological profiles: the potency of compound **6** at the 5-HT_{2C} receptor is approximately 100 times (2 log units) higher than that of compound **8**, and their Meltzer ratios vary from 1.17 in the case of compound **6** (characteristic of an atypical antipsychotic), to 0.99 in the case of **8** (characteristic of a typical antipsychotic) (see Table 2).

Several conformations for both native and recombinant human 5-HT_{2A} receptors have labelled with agonists and discriminated by antagonists [36]. In order to determine if our compounds **6** and **8**, which are potent 5-HT_{2A} antagonists (Table 2), discriminate different conformations of the 5-HT_{2A} receptor, we have used a functional method of identification of conformational differences. This method consists in assessing the functional effector response at the PLC and PLA_2 pathways activated by the 5-HT_{2A} receptor. In particular, we have studied the way our compounds inhibit the stimulation elicited by 1 μM 5-HT at human 5-HT_{2A} receptors transfected in CHO cells, by measuring simultaneously IP accumulation and AA release (Table 3 and Fig. 4). We have observed that both compounds inhibited completely, but with different potencies, the IP accumulation, being their concentration-dependent curve adjustments most compatible with monophasic sigmoids (Fig. 4a). When we measured AA release inhibition, we observed an analogous behaviour

Table 3
Potency (pIC_{50}) of compounds **6** and **8** at human 5-HT_{2A} receptors transfected in CHO cells, by measuring simultaneously IP accumulation (PLC pathway) and AA release (PLA_2 pathway)

Compound	PLC	PLA_2	
	pIC_{50}	pIC_{50} high	pIC_{50} low
6	9.96±0.26	9.83±0.12	6.76±0.31
8	6.48±0.20		6.64±0.09

Data are expressed as mean±S.E.M. of three experiments.

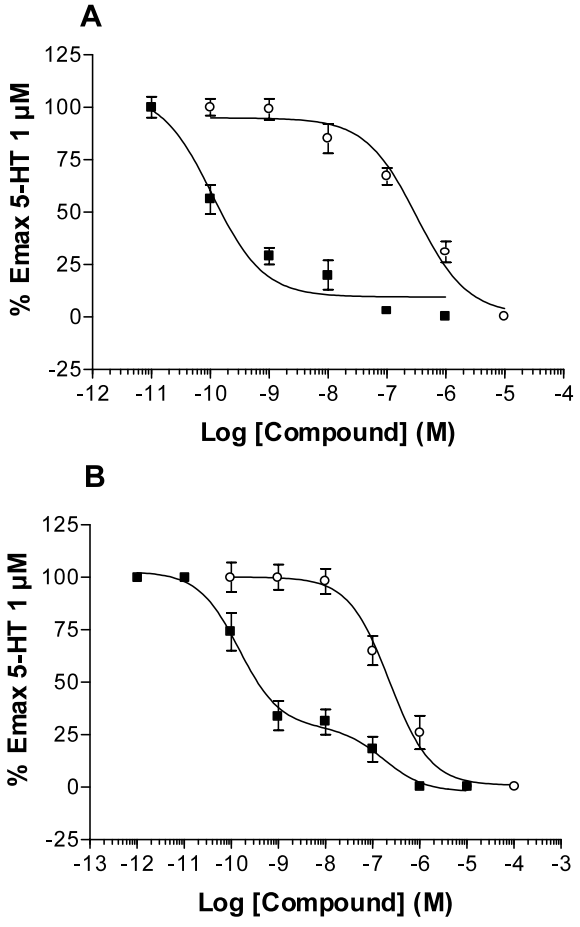


Fig. 4. Inhibition of 1 µM 5-HT -induced stimulation of IP accumulation (A) and AA release (B) by compounds **6** (■) and **8** (○) at human 5-HT_{2A} receptors transfected in CHO cells. Values represent mean± S.E.M. of three experiments.

of compound **8**. However, compound **6** inhibited AA release in a biphasic concentration-dependent way (Fig. 4b). These results suggest that both compounds disallow, but with different potency, 5-HT_{2A} receptors to adopt the serotonin-induced conformations compatible with the activation of the G protein linked to the PLC pathway. On the other hand, whilst compound **8** acts in a similar way in relation to the PLA_2 pathway, compound **6** shows a particular dual behaviour that could indicate its capability to discriminate two conformational ensembles responsible of the activation of the PLA_2 pathway with different efficacies. These results point out the usefulness of this kind of compounds and experiments to detect and discriminate pathway-dependent functional conformations of receptors. Moreover, the results show the dramatic influence that small structural differences between ligands can have in their pharmacological profile.

The alignment-free 3D-QSAR model for the 5-HT_{2C} affinities has been developed using a series of 43 compounds, which includes the reference compounds haloperidol, ketanserin and risperidone. For this series, the analysis produced a much worse PLS model (LV = 3; $r^2 = 0.79$; $q^2 = 0.36$), than that obtained for the 5-HT_{2A} series, probably because the series contains less compounds and the range of binding affinities covered is much smaller (2.22 vs. 4.43 log units). Fig. 2b shows the scatterplot of experimental vs. calculated activities. Even if the quality of the 5-HT_{2C} model is limited, the previously described structural requirement for 5-HT_{2A} receptor are qualitatively similar to those found in the present case, but with small differences that could be determinant for the selectivity.

Multiple conformational ensembles of native 5-HT_{2C} receptors have been described by means of studies made in human brain, which are related to different degrees of constitutive activity. It is known that RNA editing of human 5-HT_{2C} receptors gives rise to different isoforms, which express different degrees of constitutive activity [37]. LSD, a hallucinogenic drug that induce psychotic-like symptoms, changes its efficacy depending on the particular isoform of 5-HT_{2C} receptor, in the same way that antipsychotic drugs change its affinity and efficacy for 5-HT_{2C} receptors depending on the degree of constitutive activity.

Our selected compounds (**6** and **8**) compete for [^{3}H]mesulergine binding at human 5-HT_{2C} receptors (see Fig. 5), showing a concentration-dependent competition curves whose slopes do not significantly differ from unity. The pK_i observed for compounds **6** and **8** were 7.46 and 5.40, respectively. Moreover, we have checked their ability to withdraw constitutively active conformations from the conformational equilibrium (i.e., acting as inverse agonists), by testing their ability to reduce the basal IP accumulation of human 5-HT_{2C} receptors transfected in CHO cells (Fig. 6 and Table 4).

Table 4
Efficacy (Emax) and potency (pIC_{50}) as inverse agonists of compounds **6** and **8** at human 5-HT_{2C} receptors transfected in CHO cells

Compound	Emax (% inhibition)	pIC_{50}
6	72.15±0.67	5.97±0.30
8	0	unable to determinate

Data are expressed as mean±S.E.M. of three experiments.

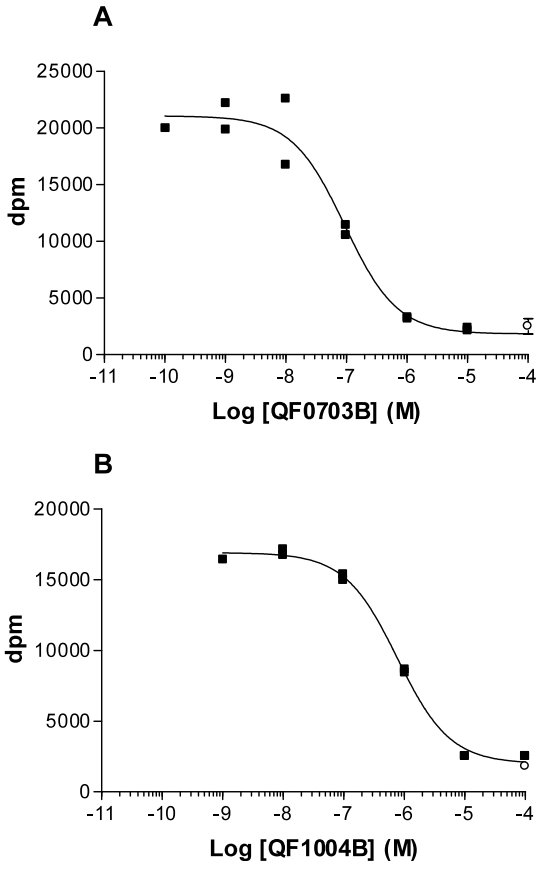


Fig. 5. Binding competition experiments at human 5-HT_{2C} receptors with **6** (QF0703B) (A) and **8** (QF1004B) (B), the radioligand used was [^{3}H]mesulergine (■), non-specific binding was measured with 10^{-6} M mianserine (○). The assay was performed with duplicate points.

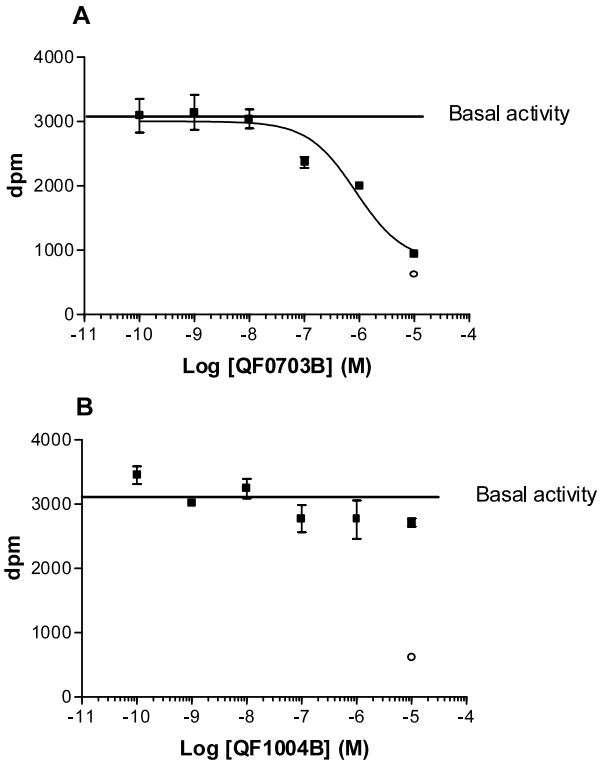


Fig. 6. Concentration–effect curves of compounds **6** (A) and **8** (B) on basal IP accumulation at human 5-HT_{2C} receptors transfected in CHO cells. 1 μM clozapine (○) was used as control. Values represent mean ± S.E.M. of three experiments.

These receptors were transfected at high expression levels (5–10 pmol mg^{-1} protein) consequently showing constitutive activity. The inverse agonist clozapine [38] was used as control in the experiments. Compounds **6** and **8** show again a different behaviour when interacting with constitutively active conformations of the 5-HT_{2C} receptor: compound **6** is able to reduce basal IP accumulation in a concentration-dependent manner revealing its inverse agonist behaviour at the 5-HT_{2C} receptor, while compound **8** is unable to reduce the 5-HT_{2C} constitutive activity showing its neutral antagonist character.

The previously described 3D-QSAR models account for the differences in affinity of compounds **6** and **8** (see Fig. 2), but such models cannot and do not aspire to explain particular pharmacological behaviours of these compounds that are related with effector pathways. 3D-QSAR models, which aim to be relevant for sizeable series of compounds, are focused on identifying major pharmacophoric patterns necessary for a certain biological activity, but such models are not relevant to identify fine-tuning dynamic structural features of individual ligand-receptor complexes that affect pharmacological behaviours related with modifications in the conformational populations of the receptors.

In summary, the present study has developed 3D-QSAR models for 5-HT_2 receptors that allow feasible predictions of activity of new compounds and reveal structural requirements for optimal affinity, particularly in the case of the 5-HT_{2A} receptor. Furthermore, two of the compounds of the series have been characterised as interesting pharmacological tools to study the influence on the antipsychotic profile of: (a) different selectivity for 5-HT_2 receptor subtypes; (b) different 5-HT_{2A}/D_2 ratios (Meltzer ratio); (c) different conformation-dependent functional behaviours at 5-HT_{2A} and 5-HT_{2C} receptors.

Acknowledgements

This work was supported in part by grants from the Spanish Ministry of Science and Technology (SAF2002-04195-C03), the Galician Government and the Fundació La Marató de TV3.

References

[1] H.Y. Meltzer, J. Clin. Psychiatry 55 (Suppl. B) (1994) 47–52.
[2] J. Horacek, Pharmacopsychiatry 33 (Suppl.) (2000) 134–142.
[3] G.K. Aghajanian, G.J. Marek, Brain Res. Rev. 31 (2000) 302–312.
[4] M.F. Green, J. Marshall-BD, W.C. Wirshing, D. Ames, S.R. Marder, S. McGurk, R.S. Kern, J. Mintz, Am. J. Psychiatry 154 (1997) 799–804.
[5] T.E. Sipes, M.A. Geyer, Brain Res. 761 (1997) 97–104.
[6] S.D. Gleason, H.E. Shannon, Psychopharmacology 129 (1997) 79–84.
[7] S. Okuyama, S. Chaki, N. Kawashima, Y. Suzuki, S. Ogawa, T. Kumagai, A. Nakazato, M. Nagamine, K. Kamaguchi, K. Tomisawa, Br. J. Pharmacol. 121 (1997) 515–525.
[8] L. Ereshefsky, C. Riesenman, J.E. True, M. Javors, Psychopharmacol. Bull. 32 (1996) 101–106.
[9] C.J. Schmidt, IBS's International Conference on: Serotonin Receptors. Targets for New Therapeutic Agents. 1996.
[10] R.A. Padich, T.C. McCloskey, J.H. Kehne, Psychopharmacology 124 (1996) 107–116.
[11] P. Martin, N. Waters, A. Carlsson, M.L. Carlsson, J. Neural Transm. 104 (1997) 561–564.
[12] H.C. Meltzer, S. Matsubara, J.C. Lee, Psychopharmacol. Bull. 253 (1989) 390–392.
[13] B.L. Roth, H.Y. Meltzer, The role of serotonin in schizophrenia, in: F.E. Bloom, D.J. Kupfer (Eds.), Psychopharmacology: The fourth generation of progress, Raven Press, 1995, pp. 1215–1227.
[14] K. Herrick-Davis, E. Grinde, M. Teitler, J. Pharmacol. Exp. Ther. 295 (2000) 226–232.
[15] H.Y. Meltzer, S.R. McGurk, Schizophrenia Bull. 25 (1999) 233–255.
[16] C. Reavill, A. Kettle, V. Holland, G. Riley, T.P. Blackburn, Br. J. Pharmacol. 126 (1999) 572–574.
[17] M.B. Knable, D.R. Weinberger, J. Psychopharmacol. 11 (1997) 123–131.
[18] D. Cussac, T.A. Newman, J.P. Nicolas, J.A. Boutin, M.J. Millan, Naunyn Schmiedebergs Arch. Pharmacol. 361 (2000) 549–554.
[19] C.M. Niswender, K. Herrick-Davis, G.E. Dilley, H.Y. Meltzer, J.C. Overholser, C.A. Stockmeier, R.B. Emeson, E. Sanders-Bush, Neuropsychopharmacology 24 (2001) 478–491.
[20] T. Kenakin, Nature Rev. Drug Discov. 1 (2002) 103–110.
[21] V.J. Hilser, E. Freire, J. Mol. Biol. 262 (1996) 756–772.
[22] K.A. Berg, S. Maayani, J. Goldfarb, C. Scaramellini, P. Leff, W.P. Clarke, Mol. Pharmacol. 54 (1998) 94–104.
[23] R.A. de Ligt, A.P. Kourounakis, A.P. IJzerman, Br. J. Pharmacol. 130 (2000) 1–12.
[24] T. Kenakin, Trends Pharmacol. Sci. 16 (1995) 232–238.
[25] P.G. Strange, Trends Pharmacol. Sci. 23 (2002) 89–95.
[26] J. Brea, J. Rodrigo, A. Carrieri, F. Sanz, M.I. Cadavid, M.J. Enguix, M. Villazón, G. Mengod, Y. Caro, C.F. Masaguer, E. Raviña, N.B. Centeno, A. Carotti, M.I. Loza, J. Med. Chem. 45 (2002) 54–71.
[27] J. Gasteiger, C. Rudolph, J. Sadowski, Tetrahedron Comp. Method 3 (1990) 537–547.
[28] ALMOND 3.0.3, Multivariate Infometric Analysis S.r.l. Perugia, Italy (2002).
[29] M. Pastor, G. Cruciani, I. McLay, S. Pickett, S. Clementi, J. Med. Chem. 43 (2000) 3233–3242.
[30] M. Baroni, G. Costantino, G. Cruciani, D. Riganelli, R. Valigi, S. Clementi, Quant. Struct. Act. Relat. 12 (1993) 9–20.
[31] E. Raviña, J. Negreira, J. Cid, C.F. Masaguer, E. Rosa, M.E. Rivas, J.A. Fontenla, M.I. Loza, H. Tristan, M.I. Cadavid, F. Sanz, E. Lozoya, A. Carotti, A. Carrieri, J. Med. Chem. 42 (1999) 2774–2797.
[32] C.F. Masaguer, E. Raviña, J.A. Fontenla, J. Brea, H. Tristan, M.I. Loza, Eur. J. Med. Chem. 35 (2000) 83–95.
[33] C.F. Masaguer, I. Casariego, E. Raviña, Chem. Pharm. Bull. 47 (1999) 621–632.
[34] C.F. Masaguer, E. Formoso, E. Raviña, H. Tristan, M.I. Loza, E. Rivas, J.A. Fontenla, Bioorg. Med. Chem. Lett. 8 (1998) 3571–3576.
[35] E. Raviña, I. Casariego, C.F. Masaguer, J.A. Fontenla, G.Y. Montenegro, M.E. Rivas, M.I. Loza, M.J. Enguix, M. Villazon, M.I. Cadavid, G.C. Demontis, J. Med. Chem. 43 (2000) 4678–4693.
[36] J.F. Lopez-Gimenez, M. Villazon, J. Brea, M.I. Loza, J.M. Palacios, G. Mengod, M.T. Vilaro, Mol. Pharmacol. 60 (2001) 690–699.
[37] C.M. Niswender, S.C. Copeland, K. Herrick-Davis, R.B. Emeson, E. Sanders-Bush, J. Biol. Chem. 274 (1999) 9472–9478.
[38] K.A. Berg, B.D. Stout, J.D. Cropper, S. Maayani, W.P. Clarke, Mol. Pharmacol. 55 (1999) 863–872.

PAPER IV