

Three-dimensional similarity of molecules with biological interest on the basis of molecular interaction potentials

Montserrat Barbany i Puig

PhD thesis

The Figure in the cover shows a superposition of the quantum mechanical MEP for the transition state and transition state analog of Claisen rearrangement of chorismate to prephenate. Isocontours in the figure represent MEP differences.

Three-dimensional similarity of molecules with biological interest on the basis of molecular interaction potentials

Montserrat Barbany i Puig

Memòria presentada per optar al grau de Doctora
per la Universitat Pompeu Fabra

Aquesta Tesi Doctoral ha estat realitzada sota la direcció del
Dr. Jordi Villà i Freixa i el Prof. Ferran Sanz i Carreras al Departament de Ciències
Experimentals
i de la Salut de la Universitat Pompeu Fabra

Jordi Villà i Freixa

Ferran Sanz i Carreras

Montserrat Barbany i Puig

Barcelona, June 2006

The research in this thesis has been carried out at the Computational Biochemistry and Biophysics Lab within the Research Unit on Biomedical Informatics (GRIB) at Institut Municipal d'Investigació Mèdica (IMIM) and Universitat Pompeu Fabra (UPF).



The research carried out in this thesis has been supported by a project of Fondo de Investigaciones Sanitarias (FIS 01/1330) of Ministerio de Sanidad y Consumo of the Spanish Government to Ferran Sanz and a travel fellowship BE from Generalitat de Catalunya to Montserrat Barbany.



La cosa més bella que podem experimentar és el misteri. És la font del veritable art i de totes les ciències. Aquell a qui aquesta emoció li és aliena, aquell que ja no pot aturar-se per meravellar-se i abstreure's amb delit, és mort: els seus ulls estan tancats.

A. Einstein

Contents

Acknowledgements	xiii
Abstract	xv
1 Introduction	1
1.1 Biomolecular interactions: energetic considerations	1
1.1.1 Energy vs free energy	2
1.1.1.1 Energy and entropy	2
1.1.1.2 The Gibbs free energy	3
1.1.1.3 Boltzmann law and partition function	4
1.1.1.4 Potential energy surfaces and free energy calculations	6
1.1.2 Transition state theory used as framework for protein-small molecule interactions	7
1.1.2.1 Ground state	8
1.1.2.2 Transition state	11
1.1.3 Molecular interaction potentials and fields	14
1.2 Protein-ligand interactions in drug design	15
1.2.1 Receptor-based drug design. Direct methodologies	16
1.2.1.1 Experimental determination of biomolecular 3D structures	16
1.2.1.2 Computational prediction of target binding modes	17
1.2.1.2.1 Sampling the orientation of the ligand in the receptor cavity.	18
1.2.1.2.2 Scoring ligand-receptor interaction.	19
1.2.2 Ligand-based drug design. Indirect methodologies	20
1.2.2.1 Correlating structural properties and biological activity	20
1.2.2.2 Biomolecular similarity	22
1.2.2.2.1 Descriptors used in biomolecular similarity calculations.	23
1.2.2.2.2 Similarity correlation coefficients.	24
1.2.2.3 The alignment problem	28
1.3 MIPSim: Molecular Interaction Potentials Similarity analysis. Overview of previous studies	30
2 Hypothesis and objectives	33

3	Methods	35
3.1	Methodologies used inside MIPSim	35
3.1.1	Calculation of MEP and MIP inside MIPSim	36
3.1.1.1	Quantum MEP	36
3.1.1.1.1	Molecular wavefunction.	36
3.1.1.1.2	Molecular electrostatic potentials.	43
3.1.1.1.3	Program GAMESS.	45
3.1.1.2	Classical MIP	45
3.1.1.2.1	Program GRID.	47
3.1.2	Optimization strategies inside MIPSim	50
3.1.2.1	Steepest descent	51
3.1.2.2	Conjugate gradients	51
3.1.2.3	Simplex	51
3.1.2.4	Genetic algorithms	52
3.2	Miscellanea of methodologies used outside MIPSim	52
3.2.1	Conformational sampling	52
3.2.1.1	Catalyst	52
3.2.1.2	Omega	53
3.2.2	Multiobjective optimization strategies. FFSQP software	53
3.2.3	Superposition of molecules and proteins	53
3.2.3.1	Superposition of molecules based on atoms. SUPERB routine	53
3.2.3.2	Multiple protein sequence alignment. STAMP package	54
3.2.4	Calculations on solvation free energy. Langevin dipoles calculations inside CHEMSOL program	54
3.2.5	Statistical tools in 3D-QSAR methodologies	58
3.2.5.1	Principal component analysis	58
3.2.5.2	Partial least squares	59
3.2.5.3	Pretreatment of data	61
3.2.5.4	Variable selection	61
3.2.6	Accuracy of a prediction. Matthews correlation coefficient	62
4	Results	63
4.1	Development of MIPSim program	63
4.1.1	Introduction of new similarity coefficients	63
4.1.2	Combination of different similarity coefficients	64
4.1.3	Introduction of new definitions of MIP	64
4.1.4	Selection of energy intervals	64
4.1.5	Conformational flexibility: TORS module	65
4.2	Benchmarking and profiling of MIPSim	67
4.3	Technological platforms development	68
4.3.1	OS platforms	68
4.3.2	User interface	68

4.3.3	Visualization tools	68
4.3.4	MIPSim web site	68
4.4	MIPSim applications. Publications	69
4.4.1	Comparison of biomolecules by MIP alignment	69
4.4.2	On the use of MIPSim for characterizing the activity of catalytic antibodies	75
4.4.3	MIPSim as scoring function in protein-ligands docking	85
4.4.4	3D-QSAR study of hERG inhibitors	99
5	Conclusions	113
A	List of abbreviations	115

Interesting links	117
B Presentations in congresses	119
B.1 Posters	119
B.2 Oral contributions	125
C Annexes	127
C.1 MIPSim installation procedure	127
C.2 Setting up MIPSim	128
C.3 MIPSim keyfile	130
C.4 Profiling of MIPSim	130
C.5 Combination of different similarity coefficients	131
C.6 Comparison of intervals of energy	131
C.7 Tors module input	131
C.8 Technical aspects of CHEMSOL	131
C.9 PDLDSMALL program	133
C.10 POLSAR program	133
C.11 SUPERB input file	134
C.12 STAMP input	134
C.13 Future developments	134
C.14 This L ^A T _E X thesis template	135
Bibliography	137
Index	149

List of Figures

1.1	Curve dose-response	10
1.2	Water-enzyme free energy profiles surfaces	12
1.3	Chemical graph	23
1.4	Pharmacophoric points	24
1.5	Molecular shape represented by vdW radii	24
1.6	Typical MEP distribution of planar molecules	25
1.7	Calculation of Gaussian index	27
3.1	Evaluation of MIPs	48
3.2	GRID files	49
3.3	Spatial representation of LDs	55
4.1	Intervals of energy for N1 probe of GRID	65
4.2	Intervals of energy for O probe of GRID	65
4.3	Intervals of energy for DRY probe of GRID	66
4.4	Cylindrical coordinates	67
4.5	Rotations generated by TORS module	67

Acknowledgements

Arribar a aquest moment, encara ara, em sembla un somni, una fita impossible d'aconseguir. Sobretot després d'aquest darrer any i mig en que posar-se a escriure quatre ratlles ha estat realment una missió impossible. La veritat és que no tinc res que retreure a la meva doble maternitat, és una de les coses més maques que et pot passar a la vida. Però sí que és veritat que és dur treure temps de petons i abraçades per dedicar-los a investigar. De la mateixa manera, el fet d'investigar em fa sentir viva i també m'omple com a persona. Per tant, estic dins d'un etern dilema.

Primer de tot m'agradaria agrair a l'Institut Municipal d'Investigació Mèdica per amablement finançar-me la reprografia i enquadernació d'aquest manuscrit.

Als meus directors de tesi: Ferran i Jordi. Gràcies per confiar en mi des d'un bon principi i durant tots aquests anys. He après molt i molt i em sento molt satisfeta del que he fet junt amb vosaltres. Ferran, tot i que la relació amb tu ha estat més distant, degut a la teva apretada agenda, cada cop que hem xerrat sobre ciència m'has encomanat el teu entusiasme. Gràcies per obrir-me les portes d'aquest món tan interessant. Jordi, et recordo una frase que em vas dir quan vaig començar la tesi amb tu: "Ja veuràs, quan acabis d'escriure la tesi hauràs acabat ben farta del teu director de tesi". Doncs puc dir ben orgullosa després de gairebé cinc anys que ha estat un grandíssim plaer treballar amb tu. No sé si el pensament és recíproc. Gràcies per tenir paciència amb mi, per no deixar-me passar les coses que faig malament, per corregir-me. I em sap greu el retard que ha tingut la finalització de la tesi. Gràcies per esperar, sé que has entès la meva situació.

Per la gent del meu grup: Jordi, t'he d'agair que mica en mica anessim ampliant el nostre bigrup fins a convertir-lo en el que és ara: un conjunt de cracks!! Per l'Alfons. Gràcies per ajudar-me amb les pàgines web i sobretot per donar-me tan ràpidament la teva amistat aquell dia a les Avellanes. Que siguis molt feliç amb la Maria Elena. A l'Adrian, per ajudar-me en tants moments crítics (et sona de res un poster?) i per ser tan simpàtic i tan bon company. To Kareem, thank you for beeing a good guy. To Michael, I'm very proud to share with you our interest for physics.

A tota la gent de modelling: a Hugo por ayudarme tanto desde un buen principio. Has estado una pieza clave en mi tesis. A Fabien, por tu humanidad. Eres de las personas más sensibles que conozco. Gracias por ayudarme en tantos momentos de dificultad científica. Gracias a Jorge i a las inolvidables Cristinas por darme vuestra amistad y comprensión y buenos ratos de sobremesa. Núria, gràcies per la teva altra manera de veure les coses. Manolo, has estado un buen profesor para mi y muy buen compañero de viaje.

Al grup del Quemogenòmica. Jordi, fas honor al teu nom. Ets tot un mestre i bon horador. N'he d'aprendre de tu. Gràcies a la Bet pels bons moments i també per ajudar-me

tant en alguns projectes. Gràcies Rut i Montse per la vostra companyia.

Al grup de bioinformàtica estructural. Baldo, has estat un bon mestre i bon company. Josep, m'has recolzat i animat molt. Sigues feliç i que res no aturi els teus somnis. Ramon, sigue escribiendo, se te da muy bien.

A la gent de genòmica...no us anomeno a tots perquè segur que no acabaria. Sobretot gràcies Pep per ajudar-me amb el Latex (hi ha alguna cosa que no sàpigues fer?) i per fer-me riure tant.

Gràcies a l'Alfons i l'Oscar, els González. Sense vosaltres no hauria pogut fer res d'això. Gràcies per solucionar-me els problemes amb Linux i sobretot per la grandíssima paciència que heu demostrat. També al Joan ("el de los Cobos") per les teves precioses converses.

To the Safety Assessment group at AstraZeneca and specially to Scott Boyer. Thank you for let me discover a new research world and for make me feel like home. Isabella my soul friend at Göteborg. Be happy with Markus and Naomi.

A la colla de física: vosaltres sou com el vi, cada any que passa sou millors i us estimo més. Roger, el meu amic de l'ànima, gràcies per transmetre'm aquest entusiasme que poses en totes les coses que fas. Saps que sempre podràs comptar amb mi. Mireia, gràcies per mirar-me l'anglès de la tesi, tot i que al final tot ha anat tan ràpid... Jordi gràcies per ajudar-me amb el Linux, sense tu tampoc això hauria estat possible. Lluís, tu també molt aviat arribaràs aquí, n'estic segura. I a tots els altres que m'heu animat tant en aquest repte que se'm presentava. M'heu fet sentir viva. Gràcies també a la Mònica i la Olga per escoltar els meus rotllos.

A tota la meva família. Sóc plenament conscient que sense vosaltres hauria estat completament impossible escriure la tesi. Mare, gràcies per l'ajuda amb les nenes, per ser-hi quan més et necessitava. No sé si algun dia et podré tornar tot aquest amor que m'has donat. Al pare, pels seus cafès, els seus suc...Sé que et sents orgullós de mi. Que sàpigues que jo me'n sento de tu. Al Pep per ajudar-me sobretot a fer la portada d'aquest manuscrit. Que la vida que comences d'aquí no res sigui enriquidora i, ja veuràs, algun dia recolliràs els fruits que estàs sembrant. A la Rosa, per recolzar-me, per animar-me per demostrar-me la teva il·lusió en el que estava fent. Toni i família, m'heu fet sentir que això que estava fent era una cosa realment important. També m'agradaria dedicar aquesta tesi a la memòria dels meus avis.

A toda la familia de Carlos y sin olvidar a mis tres maravillosos sobrinos. Gracias por quererme tanto.

Finalment al meu Carles, el meu amic i company de vida. Gràcies per animar-me en aquesta fita des d'un bon principi. Per escoltar-me quan et parlava del que feia encara que potser no entenguessis molt de què anava tot plegat. Per aguantar-me els moments baixos, d'histerisme i de mal humor. Gràcies per fer-me sentir feliç i estimada.

Per les meves petites Nora i Sílvia. Pels vostres petons i abraçades i sobretot per el vostre somriure i alegria cada matí. Vosaltres heu estat el veritable motor que m'ha ajudat a tirar endavant. La mare us estima tant...

Abstract

One of the most promising areas of biomedical and pharmaceutical research is computer assisted molecular design, which is based on the modelization of the chemical entities responsible of the pharmacological activity and the search of mathematical models describing the relationship between the physicochemical properties and the biological activity of such entities.

In general, the success of these techniques depends critically on the quality of the molecular description and, in particular, on the fact that this description should be appropriate to represent the molecular interaction phenomenon that we intend to describe. In this sense, methodologies based on the molecular interaction potential (MIP) offer important advantages with respect to other techniques. MIPs are interactions of the studied molecule with one or several selected chemical entities and they are useful tools for the comparison of series of compounds displaying related biological behaviours. As it will be shown here, structure-activity studies benefit from a detailed comparative analysis of MIP distributions of prospective drugs.

This project aims to develop tools for computer assisted molecular design based on the characterization and comparison of MIPs of different compounds. To this end, the molecular similarity program `MIPSim` (Molecular Interaction Potentials Similarity analysis) (Cáceres et al., 16, 568-569, *Bioinformatics*, 2000) has been further developed and applied to different biological and pharmacological problems.

`MIPSim` analyzes and compares MIP distributions of series of biomolecules. One of the objectives of `MIPSim` is to obtain automatic structural alignments of series of biomolecules based on their MIP distributions. This can be used to stablish hypothesis about their relative orientation at the functional site, which is sometimes non-evident when only taking into account structural features. `MIPSim` can evaluate MIPs by classical or quantum methods, thanks to its interfaces to programs `GRID` and `GAMESS` respectively.

This thesis includes four scientific studies which demonstrate the applicability of MIP similarity through `MIPSim` to study molecules of biological interest. `MIPSim` has been used to study alignments of biomolecules, to explore the electrostatic properties of enzymes and catalytic antibodies, to help in searching MIP-based docking and finally, to perform a 3D-QSAR study based on a MIP alignment.

Introduction

1.1 Biomolecular interactions: energetic considerations

Understanding how proteins work is critical to achieve an accurate understanding of biological processes(1). Proteins participate through virtually the whole cell machinery(2): proteins named enzymes catalyze chemical reactions, regulatory proteins control gene expression, proteins named hormones accept/transmit intercellular signals, immune proteins recognize/bind other molecules, etc...

Underlying every biological process there is a multitude of proteins binding to and modifying each other, forming complex frameworks and assemblies, and catalyzing reactions. Most protein functions depend on their interactions with other molecules. These may be other proteins, nucleic acids, solvent molecules such as water, metal ions or organic molecules. Life is, thus, based on molecular interactions(3). Three types of protein interactions are particularly relevant(4):

- Protein-nucleic acid interactions. Proteins that bind to DNA and RNA mediate a number of processes, including regulation of gene expression, gene transcription, DNA replication, mRNA intron splicing or mRNA translation.
- Protein-small molecule interactions. The function of some proteins is to bind a target molecule or a set of target molecules and to perform some action. Enzymes bind to substrate molecules and then catalyze chemical reactions that would otherwise occur too slowly to be biologically useful. Some proteins involved in cellular signaling bind a signal molecule and undergo a conformational change leading to further signaling or changes in cellular processes.
- Protein-protein interactions. Many proteins function by forming active complexes with each other. Protein-protein interactions are also involved in, *e.g.*, antibody-antigen binding, large scale organismal motion and cell adhesion.

The strengths of protein-molecule interactions vary widely. In some cases, these are very tight; in others, there are weak and short-lived(5). But the binding most often shows great specificity, in the sense that each protein molecule can usually bind just one or a few molecules out of the many thousands of different types it encounters. The substance that is bound by the protein is referred to as a ligand for that protein (from the Latin word *ligare*, meaning "to bind") and the protein is referred to as receptor(5). Binding interactions define

how well the compound attaches to the receptor by first being recognized as complementary to the receptor structure in shape and electronic structure.

The binding site of a protein consists of a cavity formed by a specific arrangement of amino acids. The chemical properties of a protein depend almost entirely on its exposed surface residues(5). In addition, protein-ligand binding may involve conformational changes in the protein and/or the ligand. The binding usually occurs in the presence of solvent, typically water, and interactions with solvent molecules must then be taken into account(6). All this makes the study of biomolecular interactions by means of computational methods a hard task.

In order to establish a connection between proteins and their interaction with ligands by means of the underlying physico-chemical context we need accurate enough structural information on proteins and protein-ligand complexes. The characterization of the structure and the energetics of molecular complexes is thus a key factor for understanding biological functions.

1.1.1 Energy vs free energy

In order to study protein-ligand interactions it is important to start by defining some of the terms related with our quantitative measure of the interactions.

We want to know about the relationship between the microscopic specific interactions in the active site and the macroscopic properties that our system will produce in a real situation inside a solution. In order to achieve such objective we need to characterize energetically the system and to pass from this energetic information for every conformation of the system to values comparable to experimental data. This is done through the use of statistical mechanics.

In this section we will briefly introduce the concepts of energy and free energy, which in short refer to single (microscopic) or to average (macroscopic) properties of the system being considered. This distinction is important for the proper treatment of the information obtained from the methods developed and applied in this thesis.

1.1.1.1 Energy and entropy

The thermodynamic state (macroscopic state) of a system is usually defined by a small set of parameters (thermodynamical variables), for example, temperature, T , pressure P , total energy E , number of particles N , volume V . If the thermodynamical variables are independent they are called state variables. A function that can be expressed with state variables is called a state function. The variation of a state function is independent of the process, only depends on the initial and final state.

Energy is a fundamental quantity that every physical system possesses; it allows us to predict how much work (W) the system could be made to do, or how much heat (q) it can exchange.

The first law of thermodynamics formalizes the general principle of physics that energy is conserved. It says that the total inflow of energy into a system must equal the total outflow of energy from the system, plus the change in the energy contained within the system.

$$\Delta U = q + W \quad (1.1)$$

When a system gains energy by a thermal radiation or conduction as a result of a temperature differential, it is absorbing a positive quantity of heat q . In chemistry, heat is the amount of energy which is absorbed or released by a given chemical reaction. When the system gains energy by other methods, for example, by the operation of external mechanical forces, a positive quantity of work W is being done on the system (or negative quantity of work is being done by the system).

The observed proportionality of heat and work for any cyclic process requires that the total internal energy U is a state function. The increase in such energy when a system changes from state A to state B is independent of the way in which the change is brought about. It is simply the difference between the final and the initial energy,

$$\Delta U = U_B - U_A \quad (1.2)$$

Internal energy involves energy on the microscopic scale, depending on a certain configuration of the microstate. One has to take into account the kinetic energy of the linear motion, rotational and vibrational kinetic energy and potential energy associated with the intermolecular attractive forces(7).

The second law of thermodynamics says that in any reversible process (it is possible to invent a means of restoring every system concerned to its original condition) the increase in entropy ΔS of a system, or part of a system, is equal to the heat it absorbs divided by the absolute temperature.

$$\Delta S = q_{rev}/T \quad (1.3)$$

where q is the heat received by the reservoir and T is the temperature. This property only depends on the initial and final states, then it is also a state function.

In the case of a substance in internal equilibrium and subject only to changes brought about by reversible exchange of heat with an external reservoir and reversible expansion work against an external restraining pressure, the heat absorbed is $T\Delta S$ and the work absorbed is $-P\Delta V$.

Relationship between entropy and the number of microstates compatible with a certain energy will be explained in the section 1.1.1.3.

1.1.1.2 The Gibbs free energy

Gibbs(8) in 1875 defined two new functions as follows:

Enthalpy:

$$dH \equiv dU + PdV + VdP \quad (1.4)$$

Gibbs energy or free energy:

$$dG \equiv dU - TdS + PdV + VdP = dH - TdS \quad (1.5)$$

dH and dG are state functions since all the quantities entering their definitions are so. Hence, for any constant-pressure process with no work other than that of volume change, the enthalpy increase is exactly the heat absorbed. For this reason dH is sometimes called the heat content.

Applying the definition of Gibbs free energy to an isothermal process:

$$\Delta G = \Delta H - T\Delta S \quad (1.6)$$

where ΔH is the enthalpy change and ΔS is the entropy change, both of them thermodynamic properties.

The third law of thermodynamics says that for any substance in a perfect crystalline state, the entropy is zero in the limit when T tend to zero. The third law provided a means of obtaining the absolute entropy of each substance in a reaction. If such entropy values are available for each reactant and each product, then ΔS may be calculated and combined with a calorimetrically determined ΔH according to the equation above in order to obtain ΔG . Thus, the third law opened a great new opportunity for the prediction of ΔG .

Reaction free energy (Gibbs function) is the magnitude that describes the spontaneity of thermic processes, that is, the tendency of molecular systems to associate and/or to react. The Gibbs free energy predicts whether a process carried out at constant temperature T , and constant applied pressure P can occur spontaneously ($\Delta G < 0$), cannot occur spontaneously ($\Delta G > 0$) or it is in equilibrium ($\Delta G = 0$) under the prescribed conditions.

Let us now consider the process of forming an interaction or complex from two unbound biomolecules. In particular, we will consider from this point the binding process of a protein and a small molecule. We can associate the theoretical free energy for the process of binding, $\Delta G_{binding}$, from the experimentally determined association constant (see section 1.1.2)) as:

$$\Delta G_{binding} = -RT \ln K_a \quad (1.7)$$

1.1.1.3 Boltzmann law and partition function

Predicting free energies represents a very important aspect in chemical, biological and pharmaceutical sciences. Free energy calculations are generally formulated in terms of estimating the relative free energy differences, between two equilibrium states. This is of great importance in many applications, because it is normally the difference in the thermodynamic properties between two such states that is of interest (reactants, products, transition states, association and dissociation of molecules).

But before describing in detail how we can evaluate computationally the value of ΔG , an introduction to statistical mechanics, the discipline that allows us to connect macroscopic thermodynamic properties with the microscopic behaviours of the matter, is required. Matter consist of a very large number of atoms and molecules and their behavior follows statistical factors. Statistical mechanics(7) is a tool used to interpret macroscopic properties of a system based on the statistical treatment of the microscopic states consistent with those macroscopic properties.

A system is said to be in thermodynamical equilibrium if its macroscopic variables are not dependent on the variable time. Even if a macroscopic system is in equilibrium, there

exist fluctuations at the microscopic scale. A microstate is a microscopic state of a system specified by coordinates and velocities of every particle. An ensemble is a collection of all possible systems which have different microscopic states but have an identical macroscopic or thermodynamic state. Depending on the macroscopic properties (N is the number of particles, V the volume, E the total energy, μ the chemical potential, T the temperature) we can describe different ensembles, among others:

- microcanonical (isolated system): N, V, E fixed.
- canonical (isothermal system): N, V, T fixed.
- grand canonical (open system): μ, V, T fixed.

There exists a finite number of microstates compatible with every macrostate of the system; this is the thermodynamical probability. Boltzmann hypothesis says that all accessible microstates of a system have identical probability. Let us assume that we deal with a system of noninteracting N particles. In order to make the problem tractable, we segregate the N particles into groups of N_i particles, each of which has the same energy E_i . By using probability theory one can arrive to the so-called Boltzmann distribution law:

$$P_i = \frac{\exp -\beta E_i}{\sum_i \exp -\beta E_i} \quad (1.8)$$

where E_i is the total energy of every microstate i . Boltzmann distribution is the most probable distribution of N molecules among all microstates subject to the constraints that volume V and temperature T are constant (canonical ensemble). β controls the total energy by the relative distribution over high and low energy states:

$$\beta = \frac{1}{k_B T} \quad (1.9)$$

and k_B is the Boltzmann constant.¹

The canonical partition function is:

$$Z = \sum_i \exp -\beta E_i \quad (1.10)$$

The partition function measures how the particles are distributed over the available energy states. It gives an indication of the average number of states that are thermally accessible to a molecule for a given temperature of the system. At very low temperatures, only the low energy states are accessible. We define ground state (GS) as the state with minimum E_i . Partition function Z is used to compute thermodynamical properties (the expected value for an observable) knowing the distribution of microstates:

$$\langle X \rangle = \sum_i x_i P_i \quad (1.11)$$

¹
 $k_B = \frac{R}{N_A} = 1.38 \times 10^{-23} \text{ J.K}^{-1} = 3.30 \times 10^{-27} \text{ Kcal.K}^{-1}$ where R is the ideal gas constant and N_A is Avogadro number

where $\langle X \rangle$ is the mean of a macrostate variable (thermodynamic property). x_i is the microstate variable and P_i is the probability (see equation 1.8).

Now we are able to compute macroscopic properties knowing the microscopic variables. Once we know the macroscopic properties we can compute differences between them in different states. For example, it is possible to estimate the relative free energy differences.

1.1.1.4 Potential energy surfaces and free energy calculations

We move now to develop the concepts introduced in the previous sections in the framework of protein-ligand interactions.

Initially receptor and ligand are both solvated by water molecules. When they bind together, they are stabilized by intermolecular interactions while other processes (solvent rearrangement or restriction of degrees of freedom) may occur simultaneously. A useful interpretative tool to understand how this process (and others) occurs is the definition of a unique potential energy surface (PES) for the complete systems of protein, ligand and solvent.

PES describes energy of a molecule in terms of its structure. It is generally used in quantum mechanics and statistical mechanics to model chemical reactions and interactions in simple chemical and physical systems(9).

Evaluation of free energies requires the exploration of all the points of the configurational space of the system (of all the PES). This step is called sampling. Sampling all the relevant conformations of a complex system can be challenging or even an unfeasible task in general. However, it is possible to perform conformational analysis on such complex PES by adopting some of the advanced algorithms that have been developed around molecular dynamics (MD) or Monte Carlo (MC) techniques. Among the methods developed for the calculation of relative free energies of binding we just cite here the most relevant:

- Free energy perturbation (FEP)(10) calculations are mathematical procedures to gradually convert one chemical species to another in a thermodynamic cycle. The FEP technique combined with MD or MC simulation is employed to evaluate reaction free energy profiles. However, FEP calculations become quite complicated and computationally expensive for structurally dissimilar inhibitors and for calculation of absolute free energies of binding.
- Linear Response Approximation(LRA)(11; 12) is used taking into account that the electrostatic effects in solution have a linear response approximation to the changes in polarity of the solute. This technique is very used in complexation energies of ligands with receptors and enzymes.
- The Linear Interaction Energy method(LIE)(11; 13), which develops the LRA expanding the linearity concept to the non-electrostatic interactions, only considers two states: ligand solvated with water and ligand bind to receptor and requires no transformation processes. The main idea is to consider contribution from electrostatic and non-electrostatic interactions to the total binding energy separately. The polar part can be treated using the electrostatic LRA while the non-polar contribution must be calculated using an empirical formula calibrated against a set of experimental binding data.

Both the quality of the PES and the amount of sampling performed are important factors in the quality of the evaluated free energies. However, even the most sophisticated techniques show a difficult convergence of the results, which makes them of limited use for the study of a big number of interactions at the same time, a typical case when we are screening for compounds targetting a given receptor, for example.

1.1.2 Transition state theory used as framework for protein-small molecule interactions

In this thesis we discuss interactions of small molecules with proteins. As we have introduced before, proteins are molecules that can participate in different ways in biological processes and, thus, their interaction with small molecules responds to different aims. As ligands can activate or inhibit biochemical processes, it is necessary to, first, describe the physicochemical framework we will use through the text, in order to clearly contextualize all types of interactions. A useful framework for this discussion is provided by transition state theory (TST) which we will describe shortly.

TST is one of the most successful theories in chemistry. It gives the framework of chemical reaction rate theory and today it is the general name for many theories based in whole or in part on it(14).

Classically, the fundamental assumption(15) of this theory is that there exist a hypersurface (called transition state TS) in phase space with two properties:

- it divides space into a reactant region and a product region, and
- trajectories passing through this "dividing surface" in the products direction originated at reactants and will not reach the surface again before being thermalized or captured in a product state. This second assumption is often called the no-recrossing assumption or the dynamical bottleneck assumption.

In addition to the fundamental assumption, TST invariably makes two further quantum mechanical assumptions:

- The reactants are equilibrated in a canonical (fixed-temperature) or microcanonical (fixed-total energy) ensemble.
- The reaction is electronically adiabatic (the Born-Oppenheimer separation of electronic motion from intermolecular motions is valid) in the vicinity of the dynamical bottleneck.

TST suggests that as reactant molecules approach each other closely they are momentarily in a less stable state than either the reactants or the products. In this less stable state, the atoms rearrange themselves, original bonds are weakened and new bonds are partially formed.

This increase in potential energy corresponds to an energy barrier over which the reactant molecules must pass if the reaction is to proceed. The arrangement of atoms at the maximum of this energy barrier is called the activated complex or transition state and it

is a transitory intermediate state between reactant and product. The combination can either go on to form products or fall apart to return to the unchanged reactants. The energy difference between the reactants and the potential energy maximum is referred to as the activation energy. Activation energy is the excess energy over the GS that must be acquired by a chemical system in order for the reaction to proceed.

The rate constant at a given temperature T is

$$k_T = \gamma(T) \frac{1}{\beta h} \exp -\beta \Delta G^\ddagger \quad (1.12)$$

The magnitude ΔG^\ddagger is the free energy difference between the GS and the TS. The transmission coefficient γ is a correction term that stands for all the approximations assumed in the TST. h is Planck constant.² The most expensive and problematic task is the computation of the free energy difference ΔG^\ddagger , which must be calculated along a predefined reaction coordinate.

Using TST as our framework we can distinguish two types of interactions, between a protein and a ligand: interactions related with the GS and interactions with the TS. Of course, TS interactions will occur only in enzymes, but the generalization of these two states to all types of proteins is a useful tool for the discussion in the following paragraph.

1.1.2.1 Ground state

Ground state (GS) is the lowest allowed energy state of an atom molecule in a physical system. Here we understand GS as the reactant state. Protein interactions in GS usually happen between receptors and stable small molecules called ligands. Receptor is the macromolecule of an organism that interacting with a ligand produces a biological effect. In order to understand the biological meaning of this type of interactions we need to turn to experimental methods.

Receptor is the macromolecule of an organism that interacting with a ligand produces a biological effect.

We need to take into account two fundamental properties of a ligand in order to study the receptor-ligand binding:

- **Affinity:** defined as the ability of a ligand to bind an specific receptor and form ligand-receptor complex.
- **Intrinsic activity:** defined as the ability of the ligand to induce a biological effect after binding.

Depending on these properties, we can define three types of ligands:

- **Agonist:** ligand with great affinity for the receptor and high intrinsic activity. It generates a response similar to the natural ligand.

² $h = 6.626 \times 10^{-34} \text{ J.s.}$

- Antagonist: ligand with affinity for the receptor but with no intrinsic activity (it has not pharmacological response). These ligands diminish or inhibit, depending on dose, the effect of agonists impeding the receptor-ligands union or impeding the generation of the secondary reactions to form the complex receptor-ligands. These two types of mechanisms of action are used to describe two types of antagonists:
 - Competitive antagonists: agonist and antagonist compete for the union to the same binding site at the receptor.
 - Non-competitive antagonists: the agonist binds to a receptor in a different site of the agonist, but this site is necessary for the agonist to make its effect.
- Partial agonist or mixed antagonist: it has high affinity for the receptor but its intrinsic activity is lower than the agonist or the natural ligand. It causes an agonist or antagonist response, depending on the concentration of pure agonist. Then, at low concentrations of pure agonist, the partial agonist can increment the agonist effect, while at high concentrations of pure agonist, the partial agonist act as an antagonist.

Ariens(16), Stephenson(17) and Furchgott(18) said that the effect of a ligand is proportional to the number of complexes ligand-receptors and also depend on the intrinsic activity of the ligand.

Ligand-receptor interaction follows the mass action law:



Equation of association ligand(L)-receptor(R):

$$v_{ass} = [R][L]K_{on} \quad (1.14)$$

where v_{ass} is the association rate, $[R]$ is the concentration of receptors, $[L]$ the concentration of ligands and K_{on} the reaction constant.

Equation of dissociation ligand(L)-receptor(R):

$$v_{diss} = [RL]K_{off} \quad (1.15)$$

where v_{diss} is the dissociation rate, $[RL]$ is the concentration of ligand-receptor complexes and K_{off} is the reaction constant.

At equilibrium, association rate is equal to dissociation rate.

$$[R][L]K_{on} = [RL]K_{off} \quad (1.16)$$

Defining dissociation constant K_d at equilibrium:

$$K_d = \frac{[R][L]}{[RL]} = \frac{1}{K_a} \quad (1.17)$$

where K_a is the equilibrium constant for association.

Clark's theory assumes that(19):

- Ligand-receptor binding is reversible.

- The effect of a ligand is proportional to the number of occupied receptors.
- The effect of a ligand is maximum when all the receptors are occupied.

Nowadays we know that this theory is not exactly the reality, but in certain experimental conditions, these postulates are valid and they enable to extract quantitative conditions for the interactions ligand-receptor. Assuming Clark's theory(19) we can deduce the fraction of occupied receptors:

$$\frac{[RL]}{[B_{max}]} = \frac{[L]}{K_d + [L]} \quad (1.18)$$

where B_{max} is the maximum association:

$$B_{max} = [R] + [RL] \quad (1.19)$$

Interpreting equation 1.18 we can compute the occupation at binding sites:

- If $[L]=0$; $\frac{[RL]}{[B_{max}]} = 0$
- If $[L]=K_d$; $\frac{[RL]}{[B_{max}]} = 0.5 \rightarrow 50\%$ of occupation.
- If $[L] \rightarrow \infty$; $\frac{[RL]}{[B_{max}]} = 1 \rightarrow 100\%$ of occupation.

Then, K_d , the equilibrium dissociation constant of a ligand at equilibrium is the concentration of a ligand that produces binding to 50% of receptors.

Printing $[RL]$ in function of $[L]$ we obtain figure 1.1.

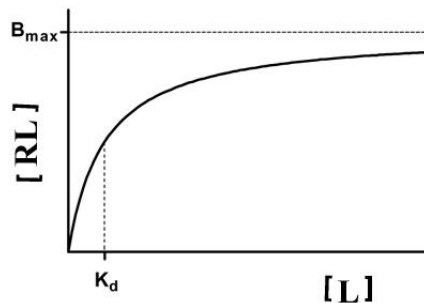


Figure 1.1: Curve dose-response. $[RL]$ in function of $[L]$ results and hyperbolic function

This model is only valid assuming that all receptors are equally accessible to ligands. Also it ignores any states of partial binding and assumes that binding is reversible. Also, the model does not take into account cooperativity: binding of a ligand to one binding site does not alter the affinity of another binding site.

In presence of competitive antagonists, the curve 1.1 moves to the right but there are not changes in the slope or in the maximum effect. Adding more agonists always we arrive

to maximum. In presence of non-competitive antagonists, the slope diminishes and also the maximum effect.

In the case of agonists, we can define $EC50$ (effective concentration) as the concentration of ligands necessary to arrive to 50% of the maximum response. It measures the activity of the ligand. For antagonist we define $IC50$ (inhibitory concentration) as the concentration necessary to block 50% of the response.

K_d is more useful than $EC50$ or $IC50$, because it is independent on the system, they are only dependent of the receptor and the ligand.

Another way to express activity is $pIC50$:

$$pIC50 = -\log IC50 \quad (1.20)$$

It is a measure for the binding affinity of the test compound to the receptors present in the cell membrane. If $pIC50$ is high, activity is higher.

However, there is another measure for binding, free energy: ΔG ,

$$\Delta G = -RT \ln K_a \quad (1.21)$$

If activity is high, binding energy decreases exponentially.

Concerning to the experimental activity, high quality and reliable biological data is required. The methods to evaluate biological activity are increasing in complexity: *in silico* methods, accounting for electronic and general molecular properties, *in vitro* methods, which provide a satisfactory description at cellular level, and *in vivo* methods, suitable to more detailed studies on specific organs and individuals. A precondition for the biological activity of a molecule is a high affinity to the binding site.

1.1.2.2 Transition state

For reactions involving more than three or four atoms, knowledge of the complete potential surface as a function of all $3N - 6$ nuclear coordinates (N is the number of atoms) is usually out of the question, and the effort is most often focused on determining special features of the PES: absolute and local minima, and saddle points that separate them.

Enzymes, catalyze biochemical reactions by binding tightly and specifically to their target molecules, called substrates, in the TS of the reaction. Almost every chemical reaction in a cell is catalyzed by these kind of proteins. Enzymatic reactions are involved in most biological processes. Enzymes accelerate a great variety of metabolic reactions allowing cells to carry out reactions that otherwise would not occur on biologically useful time scales. There is, therefore, broad interest in understanding the origin of what makes enzymes so efficient.

Many proposals have been put forward to rationalize the catalytic power of enzymes, but some of these are problematic or difficult to analyze quantitatively. Although mutation experiments have been extremely useful for identifying catalytic factors, they cannot identify the mechanism of the catalysis uniquely.

Nearly a century ago, the lock-and-key description of enzyme action was formulated by Emil Fischer(20). This simple metaphor conveys the basic principle of catalysis: that

each enzyme possess and "active site" tailored for recognition and stabilization of the rate-limiting transition state of the reaction it promotes. Enzymes just bind tightly and specifically to their target molecules called substrates. General statements suggest that the enzyme binds the TS better than the GS. But the real question is how this differential binding is accomplished and which catalytic groups are responsible.

To discuss this rate enhancement we will consider a generic enzymatic reaction and the corresponding reaction in water in Figure 1.2. To evaluate enzyme catalysis quantitatively, we first must choose a good reference. The most obvious reference is the uncatalysed reaction in water. So that the question becomes how the structured environment in the enzyme accelerates the reaction relative to the same process in a solvent cage.

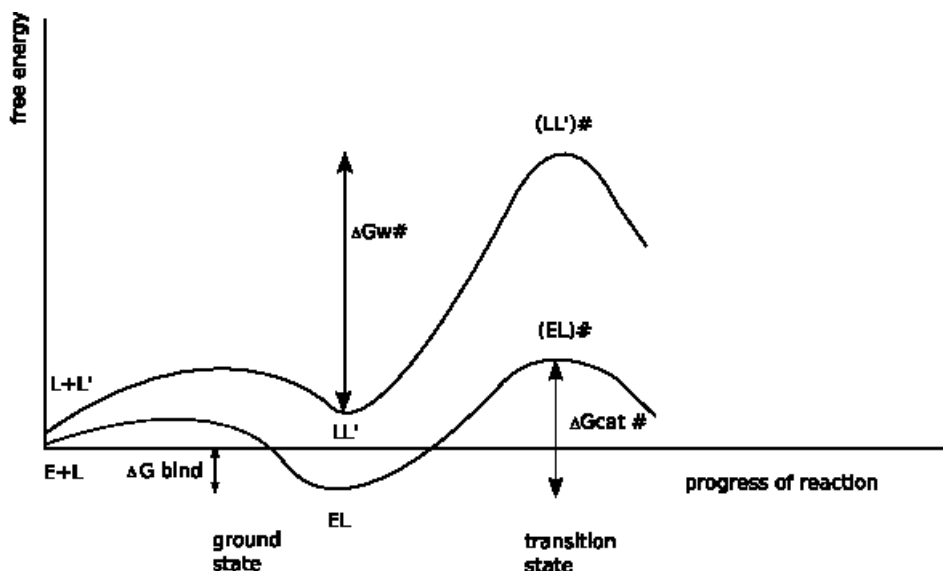


Figure 1.2: Comparing the free energy surfaces for an enzyme reaction and the corresponding reaction in solution. The substrate is designated by L (ligand) and the reactive part of the enzyme by L'

The main question addressed is the origin of the difference between $\Delta g_{\text{cat}}^{\ddagger}$ and $\Delta g_{\text{w}}^{\ddagger}$, where $\Delta g_{\text{cat}}^{\ddagger}$ and $\Delta g_{\text{w}}^{\ddagger}$ are, respectively, the activation barriers of the catalyzed reaction in enzyme and uncatalyzed reaction in water(21; 22). The fact that the activation barrier is reduced by the enzyme was stated by Pauling long ago(23).

The immune system is a rich source of highly efficient catalysts for common organic synthesis reactions. These catalysts are antibodies that have been identified in the immune system using small molecules known as haptens. Antibodies are immune responses to an antigen with high specificity. They have high affinity and specificity binding to an antigen. An antigen is a substance that provokes or is recognized by an antibody.

One generalized idea is that one could design antibodies that catalyze as enzymes creating TS analogs: stable molecules that simulates electronic structure of TS to get affinity. These TS analogs could be used to elicit antibodies with catalytic and selective function.

It is implicitly assumed that a proper transition state analog (TSA) can elicit a catalytic antibody (CA) with optimal binding to specific haptens. Then, we could create antibodies that catalyze in a similar way as enzymes. This idea was reviewed in (24; 25). To induce a binding site with a topology and stereoelectronic environment suitable for catalysis, a stable molecule which mimics the structure of the short-lived transition state of the target reaction is employed as a hapten. Transition state, ideally, it is the structure that a hapten should mimic. TSA approach employs antigens that are designed and prepared as TSAs of the target reactions. Alternatively, haptens carrying a point charge have been employed in order to recruit a complementary charged amino acid in the antibody active site to perform catalysis, which has been termed the "bait and switch" strategy, since the haptens designed according to this strategy serve as bait for eliciting catalytic functions during the immunization process, which is then switched for the substrate.

By exploiting the highly specific antigen binding properties of antibodies, experimental strategies have been devised to produce antibodies that catalyze chemical reactions. But CAs produced by TSA present low rate enhancement as compared to the corresponding enzymes. It appeared early on that antibody catalysts are limited in their capability to accelerate chemical reactions as natural enzymes in terms of efficiency(26). Best catalytic antibodies approach efficiency of the least efficient enzymes. Generating antibody catalysis that achieve enzymatic efficiency remains a challenging task, which has long been the source of a great interest both in the design of more effective haptens for immunization.

Abzymes (catalytic antibodies) are antibodies with variable regions possessing enzymatic activity. Also, a number of artificial enzymes or designer abzymes are being developed with any desired enzyme activity and specificity. There are two approaches used to develop the artificial abzyme: produce antibodies against a stabilized transition state or use molecular biology and site-directed mutagenesis.

Theozymes are theoretical catalysts, constructed by computing the optimal geometry for transition state stabilization by model functional groups. They may be used to quantify the relative stabilization of reactants and transition states by biological and non-biological catalysts.

One of the main problems is elucidating the origin of the catalytic power of enzymes is the difficulty of dissecting different energy contributions by direct experiments. The most effective way of determining the origin of enzyme catalysis is the use of computer simulation approaches. While conventional understanding of enzyme catalysis typically focuses on transition state stabilization, other proposals suggest that substrate strain or conformational effects are more important. There are two main groups of strategies to explain how enzymes work:

- Ground state destabilization (GSD).

The idea that enzymes work by increasing their GS(27) free energy has been frequently advanced(27; 28; 29). The most popular form of this proposal invokes non-polar active sites, *i.e.*, substrate destabilization by desolvation(29; 30; 31; 32; 33). It says that enzymes work by providing a non-polar environment that destabilizes highly charged GSs. Other alternatives for GSD may involve repulsive electrostatic interactions between the enzyme and the reacting region of the substrate; GSD by desolvation or Circe effect (the utilization of attractive forces to lure a substrate into a site in which it undergoes a transformation of structure). Near attack conformation (NAC) can be considered as a particular example of GSD.

- TS stabilization by electrostatic effects.

The idea that enzymes stabilize the TS primarily by electrostatic effects is consistent with numerous mutation experiments, which have shown that mutations of residues that stabilize the TS charge distribution in the native protein lead to increases in $\Delta g_{\text{cat}}^\ddagger$. In reactions in water, the solvent must pay a significant reorganization energy to orient the polar environment towards the TS charges. In enzymes, the active site dipoles associated with polar groups, internal water molecules, and ionized residues are already partially oriented toward the transition-state charge center(21; 34). Computer simulation studies have shown that the most important catalytic factor is stabilization of the TS by electrostatic preorganization of the enzyme active site, and that other effects are usually relatively small.

Non-equilibrium solvation, dynamical effects, quantum mechanical tunneling, entropic effects and other factors also have been used to explain enzyme catalysis.

1.1.3 Molecular interaction potentials and fields

Molecular interaction potentials (MIP) are scalar properties showing the potentiality of a chemical probe to interact with a given molecule. The probes typically reflect the chemical characteristics of a binding partner. The simplest probe is a proton and in such cases the potential is called molecular electrostatic potential (MEP, or MESP)(35). In more complex cases the probe can be a small molecule or a chemical group. Molecular interaction fields (MIF), on the the other hand, are vector quantities, and in this case one needs to worry about their magnitude and direction. In the case of the electrostatic field E :

$$\vec{E} = -\vec{\nabla}V \quad (1.22)$$

Which shows the relationship between the field and the potential. The electrostatic potential at field point \vec{r} due to a point charge Q at the coordinate origin is:

$$V(\vec{r}) = \frac{Q}{4\pi\epsilon_0} \frac{1}{r} \quad (1.23)$$

In our studies we have worked with MIPs, as we evaluate the capacity of interaction between a target molecule and a chemical probe and not an actual (energy) interaction between charges. In some studies both names are used indistinctively, MIP or MIF, and we will treat them as synonymous in what follows.

MIP can be computed analytically by means of quantum mechanics methodologies (see section 3.1.1.1 for more detailed information) or using the laws of molecular mechanics (see 3.1.1.2). Often, the target-probe energy is computed at regular intervals, inside a box that surround the molecule or the regions to be studied.

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.

By using computer graphics, MIP can be displayed as 3D isoenergy contours. Contours of large positive energies indicate regions from which the probe would be repelled,

while those of large negative energies correspond to energetically favorable binding regions. These regions can be exploited in the design of ligands to bind with high affinity and specificity to particular molecules.

1.2 Protein-ligand interactions in drug design

Most drugs that are nowadays used in human therapy interact with certain macromolecular biological targets, *i.e.* with enzymes, receptors, ion channels, and transporters. Molecules evolve their activity through specific binding to a macromolecular receptor. One of the main goals of drug research is to discover ligands that are predicted to interact favourably and bind strongly to its receptor active site, without interfering with the operation of other biomacromolecules in the living organism. Alternatively, this procedure can be reversed to search for hosts that interact strongly with a given ligand. Whereas most drugs are ligands, only a few ligands are drugs, because even small variations in chemical structure can influence whether the compound will be curative, physiologically inert, or toxic.

The basic aspects of ligand-protein interaction may be summarized under the term "molecular recognition" and concern the specificity as well as stability of ligand binding. Many therapeutic agents act by binding specifically and tightly to a particular macromolecular target such as a receptor protein or a nucleic acid. Nowadays, computer-aided prediction and intelligent molecular design make a large contribution to the constant search of improving protein inhibitors or activators.

Molecular modeling is one of the key techniques used during the drug discovery process. In molecular modeling, a molecule is 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 using molecular mechanics.

The goal of structure-activity modelling is analyse and detect the determining factors for the measured activity for a particular system, in order to have an insight of the mechanism and behaviour of the studied system. The factors governing the events in a biological system are represented by a multitude of physicochemical descriptors, determined empirically in the past but, more recently, they can be calculated by using computational methods. Interactions between a ligand and a molecule usually are controlled by the molecular interaction potentials.

Drug design can be approached in one of two different ways: receptor-based or non-receptor-based.

Receptor-based or direct approaches depend on the availability of 3D structure of the target protein in order to simulate *in silico* and to evaluate the most favourable conditions for interaction of ligands with their binding site. That opened an avenue to deriving computational methods to predict the binding orientation of ligands inside protein cavities, a process that is generally referred to as molecular docking. The molecular docking problem can be defined as follows: given the atomic coordinates of two molecules, predict their correct bound association. In its most general form, no additional data is provided. In practice, however, additional biochemical information may be given, in particular knowledge of the binding sites. In addition it also provided the structural basis for understanding the principles of protein-ligand recognition at a molecular level and deriving scoring schemes for the estimation of ligand binding affinities(36).

Nonreceptor-based or ligand-based approaches are used when the 3D structure is not known. Rational drug design must be achieved in an indirect way by developing an empirical model to describe the structure-activity relationships (SAR) for a data set of bioactive compounds. These model can be used to design new compounds with improved activity.

Of course, both approaches complement each other and can be combined depending on how much biostructure information is available or becomes available during the project.

1.2.1 Receptor-based drug design. Direct methodologies

The receptor-based approach is a design strategy for new chemical that applies when a reliable model of the receptor site is available. The 3D structures of protein complexes provide many insights on protein interactions, allowing more rational approaches toward drug development and the treatment of disease. Structural information from protein-ligand complexes provides key structural information on their bioactive conformation and orientation into the cavity.

The first phase is to determine the structure of the binding site using standard structural analysis from X-ray diffraction, nuclear magnetic resonance (NMR), homology modelling, or calculations involving molecular mechanics and molecular dynamics techniques. These structures represent snapshots of the protein-ligand recognition process and provide a valuable source of information to further understand the rules that govern the interaction between proteins and ligands(36).

In the absence of structural information, homology of the unknown receptor sequence with known structures that have been identified through database searches may be a good starting point. A homology model is a model of a protein, whose 3D structure is unknown, built from, e.g., the X-ray coordinate data of similar proteins or using alignment techniques and homology arguments. Then one have a model of the receptor site which identifies few specific interactions that are responsible for the binding.

The availability of the 3D coordinates of a protein opened an avenue to deriving computational methods to predict the binding orientation of ligands inside protein cavities, a process that is generally referred to as molecular docking(37; 38; 39).

The next phase is to search databases for new ligands that may bind to the chosen receptor. The results of the database search may be used directly or modified to produce candidates for further study. On the assumption that similar ligands will adopt similar binding modes, a new wave of developments in docking methods(40; 41; 42; 43) permits now the explicit incorporation of information extracted from protein-ligand complexes to actively guide the binding mode of new compounds into protein cavities (enzyme or other binding site), resulting in more accurate binding mode predictions. Prediction of the binding constants is usually performed using Gibbs free energy perturbation studies. Finally, the candidates are synthesised and tested in the laboratory.

1.2.1.1 Experimental determination of biomolecular 3D structures

Knowledge of the detailed molecular architecture of proteins has been a source of insight into how proteins recognize and bind other molecules, how they function as enzymes,

how they fold, and how they evolved. Protein structure determines function, given that the specificity of active sites and binding sites depends on the precise 3D conformation.

In the seventies, the development of X-ray crystallography and NMR provides the first 3D structures of the biological targets (hemoglobin and myoglobin (44; 45)), sometimes as complexes with a ligand bound. This new source of structural information opened the door to the structure-based drug design. In 1977 was created the Brookhaven database, known as PDB (Protein Data Bank(46)), where one can find all the published experimental structures. There is an increasing number of structures present in the PDB containing a co-crystallized drug-like molecule bound to the protein cavity. There are special databases in order to search these types of complexes, like PDBsum or Relibase. The number of proteins with a known 3D structure is increasing rapidly, and structures produced by structural genomics(47) initiatives are beginning to become publicly available.

NMR spectroscopy and X-ray crystallography are two of the most important experimental techniques for elucidating the conformation of proteins.

NMR spectroscopy technique depends on the fact that certain atomic nuclei are intrinsically magnetic. A family of structures generated from NMR structure analysis indicates the range of conformations for the protein in solution. At present, NMR spectroscopy can determine the structures of only relatively small proteins, but its resolving power is certain to increase. The NMR method is especially useful when a protein of interest has resisted attempts at crystallization, a common problem for many membrane proteins.

X-ray crystallography provides the finest visualization of protein structure currently available. This technique can reveal the precise 3D positions of most atoms in a protein molecule. The use of X-rays provides the best resolution because its wavelength is about the same length as that of a covalent bond. The technique requires that all molecules be precisely oriented, so the first step is to obtain crystals of the protein of interest. This is sometimes the great difficult to apply this methodology. There are still many proteins, especially membrane proteins, that have so far resisted all attempts to crystallize them.

1.2.1.2 Computational prediction of target binding modes

If the target has pharmacological interest probably have been cristalized joined to several ligands. In this case have have great information about the binding modes of different ligands in the active site of the protein and we can align new ligands knowing the previous information. We need to discover the biological interactions between target and ligand (with mutagenesis experiments) and using molecular modelling programs to dock both molecules. This type of programs try to join experimental data and molecular models in order to find the best model.

On the other hand exist other automatic methods to explore possible binding modes of new ligands to the target. This programs are called docking programs(37; 38; 39; 48). Docking tries to predict the binding modes (whether a given conformation and orientation of a ligand fits the active site of the protein) between the ligand and protein. The objective of molecular docking is to obtain the lowest free energy of binding. This is of fundamental importance in modern structure-based drug design.

Docking small-molecular-weight ligands to therapeutically relevant macromolecules has become a major computational method for predicting protein-ligand interactions and

guide lead optimization. From the pioneering work of Kuntz(49), numerous docking programs based on very different physicochemical approximations have been reported(37; 38; 48). They use an exhaustive exploration of different binding modes (posing), evaluating the intermolecular interactions at every position. Finally the program obtains a collection of solutions ranked based on a scoring function. This two steps have been solved in several ways in every one of the knowed docking programs.

The biological activity of a ligand also depends on its flexibility. Molecules with several rotatable bonds may adopt many different geometries. If a frozen conformation differs from the bioactive conformation of the flexible lead or if the added atoms interfere with the binding, biological activity will be more or less destroyed.

1.2.1.2.1 Sampling the orientation of the ligand in the receptor cavity. In this process one have to determine whether a given conformation and orientation of a ligand fits the active site. On finding binding modes target-ligand ideally one have to consider all the possible combinations of translation and rotation degrees. Another important issue is to consider the ligand flexibility and protein flexibility. This is a complex problem to solve in a short computational time.

The two critical elements in a search procedure are speed and effectiveness in covering the relevant conformational space.

Early docking methods were primarily based on the lock and key principle(20) and thus focused mainly on the use of geometric criteria to assess the degree of shape complementarity between ligand and binding site(49; 50). However, it was soon realized that chemical complementarity between ligand and binding site had to be also taken into account in docking approaches to reduce the number of physically unrealistic solutions being obtained on the basis of shape alone.

Nowadays there are programs that look for chemical and geometrical similarity between the ligand and a binding-site template (SLIDE) that define points for favorable interactions(51) with the protein surface atoms. Another programs(43) use and idealized active site. The construction is based on protein residues that constitute the active site.

On the beggining there were programs that did not consider any kind of flexibility in the ligand and the proteins. Then it was a rigid docking(49; 50; 52; 53). The increase in computer power has permitted recent docking methods to account for ligand flexibility(54; 55; 56; 57; 58; 59; 60; 61) and receptor flexibility(62; 63; 64; 65; 51; 66; 67; 68).

Treatment of ligand flexibility can be divided into three basic categories: systematic methods (incremental construction(49; 54; 69), conformational search in databases); random or stochastic methods (Monte Carlo(58), genetic algorithms(56), tabu search(70)); and simulation methods (molecular dynamics, energy minimization). They fall into two basic categories: those in which the ligand molecule is either incrementally built or flexed during the search and those in which rigid precomputed conformers from a database are oriented in the target binding site(71). The first category includes methods that employ Monte Carlo sampling(72), Monte Carlo simulated annealing (QXP, AUTODOCK 2.4, LigandFit(73)), genetic and evolutionary algorithms(56; 74; 58; 75), systematic search techniques(69), and incremental construction(76; 54; 49; 77). The second group includes rigid docking of flexibases, in which individual conformers are separately docked and scored and only the best-scoring conformer of each molecule is saved(78), and rigid docking of conformational

ensembles(71) generated by overlaying related conformers.

The treatment of protein flexibility is less advanced than that of ligand flexibility. Accounting for protein flexibility in protein-protein docking algorithms is challenging, and most algorithms therefore treat proteins as rigid bodies or permit side-chain motion only, as implemented in GOLD(56). Another method of treating protein flexibility is to use ensembles of protein conformations (rather than a single one) as the target for docking(63). Another strategy is to sum different conformations of the target in a unique diffuse conformation, like it is implemented in FLeX(54) or AUTODOCK(58). More recently have been introduced programs that include partially protein flexibility like QXP(72).

1.2.1.2.2 Scoring ligand-receptor interaction. A search algorithm may produce an immense number of solutions, unmanageable for any practical need. The purpose of the scoring function is to discriminate between correct native solutions with low RMSD from the crystal complex and others within a reasonable computation time.

Scoring functions(79; 80) applied to single conformations of the docked complex is a more empirical approach to affinity prediction. They are generally based on identifying individual points of intermolecular interaction such as hydrogen bonds, ionic interactions and hydrophobic interactions, as well as entropy estimates, in a given conformation of the receptor-ligand complex and assigning a binding energy score to each contributing factor. Finding the binding mode and ranking the solutions involve scoring. The pose score is often a rough measure on the fit of a ligand into the active site. To determine the rank score is necessary to estimate binding energies. Docking methodologies are designed to predict the biological activity through the evaluation of interactions between compounds and potential targets.

The scoring function should be fast enough to allow its application to a large number of potential solutions and, in principle, effectively discriminate between native and non-native docked conformations.

Free-energy simulation techniques have been developed for quantitative modeling of protein-ligand interactions and the prediction of binding affinity. However, these expensive calculations are impractical for the evaluation of large numbers of protein-ligand complexes and are not always accurate. Scoring functions implemented in docking programs make various assumptions and simplifications in the evaluation of modelled complexes. Docking methodologies try to estimate binding energy between ligand and receptor at every position. Essentially, three types or classes of scoring functions are applied: force-field-based, empirical and knowledge-based scoring functions.

A force-field is a function expressing the energy of a system as a sum of diverse molecular mechanics terms. Force fields usually quantify the sum of two energies, the receptor-ligand interaction energy and internal ligand energy. Interactions between ligand and receptor are most often described by using vdW and electrostatic energy terms. Standard force-field scoring functions have major limitations, because they were originally formulated to model enthalpic gas-phase contributions to structure and energetics, and do not include solvation and entropic terms. There are docking programs that incorporate a full force field, like GROUP (included in GRID(81)) or simplified force fields as in QXP and AUTODOCK which use AMBER(82). Also CHARMM(83) used in DARWIN(84).

The knowledge-based scoring functions estimate free energies of molecular interactions

from databases(85; 86; 87; 88). The approach essentially involves converting inter-atomic distance distributions found in protein-ligand complexes into pair-potential functions for the different pairs of protein-ligand atom types using statistical mechanics. An estimation of the free energy of interaction between a ligand and a protein is then obtained by adding the contributions from protein/ligand atom pairs within a certain distance. An example is Potential mean force (PMF)(85) and DrugScore(86). The major attraction of many knowledge-based scoring functions is their computational simplicity, which permits efficient screening of large compound databases. Furthermore, such a statistical approach implicitly incorporates physical effects not yet fully understood from a theoretical point of view, for example, solvation and polarization.

The most rapid methods for estimating binding free energies are so-called empirical scoring approaches. These are based on simple energy functions ((89; 69; 90) or on the frequency of occurrence of different atom-atom contact pair in complexes of known structure (85; 86), respectively. In statistical potentials we can assign a value to every interaction using empirical parameters. Then we sum that values for all the interactions and we obtain an empirical estimation of binding energy. This methodology is used in FlexZ(54). Pose clustering(54) is a method from pattern recognition applied to ligand orientation based on physico-chemical interactions. We have to differentiate between empirical (statistical potentials) and knowledge-based scoring functions. The term "empirical scoring function" stresses that these quality functions approximate the free energy of binding, as a sum of weighted interactions that are described by simple geometrical functions of the ligand and receptor coordinates. These approaches are very fast, but usually at the cost of accuracy.

Most empirical scoring functions are calibrated with a set of experimental binding affinities obtained from protein-ligand complexes. The docking problems are not solved yet and none of the currently available programs are perfect in predicting the correct binding modes(91).

1.2.2 Ligand-based drug design. Indirect methodologies

1.2.2.1 Correlating structural properties and biological activity

Ligand design is the design of ligands using structural information about the target to which they should bind, often by attempting to maximize the energy of the interaction. Often enough, no structural information on a particular receptor protein is available. However, frequently a considerable number of different ligands is known together with their measured binding affinities towards a receptor under consideration.

We assume that similar interaction capabilities are related with similar biological activities. Ligand-based design starts with a group of ligands that have known binding constants or biological activities. Structurally similar compounds with high activity, with no activity, and with a range of intermediate activities are required. The first phase is to determine the structure of the ligands. The next phase is to generate a query for database searching.

In drug design it is attempted to correlate structural molecular properties (descriptors) of drug molecules with their biological activity (*i.e.* physicochemical properties, biological activities, toxicity, etc.) for a set of similar compounds, by means of statistical methods. As a result of this methodology called Quantitative Structure-Activity Relationship (QSAR), a simple mathematical model that connects experimental measures with a set of chemical

descriptors for a set of compounds is established. The parameters used in QSAR should be meaningful, and easily interpretable, in physical terms. The model derived should have a good predictive capabilities as possible to predict the studied biological or physicochemical behaviour for new compounds. In QSAR there is no underlying theory about the relationship between activity and structure. For this reason, QSAR models are empirical models which provide an approximate solution. It is important to remark the difference between correlation and causation. A satisfactory QSAR correlation does not mean that a particular descriptor causes the efficient action of a compound. The lack of evidence on causation might be complemented by additional information on the various mechanisms leading to the biological activity. QSAR techniques include from chemical measurements and biological assays to the statistical techniques and interpretation of results.

Since the introduction of the Hansch equation(92) in the 1960s, the number of algorithms available for quantitative structure activity (QSAR) studies has increased explosively. Since long, medicinal chemists used this concept to modify the structures of biologically active compounds(93; 94; 95; 96). QSAR attempt to find what features of a molecule affect its activity and what can be modified to enhance their properties. Hence, for a series of biologically active molecules, any systematic variation in chemical structure from one to another is expected to be reflected in a proportional analogous variation in the biological response.

QSAR expresses a multivariate mathematical relationship between a set of physicochemical properties or descriptors x_i , and a experimental function or biological activity, y_i :

$$y_i = x_i b_i + e_i \quad (1.24)$$

where b_i are the linear slopes that express the correlation of the particular molecular property x_i with the activity y_i of the compound i and e_i is a constant.

The slopes and the constant are often calculated using regression analysis. The independent variables, so-called descriptors, are usually physicochemical properties that describe some aspects of the chemical structure, which may be either experimentally or theoretically determined. The improper choice of independent variables can result in poor QSAR models.

Many mathematical descriptions have been used in drug design, and more specifically in the field of quantitative structure-activity relationship (QSAR). Goodford(81) introduced the concept of molecular interaction field (MIF) and the work of Cramer et al.(97) who introduced the 3D chemical structure into the description of the compounds and hence developed the concept of 3D-QSAR, QSAR based on 3D models because they allow for the simulation of directional forces:hydrogen bonds, metal-ligand contacts, polarization effects, and the interaction between electric dipoles. Three-dimensional quantitative structure-activity relationships (3D-QSAR) involves the analysis of the quantitative relationship between the biological activity of a set of compounds and their 3D properties using statistical correlation methods. These new techniques, which introduce 3D parameters in the description of compounds, allow calculations extensive to the space surrounding the molecules and require the alignment of the molecules to a common pharmacophore (a 3D space representation of the collection of common functional groups within the group of active compounds, complementary to the geometry of the receptor site).

Molecular interaction potentials have been used in 3D-QSAR. MIP identify regions where certain chemical groups can interact favorably, suggesting positions where a ligand should place similar chemical groups. Regions showing favorable energy of interaction represent positions where groups of a potential receptor would interact favorably with the ligand. Using different probes, one can obtain for a certain ligand a set of such positions which defines a virtual receptor site. This abstract entity defines an ideal complementary site for a certain chemical compound and represents its potential ability to bind a biomolecule.

In summary, the objectives of QSAR models are to allow the prediction of biological activities of untested and sometimes yet unavailable compounds, and to provide insight of which relevant and consistent chemical properties are determinant for the biological activity of compounds.

These indirect methodologies have serious limitations. First, the ligands must bind to the target protein at the same location and preferably adopt the same binding mode. Second, models generated on the basis of molecular superpositions allow only to interpolate between the data, *i.e.*, a region of space which is not occupied by any of the compounds cannot be judged. Finally 3D molecular models are usually restricted to low energy conformations since the number of accessible conformers of a molecule increases dramatically with the conformational energy tolerated.

1.2.2.2 Biomolecular similarity

Assuming conservation of the binding mode, when a variety of substructures appears in a particular site of a series of active molecules for a particular target, they are normally a reflection of the characteristics of the chemical substituents allowed in that particular protein site. This property is commonly referred to as bioisosterism(98; 99). Since they interact with the same protein environment, bioisosteric chemical fragments should have a certain degree of similarity.

When an active compound is already known, similar compounds(95) can be searched in a database hopping that such compounds will have similar biological properties. Similarity methods have the ability to score target ligands on the basis of their relative superposition with respect to a reference ligand. The main applications are selection of compounds with similar activity to a given compound (similarity analysis), derivation of Structure-Activity Relations (SARs). Similarity analysis is also used in a reverse way to select the most diverse subset (diversity selection) from a given set of compounds. Sometimes exists a similarity paradox, where a small change in the chemical structure leads to a drastic change in the biochemical activity.

Similarity between chemical compounds is perceived often intuitively based on expert judgement: similar backbone and almost the same functional groups or atoms.

When similarity is measured with respect to some feature, this feature has to be relevant to the activity of interest. Some measures of similarity are more relevant than others. A vast number of methods of quantitative molecular structure description (topology, shape, physicochemical properties, quantum chemical descriptions, etc.) and comparison (similarity coefficients) have been proposed and applied to date.

1.2.2.2.1 Descriptors used in biomolecular similarity calculations. There are different features considered for quantifying the similarity between molecules:

- Topology(100; 101) These methods are based on representation of chemical compounds as molecular graphs and limited to extraction and processing of 2D topological information (molecular fingerprints). The simplest descriptors are counts of individual atoms, bonds, rings, pharmacophore points, degree of connectivity indices between atoms(102). Two-dimensional fragment descriptors (atom-centered, bond-centered, ring-centered fragments) have been studied in detail(103). The simplest of distance-based descriptors are distances between atoms or between functional groups. Angle-based descriptors are based on generalized valence angles and torsion angles. Potential pharmacophore points are a generalized mix of distance and angle descriptors(104).

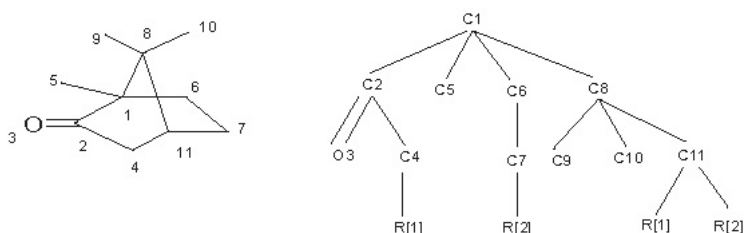


Figure 1.3: The figure shows the representation of a chemical graph of camphor

- Physicochemical properties(56; 105; 106; 107) Physicochemical properties are macroscopic descriptions of the substances. Examples are molecular weight, octanol-water partition coefficient (log P), total energy, heat of formation, ionization potential and molar refractivity. These properties are widely used in assessing similarity between chemicals(95; 108; 109; 110).
- Pharmacophoric points(111). Provides a mapping of pharmacophoric features across the set of molecules.
- Shape(112; 113; 114; 115; 116) This description is considered important because of hypothesized lock and key interaction with receptors. Descriptors such as vdW volume and surface area can reflect the size of substituents, but they contain very little information about shape. Other methods compute the common-overlap steric volume(117) between pairs of molecules. Also we can use geometrically invariant molecular surface descriptors(118). Shape calculations are slower than others, since very fine grids are required to obtain precise results.
- Field(72; 97; 119; 120; 121; 122) It is a comparison of reactive properties of molecules, like electrostatic potentials and fields(123; 124; 125; 126; 127; 128). The molecular electric field (MEF) is less frequently used, but it is important because the scalar product of the field and a dipole gives the energy of the dipole at a given point(129). Dipolar interactions are important in ligand-macromolecule binding and in solvation. Gasteiger et al.(130) describe an approach based on neural networks. Apaya et al.(131) provide an approach on the basis of the matching of local extrema of the

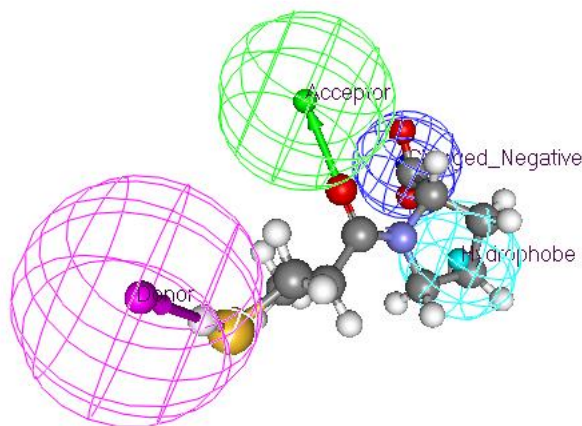


Figure 1.4: Pharmacophoric points of captopril

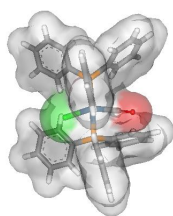


Figure 1.5: Molecular shape of tridium complex representing the vdW radii

MEP. Blaney et al.(132) calculate MEP on a regular grid and is mapped to the surface of a sphere by gnomonic projection. Two rigid structures can be compared by calculating differences of MEP values at points on the sphere. Mestres et al.(133) use two types of Gaussian-based molecular fields to evaluate molecular similarities. An atom-centered steric-volume field and an united-atom point-charge electrostatic potential(134) are used to represent the steric and electrostatic features of a molecule, respectively. Tervo et al.(135) have constructed recently fast grid-based algorithm for rigid-body molecular superposition and similarity searching. It aligns molecules using field information derived from charge distributions and van der Waals shapes of the compounds.

1.2.2.2.2 Similarity correlation coefficients. There is a number of different correlation coefficients that might be appropriate depending on the kinds of variables being studied

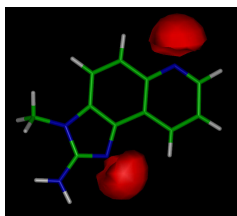


Figure 1.6: Typical MEP distribution of planar molecules

(continuous or discrete). All of them measure the degree to which two variables are related.

In our work usually we use molecular interaction potentials, then our variable should be continuous, but in some cases can be useful to transform the continuous MIP into discrete variables.

Continuum variables.

If we have continuum distributions (V_i^α, V_i^β) we can analyze their similarity with:

- PEARSON index(136). The similarity is obtained as a covariance between the two distributions normalized by the square of the variance of each one:

$$s_k^{\alpha,\beta} Pearson = \frac{\sum_{i=1}^m (V_i^\alpha - \bar{V}^\alpha)(V_i^\beta - \bar{V}^\beta)}{\sqrt{\sum_{i=1}^m (V_i^\alpha - \bar{V}^\alpha)^2 \sum_{i=1}^m (V_i^\beta - \bar{V}^\beta)^2}} \quad (1.25)$$

where \bar{V}^ξ is the mean of values V_i^ξ .

The assumption of the Pearson correlation coefficient is that there is a bivariate normal distribution. This means that for each value of V_i^α there is a normal distribution of V_i^β and for each V_i^β there is a normal distribution of V_i^α . Pearson correlation coefficient measures the strength and direction of a linear relationship between the V_i^α and V_i^β variables. This coefficient goes between -1 and 1 . If the distributions are proportional Pearson value is 1 , if they are independent, the value is 0 and -1 if they are inverse proportional.

- COSINUS index(137). Similar to Pearson, but without centering the variables:

$$S_k^{\alpha,\beta} Cosinus = \frac{\sum_{i=1}^m V_i^\alpha V_i^\beta}{\sqrt{\sum_{i=1}^m (V_i^\alpha)^2 \sum_{i=1}^m (V_i^\beta)^2}} \quad (1.26)$$

The value of the coefficient lies within the interval $[0,1]$. The involved molecules can be considered to be more similar as this index approaches to 1 , and dissimilar if it approaches to 0 .

Carbó et al.(138) substituted sumatories for integrals and this index has been used to compare electronic density distributions:

$$S_k^{\alpha,\beta} Carbo = \frac{\int \rho^\alpha \rho^\beta dv}{\sqrt{\int (\rho^\alpha)^2 dv} \sqrt{\int (\rho^\beta)^2 dv}} \quad (1.27)$$

Also the technique has been extended to compare molecular electrostatic potentials and electric fields(124; 139; 140; 141).

- HODGKIN index(124). When comparing two sets of potentials with different values but one being proportional to the other, Pearson index (eq. 1.2.2.2) would yield a perfect correlation. This can be avoided by using the Hodgkin index defined as:

$$[S_k^{\alpha,\beta}]_{Hodgkin} = \frac{2 \sum_{i=1}^m (V_i^\alpha - \bar{V}^\alpha)(V_i^\beta - \bar{V}^\beta)}{\sum_{i=1}^m (V_i^\alpha - \bar{V}^\alpha)^2 + \sum_{i=1}^m (V_i^\beta - \bar{V}^\beta)^2} \quad (1.28)$$

where m is the number of points of the tridimensional grid and \bar{V}^ξ is the mean value of the distribution V_i^ξ .

Thus the formula gives a total similarity of both shape and magnitude of the distribution. The use of Hodgkin similarity index is particularly important for calculating the MEP and MEF similarity because the shape of the distributions is similar but not the magnitudes as well. This index varies in the range of values from 0 to 1. It gets the value of 1 when the distributions in the two molecules are identical. It can be shown that the Hodgkin index will always be less than or equal to the Cosinus index (eq. 1.26).

- SPEARMAN index (136; 142). The MIPs in the grid points are ranked according to their values. The similarity is computed using these ranks:

$$s_k^{\alpha,\beta} Spearman = 1 - 6 \sum_{i=1}^m \frac{(rank_i^\alpha - rank_i^\beta)^2}{m(m^2 - 1)} \quad (1.29)$$

where m is the number of points to compute.

The nonparametric (distribution-free) rank statistic proposed by Spearman in 1904 is a measure of the strength of the associations between two non-linear variables. When normal distribution is not satisfied Spearman is the nonparametric analog of the usual Pearson correlation coefficient (eq. 1.25). It is calculated by converting each variable to ranks and calculating the Pearson correlation coefficient between the two sets of ranks. Spearman index values are between -1 and 1 . It is very useful and it is not altered significantly for big MIP values.

- GAUSSIAN index(143; 119; 126; 120; 144)

Kearsley and Smith(119) describe an alignment function for the superposition of two rigid molecules that comprises a double sum over all possible atom pairs between both molecules:

$$s_k^{\alpha,\beta} Gaussian = \frac{\sum_{i=1}^{n^\alpha} \sum_{j=1}^{n^\beta} V_i^\alpha V_j^\beta \exp(-\alpha r_{ij}^2)}{\sqrt{\sum_{i=1}^{n^\alpha} \sum_{j=1}^{n^\alpha} V_i^\alpha V_j^\alpha \exp(-\alpha r_{ij}^2)} \sqrt{\sum_{i=1}^{n^\beta} \sum_{j=1}^{n^\beta} V_i^\beta V_j^\beta \exp(-\alpha r_{ij}^2)}} \quad (1.30)$$

where n_ξ ($\xi = \alpha, \beta$) is the number of points in each grid box selected for the comparison, V_i^ξ is the potential value in the grid point for molecule ξ and r_{ij} is the distance between two points. The smoothing parameter α determines the attenuation range of this distance dependence. With small values of α also remote parts of each molecule will influence the alignment. Kearsley and Smith indicate that the parameter selection is crucial for the produced alignments and the relative ranking of the different solutions.

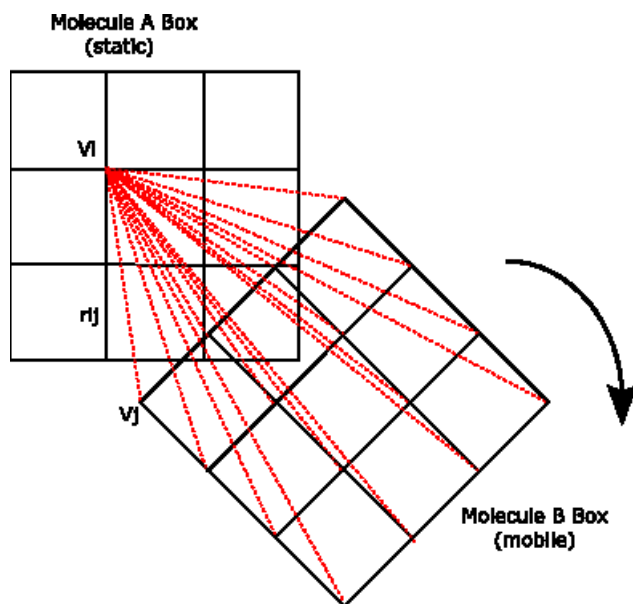


Figure 1.7: Calculation of Gaussian index. It superposes two molecules computing a double sum over all possible atom pairs

Binary variables.

Sometimes continuum variables can be transformed to dicotomic (binary variables) classifying them in two groups with a certain criterion. We can create a contingency table 2x2:

	V^α with value $< A$	V^α with value $> A$
V^β with value $< A$	a	b
V^β with value $> A$	c	d

There are a lot of similarity coefficients for binary distributions. Some of these coefficients are:

- Simple matching coefficient(145).

$$s_k^{\alpha,\beta} SM = \frac{a + d}{a + b + c + d} \quad (1.31)$$

- Russel and Rao coefficient(146).

$$s_k^{\alpha,\beta} RR = \frac{a}{a + b + c + d} \quad (1.32)$$

- Jaccard coefficient(147).

$$s_k^{\alpha,\beta} J = \frac{a}{a + b + c} \quad (1.33)$$

- Rogers and Tanimoto coefficient(148).

$$s_k^{\alpha,\beta} RT = \frac{a + d}{a + 2(b + c) + d} \quad (1.34)$$

All of them have a value that lies in the range of [0,1]

1.2.2.3 The alignment problem

When series of compounds need to be compared on the basis of their MIP, the comparison cannot be performed directly. Unfortunately, the values of each MIP are sensitive to the orientation of the structure used to generate the MIP. Then, a previous step of structural alignment (overlapping of functional groups of molecules in the space) of the compounds is required. However, 3D-QSAR methods usually require an accurate superposition of structures, which has proven to be their greatest weakness, as this procedure usually requires considerable human intervention and is generally regarded to be the most arduous and time-consuming phase of these analysis. This severely limits the efficiency of 3D-QSAR techniques when dealing with large libraries of molecules. The relevant chemical features of the ligands can be readily extracted in order to derive a pharmacophore model. Several active compounds can be aligned in order to visualize or determine a pharmacophore in the absence of a receptor model. This pharmacophore, a template of functional groups in the desired positions, may then be used to develop new compounds or to use as a receptor model. QSAR studies may provide an estimate of the binding affinity of a novel ligand towards the receptor under consideration.

Several assumptions are made in the development of alignment methodologies. Foremost is that the set of structures all interact at the same active site on the target macromolecule through the same active site groups. Another assumption is that the active site is not greatly distorted in different ways by the binding of the various structures. Moreover, structures are maintained in a rigid preselected low-energy conformations during the alignment analysis. The template molecule may be a lead compound, or a desired structure that is complementary to a receptor molecule. One want to find the conformational resemblance to the template. Possible active conformers of the molecule should be rotated to be aligned with the other molecules in the study. In applying this strategy, the

minimum energy conformers are assumed to bind most favorably in the receptor site although, in fact, there is no a priori reason to exclude higher energy conformers as the source of activity.

Structural alignment is a complex task, specially 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. The principal methods for molecule superposition mainly differ in the treatment of conformational flexibility, the optimization algorithm used and the definition of molecular similarity. Lemmen and Lengauer(149) have a good review about methods to align molecules.

Algorithmic methods for aligning molecules are:

- Directed tweak technique(150). By the use of local coordinates for the handling of rotatable bonds it is possible to formulate analytical derivatives of the objective function. With a gradient-based local optimizer flexible rms-fits are obtained extremely fast.
- Volume overlap optimization. Molecules are represented by a set of spheres(151) or Gaussians(122; 133; 152; 153) or MIP(126), and the overlap between them is quantified by means of a similarity measure.
- Clique detection(154). Structures are represented by point sets. Depending on a distance tolerance, the algorithm generates a so-called distance compatibility graph. The matching procedure uses clique detection(155) to determine overall valid distance constraints.
- Distance geometry(156). Molecules are described in a translation and rotation invariant fashion. Conformational flexibility can be addressed, as well, by providing distance intervals for all atom pairs.
- Genetic algorithms(157). GA are a class of computational problem-solving approaches that adapt the principles of biological competition and population dynamics. Model parameters are encoded in a chromosome and stochastically varied. Chromosomes yield possible solutions to a given problem and are evaluated by a fitness function. The chromosomes that correspond to the best intermediate solutions are subjected to crossover and mutation operations analogous to gene recombination and mutation to produce the next generation. The information content of the chromosomes, in this case, are the orientational degrees of freedom and a coding of the torsional degrees of freedom and a coding of the torsional degrees of freedom in the case of considered molecular flexibility. The fitness function usually is a similarity function.
- Geometric hashing(158; 159). The technique comes from the field of computer vision. It is based on the encoding of a set of geometric information in a hashtable 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.
- RMS-fitting(160) of rigid-body objects which is possible when a common structural core is shared by the compounds of the series.

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 3D 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. 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 that capture certain 3D molecular features which makes easier its analysis.

An alternative approach tries to circumvent the superposition problem by using QSAR descriptors which are sensitive to the 3D structure of the molecule but do not require structural superposition. These methodologies use autocorrelation functions and neural networks to create coordinate independent QSAR descriptor or inherently coordinate independent descriptors.

We call alignment-free descriptor any type of 3D molecular descriptor that is translational and rotational invariant, i.e. it is insensitive to the position and orientation of the molecular structures in the space and does not require the structural superimposition of the compounds studied. Such as the approaches suggested by Broto(161), Gasteiger et al.(162; 163), and Clementi et al.(164). Recently, a number of procedures have been proposed that eliminate the requirement of superposition between molecules(165; 166; 167). There are examples of alignment-independent applications for developing a virtual receptor site(168; 169; 170).

If alignment is such an important drawback, one might ask why alignment-free descriptors are not more popular. The main reason is related to the difficulties of understanding the descriptors and interpreting the results in terms that can aid in the design of novel compounds.

All the alignment methods concentrate on features that are more or less directly associated and extrapolated from the spatial positions of the individual atoms. However, molecules recognize each other by their surfaces and field properties. Thus, alignment methods that map and compare common shape and field properties appear to be better suited to reveal relevant alignments(171).

1.3 MIPSim: Molecular Interaction Potentials Similarity analysis. Overview of previous studies

Sometimes the structural similarity is not evident and sometimes the earliest procedure to relate the electrostatic pattern of molecules was the visual comparison of graphical representations of the MEP distributions(172; 173), which has been widely used in structure-activity studies in order to find similarities between molecules interacting with the same biological receptor(174; 175). This procedure had the disadvantage of its subjectiveness. Our team developed in the past several techniques to perform objective comparisons between MEP distributions(176; 177).

The development of MIPS_{im} (Molecular Interaction Potentials Similarity Analysis) in

our group began in 1988 with a program, called *MEPMIN*(177) which allowed to find automatically the MEP minima of a molecular system. Then, in 1991(139; 178) arrived the *MEPCOMP* program which provided automatically an electrostatic alignment between MEP distributions. MEP distributions were computed with *GAUSSIAN 86*. The similarity was measured with Spearman rank correlation coefficient between the MEP values of each molecule computed at the points of a 3D grid around them. The maximization procedure was driven by a gradient method. *MIPSIM* was capable to find non evident structural alignment between molecules(139), but compatible with experimental data. There were two versions of the program, the first version *MEPCOMP* that considered the molecules in fixed conformations allowing the use of any definition of MEP. At each step of the maximization procedure defined the grid of points within a volume obtained by the union of volumes defined around each molecule. The first study was to find the relative position of methotrexate and dihydrofolic acid when these molecules dock into dihydrofolate reductase, their common receptor(139). Also it was tested comparing several arylalkylamines(178; 179). The modification of this program, in order to take into account the conformational flexibility of the molecules to be compared, gave rise to the development of the *MEPCONF* program(179). *MEPCONF* took one of the molecules as reference, defined the volume to be considered around this reference molecule, and found the maximum electrostatic similarity moving either the relative position of the second molecule and a set of conformational degrees of freedom of this molecule.

The second step arrived in 1993(180). The computational package *MEPSIM* integrated the previously described programs into a common interface. *MEPSIM* included improved new versions of the programs and other modules. Module *MEPPLA* which supplied MEP values for the points of a grid defined on a plane which is specified by a set of three points. The results of this module could easily be converted into MEP maps using third-parties graphical software. Also the module *MEPPAR*(181; 180), another modification of *MEPCOMP* in order to compute the MEP similarity between two molecules, but only taking into account a particular plane and scanned relative positions of the two molecules (this was very useful in order to compare π systems in aromatic rings). *MEPSIM* created input files from output files of external programs (*GAUSSIAN 86* and *AMPAC/MOPAC*) to obtain MEP distributions. One of the most important applications of *MEPSIM* was a theoretical study on the metabolism of caffeine by cytochrome P-450 1A2 and its inhibition(182).

At 2000 arrived the new version of *MEPSIM* called *MIPSIM* (Similarity analysis of molecular interaction potentials)(126). *MIPSIM* is a computational package designed to analyse and compare 3D distributions of molecular interaction potentials (MIP) of series of biomolecules. *MIPSIM* incorporates other Molecular Interaction Potentials (MIP) in addition to the MEP. MIPs have been extensively used for the comparison of series of compounds in structure-activity relationship studies. *MIPSIM* makes calls to *GAMESS*(183) (see section 3.1.1.1.3) for the quantum mechanical computation of MEP, and the *GRID* program(81) for MIP calculations (see section 3.1.1.2.1). Also, *MIPSIM* includes an interface with the statistical package *GOLPE*(184), with the purpose of using *MIPSIM* results in the generation of 3D-QSAR models.

In the new *MIPSIM*(126) there are two principal modules: *MIPMIN* and *MIPCOMP*. *MIPMIN* characterizes the points around a molecule where MIP reaches a minimum value in comparison with those of its surroundings. It computes MIP values in the points of a homogeneous 3D grid around the molecule and looks for those points having smaller values than all the surrounding ones. Finally, an optimization algorithm refines the min-

ima values and positions. MIPMIN provides the geometrical relationships between them. Also, MIPMIN can compute several molecules in a single run. MIPCOMP compares pairs of molecules on the basis of their 3D MIP distributions computed in a common grid of points defined around the molecules. Also, MIPCOMP can perform an exploration of the maximum similarity alignments by starting from a series of randomly generated positions and optimizing them. When the relative position of the molecules changes, their grids of points fail to coincide in space. Then MIPS_{sim} incorporated a similarity index, which copes with non-coincident grids(144). It is a Gaussian correlation coefficient. In order to visualize the results, MIPS_{sim} has a tool to transform its outputs into *Insight II*(185) and *gopenmol*(186) formats. The new MIPS_{sim} was applied to agonists of the neuronal nicotine acetylcholine receptor (nAChR) and it was found that MIP superpositions were in agreement with those suggested by the pharmacophore features(187).

Hypothesis and objectives

Hypothesis:

- Methodologies based on MIP offer important advantages to explain biomolecular interactions with respect to other techniques that only take into account the structure of the ligand and not the complex interaction with the environment.
- MIP based similarity approaches appear to be well suited for obtaining relevant molecular alignments, as they are directly associated with the activity of the molecule through its interaction capacity.
- The use of MIPs allows one to take into account different types of interactions between the considered molecules and their common, but not explicitly considered, receptor.
- Electrostatic effects play an important role in enzymatic reactions. The calculation of MIPs can be a valuable tool to assess the activity of a given TS or TSA. Based on the idea that the largest catalytic effect of enzymes is related to the electrostatic complementarity of the active site we can obtain the most favourable electrostatic environment to stabilize the charge distribution of a TS. This might be eventually useful for the design of better enzyme inhibitors.
- Computational and statistical tools that study the similarity between biomolecules are useful for the analysis of the structure-activity relationships and can improve the design of new drugs, specially if we do not have cristalographic information.

Objectives:

There are two types of objectives. In one hand, the development of methodologies inside `MIPSim` and on the other hand scientific objectives, which in order to be accomplished require the above mentioned developments.

1. The development of `MIPSim`, implementing and improving computer methodologies to compute MIP and to compare them, and also to include algorithms to study the conformational flexibility.
2. Demonstrating that MIP and MEP are useful tools to perform structural alignments for molecules of pharmacological interest.

3. Extending `MIPSim` capabilities to study enzyme reactivity.
4. Exploring the origin of the low rate enhancement of CAs as compared to the corresponding enzymes examining electrostatic similarities.
5. Exploring the use of MIPs as scoring function in docking procedures, in order to find fingerprints describing the interaction between the target and the ligand.
6. Showing that 3D-QSAR studies improves by proper MIP-based alignments.

Methods

This chapter is the most important of all the thesis, because it gives a great description of the program `MIPSim`. Also, this chapter describes other methodologies used in this thesis. Some of them had to be modified or developed, some others only used as already implemented algorithms in program packages.

3.1 Methodologies used inside `MIPSim`

`MIPSim`(126) is a computational system for the automatic exploration of biomolecular similarities on the basis of molecular interaction potentials (MIP). `MIPSim` it is specially powerful when interfacing to other well-known external programs:

- The quantum package `GAMESS`(183).
- A molecular interaction potential evaluator, `GRID`(81).
- Statistical tools needed for the derivation of 3D-QSAR models, `GOLPE`(184).
- Visualization packages: `gOpenMol`(188), `InsightII`(185) and `LINK3D`(189).

One of the most useful descriptors on the ligand-protein interactions is the interaction energy of one molecule with a chemical functional group. In order to obtain the interaction energies the space around the molecule is discretized in a grid of points and at every point of the grid is evaluated the interaction energy between a chemical probe on every point of this grid and the molecule. Then it is possible to simulate interactions of this molecule with the receptor. In this section the acronym MIP refers to both classical (`GRID`(81) or point charges based) and quantum (`GAMESS`(183), MEP based) molecular interaction potentials.

`MIPSim` includes the following main modules:

- `MIN`: Automatically finds the MIP minima of a molecular system. It supplies the Cartesian coordinates of these minima, their values and all the geometrical relationships between them (distances, angles and dihedral angles). This is done by first defining a grid box containing the molecule. Then the MIP is computed on those grid points and the program looks for the positions with smaller values in the grid. These positions may be later refined by the use of some of the built-in optimization procedures. Several molecules can be processed by `MIN` in a single run. At the same time, each molecule can be analyzed using different chemical descriptors.

- **COMP:** This module compares pairs of chemical structures taking into account the MIP distribution around both molecules, giving a correlation index to show the similarity between both distributions. This similarity index can be used in combination with the built-in optimization procedures in the module to find the relative position that maximizes this MIP-based similarity. During the alignment optimization process, one molecule is kept fixed (static molecule) while the other is free to move in terms of translation and rotation (mobile molecule). The mobile molecule attains, at the end of the optimization, the best position relative to the static one. The function evaluated during the optimization is the similarity function between the distribution of all required properties (i.e., subsets of 3D descriptors, that, in addition, may be given different weights) in both molecules. As the free molecule moves, its box definition moves with it, so the two grid boxes fail to coincide in space. In this situation, recalculation of the property's value on new, coincident, points would be compulsory, as standard similarity correlation indexes correlate values from the same point. However, this strategy, that is indeed available in `MIPSim`, is too expensive in general, specially if the property to compare is the quantum MEP distribution. `MIPSim` incorporates a modified similarity index, which copes with non-coincident grids boxes(144). In such index, the sum of squares for each pair of points is weighted by its proximity(119; 120). Thus, correlation between proximate points will have more influence on the index than correlation between distant points. This strategy skips the necessity to recalculate function values and the optimizations can be faster. The optimization procedure can be carried out by means of both gradient search algorithms and a GA(190), or a combination of them. As in `MIN`, `COMP` module can handle more than two structures at a time. In such cases, this module can compare all pairs of molecules, providing as a result the corresponding distance matrix, or compare the first one to all the others. Consideration of molecular flexibility is currently under development and further development.

3.1.1 Calculation of MEP and MIP inside `MIPSim`

Molecular interaction fields can be computed with quantum methodologies (Molecular Electrostatic Potential, MEP) or using Molecular Mechanics (Molecular Interaction Potentials, MIP). In `MIPSim` we use `GRID`(81) program to compute these MIPs and `GAMESS`(183) package to compute the MEP.

3.1.1.1 Quantum MEP

3.1.1.1.1 Molecular wavefunction. Chemical properties of atoms and molecules are determined by the electronic structure(191). Quantum chemistry tries to find it studying molecular wavefunction that contains all the information about a given chemical. If we know wavefunction(192) of a molecular system is a good starting point in order to describe it (structure, activity...). This knowledge requires good mathematical tools. Using quantum mechanics 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. 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.

Using quantum mechanical postulates we can make two assumptions: consider only stationary states(193) and Born-Oppenheimer approximation(194). Using both approximations we obtain Schrödinger's equation independent of time:

$$H\Psi(r, R) = E\Psi(r, R) \quad (3.1)$$

where H is the hamiltonian operator (the sum of kinetic and potential energy), Ψ is the wavefunction and E is the energy of the system. r and R are the coordinates for electrons and for nuclei, respectively. This equation is still mathematically complex and it is necessary to make more approximations.

First approximation is to consider polielectronic wavefunction as a product of mono-electronic wavefunctions that depend explicitly on the spatial coordinates which are the molecular orbitals(195). Due to antisymmetry Pauli's principle(193), the product of simple molecular orbitals is not an acceptable function. It has to be antisymmetrized expressing it as Slater determinants(196).

Second approximation is to consider only restricted determinants. These restricted determinants at closed layer have the molecular orbitals paired, sharing both members of the same pair the same spatial part (molecular orbital) and differing on the spin part (atomic orbital). Then, every molecular orbital is occupied by two electrons.

Third approximation is LCAO (Linear Combination of Atomic Orbitals)(197). Introduced by Roothaan on 1951(198), consist on expressing molecular orbitals as linear combination of functions centered on the nucleus called atomic orbitals. It expresses the approximation of the molecular orbital function as a linear combination of atomic orbitals chosen as the basis functions.

$$\Phi_i = \sum_{\mu}^N C_{\mu i} \phi_{\mu} \quad (3.2)$$

where Φ_i is the molecular orbital, N is the number of atomic orbitals of the system, $C_{\mu i}$ are the coefficients of the linear combination and ϕ_{μ} are the atomic orbitals.

These approximations are useful for simplifying the methodology but they lower the accuracy of calculations. The principal problem is that an unique Slater determinant does not take into account the electronic correlations, interactions between electrons. These limit called Hartree-Fock's (HF) limit, is precisely the energy that we can get using a monodeterminant function. The difference between these energy and non-relativistic experimental energy is correlation energy.

Schrödinger's equation 3.1, can be split in:

$$H = T_e + T_n + V(r, R) = H_{el} + T_n \quad (3.3)$$

where T_e is the kinetic energy of electrons, T_n the kinetic energy of nuclei, V is the total potential energy of electrons and nuclei and H_{el} is the electronic hamiltonian, defined as the total hamiltonian minus the kinetic energy of nuclei T_n .

This equation cannot be solved because the nuclear and electronic coordinates are coupled. Born-Oppenheimer approximation(194) consist into separate both movements: electrons and nuclei, taking into account that nuclei are more heavier than electrons and they move slower. Empirical observations of molecular spectroscopy show that the total energy of a molecule can be viewed as a sum of several approximately non-interacting parts. So, this is a good approximation. Then it is possible to split Shrodinger's equation in two equations:

$$H_{el}\Phi_i(r; R) = U_i(R)\Phi_i(r; R) \quad (3.4)$$

$$[T_n + U_i]\Gamma_i(R) = E\Gamma_i(R) \quad (3.5)$$

The nuclear kinetic terms vanishes. We can whink of the nuclei as being fixed at arbitrary locations, and then solve the Schrdinger equation for the wavefunctions of the electrons alone. Equation3.4 is the electronic Shrodinger's equation, where Φ_i are the electronic wavefunctions that depend on the nuclear coordinates only as a parameter. U_i is the potential energy of the electronic state i , and contains the kinetic energy of the electrons and potential energy of the electrons and nuclei. Equation3.5 is the nuclear part and Γ_i is the nuclear wavefunction.

The global wavefunction can be written as a linear combination of the electronic wavefunctions and the coefficients are the nuclear wavefunctions. So, the total wave function of the system should belong to a full space created fromt he tensorial product between the nuclear space and electronic space.

$$\Psi(r, R) = \sum_i^N \Gamma_i(R)\Phi_i(r; R) \quad (3.6)$$

where N is the number of orbitals in the system.

Using this approximation we do not take into account coupled terms of nuclear and electronic movement. Then, if it is not possible to eliminate this terms, we cannot apply this approximation(199).

Resolution of equation3.4 only can be done in an approximate way, due to bielectronic terms in the electronic hamiltonian H_{el} .

The method to solve it is called Hartree-Fock(200) or Self Consistent Field (SCF). The idea is to consider that every electron can move in a field which is the sum of the field due to the nucleous and the mean field of all the other electrons. With this idea we can express H_{el} as a sum of monoelectronic hamiltonians. In order to compute the mean field of all the other electrons we need the wavefunction of the system. Then our calculation have to be iterative. Eckart's theorem(201) guarantees that if a test function that fulfil the contour conditions, we never obtain an energy less than the correct function of the fundamental state E_0 . Then the computed energy always has a value greater than the real one and its comparison with experimental is a good way to know the proximity to the correct solution.

According to the molecular orbital theory, the electrons spread throughout the whole molecule, and it is posible to define its wavefunction by a linear combination of the atomic orbitals.

$$E_0 \leq \frac{\langle \Phi | H | \Phi \rangle}{\langle \Phi | \Phi \rangle} \quad (3.7)$$

Applying the variational methodology(202), the electronic energy can be written :

$$E_{el} = \sum_i^N 2f_i H_i + \sum_i^N \sum_j^N [\alpha_{ij} \langle ii | jj \rangle + \beta_{ij} \langle ij | ij \rangle] \quad (3.8)$$

where N is the number of molecular orbitals, α_{ij} , β_{ij} , and f_i are state parameters and they only depend on the electronic state. Their value can be 0, $\frac{1}{2}$, or 1, H_i are the mono-electronic integrals, and $\langle ii | jj \rangle$ are the bi-electronic integrals of repulsion and $\langle ij | ij \rangle$ are the exchange integrals created from the antisymmetry principle.

These results are only valid considering closed layer. If we consider other cases (excited states) we have to consider open layer and to use other approximations.

Then, the general expression 3.8 can be reduced(200) to:

$$E_{el} = 2 \sum_i^N H_i + \sum_i^N \sum_j^N (2J_{ij} - K_{ij}) \quad (3.9)$$

where N is the number of molecular orbitals of the system, H_i is the mono-electronic term that represents the energy of an electron in a molecular orbital Φ_i in the field created by the nuclei, $J_{ij} = \langle ii | jj \rangle$ is the Coulomb integral or the bi-electronic term and contains the repulsive interactions between the charge distributions $\Phi_i \Phi_i$ and $\Phi_j \Phi_j$, $K_{ij} = \langle ij | ij \rangle$ is the exchange integral that takes into account the attractive interactions between the electrons of parallel spins in orbitals Φ_i and Φ_j .

Restricting that molecular orbitals have to be orthonormalized in the variational principle and using Lagrange multipliers, we arrive to the mono-electronic equation:

$$F_i \phi_i(1) = \epsilon_i \phi_i(1) \quad (3.10)$$

where ϵ_i is the energy of the electron 1 in the orbital ϕ_i and F is the Fock operator. In order to know F we have to know first the solution. Then the resolution has to be iterative using initial wavefunctions.

Fock operator is:

$$F_{\mu\sigma} = H_{\mu\sigma} + \sum_j (2 \langle \mu\sigma | jj \rangle - \langle \mu j | \sigma j \rangle) \quad (3.11)$$

where j is extended to all the occupied molecular orbitals.

Solving iteratively the Roothaan's equations we obtain coefficients for the best molecular orbitals:

$$\sum C_{\sigma\mu} [F_{\sigma\mu} - \epsilon_i S_{\sigma\mu}] = 0 \quad (3.12)$$

or written in matrixial form:

$$FC = CSE \quad (3.13)$$

where F is the Fock matrix, C is the coefficients matrix and S is the recovered matrix and E is the energies matrix. S is defined as

$$S_{\mu\sigma} = \langle \mu | \sigma \rangle \quad (3.14)$$

Using equation 3.2 to develop orbital j , equation 3.11 can be written as:

$$F_{\mu\sigma} = H_{\mu\sigma} + \sum_{\delta\tau} P_{\delta\tau} (\langle \mu\sigma | \delta\tau \rangle - \langle \mu\delta | \sigma\tau \rangle) \quad (3.15)$$

where $P_{\delta\tau}$ is an element of the density matrix defined as:

$$P_{\delta\tau} = 2 \sum_i^N c_{\delta i} c_{\tau i} \quad (3.16)$$

where N is the number of occupied orbitals.

Usually we solve the problem iteratively:

- Compute the integrals and the recovered matrix (S).
- Suppose a coefficients matrix (C) as a test.
- Compute density matrix (P) with equation 3.16.
- Find the Fock matrix (F) using equation 3.15.
- Solve equation 3.13, obtaining energies matrix (E) and a new coefficients matrix (C').
- Compute one more time density matrix (P') with new C' .
- If the differences of the new P' with P are large, we return to compute F . The process finalizes when we arrive at consistency (when the energy between two iterations is less than a certain prestablished value).

Then, using the Born-Oppenheimer(194) and Hartree-Fock approximations we can establish equations that enable knowing the description of molecular orbitals and the energy of a molecule. Exist two methodologies to compute these equations: *ab initio* and semiempirics.

Ab initio methods.

Ab initio calculations are quantum chemical calculations using exact equations with no approximations which involve the whole electronic population of the molecule. Methods of quantum mechanical calculations independent of any experiment other than the determination of fundamental constants. The methods are based on the use of the full Schrodinger equation to treat all the electrons of a chemical system. In practice, approximations are necessary to restrict the complexity of the electronic wavefunction and to make its calculation possible.

This methodology computes energy evaluating all the mono and bielectronic integrals of Roothaan 3.2. It is not introduced any experimental parameter and we only consider Born-Oppenheimer(194) and Hartree-Fock approximations.

In Roothaan's equation molecular orbitals are expressed as a linear combination of atomic orbitals, which are the basis functions. The basis will have a radial and an angular

part. The angular is almost never commented because it is always the same (s,p,d,f,...). The radial part decides how far or how close is the electron with respect to the nucleus. Depending on how many basis we compute the solution is more or less exact. If we take infinite basis functions, SCF-MO, we obtain the more exact solution. Working with infinite basis functions it is impossible, then we have to choose a finite number of basis.

- Hydrogenoid orbitals. Hydrogenoid orbitals are those obtained from the exact solution of hydrogen atom, but we do not use it due to the complexity of the integrals when someone uses a great number of the principal quantum number.
- STO Slater type orbitals. These orbitals have a radial dependency proportional to $\exp(-\delta r)$, where δ is the Slater exponent and r is the distance to the considered point, representing the electronic density in the bond zone.
- GTF Gaussian type functions. These functions can be integrated analytically. They have a radial dependency proportional to $\exp(-\alpha r^2)$, where α is the gaussian exponent and r is the distance to the considered point. These functions do not explain correctly the electronic density in the bond zone, then, we have to use a linear combination of different gaussian functions:
 - Minimum basis. We take a function or a linear combination of functions for every atomic orbital occupied. In this case we take as a standard basis $STO - NG(203)$, where every Slater's atomic orbital is represented by N gaussians.
 - Double zeta basis. Every atomic orbital is doubled in two groups of gaussians of different exponents and we get more dispersion. Standard basis of this type are $N - 21G$ and $N - 31G(204)$, where it is used a linear combination of N gaussians for all internal orbitals, and for the valence orbitals are a combination of two groups, one of two or three gaussians and the other another gaussian.
 - Triple zeta basis. To double zeta basis we introduce polarization functions. Then, we add a non-occupied exterior orbital. Standard basis of this type are: $N - 21G^*$, $N - 31G^*$, $N - 21G^{**}$, and $N - 31G^{**}$. One asterisk indicates polarization functions for heavy atoms and two asterisks polarization functions for hydrogens and for heavy atoms.

Specific implementations of this type of methodologies include: GAMESS and GAUSSIAN 98 packages. On section 3.1.1.1.3 it will be a great explanation of GAMESS(183) package.

Semiempirical methods

Ab initio methods are good but sometimes they are slow. In order to accelerate these calculations we try to simplify Fock's operator (3.11(205)), estimating some of the experimental parameters and neglecting some of them.

Semi-empirical methods are molecular orbital calculations using various degrees of approximation and using only valence electrons. In these methods, certain integrals are set equal to parameters that have been chosen to lead to the best fit to experimental quantities.

The methods which use parameters derived from experimental data to simplify computations. The simplification may occur at various levels: simplification of the Hamiltonian (e.g. as in the Extended Hückel method), approximate evaluation of certain molecular integrals (see, for example, zero differential overlap), simplification of the wave function (for example, use of p electron approximation as in Pariser-Parr-Pople).

We never consider all the electrons in this type of calculations:

- AVE methods. We consider only valence electrons. In this case, only the valence electrons are considered explicitly, so there will be only basis functions for them. Then, the nucleus in the electronic Hamiltonian has a lowered effective charge due to the screening by the core-electrons. There are the Pople methods (CNDO (Completely Neglect Differential Overlap)(192; 206), INDO(207) and NDDO(208)) and Dewar methods (MINDO(209)(Modified Intermediate Neglect of Differential Overlap), MNDO (Modified Neglect of Diatomic Overlap)(210) and AM1(Austin Model)(211)). The difference between Pople and Dewar methods is the treatment of the repulsion integrals between two centers and the attractions between the core (nucleus and internal layer electrons) and the valence electrons(210). INDO (Intermediate Neglect Differential Overlap) does not describe correctly molecules with heteroatoms. That was corrected with NDDO (Neglect of Diatomic Differential Overlap), but this one does not reproduce very well hydrogen bonds. A third generation of semiempirical programs, AM1 was born in 1985. AM1 calculations are semi-empirical molecular orbital calculations developed at the University of Austin in Texas (AM1 = Austin Model 1). These calculations involve the valence electrons of the atoms of the molecule. They are a further development of MNDO calculations. AM1 has been tested with other theoretical methods and with experimental values and it is possible to obtain very good results.
- π methods. We only consider π electrons.

Also, we can consider interelectronic repulsion (SCF) or not (independent electrons).

The easiest semiempirical approximations ignore the dependency of Fock matrix respect to the electronic wavefunctions, estimating their value with a certain algorithm and making an unique cycle of calculations (Huckel method).

All the semiempirical methods use SCF and ZDO (zero differential overlap: considered zero the product between two basis functions.) Then we do not have to calculate a large number of bielectronic integrals.

Molecular descriptors are terms that characterize a specific aspect of a molecule. Wavefunction is a good molecular descriptor in order to describe biomolecular interactions. Using wavefunction we can obtain other molecular descriptors very useful like electronic density associated to each nucleus(212). These atomic charges can be used directly or for calculating an approximate molecular electrostatic potential(213). Molecular orbitals can be used to compute dipolar moments and molecular electrostatic potential (MEP).

The total energy of the system enables to make conformational analysis(214; 215) and to search the reaction surfaces.

The quantum mechanical postulates assume that the wavefunction and the density function contain all the information of a system. The statement, applied to a chemical compound, means that all the information about any molecule could be extracted from the electron density. Bond creation and bond breaking in chemical reactions, as well as the shape changes in conformational processes, are expressed by changes in the electronic density of molecules. The electronic density fully determines the nuclear distribution and its changes account for all the relevant chemical information about the molecule. In principle, quantum chemical theory should be able to provide precise quantitative descriptions

of molecular structures and their chemical properties. The disadvantage of these methods has been that even an approximate solution of the Schrodinger equation can be extremely complex for all but the simplest systems. Long computational times have been required for meaningful calculations and the size of the system, which can be studied, has been limited. Quantum chemical methods allow derivation of molecular descriptors from the total molecular wavefunction and charge distribution(216). Other approaches include comparing electronic density between compounds or analyzing topological features of the electron density(217; 218).

In drug design, quantum mechanic calculations have a full range of application. For example a 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 simulation and molecular electrostatic potential computation.

3.1.1.1.2 Molecular electrostatic potentials. We are interested in the MEP because electrostatic interactions are especially important for molecular recognition and, thus, they are relevant in order to obtain an indirect picture of the biological receptor.

MEP are electrostatic properties of a molecule based on the charge density as calculated directly from the molecular wavefunction. The electrostatic potential (scalar with dimensions of energy) is calculated at a point in the vicinity of a molecule. The spatial derivative is the electric force (vector) acting on a unit positive charge at that point caused by the nuclei and the electrons of the molecule.

Molecular electrostatic potential (MEP) is the potential generated by the charge distribution of the molecule. Using a classical approximation, the difference of electrostatic potentials between two points A and B is

$$\int_A^B \vec{E} dr = V(r_A) - V(r_B) \quad (3.17)$$

When \vec{r} tend to infinite we assign $V(\infty) = 0$. Then the potential is the energy that we need to carry a positive charge from infinite to our point, considering an electric field.

If the system is a set of puntual charges, we use superposition principle, then:

$$V(R) = \sum_i \frac{q_i}{|\vec{r}_i - \vec{R}|} \quad (3.18)$$

where \vec{r}_i is the position vector of each puntual charge q_i .

If we consider a continuous distribution we have:

$$V(R) = \int_{Volume} \frac{1}{|\vec{r} - \vec{R}|} \Gamma(\vec{r}) dv \quad (3.19)$$

where $\Gamma(\vec{r})$ is the charge density function.

Using quantum mechanics approximation we consider wavefunctions $\psi(\vec{r})$ to describe a particle, where \vec{r} is the position vector and the product $\psi(\vec{r})^* \psi(\vec{r}) dv$ is the probability to find the particle in the volume element in the position \vec{r} . $\psi(\vec{r})^* \psi(\vec{r}) dv$ is the electronic density function.

Then, the charge density in a certain point \vec{R} is:

$$\Gamma(\vec{R}) = \langle \Psi(\vec{r}) | \delta(\vec{r} - \vec{R}) | \Psi(\vec{r}) \rangle \quad (3.20)$$

where $\delta(\vec{r} - \vec{R})$ is the Dirac's delta function, that verify the properties:

$$\delta(x - a) = 0 \text{ if } \vec{r} \neq \vec{R}$$

$$\delta(x - a) = 1 \text{ if } \vec{r} = \vec{R}$$

Then, for a polielectronic system, the electrostatic potential is:

$$V_e(\vec{R}) = \int_{Volume} \frac{1}{|\vec{r} - \vec{R}|} \langle \Psi(\vec{r}) | \sum_{i=1}^n \delta(\vec{r}_i - \vec{R}) | \Psi(\vec{r}) \rangle dv \quad (3.21)$$

This is the contribution of electrons, but we have to consider nuclei contributions. Considering Born-Oppenheimer approximation(194), nuclei do not move, then:

$$V_N(\vec{R}) = \sum_{\alpha} \frac{Z_{\alpha}}{|\vec{R}_{\alpha} - \vec{R}|} \quad (3.22)$$

where α is the number of atoms.

For a molecule the electrostatic potential will be:

$$V(\vec{R}) = \sum_{\alpha} \frac{Z_{\alpha}}{|\vec{R}_{\alpha} - \vec{R}|} - \int_{Volume} \frac{1}{|\vec{r} - \vec{R}|} \langle \Psi(\vec{r}) | \sum_{i=1}^n \delta(\vec{r}_i - \vec{R}) | \Psi(\vec{r}) \rangle dv \quad (3.23)$$

Molecules are treated as closed layer, then they have two electrons in every occupied molecular orbital and using LCAO(197). We can write:

$$\Gamma(\vec{R}) = 2 \sum_i \sum_{\delta} \sum_{\tau} C_{\delta i} C_{\tau i} \phi_{\delta}^*(\vec{r}) \phi_{\tau}(\vec{r}) \quad (3.24)$$

and defining (δ, τ) of the density matrix (P) as equation 3.16, then, the electronic density in R is:

$$\Gamma(\vec{R}) = \sum_{\delta\tau} P_{\delta\tau} \phi_{\delta}^*(\vec{r}) \phi_{\tau}(\vec{r}) \quad (3.25)$$

Then, the MEP is:

$$V(\vec{R}) = \sum_{\alpha} \frac{Z_{\alpha}}{|\vec{R}_{\alpha} - \vec{R}|} - \sum_{\delta\tau} P_{\delta\tau} \langle \phi_{\delta}(\vec{r}) | \frac{1}{|\vec{r} - \vec{R}|} | \phi_{\tau}(\vec{r}) \rangle \quad (3.26)$$

MEP value is more accurate if we use *ab initio* methods in order to compute wavefunctions. In our case we use GAMESS(183) program (see section 3.1.1.1.3) to compute MEP. These potentials consider the molecule in vacuum. If we want to consider the molecule in dissolution, the major problem is how to use the dielectric medium.

An easy way, but very rough in order to compute MEP is the puntual charge methodology(213). We consider the approximation that the molecule as a system of puntual charges is situated on the nucleous coordinates, and their value is the excess or default of charge over every center.

Then:

$$V(\vec{R}) = \sum_{\alpha} \frac{q_{\alpha}}{|\vec{R}_{\alpha} - \vec{R}|} \quad (3.27)$$

where q_{α} is the excess or default amount of charge over the nucleous α and \vec{R}_{α} is the position of nucleous α .

3.1.1.1.3 Program GAMESS. The General Atomic and Molecular Electronic Structure System (GAMESS(183)) is a general *ab initio* molecular quantum chemistry package. Also, includes semiempirical wave functions.

GAMESS package offers a wide range of quantum mechanical wave functions, capable of treating systems ranging from closed-shell molecules through bond-breaking reactions. These wave functions may be combined with various run types to perform chemically important tasks, ranging from geometry optimization to transition state location to reaction path following.

GAMESS can compute a wide range of quantum chemical computations like:

- The first and second derivatives of the potential, which are the electric field and the electric field gradient.
- Molecular properties: multipoles moments, electrostatic potentials, electron density and spin density.
- Analytic energy gradients for any of the SCF or DFT wavefunctions used for automatic geometry optimization.
- Searches for saddle points or reaction path following on the potential energy surface.
- Traces the intrinsic reaction path from the saddle point towards products, or back to reactants.
- Solvent effects may be modeled by the discrete Effective Fragment Potentials, or continuum models such as the Polarizable Continuum Model.

3.1.1.2 Classical MIP

Usually is hard to obtain the energy of a molecular system. This is the reason that in parallel to the methods of Quantum Chemistry were obtained similar results introducing qualitative chemical knowledge about molecular structure into a parametric function. That is, the strenght of a chemical bond between two atoms, dispersion forces, hydrogen bonds, electrostatics. All these interactions could be put together as a sum of analytical functions that give as a result a parametric energy function of the nuclear coordinates. This energy function is called an empirical force field and the strategy is the Molecular Mechanics (MM).

Historically there are several analytical functions that describe the inter and intramolecular interactions. However it is not until the end of sixties and the beginning of seventies with the help of the emerging computers that some useful results are obtained. These Molecular Mechanics parametric functions need to be parametrized according to the experimental results or *ab initio* calculations.

Molecular mechanics is the calculation of molecular conformational geometries and energies using a combination of empirical force fields. A force field is a method to compute geometrical and energy characteristics of molecular entities on the basis of empirical potential functions the form of which is taken from classical mechanics. Current generation force fields (or potential energy functions) provide a reasonably good compromise between accuracy and computational efficiency. They are often calibrated to experimental results and quantum mechanical calculations of small model compounds. A force field will be characterized by the number and functional type of the energy terms and by the way the parameters are obtained. 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. A force field is a set of equations representing the potential energy surface with respect to changes in the geometry of the molecule.

Molecular mechanics methods are based on the following principles:

- Nuclei and electrons are lumped into atom-like particles.
- Atom-like particles are spherical (radii obtained from measurements or theory) and have a net charge (obtained from theory).
- Interactions are based on springs and classical potentials.
- Interactions must be preassigned to specific sets of atoms.
- Interactions determine the spatial distribution of atom-like particles and their energies.

The system is described with the nuclei positions and the charge distribution is considered to remain constant. We are going to assume that the energy of the system is separable in different terms. The usual separation is the following: bonded and non-bonded interactions.

Bonded atoms interact through stretching, bending and torsion. The stretching energy estimates the energy associated with vibration about the equilibrium bond length. The bending energy the energy associated with vibration about the equilibrium bond angle. There exist several expressions for every term. Some force fields incorporate crossing terms to account for the coupling between two different interaction types.

The non-bonded energy represents the pair-wise sum of the energies of all possible interacting non-bonded atoms. Non-bonded atoms (greater than two bonds apart) interact through vdW attraction, steric repulsion, and electrostatic attraction/repulsion.

- Van der Waals energy. **Van der Waals** attraction occurs at short range, and rapidly dies off as the interacting atoms move apart by a few Angstroms. Repulsion occurs when the distance between interacting atoms becomes even slightly less than the sum of their contact radii.

These effects are often modeled using a 6-12 equation (Lennard-Jones equation):

$$E_{vdW} = \sum_i \sum_{j>i} 4\epsilon \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] \quad (3.28)$$

where r_{ij} is the distance between the two non-bonded atoms. ϵ is the well depth of the potential and σ is the collision diameter of the respective atoms i and j . The $\exp(12)$ term of the equation is responsible for small-distance repulsion, whereas the $\exp(6)$ provides an attractive term which approaches zero as the distance between the two atoms increases. When r_{ij} is small, the first term generates a dominating repulsion corresponding to a large positive value of E_{vdW} . The A parameter can be obtained from atomic polarizability measurements, or it can be calculated quantum mechanically. The B parameter is typically derived from crystallographic data so as to reproduce observed average contact distances between different kinds of atoms in crystals of various molecules.

- Electrostatic energy. **The electrostatic contribution** is modeled using a Coulombic potential.

$$E_{el} = \sum_i \sum_{j>i} \frac{1}{4\pi\epsilon} \frac{q_i q_j}{r_{ij}} \quad (3.29)$$

where r_{ij} is the distance between the two non-bonded atoms. q_i and q_j are the partial atomic charges. ϵ is the dielectric constant of the medium, which represents the environment around them (water and protein). A typical approximation is assign values to ϵ depending on the distance r_{ij} (219). The electrostatic energy is a function of the charge on the non-bonded atoms, their interatomic distance, and a molecular dielectric expression that accounts for the attenuation of electrostatic interaction by the environment (e.g. solvent or the molecule itself).

Apart from energy minimization, molecular dynamics and conformational analysis, force fields have a wide range of application in drug design. The GRID(81) force field is particularly useful to compute molecular interaction fields (MIF) (see section 3.1.1.2.1 for more detailed explanation of GRID program). 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. All the molecular mechanics calculations can be computed considering all the surroundings of the target or only taking into account a discrete number of points around it. This methodology is used in many programs.

3.1.1.2.1 Program GRID. MIPSim interacts with GRID(81) in order to compute MIPs. Programm GRID is a computational procedure for determining energetically favourable binding sites on molecules of known structure. It may be used to study individual molecules such as drugs, molecular arrays such as membranes or crystals, and macromolecules such as proteins, nucleic acids, glycoproteins or polysaccharides. It calculates interaction energies between a chemical group (probes) and another molecule (target).

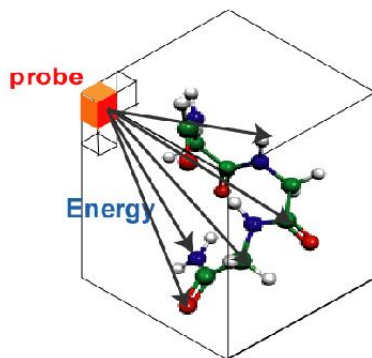


Figure 3.1: Evaluation of MIP. At every point of the grid is computed the MIP between the chemical probe and the molecular atom

A 3D grid points is established around the molecule. GRID(81) computes interaction energies between various atomic probes or functional groups and the surface of a protein at equally distributed grid points.

In the GRID force field, only non-bonded interactions are considered. The force field contains a van der Waals term (vdW), an electrostatic term, and a hydrogen bond term. Then, the non-bonded interaction energy E of the probe at each position is:

$$E = \sum E_{lj} + \sum E_{el} + \sum E_{hb} + \sum E_w \quad (3.30)$$

Each individual term in the summations relates to one pairwise interaction between the probe and a all extended atoms of the protein.

The term E_{lj} is due to vdW interactions: Lennard-Jones function (see equation 3.28)

$$E_{lj} = \frac{A}{d^{12}} - \frac{B}{d^6} \quad (3.31)$$

The interacting atoms are a distance d apart, and the energy variables A and B are calculated from the Van der Waals radius, polarizability and effective number of electrons of the atoms(219). These values are tabulated in datafile GRUB of GRID.

The term E_{el} is the electrostatic interaction energy (see equation 3.29) The electrostatic interaction for GRID program is defined as:

$$E_{el} = \frac{pqK}{M} \left[\frac{1}{d} + \frac{M - W}{(M + W)\sqrt{d^2 + 4PQ}} \right] \quad (3.32)$$

where p and q are the electrostatic charges on the probe group and the pairwise target atom, and K is a combination of geometrical factors and natural constants. The macromolecular target and the surrounding water have dielectrics of M and W respectively, and the depth of the charges p and q in the target phase is P and Q . For small molecules the target phase

is effectively absent, and P and Q are both zero. A GRID force field peculiarity is that the value of the dielectric constant changes with the local environment of the probe.

The term E_{hb} is hydrogen bond function which represents the directionality and the strenght of the hydrogen bonds. The standard hydrogen bond interaction is computed from:

$$E_{hb} = \left(\frac{C}{d^8} - \frac{D}{d^6} \right) F(U, U', U'' \dots) \cdot F'(Q) \quad (3.33)$$

where F and F' are functions; $U, U', U'' \dots$ are angles and distances defining the geometrical arrangement of the atoms engaged in hydrogen bonding and their neighbours; and Q depends on the charges of the interacting atoms. Energy variables C and D are computed from the hydrogen bond radii and hydrogen bond energies of the atoms, which are tabulated in datafile GRUB.

The term E_w takes into account interactions with water molecules. In some situations a water molecule may form a bridge between target and probe. Such water bridges can significantly stabilize the overall target-probe interaction, and a keyword may be used in order to simulate this effect.

However an entropic term was needed for the hydrophobic probe. Entropic terms are also required for conformationally flexible targets and for the detection of selectively unfavourable sites. GRID contains basic concepts to include side chain flexibility.

Program GRIN is used to prepare and check an input file, `grinkout`, needed for GRID program. A table of parameters, GRUB, is needed in order to evaluate the Lennard-Jones and other empirical energy functions, and program GRIN inside GRID appends these parameters to the atomic coordinates of the protein. The parameters in GRUB are based on the extended atom concept(83; 220). `Grinkout` file is a list of target atoms in the correct sequence with their coordinates and energy variables.

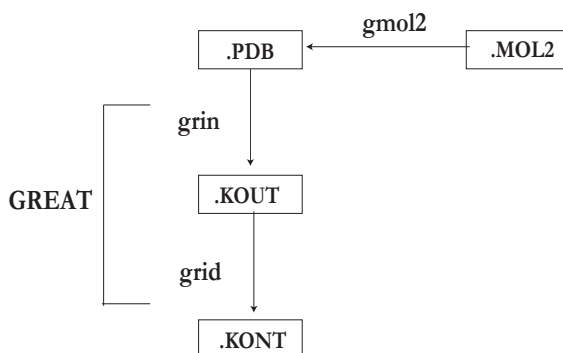


Figure 3.2: GRID files. When using MIPSIM/GRID, the user must specify the PDB in the correct format by using `gmol2`

The GRID program takes *grinkout* as an input file, which defines the properties of the target, and computes interaction energies. GRID can accept several different *grinkout* files each containing one target. GRID contains more than 60 probes of different types. Contour surfaces at appropriate energy levels are calculated for each probe. In our studies, we have used the hydrophobic probe (DRY), the amide nitrogen probe (N1, hydrogen bond acceptor) and the sp² carbonyl oxygen probe (O, hydrogen bond donor).

The results from GRID can be displayed by program GVIEW or by other computer graphics if appropriate hardware and software are available. GVIEW is the application of GRID used to visualize MIFs. The molecular shape of the target, and the interaction energies of the chosen probe, can then be viewed simultaneously. Contours at negative energy levels delineate regions at which ligand binding should be particularly favoured. Positive energy levels normally define the surface of the target. The application can be also used to export the data to standard formats and to print or save the images in Postscript or RGB format.

Program GREAT integrates programs GRIN and GRID. GREATER is a Graphical User Interface (GUI) for the GRID package. It provides GUI access to most of the functionality of programs GRIN and GRID and also to program GREAT which was the predecessor of GREATER. Furthermore the interactive GUI of GREATER is closely integrated with GVIEW, and this helps the user to visualize the structure of the target and the results of the GRID computation simultaneously.

Statistical analyses can extract other important information from the results. GOLPE(184) is one of the programs which can be used in order to analyse GRID maps statistically for QSAR or 3D-QSAR analyses(221). GOLPE (Generating Optimal Linear PLS Estimations) it is an advanced variable selection technique in partial least squares (PLS) used in 3D-QSAR studies to handle very large data sets. GOLPE can also accept *grinkout* as input files with the structure of the targets.

The best-known application of GRID in Structure-Based Drug Design(222) shows the potentiality of the approach. Moreover GRID can be used to understand the structural differences related to enzyme selectivity, a fundamental field in the rational design of drugs(223).

3.1.2 Optimization strategies inside MIPSim

Finding stationary points (minima, maxima, and saddle points) on energy surfaces is important in chemical physics because they correspond to equilibrium geometries and transition states and because the classical equations of motion connecting such points can be used to describe reactions.

Because the similarity depends on the relative arrangement of two molecules, a configurational searching method, is required to maximize the overlap and finally obtain optimized similarity indexes. The methods that require up to first derivatives of the energy with respect to the nuclear coordinates are mainly steepest descent and conjugate gradient methods. Also we can use GA.

Gradient optimization procedures have been proven to be very effective in MIN and COMP module.

3.1.2.1 Steepest descent

It was first developed by Courant(224), Curry(225) and Householder(226). Since the magnitude of the gradient indicates the steepness of the local slope, the energy of the system can be lowered by moving each atom in response to the force acting on it. This is the basis of the steepest descent methodology, where the displacement of the geometry Δq_k at iteration k may be obtained from the gradient g_k at the current geometry.

$$\Delta q_k = -\alpha_k \frac{g_k}{|g_k|} \quad (3.34)$$

where α_k is the step length determined by trust radius or line search.

This algorithm presupposes that the gradient of the function can be computed. Also, it has the severe drawback of requiring a great many iterations for functions which have long, narrow valley structures. In such cases, a conjugate gradient method is preferable.

3.1.2.2 Conjugate gradients

It was first developed by Hestenes and Stiefel(227). In conjugate gradient method the displacement is computed from the gradient at the current point plus the scaled previous displacement.

$$\Delta q_k = \alpha_k \left(-\frac{g_k}{|g_k|} + \gamma_k \Delta q_{k-1} \right) \quad (3.35)$$

where the scaling factor γ_k is computed using the previous gradient vectors. There are several expressions for this factor, the easiest form is the Fletcher-Reeves(228).

$$\gamma_k = \frac{g_k g_k}{g_{k-1} g_{k-1}} \quad (3.36)$$

It uses conjugate directions instead of the local gradient for going downhill. If the vicinity of the minimum has the shape of a long, narrow valley, the minimum is reached in far fewer steps than would be the case using the method of steepest descent. As the steepest descent method, it presupposes that the gradient of the function can be computed.

3.1.2.3 Simplex

The Nelder -Mead(229) method or Simplex method is a numerical method for minimizing an objective function in a many dimensional space.

The method uses the concept of a simplex, which is a polytope of $N + 1$ vertices in N dimensions; a line segment on a line, a triangle on a plane, a tetrahedron in 3D space and so forth. The method finds a locally optimal solution to a problem with N variables when the objective function varies smoothly. The method generates a new test position by extrapolating the behaviour of the objective function measured at each test point arranged as a simplex. The algorithm then chooses to replace one of these test points with the new test point and so the algorithm progresses. The simplest step is to replace the worst point with a point reflected through the remaining $N + 1$ points considered as a plane. If the point

is better than the best current point, then we can try stretching exponentially out along this line. On the other hand, if this new point is not much better than the previous value then we are stepping across a valley, so we shrink the simplex towards the best point.

3.1.2.4 Genetic algorithms

Genetic algorithms (GA) are computational problem-solving methods that mimic some of the principal characteristics of biological evolution and genetic reproduction(230; 231; 232; 233).

A GA creates a randomly-chosen set of individuals, known as a population, each of which contains a representation of a possible problem solution. This solution is encoded into a linear string that is normally referred to as a chromosome. The effectiveness of the solution encoded by each of the chromosomes in a population is measured by the fitness function, and the GA manipulates the chromosomes so as to maximize the value of the fitness function. Chromosomes are manipulated by mutation (where the chromosomal material may be altered slightly in a random fashion) and crossover (where new chromosomes are created by taking some chromosomal material from one parent, and some from the other) operators.

An implementation of a GA begins with a population, typically random of chromosomes. One then evaluates these structures and allocates reproductive opportunities in such a way that those chromosomes which represent a better solution to the target problem are given more chances to reproduce than those chromosomes which are poorer solutions.

3.2 Miscellanea of methodologies used outside MIPSim

Sometimes we do not know how the ligand is posed inside the protein and which is its active conformation. Then it is useful to create different conformations for every molecule in order to score, subsequently with another program, for the best possible active conformation.

3.2.1 Conformational sampling

For a ligand to bind a receptor, and thereby initiate a biological effect, it generally has to adopt a conformation which is in some way complementary to its target protein. This protein-bound conformation is termed the bioactive conformation. It is often a non-trivial task to determine it, since most drug-like molecules have numerous low-energy conformations. Today, several methods are available for generating conformational ensembles.

3.2.1.1 Catalyst

CATALYST(234) supports two methods of conformational generation termed BEST and FAST. It has an algorithm(235) designed to sample as diverse a set of conformations as possible. BEST is reported to be more thorough, in particular when handling flexible ring systems. FAST is more approximate, and therefore requires substantially less CPU time.

Catalyst uses the CHARMM force field(236), which does not include any electrostatic terms. BEST employs conjugate-gradient minimization in both torsion and Cartesian space, in conjunction with poling. The default setting is to collect a maximum of 250 conformations within an energy cut-off of 20 kcal/mol.

3.2.1.2 Omega

OMEGA(237) (Optimized Molecular Ensemble Generation Application) supports a so-called torsion-driving beam search for generating ensembles of conformers. It is a rule-based method that generates conformations extremely rapidly. By contrast with stochastic methods (GA or simulated annealing), the results are completely reproducible. OMEGA deconstructs the molecule into fragments with rotatable bonds, and uses certain build-up principles to generate a conformational ensemble. It does not minimize bond lengths or bond angles. All heavy atoms are superimposed to test for duplicate structures, and an rmsd deviation of 0.8Å is default. OMEGA includes a simple force field called the Clean force field. This force field includes torsion and non-bonded components. Any structure with an energy of more than 10 kcal/mol above the current global minimum is discarded. The default limit is to collect 75 conformations.

3.2.2 Multiobjective optimization strategies. FFSQP software

FFSQP Version 3.7(238) is a code for solving constrained nonlinear optimization problems, generating iterates satisfying all inequality and linear constraints. FFSQP(FORTRAN Feasible Sequential Quadratic Programming) is a set of FORTRAN subroutines for the minimization of the maximum of a set of smooth objective functions (possibly a single one, or even none at all) subject to nonlinear equality and inequality constraints, linear equality and inequality constraints, and simple bounds on the variables.

If there is no objective function, the goal is to simply find a point satisfying the constraints. If the initial guess provided by the user is infeasible for nonlinear inequality constraints and linear constraints, FFSQP first generates a point satisfying all these constraints by iterating on the problem of minimizing the maximum of these constraints. The user must provide subroutines that define the objective functions and constraint functions and may either provide subroutines to compute the gradients of these functions.

FFSQP implements two algorithms(239; 240; 241) based on Sequential Quadratic Programming (SQP), modified so as to generate feasible iterates.

3.2.3 Superposition of molecules and proteins

3.2.3.1 Superposition of molecules based on atoms. SUPERB routine

SUPERB is a routine written in FORTRAN based on the rotate routines by J. Villà-Freixa implemented in the rotate.f program by Corchado and Villà-Freixa, part of POLIRATE. This program used the Chen(242) algorithm for the minimums calculation to superpose two different geometries based on a group of atoms. These two groups of atoms may or not be common. The only requisite is that they are of the same size. To see an example of input file of SUPERB program see section C.11 in Annexes chapter.

3.2.3.2 Multiple protein sequence alignment. STAMP package

STAMP(243) (SStructural Alignment of Multiple Proteins) is a package for the alignment of protein sequence based on 3D structure. It provides not only multiple alignments and the corresponding "best-fit" superimpositions, but also a systematic and reproducible method for assessing the quality of such alignments.

STAMP makes extensive use of the Smith Waterman algorithm(244; 245). This is a widely used algorithm which allows fast determination of the best path through a matrix containing a numerical measure of the pairwise similarity of each position in one sequence to each position in another sequence.

At the heart of the method is the Argos and Rossmann(246) equation for expressing the probability of equivalence of residue structural equivalence.

In section C.12 in Annexes one can see more details about using STAMP.

3.2.4 Calculations on solvation free energy. Langevin dipoles calculations inside CHEMSOL program

Quantum mechanical studies of chemical processes in solutions have to take into account the effect of the environment. One can take one of the following three options:

- All-atom models or explicit solvation models. They represent explicitly all the solvent and/or proteins atoms. Some examples are free energy perturbation (FEP)(247) or thermodynamic integration(248). Such approaches require very large amounts of computer time and involve convergence problems. Also, these models depend on the chosen force field parameters.
- Continuum model(249). The system is partitioned at different shells. A core and a first sphere of the environment (directly involved in chemical changes) are modeled explicitly, while the outer environment is represented by a continuum approximation. Some disadvantages are an spherical solute-solvent boundary with uncertain radius and no reflect adequately specific features of solute-solvent interactions.
- Langevin dipoles (LD) model(250; 21; 251; 252; 253). It represents solvent molecules by a fixed cubic lattice of dipoles that would account for the main physics of the solute-solvent interaction. This approach is an intermediate between fully explicit and implicit treatments of the solvent. It does not have to assume an arbitrary dielectric constant (the dielectric is just the vacuum dielectric constant). It uses transferable atomic parameters calibrated using observed solvation energies. Charges are obtained using *ab initio* calculations. Dipolar models can capture the main physics of polar solvents and that reproducing the average polarization of the solvent should suffice for a reasonable evaluation of solvation effects. The close relationship between the LD model and more rigorous microscopic models has been demonstrated(253; 254)

Warshel and Levitt presented in 1976(251) the first practical simplified model for microscopic electrostatic calculations in proteins and solutions by representing the solvation behaviour of water by a simple cubic lattice of LD. Program CHEMSOL(255; 256) is designed

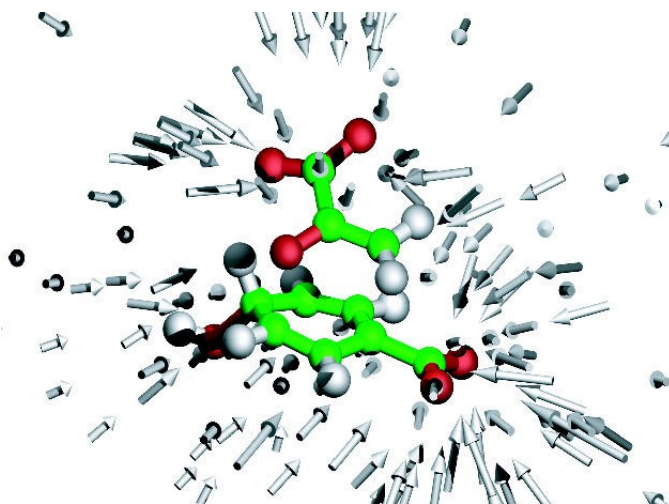


Figure 3.3: Spatial representation of LDs using solvation of TS in chorismate mutase reaction

for the calculations of solvation free energies by using LD model of the solvent and *ab initio* calculations. CHEMSOL have been used in studies of the chemical reactivity(257; 258; 259), binding(260), and conformational flexibility(261) in aqueous solution. It has been proved to be useful and robust for modeling solvation in solution and proteins.

The LD solvation model is based on the evaluation of interactions between the electrostatic field of the solute and point dipoles placed on a cubic grid placed around the solute atoms. This grid of dipoles is surrounded by a dielectric continuum. The solute electrostatic field is generated from the point charges placed at the atomic nuclei.

Point dipoles are centered at the 3Å simple cubic grid that is transformed into the denser 1Å cubic grid near the vdW surface of the solute. The boundary between the inner and outer grids is formed by points that lie at the distance of 2Å from the vdW surface of the solute. The outer grid points are constructed up to the distance of 20Å from the vdW surface of the solute. Dipoles are placed only in locations where they can really contribute to the electrostatic part of the solvation free energy. Regions of very small field are discarded. This selection is done by using the screened electric field,

$$\vec{E}_j^D = \sum_i \frac{Q_i \vec{r}_{ij}}{d(r_{ij}) r_{ij}^3} \quad (3.37)$$

where i are the solute points, j are the grid points, Q_i the atomic charges of solute points. Atomic charges are evaluated by fitting to the MEP obtained from *ab initio* calculations using GAUSSIAN 98(262) program. r_{ij} is the distance between solute points and grid points. d is a function of r_{ij} ,

$$d(r_{ij}) = \frac{\sqrt{2 + r_{ij}}}{1.7} \quad (3.38)$$

The criterion $|\vec{E}_j^D| < 0.0015e/\text{\AA}^2$ was used to select small fields.

The values of free energy of solvation, ΔG_{sol} depend upon the position of the center of the grid. It is essential to carry out LD calculations for several different grids to obtain a stable mean value of ΔG_{sol} .

$$\Delta G_{\text{sol}} = \Delta G_{ES} + \Delta G_{BULK} + \Delta G_{VDW} + \Delta G_{PHOB} + \Delta G_{RELAX} \quad (3.39)$$

- ΔG_{ES} is the electrostatic part of the solvation energy. There exist two options:
 - ILD (Iterative LD). Dipoles are allowed to interact with each other. The j th dipole, $\vec{\mu}_j$ becomes polarized along the vector of the total electrostatic field \vec{E}_j , evaluated as a sum of the unscreened contributions from the solute charges Q_i and from LD determined in the preceding $(n-1)$ th iteration.

$$\vec{E}_j = \vec{E}_j^0 + \sum_{k \neq j} \frac{[r_{jk}^2 \vec{\mu}_k^{(n-1)} - (r_{jk} \vec{\mu}_k^{(n-1)}) r_{jk}]}{r_{jk}^5} \quad (3.40)$$

where \vec{E}_j^0 is the field of the solute in vacuum,

$$\vec{E}_j^0 = \sum_i \frac{Q_i r_{ij}^{\vec{}}}{r_{ij}^3} \quad (3.41)$$

Calculations between dipoles separated by less than 2.5\AA are excluded.

- NLD (Non-iterative LD). Dipoles do not interact with each other. The field that determines the magnitude of the LD is given by equation 3.37.

ΔG_{ES} is evaluated as energy of LD in the electrostatic field generated by the solute atoms:

$$\Delta G_{ES}^{ILD} = 332K_{ES}^{ILD} \sum_j (\vec{\mu}_j^{ILD} \cdot \vec{E}_j^0) \quad (3.42)$$

$$\Delta G_{ES}^{NLD} = 332K_{ES}^{NLD} \sum_j (\vec{\mu}_j^{NLD} \cdot \vec{E}_j^D) \quad (3.43)$$

where K_{ES} is a constant needed to polarize the solvent molecules.

The extent of the dipole polarization $\vec{\mu}_j$ is given by the Langevin function(263; 264), which exhibits linear behavior for smaller fields and reaches saturation for large fields.

$$\vec{\mu}_j = \vec{\mu}_0 \left(\coth x - \frac{1}{x} \right) \quad (3.44)$$

and x is:

$$x = \frac{\vec{\mu}_0 |\vec{E}_i^j|}{K_B T} \quad (3.45)$$

where \vec{E}_i^j is the electrostatic field and is taken as equation 3.40 in the ILD approach and equation 3.41 in the NLD approach. Vector $\vec{\mu}_0$ is a constant in the direction of \vec{E}_i^j and K_B is the Boltzmann constant and T the temperature.

- ΔG_{Bulk} are the contributions from the solvent region that is outside the solvent region filled with LD. We use Born and Onsager's formulas for ionic and neutral solutes, respectively:

$$\Delta G_{Bulk}(ionic) = -166(1 - \frac{1}{\epsilon}) \frac{Q^2}{R} \quad (3.46)$$

$$\Delta G_{Bulk}(neutral) = -166 \frac{2\epsilon - 2}{2\epsilon + 1} \frac{\mu^2}{R^3} \quad (3.47)$$

where R is the radius of the LD region and Q and μ are the charge and dipole of solute, respectively. ϵ is the dielectric constant of the solvent ($\epsilon = 80$ for water).

- ΔG_{Phob} is the solvation contribution for nonpolar solutes (hydrophobic term). It is related to magnitude of the nonpolar molecular surface, which is proportional to the number of LD that lie within 1.5Å from the vdW surface of the solute.

$$\Delta G_{Phob} = K_{Phob} \sum_j f(V_j) \quad (3.48)$$

where V_j is the magnitude of the electrostatic potential at the j_{th} grid point.

- ΔG_{vdW} is the solvation contribution for nonpolar solutes (vdW term).

$$\Delta G_{vdW} = K_{vdW} \sum_i \sum_j C_i N_j [2(\frac{r_i}{r_{ij}})^9 - 3(\frac{r_i}{r_{ij}})^6] \quad (3.49)$$

where i are the solute atoms and j are the grid points. r_i are the atomic vdW radii and r_{ij} is the distance between the i_{th} atom and the j_{th} grid point. C_i are London coefficients, N_j a normalization factor and K_{vdW} is a constant parameter.

- ΔG_{Relax} is the solute-polarization term. This term takes into account the polarization of the solute electron density interacting with the LD. It uses gas-phase charges:

$$\Delta G_{Relax} = K_{relax} \sum V_i \Delta q_i \quad (3.50)$$

K_{relax} is a constant. V_i are the values of the solvent-induced electrostatic potentials evaluated as:

$$V_i = -332 \sum_j \frac{(\vec{\mu}_j \cdot \vec{r}_{ij})}{dr_{ij}^3} \quad (3.51)$$

where d is the screening factor defined in 3.38 for NLD method and 1 for the ILD method. The electrostatic potential of the solute was calculated from the HF/6-31G* wave function polarized using the polarized continuum method (PCM) of Tomasi and co-workers(265; 266) implemented in the Gaussian 94 program(262).

For more information about CHEMSOL input and output files you can see section C.8 in Annexes chapter.

3.2.5 Statistical tools in 3D-QSAR methodologies

As we have seen, QSAR are mathematical relationships linking chemical structure and pharmacological activity in a quantitative manner for a series of compounds. QSAR models must be capable not only of generalizing within a congeneric series (*i.e.* interpolate among compounds in the data set) but of correctly predicting activities for compounds outside the chemical space represented by the training set.

Once the compounds are aligned, thousands of variables are generated by the calculation of MIPs. Multivariate techniques are required to handle such amount of data, in this work Principal Component Analysis(267) (PCA) is used as a descriptive method whereas the Projection on Latent Structure or Partial Least Square (PLS) is used as a regression method. PCA and PLS project multivariate data into a space of lower dimensions, and indeed providing insight to visualize, classify, and model large sets of data.

3.2.5.1 Principal component analysis

Usually, in 3D-QSAR studies, the data file contains less than one hundred of objects and several thousands of variables. There are so many variables that by looking at them directly, no one can discover patterns, trends, clusters, etc.in the objects. The PCA is a technique extremely useful to summarize all the information contained in these variables and put it in a form understandable by human beings.

PCA is a data reduction (dimensionality reduction) method using mathematical techniques to identify patterns in a data matrix. The main element of this approach consists on the construction of a small set of new orthogonal, *i.e.*, non-correlated, variables (Principal Components PC) derived from a linear combination of the original variables that express the main information of them.

The PCA works by decomposing the original matrix of variables X as the product of two smaller matrices:

- The loading matrix (P), which contains information about the variables. It contains a few vectors, the PCs which are obtained as lineal combinations of the original X variables.
- The score matrix (T), which contains information about the objects. Each object is described in terms of their projections onto the PCs, instead of the original variables.

PCA method aims to extract the maximum amount of variance of the initial variables. To such an extent, the original descriptors are described by means of:

$$X = TP + E \quad (3.52)$$

The information not contained in matrices P and T remains as unexplained X variance in a residual matrix (E) which has exactly the same dimensionality as the original X matrix.

PC have to explain the maximum variance. They also have to be orthogonal between them. These PC describe the data in order of decreasing variance. The first axis, the so-called first PC, describes the maximum variation in the whole data set; alternatively, it

can be also pictured as the direction of greatest variance. The second PC describes the maximum remaining variance, and so forth, with each axis orthogonal, that is, linearly independent, to the preceding axis. User can decide how many PCs should be extracted (dimensionality of the model). Each PC extracted increases the amount of information (variance) explained by the model. From a practical point of view it does not matter to extract a large number of PCs if the user has no way to interpret the results.

We can represent the relative position of the objects in the space of the principal components. These plots are useful to identify clusters of objects and single objects that behave in a peculiar way (outliers). Also, the position of the objects in the plots may serve to interpret the PCs. Also we can represent the original variables in the space of the PCs. The loading of a single variable indicates how much this variable participates in defining the PC. Variables contributing very little to the PCs have small loading values and are plotted around the center of the plot. On the other hand, the variables which contribute most are plotted around the borders of the plot. The position of the observations on the new space is given by the scores and the orientation of the plane in relation to the original variables is indicated by the loadings.

In 3D-QSAR, PCA is useful to highlight the locations around the molecules or the descriptors which contain similar information or, in the opposite, which contains completely independent information. It is useful for knowing how different are one from another and why they are different. The grid-plot of loadings enables to identify the areas in the space that contribute most to a certain PC. When the meaning of the PCs is understood, the grid plot highlights the areas around the molecules associated to this meaning.

3.2.5.2 Partial least squares

PLS is used to reduce the dimensionality of the descriptor set to a small number of orthogonal latent variables (LV) correlated with the property being modeled. These variables possess the maximum amount of information relevant for the problem. In 1987, Wold(268) proposed the use of PLS analysis to correlate the field values with the biological activities. The PLS regression method(269) carries out regression using LV from the independent and dependent data that are along their axes of greatest variation and are most highly correlated.

The goal of PLS is to explain one or more Y dependent variables in terms of a number of explanatory X variables (predictors). It is typically applied when the independent variables are correlated or the number of independent variables exceeds the number of observations.

$$Y = f(X) + E \quad (3.53)$$

As for PCA the X matrix is decomposed as the product of the weight matrix W and the score matrix T . The weight matrix contains the LV, which are obtained as linear combinations of the original X variables. The loading of a single variable indicates how much of this variable is included in the LV. 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 constraints: The LV have to represent the structure fo the X matrix and the Y matrix and the LV have to maximize the fitting between X and Y .

It is possible to build many different models that fulfill the equation. Among them, the best one will be able to calculate Y values that correspond to the experimental ones, even for molecules not included in building the model. These models are predictive and can be used to calculate reliable estimations of Y values for new molecules, prior to their availability. If we try to improve too much the fitting, the model will explain also the noise. This phenomenon is called overfitting and it is very dangerous, because overfitted models seem to be very good, but they often prove to be useless to predict the Y of objects not included in the training set. Typically, the model is fit for part of the data (the training set), and the quality of the fit is judged by how well it predicts the other part of the data (the prediction set).

The best way to really evaluate the quality of the regression model is cross-validation (108; 97). In the most common Leave-One-Out (LOO) cross validation, one object (*i.e.*, one biological activity value) is eliminated from the training set and a PLS model is derived from the residual compounds. This model is used to predict the biological activity value of the compound which was not included in the model. The same procedure is repeated after elimination of another object until all objects have been eliminated once. The sum of the squared differences, between these outside-predictions and the observed Y values is a measure for the internal predictivity fo the PLS model. This is called the Standard Deviation of Error or Prediction (SDEP). SDEP and the predictive correlation coefficient (q^2) are calculated in order to avaluate the goodness of prediction of the model.

$$SDEP = \sqrt{\sum \frac{(Y - Y')^2}{N}} \quad (3.54)$$

$$q^2 = 1 - \left[\frac{\sum (Y - Y')^2}{\sum (Y - \bar{Y})^2} \right] \quad (3.55)$$

where Y is the experimental value and Y' the predicted value and N the number of objects. q^2 it is used as a diagnostic tool and is by definition smaller or equal than the overall r^2 for a model. The closer to the unity, the better predictiviness is achieved. As reference values, the commonly accepted values for a satisfactory QSAR model are r^2 greater than 0.8 and q^2 greater than 0.5.

These parameters can be used to determine the number of descriptors of the optimal model. Conversely to the classical adjustment coefficient, r^2 , which augments with the progressive addition of parameters into the regression, the q^2 coefficient presents a curve with a maximum that corresponds to the optimum number of parameters and after this maximum, the curve decreases monotonally. This means that the increase of the number of parameters of the model always improves the adjustment of data but it is not related to the predictivity of the model. As a reference value, if $r^2 - q^2$ is less than 0.3, this may indicate the presence of outliers, the selection of irrelevant descriptors, an insufficient number of data points, or the obtaining of an overfitted model, among others.

The definitive validity of the model is examined by mean of external validation, which evaluates how well the equation generalizes. The training set is used to derive an adjustment model that is after used to predict the activities of the test set members.

3.2.5.3 Pretreatment of data

Before applying multivariate analysis methods, and for the sake of quality of results, a previous treatment of the data is required. Depending on the method to be used and the amount of data available, the data set needs to be transformed by means of pre-processing methods in order to enhance the information.

The results of projection methods depend on the normalization of data. Descriptors with small absolute values have a small contribution to overall variance; this biases towards other descriptors with higher values, and leads to biased results. With appropriate scaling, equal weights are assigned to each descriptor, so that the more important variables in the model can be focussed. In order to give all variables the same importance, they are standardized by autoscaling. The standard procedure consist of normalizing each variable to mean centring and variance scaling. These transformations are recommended for ease of interpretation and numerical stability, but do not lead to changes in the coefficients or weights of variables and does not alter the interpretation of the results.

The PLS and PCA methods are sensitive to the scaling of the variables. Sometimes we have data showing very different variance. Using GOLPE(184) software, the chosen data organization allows to apply Block Unscaled Weights (BUW) scaling in order to normalize the importance of probe interactions in the final PCA model. This methods scales each single probe-protein interaction field separately, whereas the relative scales of variables within each block remain unchanged.

3.2.5.4 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. The quality of the models may be increased by the appropriate variable selection(270). Another reason for selecting variables is to simplify the PLS models, in order to make their interpretation simpler.

We have used GOLPE procedure for obtain PLS models. It involve the following steps:

- Obtain a initial PLS model. Usually it may be necessary to apply a preselection to remove the variables which contain little or redundant information for the model. In GOLPE this preselection is done choosing variables according to their positions in the loading space following a D-optimal design criterion(271; 272). This algorithm works in this space selecting the variables with higher spread and less correlation, and therefore, containin more complementarity information.
- Build the design matrix and evaluate the individual contribution of each variable to the predictivity of the model. In this step, one of the main problem, is to find the most efficient way to evaluate the individual effect of each variable in the predictivity of the models. The strategy used by GOLPE is to make a "design matrix" following Fractional Factorial Design (FFD) scheme. FFD is an experimental design technique, using a reduction factor in order to limit the number of experiments to a lower number than obtained by factorial design. The idea is to remove some variables from the model 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.

- Remove from the X-matrix the variables which do not contribute to increase the predictivity and obtain a new PLS model.

GOLPE offers the possibility to generate groups of neighbor variables in the 3D-space which represent the same chemical and statistical information. These groups can be used in the variable selection procedure, in such a way that the "groups" replace the role of the individual variables in each step of the procedure.

The SRD(221) grouping algorithm is used to group variables. In this algorithm, the variables containing more information (seeds) are extracted from the data set, following a D-optimal design criterion. Every variable is assigned to the nearest seed, following a criteria of Euclidean distance in the 3D space. Then we obtain a group, the so called Voronoi polyhedra. The Voronoi polyhedra can be used directly as groups or neighboring groups can be re-aggregated in order to merge groups containing the same information.

3.2.6 Accuracy of a prediction. Matthews correlation coefficient

Once a data set is obtained, the problem arises of defining a measure for the quality of a particular prediction. One straightforward measure of accuracy is the Matthews Correlation Coefficient (MCC)(273). MCC is defined as:

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP + FP)(TN + FN)(TP + FN)(FP + TN)}} \quad (3.56)$$

where TP are the number of true positives, TN , true negatives, FP , false positives and FN , false negatives.

It is a measure that accounts for both over and under predictions. These numbers can be counted for any given threshold value that separates the two groups. The value is close to 1 if the members of the group close mostly all above the threshold value, while the non-members score below. It is around zero, if the two groups score about equally on both sides of the threshold value (no separation). This coefficient takes into account the sensibility. It gives a value between 1 (very predictive) and -1 (anti-predictive), with a value of zero representing no useful information. Thus values of the Matthews correlation below about 0.5 are unlikely to be of great interest, and values below zero are unlikely to occur.

Results

In this section there are two important issues. The first one is the new developments implemented in MIPSim program during these three years and the second one are the scientific results derived from these new developments and applications of the program MIPSim. These results are presented chronologically and in the form and format of research papers. Thus, the contributions of this *PhD* work can directly be learned from the original publications.

4.1 Development of MIPSim program

Along these years we have developed new features inside the MIPSim program in order to make it more useful and user-friendly for the users. Also we have developed new tools that open new challenges for the future.

4.1.1 Introduction of new similarity coefficients

In Introduction chapter we have explained which kind of similarity coefficients are implemented in MIPSim. In these three years we have introduced some useful similarity coefficients.

- Hodgkin(124) similarity index. When we compare two distributions with potentials α and β with different values but with a certain proportionality, Pearson gives a perfect correlation for them. Then, it is useful to compare, in this cases, with Hodgkin similarity index defined as equation 1.28 in Introduction chapter.
- Gaussian(143; 119; 126; 144) similarity index (see section 1.2.2.2.2), now can be evaluated not only based on potential values as we have defined in equation 1.30 in Introduction chapter, also based on ordered potential ranks of these potentials.

This Gaussian similarity index based on ranks is defined as in equation 1.30, but now V_i^ξ is the ordered rank of potential value in the grid point for molecule ξ . The smoothing parameter α is taken as 0.5, value that we have found suitable in previous studies (143; 126).

We have tested if similarity matrix for a set of molecules computing all versus all comparisons is symmetric. With Spearman and Pearson index have to be exact, but not for Gaussian index.

Also we have improved the calculation algorithms of the coefficients implemented in the past in `MIPSim` like Pearson and Spearman.

We have tested the symmetry of similarity matrix with Spearman and Pearson index. With Gaussian index this symmetry is not exact for definition.

4.1.2 Combination of different similarity coefficients

Sometimes it is interesting to ponderate different values of similarity based on different electrostatic potentials derived from a set of probes computed on the molecules to compare. We can evaluate the different coefficient of similarity of every probe k we are interested in. Final similarity coefficient can be computed as:

$$S_{\alpha,\beta} = \frac{\sum_{k=1}^m w_k^{\alpha,\beta} s_k^{\alpha,\beta}}{\sum_{k=1}^m w_k^{\alpha,\beta}} \quad (4.1)$$

where $w_k^{\alpha,\beta}$ are the weights for every particular similarity coefficient $s_k^{\alpha,\beta}$ and m is the number of probes. The unique condition is that the sum of weights $w_k^{\alpha,\beta}$ is 1.

One can see in section C.5 in Annexes chapter how to compute this combination of similarity coefficients in `MIPSim`.

We have applied this weighted similarity value in studies of comparison of xantine and adenine(274).

For the evaluation of the most favourable weights to use in studies where exist structural information ligand-receptor, we have developed a protocol which obtains the optimal weights for a particular molecular type(275).

4.1.3 Introduction of new definitions of MIP

We have implemented potentials that consider solvation of molecules, using quantum mechanics (PCM) and molecular mechanics (LD) and we have obtained great results(143).

We have begun to develop, with the great contribution of H. Gutiérrez-de-Terán, `POLSAR` program (see section C.10 in Annexes chapter). It is an adaptation of `CHEMSOL`(see section 3.2.4) program designed as scoring function for screening the ability of TSA to mimic TS in order to elicit catalytic antibodies. It is possible to screen the feasible synthetic candidates on ligand databases by using LD approaches.

We have calibrated parameters with a test set of neutral and ionic solutes in aqueous solution(255). Charge distribution is computed by `GAUSSIAN 98` from the `PCM B3-LYP/HF/6-31G*`.

4.1.4 Selection of energy intervals

Sometimes is interesting to choose a certain interval of energies to compare depending on the MIP probe we are interested in: usually we work with different significative `GRID` probes that describe globally a molecule (see section 3.1.1.2.1):

DRY: probe representing hydrophobic interactions.

O: hydrogen bond acceptor carbonyl oxygen probe.

N1: hydrogen bond donor amide nitrogen probe.

These three probes are chosen because they represent the most characteristic non bonding interactions found in biological receptors. We have computed MIP for these three probes in some molecules of our studies in order to know which are the limits of the interaction energy for every probe.

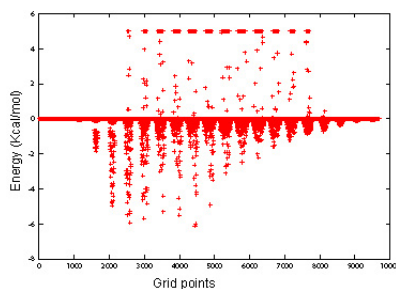


Figure 4.1: Intervals of energy for N1 probe of GRID

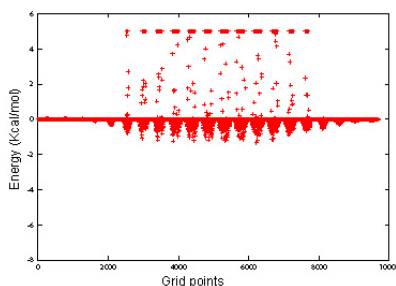


Figure 4.2: Intervals of energy for O probe of GRID

One can see in section C.6 of Annexes chapter how to compute this comparison of different intervals of energy in MIPSim.

We have applied this selection of intervals in the study of catalytic antibodies(143) and in a MIP-based alignment of HIVRT inhibitors(275).

4.1.5 Conformational flexibility: TORS module

We have designed and implemented some strategies of exploration of the conformational degrees of freedom of the molecules we want to compare. On the other hand, for the exploration of the conformational flexibility, specially in great series of molecules and without the manual intervention of the user, it is necessary the development of an algorithm of automatization of the description of the molecular structures, using the 3D coordinates of the

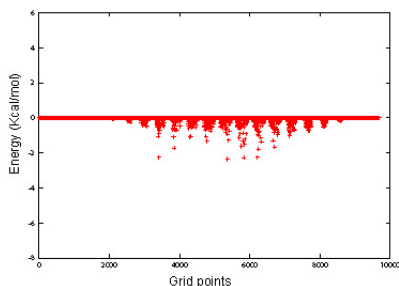


Figure 4.3: Intervals of energy for DRY probe of GRID

molecules. We try to focus on the chemistry of atoms of the ligand and not in rotamers databases to perform an automatic research of rotatable bonds.

We have introduced a preliminary option in `MIPSim` that enables to generate all possible 3D conformations of a flexible ligand. `MIPSim` does this using `TORS` module. `MIPSim` can find all the rotatable bonds in a ligand and generate conformations. User can select the bond to rotate and the rotation pass. Once the user have the conformations it is possible to explore the optimal superposition taking into account the ligand flexibility.

This is the protocol used in rotatable bonds perception and rotation:

- Perform a list with all the connection between atoms based on vdW radii and the atom type (connectivities between atoms).
- Calculate the number of fragments (different molecules) and atoms in every fragment.
- Compute the connectivity of each atom in our file and the number of atoms with each connectivity.
- Count the number of rings in each fragment.
- Detect the neighbours of every atom.

Performs a table where one can find which neighbour has every atom in every bond and a table of neighbours of every atom in our molecule depending on connectivity. Also, assign a name to every atom that tell us the connectivity of it: C4, HH, O2...

We have implemented a rudimentary routine to rotate a bond knowing an angle and both atoms. This is done in `MIPSim` using cylindrical coordinates. A point can be localized using cylindrical coordinates (r, θ, z) :

$$x = r \cos \theta \quad (4.2)$$

$$y = r \sin \theta \quad (4.3)$$

$$z = z \quad (4.4)$$

or

$$r = \sqrt{x^2 + y^2} \quad (4.5)$$

$$\theta = \operatorname{atan} \frac{y}{x} \quad (4.6)$$

$$z = z \quad (4.7)$$

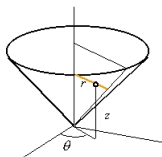


Figure 4.4: Cilyndrical coordinates

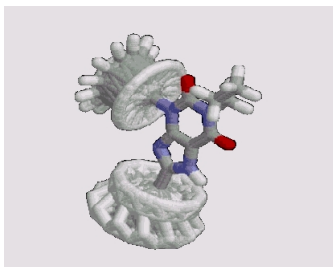


Figure 4.5: Rotations generated in two bonds for the molecule dpcpx(8-ciclopentil-1,3-dipropilxantina)

TORS finally writes all pdb files with all the different new conformations.

In section C.7 of Annexes one can find information about how to write a .key file in order to perform rotations with TORS module.

At the moment, user have to point out the atoms that form the rotatable bond in order to integrate conformational flexibility in the search of the optimal superposition of MIPS. That implies an explicit knowledge of every system and it is difficult to find similarity in big databases.

4.2 Benchmarking and profiling of MIPSim

In order to study and improve the benchmark of MIPSim we have created different test files. All of them compute important features in the program: comparison of molecules with the COMP module . Calculation of MIP with GRID with several probes and MEP with GAMESS. In these tests we perform different capabilities of MIPSim program as the calculation of MIP minima with MIN module and the comparison of MIP distributions with COMP

module. MIP are computed with classical and quantum methodologies in order to test the interface with programs `GRID` and `GAMESS`. Also there are performed comparisons of different intervals of energy depending on the probe for different probes. We test all the optimization methodologies implemented in `mipsim` and the different similarity coefficients and combinations between them. In some tests are tested the visualization tools compatible with `MIPSim`. Also is tested the interface with `GOLPE` in order to use statistical tools.

A lot of `MIPSim` algorithms have been improved in order to decrease the time of calculation and to minimize the computational cost. We have used a compiler very useful for the detection of some routines with a big computational cost. You can see C.4 section in Annexes for a more detailed information.

4.3 Technological platforms development

4.3.1 OS platforms

We have adapted `MIPSim` code in order to work in different platforms like Linux (Red-Hat 9.0), SGI (IRIX 6.5) and Alpha (True 64). The adaptation to Alpha platform enables users to work in a supercomputer center as CESCO (Centre de Supercomputació de Catalunya) with `MIPSim`.

4.3.2 User interface

We have created the first version of the graphical interface of `MIPSim` using C++ and QT libraries (in collaboration with Cristina Herraiz). Interface enables to launch calculations in an easy way for non-advanced users.

4.3.3 Visualization tools

Outputs of `MIPSim` can be visualized with different graphic programs. In particular, now it is possible, thanks to Cristina Herrantz, to use `LINK3D`(189). `LINK3D` is a visualization program and supports secure synchronous remote collaboration between scientists working in Drug Discovery and Development.

4.3.4 MIPSim web site

`MIPSim` web site has been improved in collaboration with J.A. de los Cobos. We have introduced new elements like a download section of the `MIPSim` package and detailed information of the program and its possibilities. `MIPSim` distributes two different licenses: one for the academic environment and the other for the commercial environment.

4.4 MIPSim applications. Publications

4.4.1 Comparison of biomolecules by MIP alignment

Rodrigo, J., Barbany, M. *et al.*, *J. Braz. Chem. Soc.*, **13**, 795-799 (2002) abstract, full text and pdf

One example of an application of MIPSim is the comparison of xantine and adenine. In this paper was used carbonyl oxygen and the amide nitrogen as GRID probes for MIP distributions. The most important contribution at this work was the combination of several similarities in only one global weighted similarity as we have seen in equation 4.1.

MIPSim can perform six alignments with a great similarity coefficient with biological and chemical coherence: the coincidence of the hydrogen bond acceptor and the donor bond acceptor and the heterocycles.

Rodrigo J, Barbany M, Gutiérrez-de-Terán H, Centeno NB, De-Câceres M, Dezi C, Fontaine F, Lozano JJ, Pastor M, Villà J, Sanz F.

[Comparison of biomolecules on the basis of Molecular Interaction Potentials.](#)

J. Braz. Chem. Soc. [online]. 2002, vol.13, n.6, pp. 795-799.

4.4.2 On the use of MIPSim for characterizing the activity of catalytic antibodies

Barbany, M. *et al.*, *ChemBioChem*, **4**, 277-285 (2003) abstract, full text and pdf

TSA used produces CAs with relatively low rate enhancement as compared to the corresponding enzymes, when these exist. The present work explores the origin of the problem, by developing two approaches that examine the similarity of the TSA and the corresponding TS. Both approaches focus on electrostatic effects that have been found to play a major role in enzymatic reactions(22; 276). The first method uses molecular interaction potentials to study the similarity between TS and the correspondent TSA, using MIPSim program. We analyze similarity and differences of the electrostatic distribution between TSA and TS using MIN and COMP module and computing molecular interaction potentials with GRID and molecular electrostatic potential with GAMESS. We have identified the regions where the electrostatic potentials of the TS and TSA differ, in particular, the region of the carboxylate group and of the bonds being broken and formed.

The second method, more quantitative, generates a grid of LD(21) polarized by TSA. Then, it is used to bind TS. It represents the main physics of the solute-solvent interaction (explicit solvent). This is computed by CHEMSOL program (see section 3.2.4). We use the potential generated by each grid to evaluate the solvation energy of the TS and the TSA. It was found that the environment preorganized to stabilize the TSA is not as proficient in stabilizing the TS because the environment preorganized specifically to stabilize the TS. The difference in solvation free energy of the TS charge distribution solvated by dipoles preorganized to solvate the TS in the geometry of TS, and the geometry of TSA corresponds qualitatively with the observed difference in proficiency of the enzyme and catalytic antibody. These findings suggest that catalytic antibodies raised against the TSA may not stabilize the TS effectively. TS is more stabilized by the enzyme than the corresponding TSA. The comparison of free energy of binding of TS with the grid of LD with the binding energy of TS in the enzymatic environment give a good estimation of CA efficiency.

It is demonstrated that the relatively small changes in charge and structure between the TS and TSA are sufficient to account for the difference in proficiency between the CA and the enzyme. Apparently the environment that was preorganized to stabilize the TSA charge distribution does not provide a sufficient stabilization to the TS.

Chorismate mutase (CM) catalyses the Claisen rearrangement of chorismate to prephenate. This reaction is a key step in the shikimate pathway for biosynthesis of phenylalanine and tyrosine in bacteria, fungi, and higher plants(277). This enzymatic rearrangement has been the focus of major effort in recent years, including analysis of its relationship to catalytic antibodies that catalyze the same reaction(26; 278; 279; 280; 281) and extensive simulation studies(282; 283; 284; 285; 286; 287; 288; 289; 290). The advantage of this reaction is that there is no formation of an enzyme covalent intermediate in the reaction. Thus, the reactions in water and the enzyme are both kinetically first order and directly comparable.

Barbany M, Gutiérrez-de-Terán H, Sanz F, Villà-Freixa J, Warshel A.

On the generation of catalytic antibodies by transition state analogues.

Chembiochem. 2003 Apr 4;4(4):277-85.

4.4.3 MIPSim as scoring function in protein-ligands docking

Barbany, M. *et al.*, *Proteins: Structure, Function and Bioinformatics*, **56**, 585-594 (2004) abstract, full text and pdf. Supplementary material.

MIP describe particular roles in the intermolecular recognition. MIPSim performs a mixture of MIPs to generate a smooth similarity function based on a combination of weighted MIP. We have complexes inhibitor-protein. Then we perform a superposition using STAMP. Then, we obtain ligands in the biological conformation. Then, we try to find interaction fingerprints assigning weights to every MIP. Finally we dock new ligands using MIPSim alignment procedure and similarity as scoring function.

Barbany M, Gutiérrez-de-Terán H, Sanz F, Villà-Freixa J.

[Towards a MIP-based alignment and docking in computer-aided drug design.](#)

Proteins. 2004 Aug 15;56(3):585-94.

4.4.4 3D-QSAR study of hERG inhibitors

Barbany, M. *et al.* In preparation.

One of the most important problems in structure-activity studies is to obtain good alignments of molecules that share the same binding site in order to obtain a good 3D-QSAR model. There are several ways to perform alignments: using pharmacophoric points, chemical features, shape and molecular interaction potentials (MIP). MIPSim performs structural alignments for a series of biomolecules using their MIPs.

The hERG potassium channel is expressed in the human heart. The channel is a key effector of cardiac repolarization and contributes to the QT interval measured by the electrocardiogram. Inhibition of hERG can lead to a prolongation of the QT interval, widely considered a critical risk factor for arrhythmia. Thus, hERG inhibition represents an important safety consideration in drug discovery.

In the present study we obtain structural alignments for a series of drugs that are known to inhibit the hERG potassium channel with different degrees of activity. To validate our MIP based alignments, we use them in 3D-QSAR study using the GRID/GOLPE protocol.

Using MIPSim and 3D-QSAR to study binding modes of hERG K⁺ channel inhibitors

Montserrat Barbany,^a Scott Boyer,^b Ferran Sanz,^a and Jordi Villà-Freixa,^a

Addresses:

a Computational Structural Biology Laboratory

Research Group on Biomedical Informatics (GRIB) - IMIM/UPF

Dr. Aiguader, 80 Tel: +34 93 221 1009

E-08003 Barcelona (Spain) Fax: +34 93 221 3237

b Safety Assessment Group

AstraZeneca R&D

431 83 Möndal (Sweden)

Tel: +46 31 776 2882 Fax: +46 31 776 3792

e-mail addresses: jvilla@imim.es, fsanz@imim.es, Scott.Boyer@astrazeneca.com

Abstract

The hERG potassium channel (human ether-a-go-go-related gene) is expressed in the human heart. The channel is a key effector of cardiac repolarization and contributes to the QT interval measured by the electrocardiogram. Inhibition of hERG can lead to a prolongation of the QT interval, widely considered a critical risk factor for arrhythmia. A chemically diverse series of drugs have been withdrawn from the market due to their hERG blocking properties. Thus, hERG inhibitory effects represent an important safety consideration in drug discovery.

MIPSIM is a program that analyzes and compares molecular interaction potentials (MIP) distributions for a series of biomolecules. MIPSIM performs structural alignments for a series of biomolecules using their MIPs.

In the present study we obtain structural alignments for a series of drugs that are known to inhibit the hERG potassium channel with a rank of activities. To validate our MIP based alignments, we use them in a 3D-QSAR study using the GRID-GOLPE protocol. Also, we try to test the predictivity of our model based on similarity values.

Our objective is to demonstrate that MIPs are more useful to obtain good alignments in order to create a predictivity model than a pharmacophoric manual alignment.

Keywords hERG potassium channel; structure-activity relationships; molecular interaction potentials.

1 Introduction

The potassium channel encoded by the human ether-a-go-go related gene (hERG) gives rise to the rapid component of the delayed rectifier K⁺ channel current.¹ The hERG K⁺ channel plays a crucial role for normal action potential repolarization in the heart. It has been used as a therapeutic target for anti-arrhythmic agents, but a wide range of noncardiac drugs also inhibit the hERG K⁺ channel, resulting in a drug-induced long QT syndrome (LQTS) that can cause sudden cardiac death.^{2,3} The protein product of hERG is a potassium channel that when inhibited by some drugs may lead to cardiac arrhythmia.⁴ Mutations in the hERG (Human ether-a-go-go related gene) K⁺ channel cause inherited long QT syndrome

(LQT), a disorder of cardiac repolarization that predisposes affected individuals to lethal arrhythmias.^{2,3} Acquired LQT is far more common and is most often caused by block of cardiac hERG K⁺ channels by commonly used medications, including antiarrhythmic, antihistamine,^{5,6} antibiotic⁷ and antipsychotic^{8,9} drugs. It is unclear why so many structurally diverse compounds block hERG channels. It is therefore important to assess the hERG blocking potential of novel chemical structures as early as possible during the drug discovery process. We use here a set of molecules taken from an organized list of QT-prolonging compounds for which hERG K⁺ channel inhibition had been reported and for which IC₅₀ values for inhibition expressed in mammalian cells were available.¹⁰

Table 1: QT-Prolonging Drugs

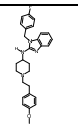
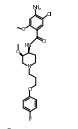
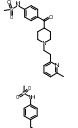
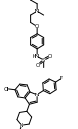
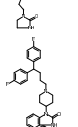

inhibitor	structure	activity IC ₅₀ (nM)
Astemizole		0.9
Cisapride		6.5
E-4031		7.7
Dofetilide		9.5-15
Sertindole		14
Pimozide		18

Table 1: QT-Prolonging Drugs

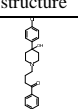
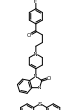
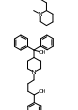
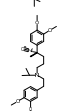
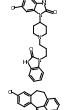
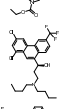
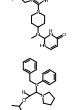
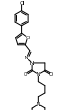



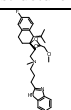













inhibitor	structure	activity IC50 (nM)
Haloperidol		28.1
Droperidol		32.2
Thioridazine		35.7
Terfenadine		56-204
Verapamil		143
Domperidone		162
Loratadine		173
Halofantrine		196.9
Mizolastine		350
Bepridil		550
Azimilide		560

Table 1: QT-Prolonging Drugs

inhibitor	structure	activity IC50 (nM)
Mibefradil		1430
Chlorpromazine		1470
Imipramine		3400
Granisetron		3730
Dolasetron		5950
Perhexiline		7800
Amitriptyline		10000
Diltiazem		17300
Sparfloxacin		18000-34400
Glibenclamide		74000
Grepafloxacin		50000-104000
Sildenafil		100000
Moxifloxacin		103000-129000
Gatifloxacin		130000

2 Objectives

Since none of the existing *in vitro* tests to assess the QT-prolonging potential of a compound has an absolute predic-

tive value,¹¹ the availability of *in silico* methods in the early phase of drug development would dramatically increase the screening rate and would also lower the costs compared to

experimental assay methods.

One of the most important problems in structure-activity relationship studies is to obtain good alignments of molecules that share the same binding site in order to obtain a good 3D-QSAR model. There are several ways to perform alignments: using pharmacophoric points, chemical features, shape and molecular interaction potentials (MIP).

MIPSIM¹² is a program that analyzes and compares MIP distributions for a series of biomolecules. MIPSIM also performs molecular alignments on the basis of MIP distributions. In this work we use MIPSIM¹² as both: a tool to optimize the alignment¹³ between molecules of interest, as a first step towards the generation of a 3D-QSAR model using tools like GRID-GOLPE¹⁴ and a method to easily correlate, for a given new molecule, similarity of the new molecule with activity. This program has been used with success in several studies^{13,15,16} showing that the MIPs are good features for performing superpositions.

For our study, first we use the conformations and final alignment performed by Cavalli et al.¹⁷ gave us with great kindness. On a different study we create conformations for every inhibitor and we perform a manual alignment. For both systems it is performed a MIPSIM alignment based on similarity of molecular interaction potentials and finally created 3D-QSAR models using GRID-GOLPE procedure.

As a final study we try to test the predictability of a model based only on similarity of MIPs.

3 Methods

3.1 Conformational exploration

As at the moment the structure of the hERG channel is unknown, not only the shape of characteristics of the binding site are missing, but also the 3D-structure of the active form of any given inhibitor.

Thus, we need a tool to explore the conformational flexibility of each ligand in table 1 prior to proceed further. In this

work we make use of program OMEGA¹⁸ to do such conformational exploration. OMEGA supports a so-called torsion-driving beam search for generating ensembles of conformers, which allows the program to generate conformations in a fast way. OMEGA generates conformations extremely fast. By contrast with stochastic methods,^{19,20} the results are completely reproducible. OMEGA deconstructs the molecule into fragments with rotatable bonds, and uses certain build-up principles to generate a conformational ensemble. OMEGA does not minimize bond lengths or bond angles. All heavy atoms are superimposed to test for duplicated structures, with a default RMSD criterion of 0.8 Å. OMEGA includes a simple force field called the Clean force field. Any structure with an energy of more than 5 kcal/mol above the current global minimum is discarded.

3.2 Superposition

In order to create a 3D-QSAR model of our series of molecules we need first to create an structural alignment of the selected conformers. In this study two approaches have been taken: a manual alignment based on pharmacophoric points and an alignment based on molecular interaction potentials (MIP).

3.2.1 Manual alignment

Using the conformations created by OMEGA we superpose all of them respect to astemizole (see table 1). We choose this template because this molecule is one of the most potent long QT-inducing drugs, and hERG channel blockers. The astemizole crystal structure was directly retrieved from the Cambridge Structure Database (CSD).²¹

Several pharmacophoric points, similars to Cavalli et al.,¹⁷ on the astemizole molecule were defined to superpose everyone of the conformations created for every inhibitor. We superpose them (using certain atoms) using the program SUPERB developed in our group (based on the Chen methodology²²) to superpose two different geometries based on a

group of atoms. We select the conformation of every compound with the minimum RMSD versus astemizole.

Initially, three pharmacophoric points on the astemizole molecule were defined, namely, the basic nitrogen of the piperidine cycle (N) and atoms of the two close aromatic moieties (called by Cavalli et al. C0 and C1). Also, a fourth pharmacophoric point (defined as the centroid C2) of the phenyl ring belonging to the N-(p-methoxyphenylethyl) substituent of astemizole. Not all the molecules displayed all four pharmacophoric points, and in such cases, the superimposition was based on the available points or it was guided by other functions present on some molecules. Particularly, in the case of compounds 2, 5-8, 12-15, 17 and 18, the halogen atom located in the para position on one phenyl ring (C0) was used to reinforce the fit. As regards the quinolones, which are the least potent hERG channel blockers considered in this study, their structures were quite different from the structure of the template, which implied that they were superposed to astemizole by first anchoring the molecular skeleton to the ba-

sic piperazine N atom. This oriented the centroid of the 4-piperidone ring onto C0.

3.2.2 MIPSIM alignment.

MIPSIM is a computational package designed to analyze and compare 3D distributions of molecular interaction potentials (MIPs) of series of biomolecules. In particular, MIPSIM can obtain similarity indices and calculate superpositions of molecules based on a single MIP or a combination of them. With a protocol based on pairwise comparisons, mipcomp module in MIPSIM can evaluate several MIPs in a grid box around each molecule and then compare the MIPs of both molecules on the basis of similarity indexes. Using the similarity index as scoring function, the program can perform an automatic search of the best relative orientation between the two molecules. At every orientation of the *mobile* molecule and for every MIP k , the similarity index $s_k^{\alpha,\beta}$ is calculated by a Gaussian coefficient^{12,15,23} in equation 1.

$$s_k^{\alpha,\beta} = \frac{\sum_{i=1}^{n_\alpha} \sum_{j=1}^{n_\beta} V_i^\alpha V_j^\beta \exp(-ar_{ij}^2)}{\sqrt{\sum_{i=1}^{n_\alpha} \sum_{j=1}^{n_\alpha} V_i^\alpha V_j^\alpha \exp(-ar_{ij}^2)} \sqrt{\sum_{i=1}^{n_\beta} \sum_{j=1}^{n_\beta} V_i^\beta V_j^\beta \exp(-ar_{ij}^2)}} \quad (1)$$

where n_ξ ($\xi = \alpha, \beta$) is the number of points in each grid box selected for the comparison, V_i^ξ is the potential value in the grid point i for molecule ξ and r_{ij} is the distance between two points. The smoothing parameter a is set to 0.5, value that has been found to work well in previous studies.^{12,15} Following this procedure we can evaluate a different similarity index for every GRID probe we are interested in and obtain the global similarity index using Equation 2.

$$S_{\alpha,\beta} = \frac{\sum_{k=1}^m w_k^{\alpha,\beta} s_k^{\alpha,\beta}}{\sum_{k=1}^m w_k^{\alpha,\beta}} \quad (2)$$

where $w_k^{\alpha,\beta}$ are the weights of every particular similarity index $s_k^{\alpha,\beta}$ and m is the number of GRID probes.

Using the conformers selected in the last calculation and performing an exploration of the best superposition using

astemizole as the static molecule, we obtain new positions for every compound.

We chose three representative probes that cover a wide range of possible protein-ligand interactions with weights of 0.33 for every probe: O (hydrogen bond acceptor group), N1 (hydrogen bond donor group), and DRY (hydrophobic interactions). For each comparison vs astemizole several initial random orientations were calculated and from each of them the conjugated gradient optimization protocol in MIPSIM found what it considered the best orientation. In this study, ten tests starting from, respectively, ten initial random relative orientations between static and mobile molecules were performed for each comparison. MIPSIM selects the best compound that makes similarity higher.

3.3 3D-QSAR models

The next step is to create a 3D-QSAR model for this set of inhibitors. After the two alignments in previous section have been generated, we create two different GRID-GOLPE models depending of the alignment used. The first one is based on pharmacophoric points and the other one is based on Molecular Interaction Potentials. This procedure is followed for the two alignments.

We use GOLPE 4.5 in IRIX. We performed a Principal Component Analysis (PCA) and Partial Least Squares (PLS). Data pretreatment included zeroing of the values between -0.1 and 0.1, exclusion of variables with standard deviance less than 0.1 and exclusion of N-level variables. Also, we perform an scaling BUW of variables.

We compute Smart Region Definition (SRD).²⁴ SRD procedure works by extracting a subset of highly informative X variable and then partitioning the space around the molecules amongs them. Then we compute a fractional factorial design (FFD) variable selection using leave-one-out procedure as PLS validation.

3.4 Predictivity of the model

The data test set used in this paper is taken from Aronov et al.²⁵ We choose six of the positive inhibitors and four of the negative inhibitors of hERG K+ channel. In order to take into account the flexibility of them, conformers are created using the Confirm module of Catalyst 4.9.²⁶ Fast conformation generation uses a heuristic method that quickly builds a geometrically diverse conformatinal model for the molecule.

Comparing every one of the conformers for every molecule in the test set we compare them using MIPSIM with the 10 first more potent inhibitors from Cavalli et al.¹⁷ Performing a PLS analysis we try to create a model that is useful in order to detect positive inhibitors and negative inhibitors and negative inhibitors of hERG K+ channel.

4 Results

4.1 Conformational exploration

Using OMEGA, we collect a maximum of 20 conformations for every compound using the defaults value of the program.

In figure 1 we can see the number of conformations created for every one of the compounds.

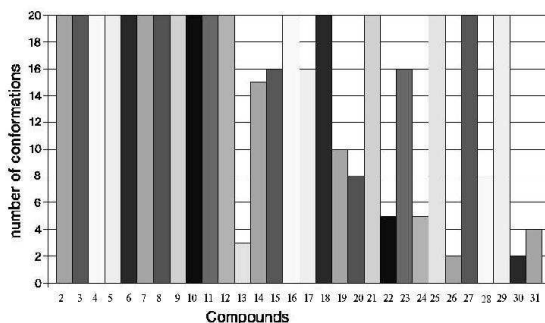


Figure 1: Number of conformations for every one of the hERG inhibitors

4.2 Alignment procedures

4.2.1 Alignment based on pharmacophoric points

Using pharmacophoric points cited in the paper of Cavalli et al.¹⁷ we have aligned all hERG inhibitors (figure 2).

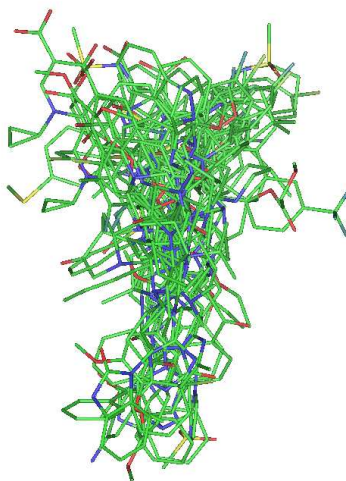


Figure 2: Best superposition of all hERG inhibitors based on pharmacophoric points

4.2.2 Alignment of Cavalli refined with MIP based superposition using MIPSIM

Using the conformers selected in the last calculation and performing an exploration of the best superposition using astem-

izole as the static molecule, we obtain new positions for every compound. The alignment can be seen in figure 3.

4.3 3D-QSAR study

4.3.1 3D-QSAR model created with the superposition based on pharmacophoric points

We create a 3D-QSAR model using GOLPE and the superposition based on pharmacophoric points as the training set. Ac-

cording to the SRD methodology we use in this case 2 components and computing PLS model we obtain results in table 2 for r^2 . Performing the leave-one-out validation for PLS we obtain the q^2 shown in table .

We can observe that q^2 is very low. This means that the

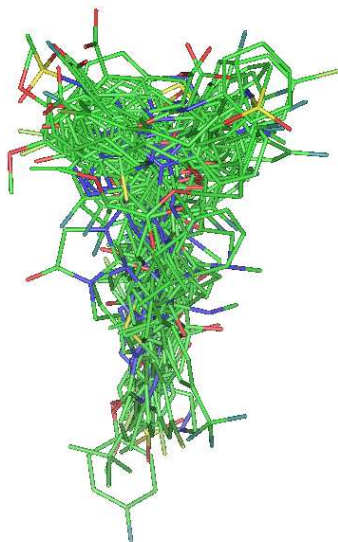


Figure 3: Best superposition of all hERG inhibitors based on MIPs computed by MIPSIM

model is not very good to predict the activity of the compounds.

These results are not directly comparable with those in Cavalli et al., due to the fact that, even following a similar protocol for the generation of conformers and their superpo-

sitions in both Cavalli et al. and this paper, the details of such a manual superposition can provoke (as in fact they do) a very different final 3D-QSAR model. This is the reason why it is needed a more automatic and reproducible way to perform such an alignment. In the next section it is shown how MIPSIM can help in this step.

Table 2: PLS results

LV	XAcc	SDEC	r^2
0	0	1.44	0
1	18.09	1.09	0.43
2	40.95	0.97	0.55

Table 3: PLS results

LV	SDEP	q^2
0	1.49	-0.07
1	1.33	0.15
2	1.17	0.34

4.3.2 3D-QSAR model using a MIP based superposition model

Then we create a 3D-QSAR model using PLS implemented in GOLPE. We obtain a r^2 that can be shown in table 4. Per-

forming the leave-one-out validation we obtain the q^2 shown in table 5. We can observe that q^2 is higher than the last GRID-GOLPE calculation. This means that the model is better to predict the activity of the compounds.

Table 4: PLS results

LV	XAcc	SDEC	r^2
0	0	1.44	0
1	11.25	0.53	0.86
2	18.06	0.28	0.96

Table 5: PLS results

LV	SDEP	q^2
0	1.49	-0.07
1	0.79	0.70
2	0.62	0.81

5 Conclusions

We have demonstrated that a good alignment of biomolecules is a crucial step in 3D-QSAR studies in order to obtain a good predictable model. In this sense, the execution of detailed and objective comparative analysis of MIP distributions is useful to obtain good alignments. We can observe the improvement in PLS study when using MIPSIM alignment.

One of the possible studies to perform would be to perform all pairwise comparisons of all the conformations for every compound in order to find the active conformation of

hERG inhibitors. We assume that the active conformation will be those who have the best similarity of MIPs between the most active compounds. Once we know the active conformation for every compound we can perform a 3D-QSAR modeling. It could detect side effects as early as possible during drug development.

6 Acknowledgements

We are grateful to the Fondo de Investigaciones Sanitarias for financial support of this research (FIS 01/1330) and to travel fellowship BE of Generalitat de Catalunya. Also, we are grateful to the Safety Assessment group at AstraZeneca in Sweden. We want to thank to the Centre de Computació i Comunicacions de Catalunya (C4) for computer time allocation. We also thank Dr. Manolo Pastor and Fabien Fontaine for helpful discussions.

References and Notes

1. Roche, O.; Truble, G.; Zuegge, J.; Pflimlin, P.; Alaine, A.; Schneider, G. *ChemBioChem* **2002**, *3*, 455-459.
2. Keating, M.T. and Sanghinetti, M. *Cell* **2001**, *104*, 569-580.
3. Trudeau, M.; Warmke, J.; Ganetzky, B.; Robertson, G. *Science* **1995**, *269*, 92-95.
4. Ben-David, J.; Zipes, D.
5. Roy, M.-L.; Dumaine, R.; Brown, A. *Circulation* **1996**, *94*, 817-823.
6. Zhou, Z.; Vorperian, V.; Gong, Q.; Zhang, S.; C.Y., J. *J. Cardiovasc. Electrophysiol.* **1999**, *10*, 836-843.
7. Vandenberg, J.; Walker, B.; Campbell, T. *Trends Pharmacol. Sci.* **2001**, *22*, 241-246.
8. Rampe, D.; Roy, M.-L.; Dennis, A.; Brown, A. *FEBS Lett.* **1997**, *417*, 28-32.
9. Kahn, J.; Wang, L.; Cai, F.; Rampe, D.
10. De Ponti, F.; Poluzzi, E.; Montanaro, N. *Eur. J. Clin. Pharmacol.* **2001**, *57*, 185-209.
11. De Ponti, F.; Poluzzi, E.; Cavalli, A.; Montanaro, N. *Drug. Saf.* **2002**, *25*, 263-286.
12. de Cáceres, M.; Villà, J.; Lozano, J. J.; Sanz, F. *Bioinformatics* **2000**, *16*, 568-569.
13. Rodrigo, J.; Barbany, M.; Gutiérrez-de Terán, H.; Centeno, N. B.; de Cáceres, M.; Dezi, C.; Fontaine, F.; Lozano, J. J.; Pastor, M.; Villà, J.; Sanz, F. *J. Braz. Chem. Soc.* **2002**, *13*, 795-799.
14. Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9-20.
15. Barbany, M.; Gutiérrez-de Terán, H.; Sanz, F.; Villà-Freixa, J.; Warshel, A. *ChemBioChem* **2003**, *4*, 277-285.
16. Barbany, M.; González-de Terán, H.; Sanz, F.; Villà-Freixa, J. *Proteins: Struct. Func. Gen.* **2004**, *56*, 585-594.
17. Cavalli, A.; Poluzzi, E.; De Ponti, F.; Recanatini, M. *J. Med. Chem.* **2002**, *45*, 3844-3853.
18. Böstrom, J. *Comput. Aid. Mol. Des.* **2001**, *15*, 1137-1152.
19. C., M.; Bohacek, R. *J. Comput. Aid. Mol. Des.* **1997**, *11*, 333-344.
20. Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley: New York, 1989.
21. Allen, F. *Acta Crystallogr.* **2002**, *B58*, 380-388.
22. Chen, Z. *Theoretica Chimica Acta* **1989**, *75*, 849-857.
23. Bagdassarian, C. K.; Schramm, V. L.; Schwartz, S. D. *J. Am. Chem. Soc.* **1996**, *118*, 8825-8836.
24. Pastor, M. *J. Med. Chem.* **1997**, *40*, 1455-1464.
25. Aronov, A.; Goldman, B. *Bioorg. Med. Chem.* **2004**, *12*, 2307-2315.
26. Sprague, P. *Perspect. Drug Disc. Des.* **1995**, *3*, 1-20.

Conclusions

1. We have further developed critical aspects of MIPS_{im}. It is possible now to compute new similarity coefficients in MIPS_{im} as well as new MIP descriptors. It is now possible to obtain global similarity coefficients based on weighted similarity for different potentials derived from a set of probes. It has also become possible to select different energy intervals depending on the MIP considered. Finally, a preliminary version of the flexible similarity tool has been introduced.
2. The development of a conveniently weighted combination of similarity indexes has proved to be useful to perform similarity alignments between adenine and xantine using this combination of coefficients with different probes chosen from a priori chemical and biological knowledge.
3. We have shown that the electrostatic preorganization theory in enzyme reactivity can be demonstrated by means of the use of simple MEPs.
4. The use of MEP in reactivity studies allows us to take into account the features of the potential in cases where classical probes cannot be used (bonds that are being broken or formed). We have illustrated the origin of the low enhancement of CAs compared to the corresponding enzymes. We demonstrated that a relative small difference between the charge distributions of the TSA and TS leads to significant differences in the polarization of the corresponding complementary environment.
5. We have developed a methodology that tries to find the structural alignment of series of molecules in their biological conformation using protein-ligand structural information and MIP. MIPS_{im} was capable to find biological alignments with good experimental agreement in a series of DHFR and non nucleoside HIV-1 retrotranscriptase inhibitors using similarity of MIPs as scoring function.
6. In series of hERG potassium channel inhibitors, we have confirmed that the best correlation between the structure and function of biological molecules is provided by the comparison of molecules aligned based on their MIPs. The subsequent 3D-QSAR study by means of the GRID/GOLPE protocol has shown the improvement of the predictivity of the model when using such consistent approach.

List of abbreviations

3D Three-Dimensional

BUW Block Unscaled Weights

CA Catalytic Antibody

CoMFA Comparative Molecular Fields Analysis

DNA Deoxyribonucleic Acid

FEP Free Energy Perturbation

FFD Fractional Factorial Design

GA Genetic Algorithm

GOLPE Generating Optimal Linear PLS Estimation

GRIND Grid-INdependent Descriptors

GS Ground State

GSD Ground-State Destabilization

GUI Graphical User Interface

HF Hartree Fock

ILD Iterative Langevin Dipoles

LCAO Linear Combination of Atomic Orbitals

LD Langevin Dipoles

LIE Linear Interaction Energy

LOO Leave-One Out

LRA Linear Response Approximation

LV Latent Variable

mRNA Messenger Ribonucleic Acid

MC Monte Carlo
MCC Matthews Correlation Coefficient
MD Molecular Dynamics
MEF Molecular Electric Field
MEP Molecular Electrostatic Potential
MIF Molecular Interaction Field
MIP Molecular Interaction Potential
MM Molecular Mechanics
NAC Near Attack Conformation
NLD Non-iterative Langevin Dipoles
NMR Nuclear Magnetic Resonance
PC Principal Component
PCA Principal Component Analysis
PCM Polarized Continuum Model
PDF Portable Document Format
PES Potential Energy Surface
PLS Partial Least Square
QM Quantum Mechanics
QSAR Quantitative Structure-Activity Relationship
RMSD Root Mean Square Deviation
RNA Ribonucleic Acid
SAR Structure-Activity Relationship
SCF Self Consistent Field
SDEP Standard Deviation of Error of Prediction
STO Slater Type Orbitals
TS Transition State
TSA Transition State Analog
TST Transition State Theory
vdW van der Waals

Interesting links

- PDBsum <http://www.biochem.ucl.ac.uk/bsm/pdbsum>
- Relibase <http://relibase.ebi.ac.uk/>
- Protein Data Bank (PDB) <http://www.rcsb.org/pdb/>
- InsightII <http://www.accelrys.com/insight>
- Link3D <http://www.tecn.upf.es/prj/link3d/>
- MIPSim <http://diana.imim.es/software/mipsim/index.html>
- GAMESS <http://www.msg.ameslab.gov/GAMESS/GAMESS.html>
- GRID http://www.moldiscovery.com/soft_grid.php
- STAMP <http://barton.ebi.ac.uk/>
- FFSQP <http://www.aemdesign.com/FSQPwhatis.htm>
- PGI http://www.pgroup.com/hpf_docs/pghpf_ug/hpfug11.htm

Presentations in congresses

B.1 Posters

M. Barbany, J. Villà-Freixa, H. Gutiérrez-de-Terán and F. Sanz

Comparison of biomolecules on the basis of molecular interaction potentials.

The new version of the MIPSim package

XVIIth International Symposium on Medicinal Chemistry, Barcelona (Spain) (2002)

Molecular Interaction Potentials (MIP) are useful tools for the comparison of series of compounds displaying related biological behaviours. Structure-activity studies need a detailed comparative analysis of MIP distributions for the pharmaceutical interest in the design of new drugs.

The MIPS_{im} package(126) allows the automatic analysis of the similarity between molecular interaction potentials distributions in series of biomolecules. MIPS_{im} can evaluate MIPs by classical or quantum methods, by interfacing with GRID and GAMESS programs, respectively. MIPS_{im} can perform an automatic search of the relative position of series of compounds that maximize their similarity. This can be used to stablish hyphotesis about their relative orientation at the receptor site, which is sometimes non-evident when only taking into account structural features. The new version of MIPS_{im} presented here, incorporates several definitions of similarity coefficients and allows the combination of different similarity measures into a single one. In addition, new tools for automatic exploration of conformational flexibility during the similarity search have been implemented.

The new features of the program are tested on a series of HIV-1 reverse transcriptase inhibitors.



COMPARISON OF BIOMOLECULES ON THE BASIS OF MOLECULAR INTERACTION POTENTIALS. THE NEW VERSION OF THE MIPSIM PACKAGE.

Montserrat Barbany, Jordi Villà, Hugo Gutiérrez-de-Terán, Ferran Sanz
Computational Structural Biology Laboratory
Research Group on Biomedical Informatics (GRIB)-IMIM/UPF
Pg. Marítim, 37-49, 08003 Barcelona (Spain)



INTRODUCTION

The current version of the MIPSim package (1):

1) Allows the computation of Molecular Interaction Potentials (MIPs) distributions by classical or quantum methods, by interfacing with GRID(2) and GAMESS(3) programs respectively.

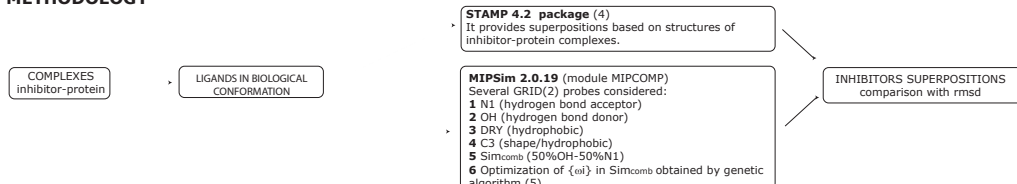
2) Performs an automatic search of the relative position of series of compounds that maximizes the similarity of their MIPs.

3) Allows the combination of several similarity measures into a single one: $Sim_{comb} = \frac{\sum_{i=1}^k \omega_i Sim_i}{\sum_{i=1}^k \omega_i}$

MIPSim is a useful tool for the comparison of series of compounds displaying related biological behaviours. It provides hypothesis about the sometimes non evident orientations at the receptor site.

We present here several tests using the well-known series of HIV-1 reverse transcriptase inhibitors.

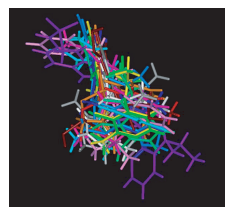
METHODOLOGY



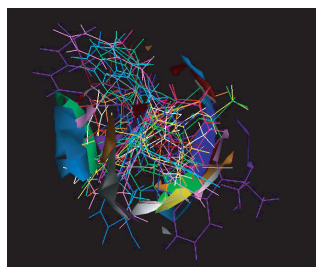
RESULTS

Inhibitors (PDB entry)	RMSD					
	1	2	3	4	5	6
1BQM (*)	0.00	0.00	0.00	0.00	0.00	0.00
1DTQ	5.73	5.38	4.51	5.01	4.99	5.42
1DTT	5.36	3.67	4.23	5.35	5.30	4.77
1EET	5.70	5.69	5.72	5.75	5.74	5.77
1EP4	2.15	2.21	1.53	1.67	2.17	1.63
1IKY	6.34	6.18	6.12	6.27	6.26	1.70
1ILQ	4.90	4.71	5.94	4.07	6.03	4.06
1KLM	4.19	4.49	7.08	1.70	4.17	5.27
1RT1	5.61	2.39	3.02	3.36	2.95	3.35
1RT4	4.09	4.54	4.90	4.44	4.59	4.55
1RT7	3.73	3.14	4.99	5.05	3.90	5.06
1COT	6.03	4.85	4.13	4.01	4.50	4.23
1COU	4.61	4.48	4.84	4.19	4.61	4.04
1RT2	7.08	2.95	6.84	2.61	2.73	2.36
Averages	5.07	4.20	4.69	4.11	4.39	4.09

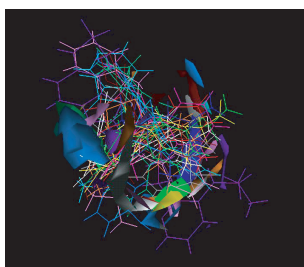
*1BQM is taken as reference



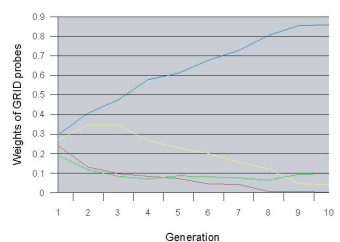
Inhibitors superposition generated by STAMP



STAMP superposition displaying the C3 MIP contour (-2 kcal/mol).



MIPSIM superposition displaying the C3 MIP contour (-2 kcal/mol).



Evolution of GRID probes throughout the generations of GA

CONCLUSIONS

MIPSim generates similar superpositions (see rmsd values in the above table) to those obtained by STAMP using inhibitor-protein complexes.

In the present series, the best solution is obtained with a combination of 85.9% C3, 9.7% DRY, 4% OH and 0.4% N1. This combination is consistent with the importance of the shape/hydrophobic interactions in the HIV-1 reverse transcriptase inhibitors.

REFERENCES

- (1) MIPSIM: <http://www1.imim.es/modeling/mipsim/index.html> de Cáceres, M.; Villà, J.; Lozano, J.J.; Sanz, F. *Bioinformatics* 2000
- (2) GRID: <http://www.moldiscovery.com> Goodford, P. *J. Med. Chem.*, 28, 849-857, 1985
- (3) GAMESS: <http://www.msg.ameslab.gov/GAMESS/GAMESS.html>
- (4) STAMP: <http://www.compbio.dundee.ac.uk/manuals/stamp.4.2/stamp.html> Russell, R.B.; Barton, G.J. *Proteins: Struct. Funct. Genet.*, 14, 309-323, 1992
- (5) Carroll, D.L. *Fortran Genetic Algorithm* <http://cucaerspace.com/carroll/ga.html>

ACKNOWLEDGEMENTS

We are grateful to the Fondo de Investigaciones Sanitarias for the financial support of this research.

M. Barbany, H. Gutiérrez-de-Terán, F. Sanz and J. Villà-Freixa

Similarity between transition state analogs and the corresponding transition states on the basis of Molecular Interaction Potentials and Langevin Dipoles.
Theoretical biophysics symposium, Donostia (Spain) (2003)

It is implicitly assumed that a proper transition state analog (TSA) can elicit a catalytic antibody (CA) with optimal binding to specific haptens. In most cases it was found that these CAs produced by TSA, present low rate enhancement as compared to the corresponding enzymes. This poster illustrates the origin of this problem, applying two methodologies that examine the similarity of the TSA and the corresponding transition state (TS).

Both approaches focus on electrostatic effects, that have been found to play a major role in enzymatic reactions(291). The first method makes use of molecular interaction potentials for computing the similarity between the TSA and the TS using `MIPSim` package(126). The second generates a grid of Langevin dipoles(21), which are polarized by the TSA, and then uses this grid to bind the TS. The comparison of the resulting binding energy with the binding energy of the TS in the enzyme environment, provides an estimate of the proficiency of the given CA.

These methods have been used to examine the origin for the difference between the catalytic power of the 1F7 CA and the enzyme chorismate mutase.



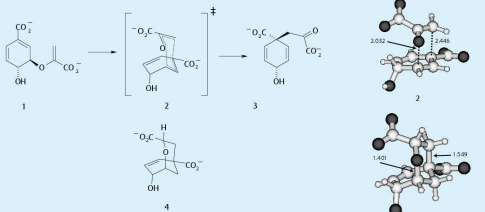
SIMILARITY BETWEEN TRANSITION STATE ANALOGS AND THE CORRESPONDING TRANSITION STATES ON THE BASIS OF MOLECULAR INTERACTION POTENTIALS AND LANGEVIN DIPOLES

Montserrat Barbany, Hugo Gutiérrez-de-Terán, Ferran Sanz and Jordi Vilà
Computational Structural Biology Laboratory
Research Group on Biomedical Informatics (GRIB) - IMIM/UPF



INTRODUCTION

It is implicitly assumed that a proper transition state analog (TSA) can elicit a catalytic antibody (CA) with optimal binding to specific haptens. In most cases it was found that these CAs produced by TSA, present low rate enhancement as compared to the corresponding enzymes. This poster illustrates the origin of this problem, applying two methodologies that examine the similarity of the TSA and the corresponding transition state (TS). Both approaches focus on electrostatic effects, that have been found to play a major role in enzymatic reactions [1]. These methods have been used to examine the origin of the difference between the catalytic power of the 1F7 CA and the enzyme chorismate mutase.



Claisen rearrangement of chorismate 1 to prephenate 3, showing the transition state (TS) for the reaction 2 and the transition state analog (TSA) 4 chosen for this study.

Relevant geometric parameters obtained at the HF/6-31+G* level for the TS and the TSA.

LANGEVIN DIPOLES

The second approach evaluates semiquantitatively the difference between proficiency of the CA and the corresponding enzyme. We generated relaxed LD grids [5] for the TS and TSA and then used the potential generated by each grid to evaluate the solvation energy of the TS and TSA.

charge distribution	TSA	TS	TS
structure			
solvent polarization	TSA	TSA	TS
ΔG_{sol}	0.0	-0.8	-4.4

Charge distribution	Solvent polarization	Structure	ΔG_{sol}
TSA	TSA	TSA	0.0
TS	TSA	TSA	-0.8
TS	TS	TS	-4.4

As seen from the table, TS is more stabilized by the enzyme than TSA. Computing the difference between both values we obtained a value $\Delta G_{TS}^{\dagger} - \Delta G_{TSA}^{\dagger} = 3.6 \text{ kcal mol}^{-1}$. This value is in qualitative agreement with the observed difference in proficiency (6 kcal mol^{-1}).

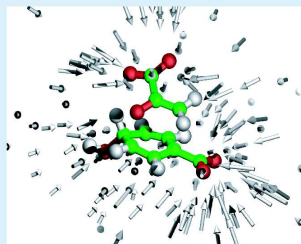
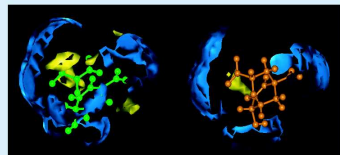


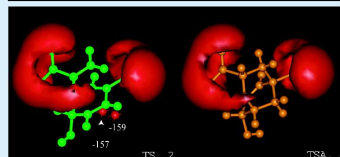
Figure representing the grid of Langevin dipoles created by solvent distribution of TS oriented towards the TS structure.

MIPSim

The first method makes use of molecular interaction potentials (MIP) to describe both TS and TSA. This property was computed with the MIPSim module of MIPSim package [2]. This software uses GRID [3] for classical MIP and GAMESS [4] for quantum molecular electrostatic potential (MESP) calculations and finds local minima performing a conjugated gradient optimization.



Classical molecular fields obtained with GRID probes OH (blue) and DRY (yellow) for TS and TSA. The isocountours are shown for energies of -5 kcal mol^{-1} for the OH probe and $-0.2 \text{ kcal mol}^{-1}$ for the DRY probe.



MESP of TS and TSA computed at HF/6-31+G* level of theory. Isocontour at -180 kcal/mol is shown. Isocontour at 180 kcal/mol is shown. Some minima (in kcal mol^{-1}) that appear in the MESP of TS and not in TSA are also depicted in the figure.

General electrostatic trends of the TS and TSA are apparently very similar.

In order to better describe the differences we aligned both molecules using module MIPComp of MIPSim package. Similarity index is computed by a Gaussian coefficient of the form:

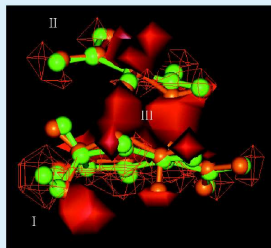
$$s_{ij} = \frac{\sum_{\alpha} \sum_{\beta} V_{\alpha}^i V_{\beta}^j \exp(-\alpha r_{ij}^2)}{\sqrt{\sum_{\alpha} \sum_{\beta} V_{\alpha}^i V_{\beta}^i \exp(-\alpha r_{ij}^2)} \sqrt{\sum_{\alpha} \sum_{\beta} V_{\alpha}^j V_{\beta}^j \exp(-\alpha r_{ij}^2)}}$$

where V_{α}^X is the α th potential value for molecule X and r_{ij} is the distance between the two points.

The parameter α is taken as 0.5. Superposition of the quantum mechanical MESP grid for TS and TSA. Only the very negative values of the MESP ($< -200 \text{ kcal mol}^{-1}$) where selected for the similarity calculations in order to superpose the MESP around the carboxylates.

Isocontours in the figure represent the differences between the potential fields for the actual TS and the TSA. Solid surfaces are regions at $-100 \text{ kcal mol}^{-1}$ and wire frame surfaces are regions at $100 \text{ kcal mol}^{-1}$.

Region III corresponds to differences of electron density in the bond-forming pattern in the TS respect to TSA.



CONCLUSIONS

The present study [6] illustrates the difficulties in the use of TSAs to elicit CAs. Nevertheless, the search for improved TSAs can be helped by the present approaches. It is possible to screen the feasible synthetic candidates of already existing molecules with the GRID and LD approaches and assess which candidate will elicit the best complementary environment.

REFERENCES

- [1] Warshel, A. Electrostatic origin of the catalytic power of enzymes and the role of preorganized active sites. *J. Biol. Chem.* 1998; 273, 27035-27038; b) Vilà, J, Warshel, A. Energetics and Dynamics of Enzymatic Reactions. *J. Phys. Chem. B* 2001, 105, 7987-7997.
- [2] De Cáceres M, Vilà J, Lozano JJ, Sanz F. MIPSIM: Similarity analysis of molecular interaction potentials. *Bioinformatics* 2003; 16: 568-569.
- [3] Goodford P. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* 1985, 28, 849-857.
- [4] Schmidt, M et al. General Atomic and Molecular Electronic Structure System. *J. Comput. Chem.* 1993, 14, 1347-1363.
- [5] Warshel A. *Computer modeling of Chemical Reactions in Enzymes and Solutions*, John Wiley & Sons, New York, 1991
- [6] Barbany M, Gutiérrez-de-Terán H, Sanz F, Vilà-Freixa J and Warshel A. On the generation of catalytic antibodies by transition state analogs. *ChemBioChem*, in press.

ACKNOWLEDGEMENTS

We are grateful to the Fondo de Investigaciones Sanitarias for financial support of this research (FS 01/1330) and to the Centre de Computació i Comunicacions de Catalunya (C3) for computer time allocation.

M. Barbany, J. Villà-Freixa, F. Sanz and S. Boyer

Using MIPSim and 3D-QSAR to study binding modes of HERG K+ Channel Inhibitors.
EuroQSAR 2004, Istanbul (Turkey) (2004)

The HERG potassium channel (human ether-a-go-go-related gene) is expressed in the human heart. The channel is a key effector of cardiac repolarization and contributes to the QT interval measured by the electrocardiogram. Inhibition of HERG can lead to a prolongation of the QT interval, widely considered a critical risk factor for arrhythmia. A chemically diverse series of drugs have been withdrawn from the market due to their HERG blocking properties. Thus, HERG inhibitory effects represent an important safety consideration in drug discovery.

MIPS_{im}(126) is a program that analyzes and compares molecular interaction potential (MIP) distributions for a series of biomolecules. MIPS_{im} performs structural alignments for a series of biomolecules using their MIPs. This program has been used with success in both the alignment of pharmacologically relevant series of molecules(274) and even in studies of the role of electrostatics in catalytic antibodies action(143). Recently we have developed a methodology(275) that tries to find the structural alignment of series of biomolecules in their biological conformation using protein-ligand structural information. MIPS_{im} was capable to find biological alignments with good experimental agreement in a series of non nucleoside HIV-1 retrotranscriptase inhibitors.

In the present study we obtain structural alignments for a series of drugs(292) that are known to inhibit the HERG potassium channel with a rank of activities. To validate our MIP based alignments, we use them in a 3D-QSAR study using the GRID/GOLPE protocol.

Using MIPSim and 3D-QSAR to study binding modes of HERG K⁺ Channel Inhibitors

MONTserrat BARBANY¹JORDI VILLÀ-FREIXA^{1,*}FERRAN SANZ¹SCOTT BOYER^{2,3}

¹ Research Unit on Biomedical Informatics (GRIB), Institut Municipal d'Investigació Mèdica (IMIM), Universitat Pompeu Fabra (UPF), Barcelona, Spain
² Safety Assessment Group at AstraZeneca, Mönklad, Sweden



IMIM
Institut Municipal
d'Investigació Mèdica



Summary

MIPSim(1) is a program that analyzes and compares molecular interaction potential (MIP) distributions for a series of biomolecules. MIPSim performs structural alignments using their MIPs.

The HERG potassium channel (human ether-a-go-go-related gene) is expressed in the human heart. The channel is a key effector of cardiac repolarization and contributes to the QT interval measured by the electrocardiogram. Inhibition of HERG can lead to a prolongation of the QT interval, widely considered a critical risk factor for arrhythmia. Thus, HERG inhibition represents an important safety consideration in drug discovery.

In the present study we obtain structural alignments for a series of drugs(5) that are known to inhibit the HERG potassium channel with different degrees of activity. To validate our MIP based alignments, we use them in a 3D-QSAR study using the GRID/GOLPE protocol.

The alignment problem

One of the most important problems in structure-activity studies is to obtain good alignments of molecules that share the same binding site in order to obtain a good 3D-QSAR model.

There are several ways to perform alignments: using pharmacophoric points, chemical features, shape and Molecular Interaction Potentials (MIPs).



Alignment using type of atoms.

Alignment using MIPs.

MIPSim(1) is a program that analyzes and compares MIP distributions for a series of biomolecules. MIPSim also performs molecular alignments on the basis of MIP distributions. This program has been used with success in several studies(2,3,4) showing that the MIPs are good features for performing superpositions.

The biological system

A chemically diverse series of drugs has been withdrawn from the market due to their HERG blocking properties. Thus, HERG inhibitory effects represent an important safety consideration in drug discovery. In the present study we try to obtain structural alignments for a series of drugs(5) that are known to inhibit the HERG potassium channel with different degrees of activity.

Compound	HERG Blocker	Activity	IC50 (nM)	IC90 (nM)	IC95 (nM)
1	Yes	High	100	1000	10000
2	Yes	Medium	1000	10000	100000
3	Yes	Low	10000	100000	1000000
4	No	None	>100000	>1000000	>10000000
5	No	None	>1000000	>10000000	>100000000

HERG K⁺ Channel Blocking Activity of Compounds.

Methodology

1. Generation of conformations using OMEGA.

We created a maximum of 20 conformations for every compound using a minimum RMSD of 0.8 Å and not including any structure with an energy of more than 5 kcal/mol above the current global minimum.

2. Structural alignment using pharmacophoric points.

We superimposed the conformations of every compound with respect to the crystal structure of Astemizole, retrieved from the Cambridge Structure Database using our program SUPERB (based on Chem7) algorithm to superpose two different geometries based on a group of atoms. We choosed the superpositions having lower RMSD vs. Astemizole.

3. MIP based alignment using MIPSim(1).

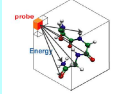
Using the similarity index as scoring function, MIPSim can perform an automatic search of the best relative orientation between two molecules. At every orientation of the molecule and for every MIP k , the similarity index $s_i^{k,j}$ is calculated by a Gaussian coefficient (1):

$$s_i^{k,j} = \frac{\sum_{\xi \in \alpha} \sum_{\beta \in \beta} V_{\xi}^k V_{\beta}^j \exp(-\alpha r_{\xi\beta}^2)}{\sqrt{\sum_{\xi \in \alpha} \sum_{\beta \in \beta} V_{\xi}^k V_{\beta}^j \exp(-\alpha r_{\xi\beta}^2)} \sqrt{\sum_{\xi \in \alpha} \sum_{\beta \in \beta} V_{\xi}^k V_{\beta}^j \exp(-\alpha r_{\xi\beta}^2)}} \quad (1)$$

where n_{ξ} ($\xi \in \alpha, \beta$) is the number of points in each grid box selected for the comparison, V_{ξ}^k is the potential value in the grid point ξ for molecule k and $r_{\xi\beta}$ is the distance between two points. The smoothing parameter α is set to 0.5, value that has been found to work well in previous studies(2,3,4). Following this procedure we can evaluate a different similarity index for every GRID(m) probe we are interested in and obtain the global similarity:

$$S_{\alpha,\beta} = \frac{\sum_{i=1}^m w_i^{k,j} s_i^{k,j}}{\sum_{i=1}^m w_i^{k,j}} \quad (2)$$

where $w_i^{k,j}$ are the weights of every particular similarity index $s_i^{k,j}$ and m is the number of GRID probes.



GRID interaction energies for a certain probe.

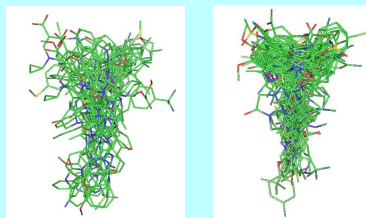
In this study we choosed $w_i^{k,j}$ of 0.33 for three GRID probes: O (hydrogen bond donor), N1 (hydrogen bond acceptor), DRY (hydrophobic).

4. GRID/GOLPE studies.

Data pretreatment included zeroing of the values between -0.1 and 0.1, exclusion of variables with stand deviance less than 0.1 and exclusion of N-level variables. We performed a Principal Component Analysis (PCA) and Partial Least Squares (PLS modelling).

Results

- Structural alignment using the same pharmacophoric points of Cavalli et al.(5) GRID/GOLPE model using the structural superposition gives an r^2 of 0.92 choosing two components and a maximum q^2 (leave-one-out validation) of 0.22.



Best superposition of HERG inhibitors based on pharmacophoric points.

Best superposition of HERG inhibitors based on MIPs computed by MIPSim.

- Structural alignment using MIPSim alignment. Using the conformers selected in the last calculation and performing an exploration of the best superposition using Astemizole as the static molecule, we obtained new positions for every compound. GRID/GOLPE model using the structural superposition gives an r^2 of 0.95 choosing two components and a maximum q^2 (leave-one-out validation) of 0.48.

References

- De Cáceres M, Villà J, Llanos, J.L., Sanz F. *Bioinformatics* 16, 568-569 (2000). MIPSim is available at <http://ollama.imim.es/software/mipsim/index.html>
- Rodrigo J, Barbany M, Gutiérrez-de-Terán H, Centeno N.R., de Cáceres M, Deol C, Fontaine F, Lozano J.J., Pastor M, Villà-Freixa J, Sanz F. *Drug Discov. Chem. Sci.* 13, 795-799 (2012)
- Barbany M, Gutiérrez-de-Terán H, Sanz F, Villà-Freixa J, Warshel A. *ChemBioChem* 4, 277-285 (2003)
- Barbany M, Gutiérrez-de-Terán H, Sanz F, Villà-Freixa J. *PROTEINS: Struct., Funct. and Bioinformatics* 56:S55-S54 (2004)
- Cavalli A, Poluzzi F., De Paulis F., Reanami, M.J. *Med. Chem.* 48,384-3853 (2002)
- Goodford P.J. *Med. Chem.* 28, 849-857 (1985)
- Chen Z. *Theoretica Chimica Acta* 75(6), 481-484 (1989)
- Ingólfsson C., Schramm V.L., Schwartz S.D. *J. Am. Chem. Soc.* 118, 8825-8836 (1996)

Conclusions

- A good alignment of biomolecules is a crucial step in 3D-QSAR studies in order to obtain a good predictable model.
- In this sense, the execution of detailed and objective comparative analysis of MIP distributions is useful to obtain good alignments.
- Limited results in predictability are due to a not optimal generation of the set of conformers, but we can observe the improvement in PLS study when using a MIPSim alignment.

Next future

- To perform all pairwise comparisons of all the conformations for every compound in order to find the active conformation of HERG inhibitors. We assume that the active conformations will be those who have the best similarity of MIPs between the most active compounds.
- Once we know the active conformation for every compound we can perform a 3D-QSAR modeling. It could detect side effects as early as possible during drug development.

Acknowledgments

We are grateful to the Fondo de Investigaciones Sanitarias for financial support of this research (FIS 01/1330) and to Generalitat de Catalunya for the travel fellowship BE. We want to thank Fabien Fontaine for his help in GRID/GOLPE calculations.

*To whom correspondence should be addressed:

- Research Group in Computational Biochemistry and Biophysics (FIS 01/1330) and to Generalitat de Catalunya for the travel fellowship BE. We want to thank Fabien Fontaine for his help in GRID/GOLPE calculations.
 Research Group on Biomedical Informatics (GRIB) - IMIM/UPF
 Doctor Aiguader, 86, 08035 Barcelona (Spain)
 Tel.: +34 93 221 1909 ext. 2919 f. Fax: +34 93 221 3237
 e-mail: jvillafreixa@imim.es fabien.fontaine@imim.es
- Safety Assessment, AstraZeneca Research and Development, Pepparedsleden 1, 431 83 Mölndal (Sweden)
 Tel.: +46 31 776 2823 f. Fax: +46 31 776 3792
 e-mail: Scott.Boyer@astrazeneca.com

B.2 Oral contributions

Estudi de la similaritat entre estats de transició anàlegs i els corresponents estats de transició basat en potencials d'interacció molecular i dipols de Langevin
XIX Reunió de la Xarxa de Química Teòrica, Universitat de Girona (Spain) (July 2003)

Annexes

C.1 MIPSim installation procedure

Current version of MIPSim use GAMESS as a function evaluator for quantum MEP. Similarly MIPSim uses GRID as a function evaluator for classical MIP. Users have to acquire a copy of both programs. MIPSim is distributed as executable files and GAMESS and GRID are used as external programs.

MIPSim is distributed as a compressed file, called MIPSIM_‘version’_‘uname’.tar.gz ‘version’ is the version of the program and the ‘uname’ stands for the operating system being used. This file must be uncompressed using:

```
gzip -d MIPSIM_2.4_‘uname’.tar.gz
```

and untarred by using:

```
tar xvf MIPSIM_2.4_‘uname’.tar
```

Once the file has been uncompressed and untarred, a tree of directories will be created with the following structure:

```
MIPSIM/bin  
MIPSIM/testfiles  
MIPSIM/doc
```

These directories include:

- Test-run files and the needed utilities for running and checking them:

```
MIPSIM/testfiles/shrunall  
MIPSIM/testfiles/shchecktests  
MIPSIM/testfiles/shcleanalldemo  
MIPSIM/testfiles/shclean  
MIPSIM/testfiles/shsavetests
```

- Execution scripts:

```
MIPSIM/bin/mipsim_rc  
MIPSIM/bin/mipsim  
MIPSIM/bin/shmipsim  
MIPSIM/bin/shgame
```

- MIPS*im* executables :

```
MIPSIM/bin/mipmin.2.4.x
MIPSIM/bin/mipcomp.2.4.x
```

- A series of README files with further useful information.

C.2 Setting up MIPS*im*

The first step in setting up the program is to add the path for the binary files in your .login file. In csh or tcsh this is done by adding the lines:

```
setenv MIPSIM_PATH "<where_mipsim_has_been_installed>"
setenv PATH "$MIPSIM_PATH/bin:$PATH"
```

If the sh or ksh is used, the following lines are needed

```
MIPSIM_PATH="<where_mipsim_has_been_installed>"
export $MIPSIM_PATH
PATH=$MIPSIM_PATH/bin:$PATH
export $PATH
```

In order for the program to run it must know the location of the external software. This information is read in the form of environmental variables. In particular, MIPS*im* will look for the file

```
$MIPSIM_PATH/bin/mipsim_rc
```

that will contain the definition of these environmental variables. This file is provided in the distribution and may require modifications by the system administrator in order to properly run MIPS*im*. Next, the program will source the file called (if it exists):

```
$HOME/.mipsim_rc
```

Finally, MIPS*im* will source the:

```
./mipsim_rc file
```

This is, in the current working directory, in case this file has been created. Note that the parent mipsim_rc file does not start with ".", and the opposite occurs with the user-specified .mipsim_rc files.

```
$MIPSIM_PATH/bin/mipsim_rc
```

A complete list of environmental variables is in the file distributed with the program. A typical mipsim_rc is given below:


```
# directory for output of \mipsim(OUTDIR).
setenv MIPSIM_OUTPUT .
# directory for \gamess\ temporary files
setenv GAMESS_TMP_DIR .
# \gamess\ complete path
setenv GMSDIR /usr/local/modelling/mipsim_tests/gamess
setenv GMSVERNO 01 #executable is gamess.$VERNO.x
# directory for scratch information for program \mipsim.
setenv MIPSIM_SCRATCHDIR /tmp/$USER
#GRID, GRIN and GRUB complete path
setenv GRDDIR /usr/local/modelling/grid21
setenv GRDCOMMAND $GRDDIR/grid
setenv GRNCOMMAND $GRDDIR/grin
setenv GRUCOMMAND $GRDDIR/grub.dat
```

Requirements to run MIPSim calculations:

- Directory MIPSIM_PATH/bin must be in the user's PATH environment variable.
- There must be a KEYFILE (*.key) in the current directory. This is the input file for MIPSim, where the user must specify, among other information, which MODULE wants to run. A simple example of KEYFILE is given below.
- The coordinate files listed in the KEYFILE.

Suppose we have written a KEYFILE named 'example.key' as this:

```
module=min property=gms_mep
Title of the work
molecule1.pdb
molecule2.xyz
```

We then will just type:

```
$ mipsim [-ds] example
```

This will launch MIPSim. Initially, MIPSim creates a new directory called, in this case, 'example' where it will copy all the relevant files and run the calculations. This directory will be called OUTDIR in this document. At the end of the run, the main output files are located in OUTDIR, while the scratch files will be located in the directories specified in the mipsim_rc configuration files.

MIPSim includes some utilities to transform output data to other molecular modeling and QSAR packages. The key files are the PTS files (.pts) and the SUMMARY file (.sum). Currently MIPSim supports format conversions to:

- Insight II: MIPSim writes ASCII '.grd' potential files.
- gOpenMol: MIPSim writes binary '.plt' potential files.

On the other hand, one can obtain potential files converted to Insight II, gOpenMol or GOLPE formats directly from MIPSim runnings.

C.3 MIPSim keyfile

Syntax in KEYFILE must be as follows:

```
1 ROUTE BLOCK: One or more lines for the keywords
-- blank line --
2 TITLE BLOCK: Title of the job.
-- blank line --
3 LIST BLOCK: list of molecules to be computed (one per line)
-- blank line --
4 SPECIFIC BLOCK: list of specific keywords per molecule (optional, one per line).\
It must start with a number identifying the molecule with specific keywords.
```

It is mandatory to specify the keyword `MODULE` in the ROUTE BLOCK. The last block in the input file is used to specify keywords for each particular molecule. These lines are optional and the effect of using them is overwriting the information in the main block of keywords. Finally, an example of input file could be:

```
module=comp property=(gms_mep,grd_oh)
out_link3d
cmp_allvsfirst
cmp_interval=(-10,10)
comparison of two water molecules and a OH- with \gamess.
wat1.pdb
wat2.pdb
oh.pdb
1 gms_pcm
3 gms_icharg=-1
```

C.4 Profiling of MIPSim

Profiler used in MIPSim is PGI. Profiling is a three step process:

- **Compilation.** Compiler switches cause special profiling calls to be inserted in the code and data collection libraries to be linked in.
- **Execution.** The profiled program is invoked normally, but collects call counts and timing data during execution. When the program terminates, a profile data file is generated (pgprof.out).
- **Analysis.** The PGPROF tool interprets the pgprof.out file to display the profile data and associated source files. The PGPROF profiler is invoked as follows:

```
pgprof[options] [-I srcdir] [datafile]
```

The following list shoes driver switches that cause profile data collection calls to be inserted and libraries to be linked in the executable file:

- `-Mprof=func` Insert calls to produce a pgprof.out file for function level data.
- `-Mprof=lines` Insert calls to produce a pgprof.out file which contains both function and line level data.

C.5 Combination of different similarity coefficients

We can see in an example .key file of MIPS_{im} how to use these keywords:

```
cmp_weight=(0.3,grd_oh)
cmp_weight=(0.2,grd_n2)
```

In the last example, it is chosen automatically for the third probe the weight value of 0.5 in order to obtain a sum of weights of value 1.

C.6 Comparison of intervals of energy

MIPS_{im} now enables the comparison of different intervals of energy for every probe in the comparison. The keywords used in a .key file of MIPS_{im} are:

```
cmp_interval=(-4,-1,grd_oh)
cmp_interval=(-3,-1,grd_n1)
cmp_interval=(-6,-2,grd_n2)
```

C.7 Tors module input

In TORS module rotatable bonds and the angle of rotation can be selected in input file .key:

```
module=tors
property\_\_ptc\_mep

TST tors.

1 trs\_conform trs\_conform\_at= (1-2,3-6) trs\_conform\_angl=(45-30)
```

where 1 is the molecule one want to rotate (in this case the number 1), trs_conform enable or not to perform rotation, trs_conform_angl tell us the angle of rotation (in degrees) and trs_conform_at are the limits of the rotatable bond.

C.8 Technical aspects of CHEMSOL

CHEMSOL (see section 3.2.4) is a program designed for calculations of solvation free energies using Langevin Dipoles. It contains the files:

- cs, simple script to run the .ps file
- cs21.f, source file
- cs21_manual.ps, a manual in postscript.
- cs.arc, the archive file
- test1.cs, input file with one sets of charges and default vdW radii (protonated cytosine)

- test2.cs, input file with two sets of charges and user-defined vdW radii (methanol)
- vdw.par, standar parameter file

- Atomic charges

Atomic charges are obtained by fitting to the electrostatic potential of the solute calculated from the PCM B3-LYP/6-31*//HF/6-31G* wavefunction. A possible command line for the calculation of atomic charges using the GAUSSIAN 94 program is

```
B3LYP 6-31G* scrf=tomasi iop(1/11=200) population=(mk,dipole)
...
80. 400
```

and for GAUSSIAN 98:

```
B3LYP 6-31G* scrf=oldpcm iop(1/11=200) population=(mk,dipole)
....
80. 400
```

- The parameter input file (vdw.par)

The file contains the values of the selected parameters of the LD model. The first number on the left indicate if we want NLD or ILD. 1 for NLD and 0 for ILD. Another parameters are Van der Waals radii for selected atoms and London coefficients.

- The CHEMSOL input file (example.cs)

The solute structure (Cartesian coordinates) and charge distribution (atomic point charges) need to be specified on input. An example input file for methanol:

```
title
- number of atoms in the solute molecule, number of different structures
- empty line
- title for the substructure (only if there are more than one different
  structure, otherwise empty line)
- empty line
- atom name, nuclear charge, atomic charge (gas phase), cartesian coordinates
  (x,y,z) (format (1x,a8,f8.1, 2f10.4,3f9.4))
- empty line
- keyword (pcm) It is used if corrections for solute polarization are
  to be evaluated explicetely.
- atom name, nuclear charge, atomic charge (pcm), cartesian coordinates
  (x,y,z) (format (1x,a8,f8.1, 2f10.4,3f9.4))
- empty line
```

- Output data

A single output line is appended in the end of the output file called cs.arc.

- *l_{gvn}* Electrostatic part of the solvation free energy in kcal/mol.
- *vdW* London dispersion part of the free energy of solvation (kcal/mol).

- $-TdS$ Hydration entropy (kcal/mol).
- *Relax* Contribution of solute polarization to the solvation free energy.
- *Born* Continuum correction for a finite size of the sphere of point dipoles.
- dH_{solv} Hydration enthalpy.
- dG_{solv} Hydration free energy.

The total solvation free energy:

$$\Delta G = \Delta H_{solv} - T\Delta S = lgvn + vdW + Born - T\Delta S \quad (C.1)$$

All the energies are expressed in kcal/mol

C.9 PDLDSMALL program

This program performs simple Langevin Dipoles (LD) calculations and provides a simple estimate of solvation free energies.

In this program, the subroutine *solvate* evaluates solvation energies using the fixed centered langevin dipoles method, builds the grid of dipoles, places langevin dipoles at grid points, stores the langevin dipoles, calculates dipole-field interaction and the energy spent in polarizing the solvent.

C.10 POLSAR program

Adaptation of program *CHEMSOL* designed as a scoring function for screening the ability of TSA to mimic TSs in order to elicit catalytic antibodies. Parameters are calibrated with a test set of neutral and ionic solutes in aqueous solution.(255). Developed by H.Gutiérrez-de-Terán.

Implement simple *CHEMSOL* for fast evaluation of activities in a series of ligands. The idea is to use the same trick we have done for the CA paper as a tool for fast evaluation of differences from a given target. The study should be complemented with vdW interactions and/or with the definition of new vdW "vectorial field" that may be added to the Langevin term. If this works it can be an extremely interesting improvement, providing a very efficient tool to discriminate between initial molecules by means of energy criteria.

ΔG solvation:

$$\Delta G_{solv} = \Delta G_{ES} + \Delta G_{BULK} + \Delta G_{vdW} + \Delta G_{vdW_{phob}} \quad (C.2)$$

ΔG_{ES} is non-iterative langevin dipoles (NILD).

ΔG_{BULK} is Born-Onsager's approximation due to outer continuum dielectric.

$\Delta G_{vdW-phob}$ Contribution due to vdW and hydrophobic interactions.

Charge distribution computed by Gaussian 94 from the PCM B3-LYP/6-31G**//HF/6-31G*

C.11 SUPERB input file

This is an example of SUPERB input file:

```
fixedmol.xyz
mobilmol.xyz
list of atoms in molec 1 to superpose
list of atoms in molec 2 to superpose
```

C.12 STAMP input

In order to superpose pdbs with STAMP we proceed as follows:

First, be sure you have the correct definition of environmental variables:

```
setenv STAMPDIR /usr/local/modelling/structure/stamp/stamp.4.3/defs
setenv PATH "$STAMPDIR/../bin/linux:$PATH"
```

Then, create the STAMP input file using PDBC note that first it is a good idea to locate in the PDB the chains we are really interested in. PDBC finds and reports information about PDB files given a chain identifier. Note too that the first structure is the one that is going to remain unrotated. This can be important in some cases.

```
pdbc -d lhcl >! cdk2.domains
pdbc -d lfina >> cdk2.domains
pdbc -d lfinc >> cdk2.domains
pdbc -d lhck >> cdk2.domains
pdbc -d lb38 >> cdk2.domains
```

Then one have to run STAMP:

```
stamp -l cdk2.domains -rough -n 2 -prefix cdk2
```

Then perform the tranformations to the original PDB files. Tranform outputs the corresponding set of coordinates given a list of transformations.

```
transform -f cdk2.4 -het -g -o superb.pdb
```

C.13 Future developments

Tasks to be developed/implemented in MIPSim:

- Modify algorithms to add Fast Fourier Transform (FFT)(144) methods in COMP module, cause time is critical in some of the routines.
- Implement an eigenvalue following routine for optimization in MIN in order to locate all stationary points (minima and higher order stationary points) in the electrostatic potential map of a given molecule.
- Implement similarity based on the electric field(124).

- Add multipole analysis and introduce implementations of multipole-based similarities.
- Fix de use of static memory in MIPS_{im} calculations. Now, the size of the program is still big due to the arrays depending on the number of field points.

Applications:

- Study of MAO inhibitors.
- Systematic study of catalytic antibodies.

C.14 This L^AT_EX thesis template

This thesis layout largely derives from the L^AT_EX_{2 ϵ} template created by Robert Castelo. However, it has been extensively modified and, maybe, improved. Here, I provide some comments on it and the source code for download. R. Castelo wrote his own thesis style file (`mythesis.sty`) to handle fonts and control section title layout. L^AT_EX_{2 ϵ} is a document preparation system, powerful, robust and able to achieve professional results. However, the learning curve may be stiff, thus, an initial template is given here for your convenience.

This file (`Makefile`), also derived from R. Castelo's equivalent, is read to produce a PDF version of your thesis just by typing `make pdf` in the command line. It is aware of any change you make in any of the child directories and L^AT_EX files that compose your document. It also reruns itself to update the references section. It needs the `make` program in your system, though it is usually installed by default.

This body section of the document (`mythesis.tex`) is simply a call of the independent chapter files (`*.tex`) your thesis consist of (each of them nicely placed in its own directory). It also calls, at the very beginning, the preamble file.

Bibliography

- [1] Joshi-Tope, G.; Gillespie, M.; D'Eustachio, P.; Schmidt, E.; de Bono, B.; Jassal, B.; Gopinath, G.; Wu, G.; Matthews, L.; Lewis, S.; Birney, E.; Stein, L. *Nucleic Acids Res.* **2005**, *33*, 428-432.
- [2] A. Buckingham, A. L.; Robert, S. *Principles of Molecular Recognition*; Blackie Academic Professional: London, 1993.
- [3] Xenarios, I.; Eisenberg, D. *Curr. Opin. Biotechnol* **2001**, *12*, 334-339.
- [4] Pham, H. *Phd: Support Vector Learning and Rule Induction for Knowledge Discovery in Biological Data*; Japan Advanced Institute of Science and Technology, 2005.
- [5] Alberts, B.; Bray, D.; Lewis, J.; Raff, M.; Roberts, K.; Watson, J. D. *Molecular Biology of the Cell. 3rd Ed.*; Garland Publishing: New York and London, 1994.
- [6] Mitchell, J. "Protein interaction notes", 2003.
- [7] McQuarrie, D. A. *Statistical Mechanics*; Harper and Row: New York, 1976.
- [8] Gibbs, G. *Trans. Connect. Acad.* **1875**, *3*, 108-248.
- [9] Schlick, T. Geometry optimization. In *Encyclopedia of Computational Chemistry*; Schleyer, P., Ed.; John Wiley and Sons: New York, 1998.
- [10] Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420.
- [11] Lee, F. S.; Chu, Z. T.; Bolger, M. B.; Warshel, A. *Prot. Eng.* **1992**, *5*, 215-228.
- [12] Pratt, L.; Hummer, G. *Simulation of Electrostatic Interactions in solution*; AIP: New York, 1991.
- [13] Åqvist, J.; Medina, C.; Samuelsson, J. *Prot. Eng.* **1994**, *7*, 385-391.
- [14] Truhlar, D. G.; Garrett, B. C.; Klippenstein, S. J. *J. Phys. Chem.* **1996**, *100*, 12771-12800.
- [15] Wigner, E. *Trans. Faraday Soc.* **1938**, *34*, 29.
- [16] Ariens, E. *Arch. Intern. Pharmacodyn.* **1954**, *99*, 32-49.
- [17] Stephenson, R. *Br. J. Pharmacol.* **1954**, *11*, 379-393.
- [18] Furchgott, R. *Ann. Rev. Pharmacol.* **1956**, *4*, 21-50.
- [19] Clark, A. *The mode of action of drugs on cells.*; E. Arnold and Co.: London, 1933.

- [20] Fischer, E. *Ber. Dtsch. Chem. Ges* **1894**, *27*, 2985-2993.
- [21] Warshel, A. *Computer Modeling of Chemical Reactions in Enzymes and Solutions*; John Wiley & Sons: New York, 1991.
- [22] Warshel, A. *J. Biol. Chem.* **1998**, *273*, 27035-27038.
- [23] Pauling, L. *Chem. & Eng. News* **1946**, *24*, 1375-1377.
- [24] Radzicka, A.; Wolfenden, R. *Methods Enzymol.* **1995**, *249*, 284-312.
- [25] Schramm, V. *Annu. Rev. Biochem.* **1998**, *67*, 693-720.
- [26] Mader, M.; Bartlett, P. *Chem. Rev.* **1997**, *97*, 1281-1301.
- [27] Cohen, S. G.; Vaidya, V. M.; Schultz, R. M. *Proc. Natl. Acad. Sci. USA* **1970**, *66*, 249-256.
- [28] Jencks, W. P. Binding Energy, Specificity, and Enzymic Catalysis: The Circe Effect. In *Advances in Enzymology and Related Areas of Molecular Biology*, Vol. 43; Meister, A., Ed.; J. Wiley & Sons, Inc.: New York, 1975.
- [29] Dewar, M.; Storch, D. *Proc. Natl. Acad. Sci. USA* **1985**, *82*, 2225-2229.
- [30] Crosby, J.; Stone, R.; Lienhard, G. E. *J. Am. Chem. Soc.* **1970**, *92*, 2891-2900.
- [31] Dewar, M. J. S.; Dieter, K. M. *Proc. Natl. Acad. Sci. USA* **1988**, *82*, 2225-2228.
- [32] Lee, J. K.; Houk, K. N. *Science* **1997**, *276*, 942-945.
- [33] Lightstone, F.; Zheng, Y.-J.; Maulitz, A.; Bruice, T. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 8417-8420.
- [34] Warshel, A. *Proc. Natl. Acad. Sci. USA* **1978**, *75*, 5250-5254.
- [35] Hinchliffe, A. *Molecular Modelling for Beginners*; Wiley: New York, 2003.
- [36] Gohlke, H.; Klebe, G. *J. Med. Chem.* **2002**, *45*, 4153-4170.
- [37] Halperin, I.; Ma, B.; H., W.; Nussinov, R. *Proteins: Struct. Func. Gen.* **2002**, *47*, 409-443.
- [38] Taylor, R.; P.J., J.; Essex, J. *J. Comput. Aid. Mol. Des.* **2002**, *16*, 151-166.
- [39] Glen, R.; Allen, S. *Curr. Med. Chem.* **2003**, *10*, 763-777.
- [40] Fradera, X.; Knegtel, R.; Mestres, J. *Proteins: Struct. Func. Gen.* **2000**, *40*, 623-636.
- [41] Hindle, S.; Rarey, M.; Buning, C.; Lengaue, T. *J. Comput. Aid. Mol. Des.* **2002**, *16*, 129-149.
- [42] D., M.; B.E., T.; Belmarsh, M.; Moustakas, D.; Alvarez, J. *Proteins: Struct. Func. Gen.* **2003**, *51*, 172-188.
- [43] Jain, A. *J. Med. Chem.* **2003**, *46*, 499-511.
- [44] Perutz, M. *Brookhaven Symp. Biol* **1960**, *13*, 165-183.

- [45] Kendrew, J.; Bodo, G.; Dintzis, H.; Parrish, R.; Wyckoff, H.; Phillips, D. *Nature* **1958**, *181*, 662-666.
- [46] Berman, H.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I.; Bourne, P. *Nucleic Acids Research* **2000**, *28*, 235-242.
- [47] Westbrook, J.; Feng, Z.; Chen, L.; Yang, H.; Berman, H. *Nucleic Acid Res.* **2003**, *31*, 489-491.
- [48] Kitchen, D.; Decornez, H.; Furr, J.; Bajorath, J. *Nature Reviews* **2004**, *3*, 935-949.
- [49] Kuntz, I.; Blaney, J.; Oatley, S.; Langridge, R.; Ferrin, T. *J. Mol. Biol.* **1982**, *161*, 269-288.
- [50] Gabb, H.; Jackson, R.; Sternberg, M. *J. Mol. Biol.* **1997**, *272*, 106-120.
- [51] Schnecke, V.; Swanson, C.; Getzoff, E.; Tainer, J.; Kuhn, L.
- [52] Meng, E.; Kuntz, I.; Abraham, D.; Kellogg, G. *J. Comput. Aid. Mol. Des.* **1993**, *8*, 299-306.
- [53] Gschwend, J. *Comput. Aid. Mol. Des.* **1996**, *10*, 123-132.
- [54] Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. *J. Mol. Biol.* **1996**, *261*, 470-489.
- [55] Welch, W.; Ruppert, J.; Jain, A. *ChemBiol* **1996**, *3*, 449-462.
- [56] Jones, G.; Willett, P.; Glen, R.; Leach, A.; Taylor, R. *J. Mol. Biol.* **1997**, *267*, 727-748.
- [57] Ewing, T.; Kuntz, I. *J. Comput. Chem* **1997**, *18*, 1175-1189.
- [58] Morris, G.; Goodsell, D.; Halliday, R.; Huey, R.; Hart, W.; Belew, R.; Olson, A. *J. Comput. Chem* **1998**, *19*, 1639-1662.
- [59] Makino, S.; Kuntz, I. *J. Comput. Chem.* **1998**, *19*, 1839-1852.
- [60] Baxter, C.; Murray, C.; Clark, D.; Westhead, D.; Eldridge, M. *Proteins: Struct. Func. Gen.* **1998**, *33*, 367-382.
- [61] Wang, J.; Kollman, P.; Kuntz, I. *Proteins: Struct. Func. Gen.* **1999**, *36*, 1-19.
- [62] Leach, A. R. *J. Mol. Biol.* **1994**, *235*, 345-356.
- [63] Knegtel, R.; Kuntz, I.; Oshiro, C. *J. Mol. Biol.* **1997**, *266*, 424-440.
- [64] Apostolakis, J.; Pluckthun, A.; Caflisch, A. *J. Comput. Chem.* **1998**, *19*, 21-37.
- [65] Sandak, B.; Wolfson, H.; Nussinov, R. *Proteins: Struct. Func. Gen.* **1998**, *32*, 159-174.
- [66] Zacharias, M.; Sklenar, H. *J. Comput. Chem.* **1999**, *20*, 287-300.
- [67] Trosset, J.; Scheraga, H. *J. Comput. Chem.* **1999**, *20*, 412-427.
- [68] Mangoni, M.; Roccatano, D.; Di Nola, A.
- [69] Eldridge, M.; Murray, C.; Auton, T.; Paolini, G.; Mee, R. *J. Comput. Aid. Mol. Des.* **1997**, *11*, 425-445.

- [70] Westhead, D.; Clark, D.; Murray, C. J. *Comput. Aid. Mol. Des.* **1997**, *11*, 209-228.
- [71] Thomas, B.; Joseph-McCarthy, D.; Alvarez, J. *Pharmacophore-based molecular docking*; University Press: International La Jolla, 2000.
- [72] C., M.; Bohacek, R. J. *Comput. Aid. Mol. Des.* **1997**, *11*, 333-344.
- [73] Venkatachalam, C.; Jiang, X.; Oldfield, T.; Waldman, M. J. *Mol. Graphics and Modelling* **2003**, *21*, 289-307.
- [74] Gehlhaar, D.; Verkhivker, G.; Rejto, P.; Sherman, G.; Fogel, D.; Fogel, L.; Freer, S. *Chem. Biol* **1995**, *2*, 317-324.
- [75] Verdonk, M.; Cole, J.; Hartshorn, M.; Murray, C.; Taylor, R. *Proteins: Struct. Func. Gen.* **2003**, *52*, 609-623.
- [76] Makino, S.; Kuntz, I. J. *Comput. Chem.* **1997**, *18*, 1812-1825.
- [77] Kramer, B.; Rarey, M.; Lengauer, T. *Proteins: Struct. Func. Gen.* **1999**, *37*, 228-241.
- [78] Kearsley, S.; Underwood, D.; Miller, M. J. *Comput. Aid. Mol. Des.* **1994**, *8*, 565-582.
- [79] Charifson, P.; Corkery, J.; Murcko, M.; Walters, W. J. *Med. Chem.* **1999**, *42*, 5100-5109.
- [80] Bissantz, C.; Folkers, G.; Rognan, D. J. *Med. Chem.* **2000**, *43*, 4759-4767.
- [81] Goodford, P. J. *Med. Chem.* **1985**, *28*, 849-857.
- [82] Cornell, W.; Cieplak, P.; Bayly, C.; Gould, I.; Merz, K.M., J.; Ferguson, D.; Spellmeyer, D.; Fox, T.; Caldwell, J.; Kollman, P. J. *Am. Chem. Soc.* **1995**, *117*, 5179-5197.
- [83] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; SwamiNaturehan, S.; Karplus, M. J. *Comput. Chem.* **1983**, *4*, 187-217.
- [84] Taylor, J.; Burnett, R. *Proteins: Struct. Func. Gen.* **2000**, *41*, 173-191.
- [85] Muegge, I.; Martin, Y. J. *Med. Chem.* **1999**, *42*, 791-804.
- [86] Gohlke, H.; Hendlich, M.; Klebe, G. J. *Mol. Biol.* **2000**, *295*, 337-356.
- [87] Nobeli, I.; Mitchell, J.; Alex, A.; Thornton, J. J. *Comput. Chem.* **2001**, *22*, 673-688.
- [88] Ishchenko, A.; Shakhnovich, E. J. *Med. Chem.* **2002**, *45*, 2770-2780.
- [89] Jain, A. J. *Comput. Aid. Mol. Des.* **1996**, *10*, 427-440.
- [90] Bohm, H. J. J. *Comput. Aid. Mol. Des.* **1994**, *3*, 243-256.
- [91] Wang, R.; Lu, Y.; Wang, S. *JMC* **2003**, *46*, 2287-2303.
- [92] Hansch, C.; T., F. J. *Am. Chem. Soc.* **1964**, *86*, 1616-1626.
- [93] Hansch, C.; Sammes, P.; Taylor, J. . In *Comprehensive Medicinal Chemistry*; Pergamon Press: Oxford, 1990.

- [94] Wolff, M. . In *Burger's Medicinal Chemistry 5th ed.*; John Wiley and Sons: New York, 1995.
- [95] Dean, P. *Molecular Similarity in Drug Design*; Blackie Academic and Professional: London, 1995.
- [96] Wermuth, C. *The Practice of Medicinal Chemistry*; Academic Press: London, 1996.
- [97] Cramer, R.; Petterson, D.; Bunce, J. J. *Am. Chem. Soc.* **1998**, *110*, 5959-5967.
- [98] Thornber, C. *Chem. Soc. Rev.* **1979**, *8*, 563-583.
- [99] Sheridan, R.; Kearsley, S. *Drug Discov. Today* **2002**, *7*, 903-911.
- [100] Jakes, S.; Willet, P. J. *Mol. Graphics* **1986**, *4*, 12-20.
- [101] Perkins, T.; Dean, P. J. *Comput. Aid. Mol. Des.* **1993**, *7*, 155-172.
- [102] Hall, L.; Kier, L. *Reviews of Computational Chemistry* **1991**, *2*, 367-422.
- [103] Willett, P. *Analytical Biotechnology* **2000**, *11*, 85-88.
- [104] Fisanick, W.; Cross, K.; Rusinko, A. J. *Chem. Inf. Comput. Sci.* **1992**, *32*, 664-674.
- [105] Marshall, G.; Barry, C.; Bosshard, H.; Dammkoehler, R.; Dunn, D. *Trends Biochem. Sci.* **1979**, *112*, 205-222.
- [106] Van Drie, J.; Weininger, D.; Martin, Y. J. *Comput. Aid. Mol. Des.* **1989**, *3*, 225-251.
- [107] Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. J. *Chem. Inf. Comput. Sci.* **1996**, *36*, 563-571.
- [108] Kubinyi, H. *3D QSAR in Drug Design: Theory, Methods and Applications*; ESCOM: Lenden, 1993.
- [109] Johnson, M.; Maggiora, G. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990.
- [110] Sen, K. *Topics Curr. Chem.* **1995**, *173*,.
- [111] Crippen, G. J. *Comput. Chem.* **1995**, *16*, 486-500.
- [112] Meyer, A. Y.; Richards, W. J. *Comput. Aid. Mol. Des.* **1991**, *5*, 427-439.
- [113] Tokarski, J. S.; Hopfinger, A. J. *Med. Chem.* **1994**, *37*, 3639-3654.
- [114] Jain, A.; Dietterich, T.; Lathrop, R.; Chapman, D.; Critchlow, R.; Bauer, B.; Webster, T.; Lozano-Perez, T. J. *Comput. Aid. Mol. Des.* **1994**, *8*, 635-652.
- [115] Mezey, P. *Shape in Chemistry: An introduction to Molecular Shape and Topology*; VCH: New York, 1993.
- [116] Chau, P.; Dean, P. J. *Mol. Graphics* **1987**, *5*, 97-100.
- [117] Walters, D.; Hopfinger, A. J. *Mol. Struct.: THEOCHEM* **1986**, *134*, 317-323.

- [118] Goldman, B.; Wipke, W. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 644-658.
- [119] Kearsley, S.K., S. G. *Tet.Comp.Met* **1990**, *3*, 615-633.
- [120] Klebe, G.; Abraham, U.; Mietzner, T. *J. Med. Chem.* **1994**, *37*, 4130-4146.
- [121] Perkins, T.; Mills, J.; Dean, P. *J. Comput. Aid. Mol. Des.* **1995**, *9*, 479-490.
- [122] Grant, J.; Gallardo, M.; Pickup, B. *J. Comput. Chem.* **1996**, *17*, 1653-1666.
- [123] Richard, A. *J. Comput. Chem.* **1991**, *12*, 959-969.
- [124] Hodgkin, E. E.; Richards, W. G. *Int. J. Quan. Chem.* **1987**, *14*, 105-110.
- [125] Burt, C.; Richards, W.; Huxley, P. *J. Comput. Chem.* **1990**, *11*, 1139-1146.
- [126] de Cáceres, M.; Villà, J.; Lozano, J. J.; Sanz, F. *Bioinformatics* **2000**, *16*, 568-569.
- [127] Petke, J. *J. Comput. Chem.* **1993**, *14*, 928-933.
- [128] Thorner, D. A.; Willett, P.; Wright, P.; Taylor, R. *J. Comput. Aid. Mol. Des.* **1997**, *11*, 163-174.
- [129] Pullman, B. *J. Biomolec. Str. Dynam.* **1986**, *1*, 773.
- [130] Gasteiger, J.; Li, X. *Angew. Chem. Int. Ed. Engl.* **1994**, *33*, 643-646.
- [131] Apaya, R.; Lucchese, B.; Price, S.; Vinter, J. *J. Comput. Aid. Mol. Des.* **1995**, *9*, 33-43.
- [132] Blaney, F.; Edge, C.; Phippen, R. *J. Mol. Graphics* **1995**, *13*, 165-174.
- [133] Mestres, J.; Rohrer, D.; Maggiora, G. *J. Comput. Chem.* **1997**, *18*, 934-954.
- [134] Good, A.; Hodgkin, E.; Richards, W. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 188-191.
- [135] Tervo, A.; Ronkko, T.; Nyronen, T.; Poso, A. *J. Med. Chem.* **2005**, *48*, 4076-4086.
- [136] Armitage, P.; Berry, G. *Statistical Methods in Medical Research. 3rd edition*; Blackwell Science Ltd.: Oxford, 1994.
- [137] Driver, H. *The University of California Publications in American Archaeology and Ethnology* **1932**, *31*, 211-256.
- [138] Carbó, R.; Leyda, L.; Arnau, M. *Intl. J. Quantum Chemistry* **1980**, *17*, 1185-1189.
- [139] Manaut, F.; Lozoya, E.; Sanz, F. Automatic determination of maximum electrostatic alignment between methotrexate and dihydrofolic acid. In *QSAR: Rational Approaches to the Design of Bioactive Compounds*; Silipo, C.; Vittoria, A., Eds.; Elsevier Science: Amsterdam, 1991.
- [140] Gillet, V.; Wild, D.; Bradshaw, J. *Comp. J.* **1998**, *41*, 547-558.
- [141] Cioslowski, J.; Fleischmann, E. *J. Am. Chem. Soc.* **1991**, *113*, 64-67.
- [142] Hollander, M. *Nonparametric statistical methods*; Wiley: New York, 1973.

- [143] Barbany, M.; Gutiérrez-de Terán, H.; Sanz, F.; Villà-Freixa, J.; Warshel, A. *Chem-BioChem* **2003**, *4*, 277-285.
- [144] Bagdassarian, C. K.; Schramm, V. L.; Schwartz, S. D. *J. Am. Chem. Soc.* **1996**, *118*, 8825-8836.
- [145] Sokal, R.; Michener, C. *Univ. Kans. Sci. Bull.* **1958**, *38*, 1409-1438.
- [146] Russel, P.; Rao, T. J. *Malaria Inst. India* **1940**, *3*, 153-178.
- [147] Jaccard, *Bull. Soc. Vaudoise Sci. Nat.* **1901**, *37*, 547-579.
- [148] Rogers, J.; Tanimoto, T. *Science* **1960**, *132*, 1115-1118.
- [149] Lemmen, C.; Lengauer, T. *J. Comput. Aid. Mol. Des.* **2000**, *14*(3), 215-232.
- [150] Hurst, T. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 190-196.
- [151] Masek, B.; Merchant, A.; Matthew, J. *J. Med. Chem.* **1993**, *36*(9), 1230-1238.
- [152] Lemmen, C.; Hiller, C.; Lengauer, T. *J. Comput. Aid. Mol. Des.* **1998**, *12*(5), 491-502.
- [153] Nissink, J.; Verdonk, M.; Kroon, J.; Mietzner, T.; Klebe, G. *J. Comput. Chem.* **1997**, *18*(5), 638-645.
- [154] Martin, Y.; Bures, M.; Danaher, E.; DeLazzer, J.; Lico, I.; Pavlik, P. *JCAMD* **1993**, *7*, 83-102.
- [155] Bron, C.; Kerbosch, J. *Communications of the ACM* **1973**, *16*, 575-577.
- [156] Crippen, G.; Havel, T. *Distance geometry and Molecular Conformation.*; Research Studies Press: Taunton, UK, 1988.
- [157] Devillers, J. *Genetic Algorithms in Molecular Modelling.*; Academic Press: London, 1996.
- [158] Laindan, Y.; Wolfson, H. *Proceedings of the nd International Conference on Computer Vision* **1988**, 238-251.
- [159] Nussinov, R.; Wolfson, H. *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 10495-10499.
- [160] Kabsch, W. *Acta Crystallographica A* **1976**, *32*, 922-923.
- [161] Broto, P.; Moreau, G.; Vandycke, C. *Eur. J. Med. Chem.* **1984**, *19*, 66-70.
- [162] Wagener, M.; Sadowski, J.; Gasteiger, J. *J. Am. Chem. Soc.* **1995**, *117*, 7769-7775.
- [163] Anzali, S.; Barnickel, G.; Krug, M.; Sadowski, J.; Wagener, M.; Gasteiger, J.; Polanski, J. *J. Comput. Aid. Mol. Des.* **1996**, *10*, 521-534.
- [164] Clementi, S.; Cruciani, G.; Riganelli, D.; Valigi, R.; Constantino, G.; Baroni, M.; Wold, S. *Pharm. Pharmacol. Lett.* **1993**, *3*, 5-8.
- [165] Silverman, B. D.; Platt, D. E. *J. Med. Chem.* **1996**, *39*, 2129-2140.
- [166] Bursi, R.; Dao, T.; van Wijk, T.; de Gooyer, M.; Kellenbach, E.; Verwer, P. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 861-867.

- [167] Turner, D.; Willett, P.; Ferguson, A. *J. Comput. Aid. Mol. Des.* **1999**, *13*, 271-296.
- [168] Benedetti P, Mannhold R, C. G. P. M. *JMC* **2002**, *45*, 1577-1584.
- [169] Gratteri, P.; Romanelli, M.; Bonaccini, C. *Grid Independent descriptors (GRIND) in the study of the s-receptor subtype selectivity. Drug Future. (Suppl. A) 257*; Prous Science, SA: Barcelona, Spain, 2002.
- [170] Fontaine, F.; Pastor, M.; Sanz, F. *Drugs Future* **2002**, *27*, 232.
- [171] Klebe, G. *Structural alignment of molecules. In: 3D QSAR in Drug Design. Theory, Methods, and Applications*; Escom Science: Leiden, 1993.
- [172] Martin, M.; Sanz, F.; Campillo, M.; Pardo, L.; Pérez, J.; Turmo, J. *Int. J. Quantum Chem.* **1983**, *23*, 1627-1641.
- [173] Murray, J.; Brinck, T.; Grice, M.; Politzer, P. J. *Mol. Struct. (Theochem)* **1992**, *256*, 29-45.
- [174] Scrocco, E.; Tomasi, J. . In *Advances in Quantum Chemistry*, Vol. 11; Academic Press: New York, 1978.
- [175] Dean, P. *Molecular foundations of drug-receptor interaction*; Cambridge University Press: Cambridge, 1987.
- [176] Sanz, F.; Martin, M.; Lapena, F.; Manaut, F. *Quant. Struct. Act. Relat.* **1986**, *5*, 54-57.
- [177] Sanz, F.; Manaut, F.; Segura, J.; Carbó, M.; de la Torre, R. *TEOCHEM* **1988**, *170*, 171-180.
- [178] Manaut, F.; Sanz, F.; José, J.; Milesi, M. *J. Comput. Aid. Mol. Des.* **1991**, *5*, 371-380.
- [179] Sanz, F.; Manaut, F.; Sánchez, J.; Lozoya, E. *TEOCHEM* **1991**, *230*, 437-446.
- [180] Sanz, F.; Manaut, F.; Rodríguez, J.; Lozoya, E.; López de Briñas, E. *J. Comput. Aid. Mol. Des.* **1993**, *7*, 337-347.
- [181] Sanz, F.; Manaut, F.; Dot, T.; López de Briñas, E. *TEOCHEM* **1992**, *256*, 287-293.
- [182] Sanz, F.; López de Briñas, E.; Rodríguez, J.; Manaut, F. *Quant. Struct.-Act. Relat.* **1994**, *13*, 281-284.
- [183] Schmidt, M.; Baldrige, K.; Boatz, J.; Elbert, S.; Gordon, M.S. Jensen, J.; Koseki, S.; Matsunaga, N.; Nguyen, K.; Su, S. J.; Windus, T.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347-1363.
- [184] Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9-20.
- [185] "InsightII", 1995.
- [186] Laaksonen, L. *gOpenMol*; Helsinki, Finland, 1999.
- [187] Nicolotti, O.; Pellegrini-Calace, M.; Carrieri, A.; Altomare, C.; Centeno, N. B.; Sanz, F.; Carotti, A. *J. Comput. Aid. Mol. Des.* **2001**, *15*, 859-872.

- [188] Laaksonen, L. J. *Mol. Graphics* **1992**, *10*, 33-34.
- [189] Pastor, M.; Benedetti, P.; Carotti, A.; Carrieri, A.; Díaz, C.; Herráiz, C.; Höltje, H.-D.; Loza, M.I., O. T.; Padín, F.; Pubill, F.; Sanz, F.; Stoll, F.; the LINK3D Consortium, *J. Comput. Aid. Mol. Des.* **2003**, *16*, 809-818.
- [190] Carroll, D. L. "GA program, version 1.6.4", 1997.
- [191] Richards, W. . In *Quantum pharmacology*; Butterworths: London, 1983.
- [192] Pople, J.; Santry, D.; Segal, G. J. *Chem. Phys* **1965**, *43*, 129.
- [193] Levine, I. . In *Quantum chemistry*; Allyn and Bacon: Boston, 1984.
- [194] Born, M.; Oppenheimer, J. *Ann. Physik* **1927**, *84*, 457-484.
- [195] Hartree, D. *Proc. Cambridge Phil. Soc.* **1928**, *24*, 89-111.
- [196] Slater, J. *Phys. Rev.* **1930**, *35*, 509-529.
- [197] Parr, R. *The Quantum Theory of Molecular Electronic Structure*; W.A. Benjamin Inc.: New York, 1963.
- [198] Roothan, C. *Rev. Mod. Phys.* **1951**, *23*, 69-89.
- [199] Nesbet, R. *Proc. Roy. Soc.* **1955**, *A230*, 312-322.
- [200] Fock, V. *Physik* **1939**, *61*, 126-148.
- [201] Wold, S.; Dunn, W. J. *Chem. Inf. Comput. Sci.* **1983**, *23*, 6-13.
- [202] Szabo, A.; Ostlund, N. *Modern Quantum Chemistry*; McGraw-Hill: New York, 1982.
- [203] Hehre, W.; Stewart, R.; Pople, J. J. *Chem. Phys.* **1969**, *51*, 2657-2664.
- [204] Diechfield, R.; Hehre, W.; Pople, J. J. *Chem. Phys.* **1970**, *54*, 724.
- [205] Alsenoy, C. J. *Comp. Chem.* **1988**, *9*, 620-626.
- [206] Pople, J.; Segal, G. J. *Chem. Phys* **1965**, *44*, 3289-3296.
- [207] Parr, R. J. *Chem. Phys.* **1952**, *20*, 1499.
- [208] Pople, J.; Beveridge, D.; Dobash, P. J. *Chem. Phys* **1967**, *47*, 2026.
- [209] Dewar, M.; Shanshal, M. J. *Am. Chem. Soc.* **1969**, *91*, 3654-3655.
- [210] Dewar, M.; Thiel, W. J. *Am. Chem. Soc.* **1977**, *99*, 4899-4907.
- [211] Dewar, M.; Zoebish, E.; Healy, E.; Stewart, J. J. *Am. Chem. Soc.* **1985**, *107*, 3902-3909.
- [212] Cammarata, A. . In *Molecular Orbital Studies in Chemical Pharmacology.*; Springer-Verlag: New York, 1970.
- [213] Nakayama, A.; Richards, W. *Quant. Struct. Act. Relat.* **1987**, *6*, 153-157.
- [214] Venanzi, T.; Venanzi, C. J. *Comp. Chem.* **1988**, *9*, 67-74.

- [215] Grunewald, G.; Creese, M. *Drug Design and Delivery* **1986**, *1*, 23-37.
- [216] Karelson, M.; Lobanov, V.; Katritzky, A. *Chem. Rev* **1996**, *96*, 1027-1043.
- [217] Bader, J. S.; Kuharski, R. A.; Chandler, D. J. *Chem. Phys.* **1990**, *93*, 7213-7224.
- [218] Popelier, P. *Atoms in Molecules. An Introduction*; Pearson Education: Harlow, 2000.
- [219] Hopfinger, A. *Conformational properties of Macromolecules*; Academic Press: New York, 1973.
- [220] McCammon, J. A.; Wolynes, P. G.; Karplus, M. *Biochemistry* **1979**, *18*, 927-942.
- [221] Pastor, M. *J. Med. Chem.* **1997**, *40*, 1455-1464.
- [222] von Itzstein, M.; Wu, W.; Kok, G.; Pegg, M.; Dyason, J.; Jin, B.; Van Phan, T.; Smythe, M.; White, H.; Oliver, S.; et, a. *Nature* **1993**, *363*, 418-423.
- [223] Kastenholz, M.; Pastor, M.; Cruciani, G.; Haaksma, E.; Fox, T. *J. Med. Chem.* **2000**, *43*, 3033-3044.
- [224] Courant, R. *Bull. Amer. Math. Soc.* **1943**, *49*, 1-23.
- [225] Curry, D. *Qu. App. Maths.* **1944**, *2*, 258-261.
- [226] Householder, A. *Principles of Numerical Analysis*; McGraw-Hill: New York, 1953.
- [227] Hestenes, M.; Stiefel, E. *J. Res. N.B.S.* **1952**, *49*, 409-436.
- [228] Fletcher, R.; Reeves, C. *Comp. J.* **1964**, *7*, 149-154.
- [229] Nelder, J.; Mead, R. *Comput. J.* **1965**, *7*, 308-313.
- [230] Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley: New York, 1989.
- [231] Davis, L. *Handbook of Genetic Algorithms*; Van Nostrand Reinhold: New York, 1991.
- [232] Forrest, S. *Science* **1993**, *261*, 872-878.
- [233] Goldberg, D. E. *Commun. ACM* **1994**, *37*, 113-119.
- [234] Sprague, P. *Perspect. Drug Disc. Des.* **1995**, *3*, 1-20.
- [235] Smellie, A.; Teig, S.; Towbin, P. J. *Comput. Chem.* **1995**, *16*, 171-187.
- [236] MacKerell, A. e. J. *Phys. Chem. B* **1998**, *102*, 3586-3616.
- [237] Böstrom, J. *Comput. Aid. Mol. Des.* **2001**, *15*, 1137-1152.
- [238] Zhou, J.; Tits, A.; Lawrence, C. *FFSQP program, version 3.7*; AEM Design, University of Maryland: Maryland, 1997.
- [239] Panier, E.; Tits, A. *Math Programming* **1993**, *59*, 261-276.
- [240] Bonnans, J.; Panier, E.; Tits, A.; Zhou, J. *SIAM J. Numer. Anal.* **1992**, *29*, 1187-1202.

- [241] Zhou, J.; Tits, A. J. *Optim. Theory Appl.* **1993**, *76*, 455-476.
- [242] Chen, Z. *Theoretica Chimica Acta* **1989**, *75*, 849-857.
- [243] Russell, R. B.; Barton, G. J. *Proteins: Struct. Func. Gen.* **1992**, *14*, 309-323.
- [244] Smith, T.; Waterman, M. J. *Mol. Biol.* **1981**, *147*, 195-197.
- [245] Barton, G. *Comp.App.Biosci* **1993**, *9*, 729-734.
- [246] Argos, P.; Rossmann, M. J. *Mol. Biol.* **1976**, *105*, 75-95.
- [247] Kollman, P. *Chem. Rev.* **1993**, *93*, 2395-2417.
- [248] Straatsma, T.; Berendsen, H.; Postma, J. J. *Chem. Phys.* **1986**, *85*, 6720-6727.
- [249] Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027 - 2094.
- [250] Russell, S.; Warshel, A. J. *Mol. Biol.* **1985**, *185*, 389-404.
- [251] Warshel, A.; Levitt, M. J. *Mol. Biol.* **1976**, *103*, 227-249.
- [252] Warshel, A. J. *Phys. Chem.* **1979**, *83*, 1640-1650.
- [253] Warshel, A.; Russell, S. T. *Q. Rev. Biophys.* **1984**, *17*, 283-421.
- [254] King, G.; Warshel, A. J. *Chem. Phys.* **1989**, *91*, 3647-3661.
- [255] Florián, J.; Warshel, A. J. *Phys. Chem. B* **1997**, *101*, 5583-5595.
- [256] Florián, J.; Warshel, A. J. *Phys. Chem. B* **1999**, *103*, 10282-10288.
- [257] Florián, J.; Warshel, A. J. *Am. Chem. Soc.* **1997**, *119*, 5473-5474.
- [258] Florián, J.; Warshel, A. J. *Phys. Chem. B* **1998**, *102*, 719-734.
- [259] Florián, J.; Åqvist, J.; Warshel, A. J. *Am. Chem. Soc.* **1998**, *120*, 11524-11525.
- [260] Florián, J.; Sponer, J.; Warshel, A. J. *Phys. Chem. B* **1999**, *103*, 884-892.
- [261] Florián, J.; Štrajbl, M.; Warshel, A. J. *Am. Chem. Soc.* **1998**, *120*, 7959-7966.
- [262] Frisch, M. J. *et al.* "Gaussian 98, Revision A.7, Gaussian, Inc., Pittsburgh, PA, USA", 1998.
- [263] Debye, P. J. W. *Polar Molecules*; Chemical Catalog Co.: New York, 1929 USC library QC585.D413 1929.
- [264] Langevin, P. *Annales de Chimie Physique* **1905**, *5*, 70.
- [265] Miertus, S.; Scrocco, E.; Tomasi, J. J. *Chem. Phys.* **1981**, *55*, 117-129.
- [266] Miertus, S.; Tomasi, J. *Chem. Phys.* **1982**, *65*, 239-245.
- [267] Jolliffe, I. *Principal Component Analysis*; Springer-Verlag: New York, 1986.
- [268] Wold, S.; Hellberg, S.; Lundstedt, T.; Sjostrom, M.; Wold, H. *Proc. Symp. on PLS Model Building: Theory and Application*; Frankfurt am Main, 1987.

- [269] Wold, S.; Sjoström, M.; Eričsson, L. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109-130.
- [270] Cruciani, G.; Watson, K. J. *Med. Chem.* **1994**, *37*, 2589-2601.
- [271] Mitchell, T. *Technometrics* **1974**, *16*, 203-210.
- [272] Steinberg, D.; Hunter, W. *Technometrics* **1984**, *26*, 71-76.
- [273] Matthews, B. *Biochim. Biophys. Acta* **1975**, *405*(2), 442-51.
- [274] Rodrigo, J.; Barbany, M.; Gutiérrez-de Terán, H.; Centeno, N. B.; de Càceres, M.; Dezi, C.; Fontaine, F.; Lozano, J. J.; Pastor, M.; Villà, J.; Sanz, F. J. *Braz. Chem. Soc.* **2002**, *13*, 795-799.
- [275] Barbany, M.; González-de Terán, H.; Sanz, F.; Villà-Freixa, J. *Proteins: Struct. Func. Gen.* **2004**, *56*, 585-594.
- [276] Villà, J.; Warshel, A. J. *Phys. Chem. B* **2001**, *105*, 7887-7907.
- [277] Haslam, E. *Shikimic Acid: Metabolism and Metabolites*; John Wiley & Sons: New York, 1993.
- [278] Cload, S.; Liu, D.; Pastor, R.; Schultz, P. J. *Am. Chem. Soc.* **1996**, *118*, 1787-1788.
- [279] Weist, O.; Houk, K. J. *Am. Chem. Soc.* **1995**, *117*, 11628-11639.
- [280] Hilvert, D. *Annu. Rev. Biochem.* **2000**, *69*, 751-793.
- [281] Kienhofer, A.; Kast, P.; Hilvert, D. J. *Am. Chem. Soc.* **2003**, *125*, 3206-3207.
- [282] Khanjin, N.; Snyder, J.; Menger, F. M. J. *Am. Chem. Soc.* **1999**, *121*, 11831-11846.
- [283] Copley, S. D.; Knowles, J. R. J. *Am. Chem. Soc.* **1987**, *109*, 5008-5013.
- [284] Lee, A.; Stewart, J.; Clardy, J.; Ganem, B. *Chem. & Biol.* **1995**, *2*, 195-203.
- [285] Lyne, P.; Mulholland, A.; Richards, W. J. *Am. Chem. Soc.* **1995**, *117*, 11345-11350.
- [286] Guo, H.; Cui, Q.; Lipscomb, W.; Karplus, M. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 9032-9037.
- [287] Martí, S.; Andrés, J.; Moliner, V.; Silla, E.; Tuñón, I.; Bertrán, J. J. *Phys. Chem. B* **2000**, *104*, 11308.
- [288] Martí, S.; Andrés, J.; Moliner, V.; Silla, E.; Tuñón, I.; Bertrán, J. **2001**, *3*, 207-212.
- [289] Hur, S.; Bruice, T. C. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 1176-1181.
- [290] Hur, S.; Bruice, T. C. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 12015-12020.
- [291] Warshel, A.; Florián, J.; Štrajbl, M.; Villà, J. *ChemBioChem* **2001**, *2*, 109-111.
- [292] Cavalli, A.; Poluzzi, E.; De Ponti, F.; Recanatini, M. *J. Med. Chem.* **2002**, *45*, 3844-3853.

Index

- CATALYST, 52
CHEMSOL, 55, 64, 75, 131, 133
GAMESS, 31, 35, 36, 41, 44, 45, 67, 68, 75, 127
GOLPE, 31, 35, 50, 61, 62, 68, 129
GRID, 19, 31, 35, 36, 47, 64, 67, 69, 75, 127
MIPSim development, 31, 36, 63, 134
MIPSim installation, 127
MIPSim profiling, 67, 130
FFSQP, 53
MIPSim interface, 68
MIPSim methods, 35
MIPSim previous studies, 30
MIPSim protocols, 130
MIPSim web site, 68
OMEGA, 53
PDLDSMALL, 133
POLARSAR, 133
STAMP, 54, 134
SUPERB, 53, 134
TORS module, 66, 131
3D-QSAR, 21, 22, 28, 31, 35, 50, 58, 59, 99
- alignment, 16, 21, 27, 28, 31, 36, 65, 85
- binary variables, 27
binding modes, 16–18, 20
binding site, 2, 9–11, 13–18, 47
biomolecular interactions, 1
Boltzmann law, 4, 5
- Carbó index, 25
catalytic antibodies, 13, 64, 65, 75, 133
classical fields, 45
comparison of biomolecules, 27, 28, 30, 31, 63, 67, 69
conformational sampling, 52, 53
conjugate gradients, 51
continuum variables, 25, 27
Cosinus index, 25, 26
- docking, 15–18, 85
drug design, 15–17, 20, 21, 43, 47
- energy, 2
energy intervals, 64, 131
enthalpy, 3, 4
entropy, 2–4, 19
experimental structure, 16, 21
- first law of thermodynamics, 2
free energy, 2–4, 6, 8, 11, 16, 17, 20, 54, 56, 75, 132
- Gaussian index, 26, 32, 63
genetic algorithms, 18, 29, 36, 52, 53
ground state, 5, 7, 8
- heat, 2–4
Hodgkin index, 26, 63
- internal energy, 3
- Langevin dipoles, 54, 55, 64, 75, 132, 133
ligand, 1
ligand-based design, 16, 20, 47
- macroscopic state, 2, 5
Matthews coefficient, 62
microstate, 3, 5
molecular electrostatic potentials, 42, 43, 75
molecular interaction fields, 14, 21, 36
molecular interaction potentials, 64
molecular superposition, 22, 28–30, 53
molecular wavefunction, 36, 37
- optimization, 29, 31, 35, 50, 53, 68
- partial least squares, 50, 58, 59, 61
partition function, 4, 5
Pearson index, 25, 26, 63

- potential energy surface, 6, 7, 45, 46
- pretreatment of data, 61
- principal component analysis, 58, 61
- protein-small molecule interactions, 1, 2, 6, 15, 17, 19

- quantum fields, 36

- receptor, 1
- receptor-based design, 15, 16

- second law of thermodynamics, 3
- sequence alignment, 54
- similarity, 22, 29, 30, 36, 50, 64, 75
- similarity coefficients, 27, 50, 63, 64, 131
- simplex, 51
- Spearman index, 26, 31, 63
- state function, 2, 3
- steepest descent, 51

- third law of thermodynamics, 4
- transition state, 7, 11–13
- transition state theory, 7, 8

- variable selection, 50, 61
- visualization tools, 68

- work, 2