UNIVERSITA' DEGLI STUDI DI TRIESTE

Sede amministrativa del dottorato di ricerca

DEVELOPMENT OF NEW COMPUTATIONAL METHODS FOR THE SIMULATION OF ENZIMES UNDER OPERATIONAL CONDITIONS

Dottorando
Valerio Ferrario

Direttore della scuola di dottorato
Chiar. Mo Prof Enzo Alessio

Tutore e relatore
Prof.Ssa Lucia Gardossi

Correlatore
Dr, Paolo Braiuca

INDEX

“Con la forza della verità, in vita, ho conquistato l’universo”

Faust

CHAPTER 1

INTRODUCTION

1.1 Introduction

At the beginning of the XXIst century every kind of industry is worried about making their products more competitive in the global market. The competition concerns different aspects of the production processes, from the starting material to the products purity.

Going through this direction it is essential to think back to the traditional synthetic route mostly based on classical chemical synthesis. In order to perform more efficient processes in terms of reaction selectivity, products purity, operational conditions, etc. biocatalytic reactions represent the most promising solution. In this sense they are covering an even more important role every day.

The conversion of most industrial chemical processes into biochemicals requires a huge number of well known enzymes able to catalyse almost all the necessary chemical reactions. The industrial evolution in this direction is braked by a cultural factor against the enzymatic reactions, as well as, by a lack of knowledge of enzymatic mechanisms which limits the enzymatic process optimization.

Recent advantages in computational sciences have led to novel sophisticated and refined computational methods that are able to describe the biocatalyst machinery in detail.[1]

In this field molecular dynamics simulations (MD) cover an important role since these techniques are able to simulate complex systems with high accuracy and to investigate conformational possibilities of the simulated elements.[2] The application of MD simulations to biocatalysis problems allows to investigate

[1] P Braiuca, C Ebert, A Basso, P Linda, L Gardossi, *Trends Biotechnol*, **2006**, 24, 419.

[2] J P Mesirov, K Schulten, D W Sumners, *Mathematics and Its Applications*, **1996**, New York, Springer-Verlag.

enzymatic systems in real operational behavior and to investigate, at the molecular level, enzymes-substrates interactions as well as enzyme stability and all the other important characteristics that a good biocatalyst needs.[3]

Unfortunately MD simulations are not able to simulate too complicate problems; in these cases other computational approaches, such as chemometrics, can be of aid because they are able to extract relevant information from experimental data and eventually they enable the construction of predictive models of the enzyme behavior.[1] Chemometrics is a set of statistical approaches which are able to extract relevant and latent information from experimental data sets. Moreover extensive statistical investigations of various data sets, coming both from MD simulations and experiments, can lead to the creation of predicting models based on empirical equations.

Joint applications of the two computational approaches have been already employed to successfully solve several complex issues, especially in cases of enzyme selectivity previsions.[4]

The combination of computational and experimental research in biocatalysis offers the possibility to greatly enhance the level of knowledge of a biocatalytic system and reduce experimental efforts and costs. The solution of research problems can be found within different time frames and accuracy levels as a function of the computational techniques chosen.[1]

As a matter of fact, computational methods have been developed just for the last twenty years; nevertheless the recent evolution of modern calculators offers new and exciting opportunities to explore this research field.

The further development of these promising combined experimental and computational approaches require the

[3] R J Kazlauskas, *Curr Opin Chem Biol*, **2000**, 4, 81.

[4] P Braiuca, L Knapic, V Ferrario, C Ebert, L Gardossi, *Adv Synth Catal*, **2009**, 351, 1293.

improvement and invention of new and more efficient computational techniques as well as the creation of new hybrid strategies in order to achieve the full comprehension of bicatalytic problems.

1.2 Molecular modelling

The possibility to simulate real systems, to predict properties and to explain experimental data, makes molecular modelling a powerful complementary tool of experimental research. It is able to offer new research opportunities, improve the activity of already known compounds and reduce the rate of the failures.[5] The molecular modelling opens the perspective to calculate several molecular properties from three-dimensional models of chemical systems. It comprises a plethora of methods and approaches, relying on algorithms of different complexity and it is applicable to almost any (bio-)chemical field.

The interest in these techniques is long-term established in some areas, such as drug design, but the interest is increasing also in all the other chemistry related areas, because they can fill the lack of routine experimental procedures allowing to observe phenomena at molecular level.

The increasing computational power (and the subsequent decrease of costs) makes complex simulation methods more and more widely available, increasing their application scope and utility.

Despite the huge power of the modern CPUs, often calculations still require a significant computational cost, especially for the most interesting applications, where the complexity of the systems increase enormously the number of of variables of the system. The need to apply computational chemistry to complex systems pushed the development of simplified methods of empirical

[5] M Ferameglia, S Pricl, G Longo, *Chem Biochem Eng Q*, **2003**, 17, 69.

nature, that despite their fundamental simplicity and intrinsic limitations, are of impressive utility and are undoubtedly considered as potent tools in the hand of chemists. Molecular mechanics (MM) belongs to this ensemble of empirical methods and comprises the vast majority of applications in chemistry-biology interface fields, such as biocatalysis.

Quantum mechanics represents the rigorous approach, able to simulate virtually any chemical system and process, since it relies on the solution of the Schroedinger equation. Nevertheless it is severely hampered by the complexity if its mathematics and its application is still limited to small molecular systems (up to 200 atoms).

1.2.1 Quantum mechanics and molecular mechanics

The construction of a chemical model in a virtual three-dimensional space, able to be treated computationally, requires the transformation of chemical information into mathematical information. To achieve this, the concept of chemical bond has to be redefined into a set of atoms in a Cartesian space. Each atom is then described by a triplet of numbers, representing its spatial coordinates in the Cartesian space, perfectly saving the geometrical information such as the bond distances and angles. From the structure it is possible to derive energy information, applying an appropriate law. From the energy of a system it is possible to derive many other properties.

Molecular modelling decomposes the energy into two principal components, potential energy and kinetic energy. This is what is currently defined as dynamic energy:

$$E_{dynamic} = E_{potential} + E_{kinetic}$$

If the molecule is completely still (0 K) or if the system is relativistic (the observer is moving together with the molecule), the kinetic energy is null and therefore the total energy comes from the potential energy only. This is what is usually defined as mechanic energy.

The correlation of the structure with its energy is the fundamental idea of molecular modelling. All the molecular modelling methods, either belonging to MM or QM, are based on the definition of a set of equations aiming at correlating structure and energy:

$$E = f\begin{pmatrix} H_{(x,y,z)} \\ C_{(x,y,z)} \\ O_{(x,y,z)} \\ \dots \end{pmatrix}$$

To achieve this, the application of several principles is necessary:

- **QM laws**: complicated and expensive in terms of time and resources, they include the quantistic mechanic (QM) and the quantistic dynamic (QD);
- **Classical mechanic laws or Newtonian laws:** simplistic and not sufficient to explain all the molecular properties of interest, they include the classical molecular mechanic (MM) and the classical molecular dynamic (MD).

Efforts to combine the flexibility of MM to the precision of QM gave origin to hybrid systems (MM/QM), a compromise that allows to define a part of a molecule, usually the most important (i.e. the active site), in a quantistic mode, and the rest of the molecule, less important, in a classical mode. It is important to specify that the hybrid systems are complicated by the necessity to parametrize properly the interface between the QM and the MM part. Complexity of this aspect sometimes overcome the advantages of the general idea, thus making application of

QM/MM methods still questionable in many cases.

1.2.2 Quantum mechanics and molecular orbital method (MO)

All the systems using quantomechanic principles, combine the *ab-initio* methods and the semi-empirical methods, are based on the approximate solution of the Schrödinger equation:

$$H\psi = E\psi$$

where *H* is the Hemiltonian operator which describes the kinetic energy of the nucleus and of the electrons of the molecule. *E* is the total energy of the system and ψ represent the wave function that describes the motion of the particles.

The resolution of this equation, valid for polielectronic atoms and molecules, is possible using approximate solutions only.

- Born-Oppenehimer approximation: the nucleus are fixed and only the electrons movement are considered.
- Hartree-Fock approximation: the electrons movement is described only by monoelectronic wave functions (the spatial part is called molecular orbital) and not by polielectronic wave functions, with the aim of following the movement of a single electron in an electromagnetic field generated by the other electrons.
- LCAO method: the wave function ψ, named also molecular orbital, is expressed by a linear combination of atomic orbitals Φ:

$$\psi_i = \sum_{i=1}^{N} C_{ij}\phi_i$$

The final quality of the ab-initio calculation is strictly dependent from the base set, namely the set of Φ_i used.[6] The C_{ij} coefficients are calculated by the same algorithm used for the *ab-initio* calculation that leads to an energy minimum where the orbitals assume a constant value by iterative variations.[7]

This method is affected by a disadvantage: the calculation time increases with the fourth potency of the number of the basis sets used;[8] this is translated into a big computational cost, limiting the application to systems with less than, or with a few hundred atoms only.

Semiempirical methods introduce simplifications and empiric parameters in the molecular orbital calculations, leading to a considerable gain in terms of calculation time, but also to an unavoidable loss in precision.

Only a compromise between the calculation rate and the accuracy of the results can expand the application scope of QM methods to system of several hundred atoms.

1.2.3 Molecular mechanics: atom types and force field

In molecular mechanics (MM) the atom loses its quantomechanic characteristics and it is simply described as a sphere with a certain mass, a volume and a point charge on the basis of the atom it represents. Calculation of molecular electronic state is completely avoided. To recover the concept of valence, a concept lost in this approximation, it is necessary to introduce the concept of atom type which correlates every sphere to all the properties describing each atom. In molecular mechanic there are as many atom types

[6] D De Frees, B Levi, S Pollak, W Hehre, S Binkley, J Pople, *J Am Chem Soc,* **1979**, 6, 2.

[7] T Clarke, *Handbook of Computational Chemistry*, New York, USA, **1985**.

[8] D Boyd, K Lipkowitz, *J Chem Educ,* **1982**, 59, 269.

as the number of possible chemical situations, in different molecules.

For the description of the interactions of a molecule it is necessary to use a mathematical function called force field which based on classical mechanics laws. The force field has to describe as a simple mathematical function, continues and differentiable functions which define the potential energy in relation to the coordinates of all the atoms that belong to the molecule.

It is fundamental, for its applicability, that the potential energy associated to the force field balances opportunely its simplicity with its accuracy in the description of its energetic and structural properties among the different analysed molecules.

Therefore, there are different force field specialized for different molecule types and applications (Table 1.1).

MM2, MM3, MM4 (Allinger, 1977, 1988, 1989, 1996, 1997; Lii, 1989a, 1989b, 1989c, 1991, 1998; Nevins, 1996; Hay, 1998) **CFF93** (“Central Force Field”, Levy, 1979) **MMFF** (“Merck Molecular Force Field”, Halgren, 1992, 1996a, 1996b, 1996c, 1996d, 1996e)	**Small Organic Molecules**
PEF95SAC (Rasmussen, 1997)	**Polysaccharides**
SHAPES (Allured, 1991)	**Metallic Compounds**
ECEPP (“Empirical Conformational Energy Program for Peptides”, Momany, 1975; Nemethy, 1983; Sippl, 1984)	**Proteins and Nucleic Acids**

CHARm (“Chemistry at Harvard Macromolecular Mechanics”, Brooks, 1983; MacKerell, 1998, 2004) **AMBER** (“Assisted Model Building with Energy Refinement”, Weiner, P.K. 1981; Weiner, S.J. 1984; Kollman, 1986, 1995; Pearlman, 1991; Ponder, 2003) OPLS (“Optimised Potentials for Liquid Simulations”, Jorgensen, 1988, 1996; Kaminski, 1994; Damm, 1997) **GROMOS** (“Groningen Molecular Simulation”, Hermans, 1984; Ott, 1996)

Table 1.1: Examples of specialized force fields.

The molecular mechanics is anyway affected by some intrinsic limits due to the theory at its basis:

- The deletion from the mathematical treatment of the intrinsic atomic structures, and therefore, the implicit representation of its electronic configuration, limits these methods to the study of the fundamental molecular structures. The accurate description of every process that implicates the formation or the breaking of chemical bonds is not possible.
- The obtained results are strictly related to the quality of the potential energy function (force field) and to the parameters set of every atom type.
- The potential energy function described by the force field has scarce chemical meaning, except for structures associated with stable thermodynamical conformations and in some cases for energetic rotational barriers.

The force field is therefore an empiric function of the potential energy. Force fields are created to be applied mainly to conformational analysis and these force field-based methods originated in the same period as the development of the applications of the quantomechanical methods. These came from the vibrational spectroscopy in which it is necessary to build up particular potential energy functions to use the spectroscopic information for the description of the global molecular behaviour. In this field the potential energy function, used to describe molecular vibrations, is simulated by:

- A mathematical function which is the sum of all the internal interactions among atoms, without any precise correlation with the covalent structure of the molecule.[9]
- A mathematical function correlated to the values of the distances and the values of the interatomic angles;[10] the difference from the CFF model is that this model is tightly dependent on the molecule.

The modern molecular mechanics, based on the force field concept, was developed from those different approaches; this methods allow very important calculations for the modern organic chemistry, from the thermodynamical properties to the vibrational spectra.

These methods, as explained before, treat the molecule as a set of particles joined by simple harmonic forces described in terms of potential energy, adding all the steric factors that have a contribution. The results is the following equation:[11]

$$E_p = E_{str} + E_{bend} + E_{tors} + E_{nb} + E_{H-bonds} + E_i$$

Where E_{str} is the energy due to the bonds deformation along the

[9] J Maple, M Hwang, T Stockfish, U Dinur, M Waldman, C Ewig, A Adler, *J Comput Chem*, **1994**, 15, 161.

[10] J Martins, A Zunger, *Phys Rev*, **1984**, 30, 6217.

[11] P Cox, *J Chem Educ*, **1982**, 59, 275.

axis (stretching or compression), E_{bend} is the energy due to the bending, E_{tors} is the energy due to the bonds torsion, E_{nb} is the energy due to non bonding interactions such as electrostatics or Van der Waals, $E_{H\text{-}bonds}$ is the energy due to the hydrogen bonds formation and E_i is the term that includes the solvent effect or other particular contributions. Each of these terms represents a possible molecule deformation from a hypothetical reference geometry.

If the length of the bond C_{sp3}-C_{sp3} free from any stress is about 1.5-2.0 Å, every deviation from this value causes an increase of the potential energy. This factor describes the bond deformation and can be expressed with the following formula:

$$E_{str} = \sum_{bonds} K_l (l - l_0)^2$$

Where K_l (kcal/mol·Å^2) is the force elastic constant, l is the bond length (Å), l_0 is the bond length of the same bond free from any perturbation and the summation is for every bond of the molecule.

The potential energy for the valency angles is described by the expression:

$$E_{bend} = \sum_{angles} K_\theta (\theta - \theta_0)^2$$

Where K_θ is the bending constant [kcal/mol·(°)^2], θ is the angle value between two next atoms (°) and θ_0 is the value of the same angle free from any perturbation.

Concerning torsion angles, the energetic contribution is described from the following formula:

$$E_{tors} = \sum_{dihedrals} K_\omega (1 + s \cos n\omega)$$

Where k_ω is the constant force which expresses the free rotation energetic hindrance (kcal/mol), ω represents the torsion angle (°), n is the periodicity of k_ω and s can assume values of +1 (minimal energy) 0 and -1 (maximum energy).

The potential energy due to the non bonding interaction (electrostatics or Van der Waals) is dependent by the distance r, and can be expressed by the formula:

$$E_{nb} = \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} + \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}}\right]$$

Where A_{ij} is the coefficient that describes the atomic repulsive interactions $(A_iA_j)^{1/2}$, B_{ij} is the coefficient that describes the atomic attractive interactions $(B_iB_j)^{1/2}$, q_i and q_j are the net charges on the atom *i* and on the atom *j*, ε is the dielectric constant of the media and R_{ij} is the distance between the atom *i* and the atom *j* (Å).

At the end, the contribution of the hydrogen bonds formations:

$$E_{H-bonds} = \sum_{i<j}\left[\frac{C_{ij}}{R_{ij}^{12}} - \frac{D_{ij}}{R_{ij}^{10}}\right]$$

Where C_{ij} is the coefficient that describes the repulsive interaction between the hydrogen atom and the acceptor, $(C_iC_j)^{1/2}$, D_{ij} is the coefficient that describes the attractive interactions between the hydrogen atom and the acceptor $(D_iD_j)^{1/2}$ and R_{ij} is the distance between the atoms (Å).

If necessary, other factors can be used to take into account the deformations outside Coulomb and solvent interactions plane.

Several empirical parameters are necessary for the force field, such as force constants and geometrical values without perturbations. These parameters can be determined by thermodynamical experiments or by diffraction experiments performed on a statistical significant number of appropriate molecules. The initial values obtained this way are often a coarse estimation and they need to be improved by a trial and error approach or by a least square method. The quality of the derived force field is defined by its ability to reproduce the data with a higher or at least equal accuracy compared to the experimental methods.[11]

For an optimal use, a set of three-dimensional coordinates of the molecule's atoms under examination is necessary, this set will be progressively modified during the calculation to reduce the energy penalty with respect to an “ideal” situation. It’s possible to assume that from all the possible conformations of the molecule, that one having the lowest energy level represents the most favourable conformation for the isolated molecule.

It is important to notice that molecular mechanics is basically an empiric method and the model obtained by using this technique is referred to an hypothetical (defined as ideal, as best compromise to fulfil the atom types definitions) state of immobility at the absolute zero K.

1.2.4 Conformation analysis

The conformational transformations a molecule can incur are mainly due to torsion angles variations because changes in angles and bond lengths are usually associated with severely higher energy penalties, except the normal vibrations, constrained within very limited ranges.

Conformational changes of a molecule can be considered as a movement in a multidimensional surface which describes the relationship between the potential energy and the molecule geometry. This surface is generally called potential energy hypersurface, or simply potential energy surface. Each point on it represents the potential energy of a given conformation of the molecule. Energetically stable conformations corresponds to local minima and the peaks on the surface the transition energy to pass from one conformation to the other.

The relative population of a conformation depends on its statistical weigh, which is influenced not only by the single value of potential energy but also by the energy barriers that separate it from all the rest of the conformational space. Consequently the absolute minima of the potential energy surface is not always correspondent to the structure having the major statistical incidence.[12]

In the "real world" the conformational change is a dynamic process heavily affected by the entropic contribution, often dominating the potential energy contribution. This is not necessarily taken into account by computational methods for conformational search and this potential limitation has to be taken into account when using information gathered by computational methods into the experimental practice.

Experimental techniques like NMR are able to supply information about one or few conformations of a molecule. A comprehensive analysis of the conformational space can be accomplished by theoretical calculations only.

For this reason a lot of theoretical methods for the conformational analysis have been developed.

The most general methods are able to identify all the minima on the potential energy surface, but their computational cost is directly dependent on the number of rotatable bonds and the angle steps considered in the simulation, thus creating a huge number of conformations as result for most molecules and requiring unacceptably long simulation time. The time necessary for a conformational analysis depends also by the method used for the energy evaluation. The quantum mechanical methods are very

[12] H Holtje, G Folkers, *Molecular Modeling, Basic Principles and Applications*, **1997**, Wiley VCH ed, Weinheim, Germany.

expensive in this sense. For these reasons most of the conformational search software are based on molecular mechanic methods for the energy calculation.

1.2.5 Systematic conformational search procedures

The systematic search represents the most simple and natural method for the conformational search, as far as it generates all the possible conformations going through the systematic variations of every torsion angle of the molecule.[12]

If the value selected for the increase of every single angle selected by this method is small enough, a complete exploration of the conformational space of the molecule is possible.

The number of the generated conformers depends on the value of the angle increase chosen, but also by the number of rotatable bonds: if *n* is the numbers of these bonds, the number of conformations grows with the nth potency, as it is easily understandable from the following formula where *N* is the number of generated conformers and *δ* is the angle increase.

$$N = \left(\frac{360°}{\delta} \right)^{n}$$

In some cases, the result in terms of different conformations is so high that it can not be easily analysed. In these cases some procedures are necessary for the reduction of obtainable conformers. The first step, in this sense, is applied before the potential energy calculation, analysing the structures and discarding the ones in which there are non bonded atoms overlapping (Van der Waals screening or bump check).

The remaining conformers undergo to potential energy calculation by molecular mechanics method, with the possibility to discard more conformers considering an energy space which excludes automatically the conformers with inappropriate energy values.

The resulting conformations represent a complete set of energetically acceptable conformers for the molecule under examination.[12]

1.2.6 Monte Carlo method or random search

The Monte Carlo method,[13] named by its creator von Neumann alluding to the capital of Monaco, executes the search of the possible structure conformations using statistical techniques. Every generated structure is randomly modified in the step after step in order to obtain a new one. The search has to start from a pre-optimized structure, this is a fundamental requirement to increment the validity of results.

On every iteration, the new torsion angles and the new cartesian coordinates are placed randomly. The resulting conformations are then minimized with a molecular mechanics method and this random process is then repeated. Every generated conformation is compared with the previous one and kept only if it is different.

This process assures, in principle, a complete exploration of all the possible regions of the conformational space, but practically this is possible only if the process is performed for a sufficient time, which can become very long, because the possibility to find a new and unique conformation decreases with the increase of the conformers number already discovered. In reality, if a conformation has already been found for n times, the probability that all the searched conformations were found is [1-(1/2)n]. Numerically, the algorithm used by this method stops the search when the same structure is found for 8 times, assuring in this way

[13] W Hastings, *Monte Carlo sampling methods using Markov chains and their applications*, Biometrikam, **1970**, 57, 97.

the theorical exploration of the 99.6 % of the conformational space.

The major advantages of this technique consists in the possibility of processing structures of any dimensions, even if molecules with high flexibility give just occasionally converging results due to the huge dimension of the conformational space.

Another advantage is the possibility of analysis of cyclic systems that are usually hardly analysable with the systematic search.

Concerning this technique, it is important to highlight that the method is able to perform conformational search with good quality for different kind of molecules, but it can require too much computational time to assure a completeexploration of the whole conformational space.

1.2.7 Molecular dynamics

This method is able to study and to explore the conformational space of a molecular structure, also a complex one, without the limitations of the systematic search and without the high computational cost of the Monte Carlo method. The principle of this method is based on the integration of the classical motion equations derived from the second Newton low.[12]

$$F_i(t) = m_i a_i(t)$$

Where F_i is the force acting on the atom *i* at the time *t*, m_i is the mass of the atom *i* and a_i is the acceleration of the atom *i* at the time *t*.

The force acting on the atom *i* can be calculated directly deriving the potential energy function *E* relative to its coordinates r_i.

$$-\left(\frac{dE}{dr_i}\right) = m_i\left(\frac{d^2 r_i}{dt^2}\right)$$

This classical motion equation allows to determine the coordinates and the velocity after a given time (trajectory). The potential energy gradient, computed by the force field, is used for the determination of the forces that act on every single atom, while the starting velocities of the atoms are generated randomly at the beginning of the simulation. This simulation, which duration can be decided by the scientist, generates a series of energetically accessible conformations for the molecule under examination.

Compared to the other techniques for performing conformation analysis, molecular dynamic has the advantage that the energetically non accessible conformers are automatically deleted, but it has also the limit that the energy barrier that separates the different conformations can be hard to overcome, thus excluding some conformational space regions. This disadvantage can be overcome using suitable temperatures that allows to overcome these barriers, increasing kinetic energy.

The atoms movement calculations are performed every discreet interval (time step) which is defined by the operator. The movements of the atoms during the time step is calculated using the Varlet[14] method, which uses the speed in the average point of the time step. Since the atom velocity is not constant during the time step, this speed is extrapolated from the speed and the acceleration values of the previous step, using the following algorithm:

$$r(t+\Delta t) = r(t) + \Delta t v(t) + \frac{1}{2}\Delta t^2 a(t)$$

[14] L Verlet, *Phys Rev*, **1968**, 165, 201.

$$a(t+\Delta t)=\frac{F(t+\Delta t)}{m}$$

$$v(t+\Delta t)=v(t)+\frac{1}{2}\Delta t\left[a(t)+a(t+\Delta t)\right]$$

Where *F(t+Δt)* is the force that acts on the atom and is calculated deriving the potential energy function at the position *r(t+Δt).*

To assure a correct integration of the equation, and reduce mistakes of system energy calculations, it is necessary that the integration interval is between 1/1000 and 1/20 of the time associated to the fastest movement which the system is subjected to.

In classical molecular dynamics, the fastest movement is associated to the bond vibration (10/100 fs). If the time steps are too high the resulting atom movement is too enhanced, on the other hand a too short time step involves the analysis of a major number of conformations that lead to an increase of the calculation time.

A good compromise can be achieved by using the shake algorithm,[15] which allows to freeze the bonds vibration movements with the hydrogen atoms (C-H, N-H, O-H, etc.). This algorithm has to be used taking care of the parameter settings, a too high number of iterations can cancel the advantage derived from using bigger time steps. Usually in protein-ligand interaction studies the durations of ps units are pre-set,[16] while the time steps used are usually from 1 to 5 fs, using the shake algorithm. In a molecular dynamics simulation the temperature is initially set to 0 K and that leads to the desired temperature (equilibration stage); then this desired temperature is kept during all the dynamics.

During the equilibration stage the velocity of all the atoms is equally modified in order to keep the population kinetically

[15] J Ryckaert, G Cicciotti, H Berendsen, *J Comput Phys,* **1977**, 23, 327.

[16] M Norin, K Hult, *Biocatalysis,* **1993**, 7, 131.

homogeneous:

$$\left(\frac{v_{new}}{v_{old}}\right)^2 = \frac{T_0}{T}$$

Where T_0 represents the temperature of experiment and T is the temperature of the system. After the equilibration, the system is kept at the constant temperature using the Berendsen method,[17] in which the velocity is put in relation with the little oscillation of the temperature by a factor λ, given from the formula:

$$\lambda = \left[1 - \frac{\Delta t}{\tau}\left(\frac{T - T_0}{T}\right)\right]^{\frac{1}{2}}$$

Where Δt is the size of the time step, τ is the relaxing time, T_0 is the simulation temperature and T is the instantaneous temperature. The temperature achieved after the equilibration step incur in some oscillation, for this reason frequently adjustments are necessary; these continuous temperature oscillations are due to the energy of the system which is affected by the atoms positions. Concerning the pressure, the procedure is analogous to the temperature.

A frequently used technique is the blocks definition: groups of atoms which positions are keep fixed during the dynamics, while the energy contributions is calculated anyway and considered in the system total calculations. This strategy is manly used in the enzyme-substrate interactions evaluation, where the aminoacids of the active site are more interesting. The main advantage is obviously the reduction of the calculation time.

[17] H Berendsen, J Postma, W van Gunsteren, A Di Nola, J Haak, *J Chem Phys*, **1984**, 81, 3684.

The chemical-physical phenomena that occur at atomic and molecular level can require different time, from few fentoseconds to several tens of seconds, to happen. Therefore the simulation time must be tuned on the basis of the phenomena to be studied.
As reported in the Figure 1.1 it is possible observe that:

- The fastest phenomenon is the electrons transfer (1fs). For an accurate description quantum mechanics equations are necessary, extremely costly in terms of computational time. Nevertheless, if the system is made by a limited numbers of atoms, this phenomenon can be simulated in a reasonable time.
- The movement of the aminoacidic chains of a protein takes picoseconds. For a very accurate simulation of this phenomenon a quantum mechanical simulation is necessary, anyway the system in this case would be too complicated. For this reason the system is simplified and simulated by molecular mechanics.
- Conformational changes are more complex phenomena and require from tens to hundreds of nanoseconds; often some approximations are necessary to simulate this phenomena in a reasonable time.
- Folding of peptides, enzymes, or nucleic acids is the most difficult phenomenon to simulate and its duration depends on the number of atoms of the system. Several hundreds of aminoacids sequences take some microseconds to fold normally, while the folding of complex system, like DNA or enzymes, takes some seconds. For this type of simulations several hundreds of processors are necessary but themost of them are still unreachable with the current computational facilities.

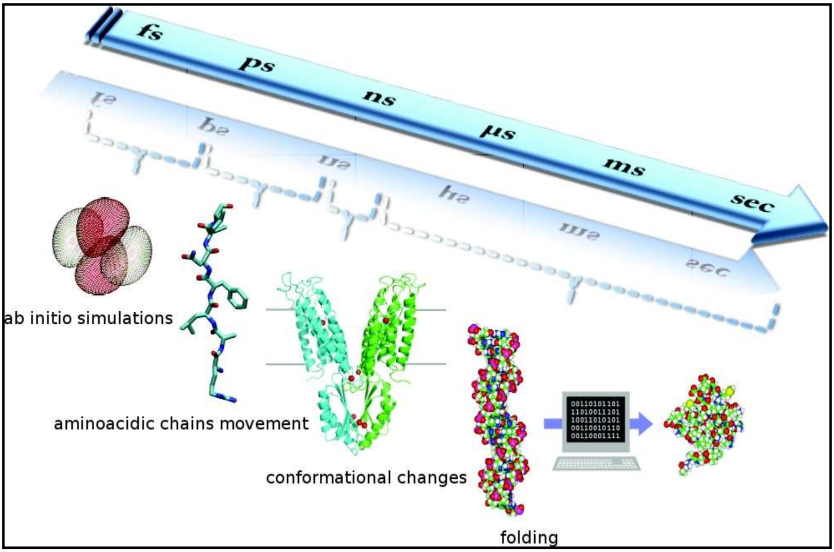


Figure 1.1: Timescale of the principle atomic and molecular movement.

1.2.8 GROMACS

GROMACS (Groningen Machine for Chemical Simulation) is an engine to perform molecular dynamics simulations and energy minimization developed by the research group of the Professor Berendsen and Professor Van Gunsteren of the Chemical department of the Groningen University during the second half of the eighties. It is a collection of libraries for molecular dynamics simulations (MD) and data analysis of the trajectories.[18, 19]

Even if the software was developed for biological molecules with complex binding interactions, the implementation of the non-

[18] H Berendsen, D van der Spoel, R van Drunen, *Comp Phys Comm*, **1995**, 91, 43.

[19] D van der Spoel, E Lindahl, B Hess, G Groenhof, A Mark, H Berendsen, *J Comput Chem*, **2005**, 26, 1701.

bounded interactions calculations makes the software suitable for any kind of molecular dynamic simulation.

GROMACS is essentially based on the GROMOS package, which was developed for the simulation of bio(macro)molecules in solution. The planned choices for the GROMACS development are:

- There are three bond type: *bond forces* based on pre-fixed lists that include four-body interactions which allow to describe more appropriately torsion angles; *non-bonding forces* based on dynamic list of of particle couples; *external forces* which take into account non-equilibrium forces.
- The calculation of the non-bonding forces is based on a couple lists which are updated every n steps. Particles are divided in charged groups, the charged group is included in the list if it is positioned within a *cut-off* radius. This procedure avoids the charge creation on a neutral group.
- Optionally a *twin cut-off* range can be used: two list are prepared, *Rshort* and *Rlons cut-off*, when the *Rshort cut-off* list is prepared the coulomb forces between particles and charged groups placed in a distance between *Rshort* and *Rlons* are calculated. These *long-range* forces are keep constant for n steps and added to the *short-range* forces.
- *Leap-frog* algorithm, which is equal to the *Verlet* algorithm,[14] is used to solve the motion equation. This involves the position in discreet time intervals, measured in time steps, and the velocities. The system is keep in constant condition of temperature and pressure.
- The length of the covalent bonds and bond angles can be limited. The resulting *constrain* equation are solved by the *Shake* algorithm,[20] which changes the constrain-free configuration in a constrained configuration moving the

[20] S. Miyamoto, P A Kollman, *J Comput Chem*, **1992**, 13, 952.

vectors trough a new direction based on a reference structure.

It is useful at this point to consider the limitations of MD simulations. The user should be aware of those limitations and always perform checks on known experimental properties to assess the accuracy of the simulation. The list of approximations can be found below.

1 - The simulations are classical

Using Newton's equation of motion automatically implies the use of classical mechanics to describe the motion of atoms. This is all right for most atoms at normal temperatures, but there are exceptions. Hydrogen atoms are quite light and the motion of protons is sometimes of essential quantum mechanical character. For example, a proton may tunnel through a potential barrier in the course of a transfer over a hydrogen bond. Such processes cannot be properly treated by classical dynamics. Helium liquid at low temperature is another example where classical mechanics breaks down. While helium may not deeply concern us, the high frequency vibrations of covalent bonds should make us worry! The statistical mechanics of a classical harmonic oscillator differs appreciably from that of a real quantum oscillator, when the resonance frequency υ approximates or exceeds k_BT/h. Now at room temperature the wavenumber $\sigma = 1/\lambda = \upsilon/c$ at which $h\upsilon = k_BT/h$ is approximately 200 cm^{-1}. Thus all frequencies higher than, say, 100 cm^{-1} may misbehave in classical simulations. This means that practically all bond and bond-angle vibrations are suspect, and even hydrogen-bonded motions as translational or librational H-bond vibrations are beyond the classical limit. What can we do? Well, apart from real quantum-dynamical simulations, we can do one of two things: (a) If we perform MD simulations using harmonic oscillators for bonds, we should make corrections to the total internal energy $U = E_{kin} + E_{pot}$ and specific heat C_V (and to entropy S and free energy A or G if those are calculated). The corrections to the energy and specific heat of a one-dimensional oscillator with frequency *υ* are:

$$U^{QM} = U^{cl} + kT\left(\frac{1}{2}x - 1 + \frac{x}{e^{x}-1}\right)$$
$$C_v^{QM} = C_v^{cl} + k\left(\frac{x^2 e^x}{(e^x-1)^2} - 1\right)$$

where $x = h\upsilon/kT$. The classical oscillator absorbs too much energy (kT), while the highfrequency

quantum oscillator is in its ground state at the zero-point energy level of *1/2hυ*. (b) We can treat the bonds (and bond angles) as constraints in the equation of motion. The rational behind this is that a quantum oscillator in its ground state resembles a constrained bond more closely than a classical oscillator. A good practical reason for this choice is that the algorithm can use larger time steps when the highest frequencies are removed. In practice the time step can be made four times as large when bonds are constrained than when they are oscillators. GROMACS has this option for the bonds and bond angles.

The flexibility of the latter is rather essential to allow for the realistic motion and coverage of configurational space.[21]

2 - Electrons are in the ground state

In MD we use a conservative force field that is a function of the positions of atoms only. This means that the electronic motions are not considered: the electrons are supposed to adjust their dynamics instantly when the atomic positions change (the Born-Oppenheimer approximation),[22] and remain in their ground state. This is really all right, almost always. But of course, electron transfer processes and electronically excited states can not be

[21] W van Gunsteren, H Berendsen, *Mol Phys*, **1977**, 34, 1311.

[22] M Born, J Oppenheimer, *Ann Phys*, **1927**, 84, 457.

treated. Neither can chemical reactions be treated properly, but there are other reasons to shy away from reactions for the time being.

3 - Force fields are approximate

Force fields provide the forces. They are not really a part of the simulation method and their parameters can be user-modified as the need arises or knowledge improves. But the form of the forces that can be used in a particular program is subject to limitations. The force field that is incorporated in GROMACS is described in Chapter 4. In the present version the force field is pair-additive (apart from long-range coulomb forces), it cannot incorporate polarizabilities, and it does not contain fine-tuning of bonded interactions. This urges the inclusion of some limitations in this list below. For the rest it is quite useful and fairly reliable for bio macro-molecules in aqueous solution!

4 - The force field is pair-additive

This means that all non-bonded forces result from the sum of non-bonded pair interactions. Non pair-additive interactions, the most important example of which is interaction through atomic polarizability, are represented by effective pair potentials. Only average non pairadditive contributions are incorporated. This also means that the pair interactions are not pure, i.e., they are not valid for isolated pairs or for situations that differ appreciably from the test systems on which the models were parametrized. In fact, the effective pair potentials are not that bad in practice. But the omission of polarizability also means that electrons in atoms do not provide a dielectric constant as they should. For example, real liquid alkanes have a dielectric constant of slightly more than 2, which reduce the long-range electrostatic interaction between (partial) charges. Thus the simulations will exaggerate the long-range Coulomb terms. Luckily, the next item compensates this effect a bit.

5 - Long-range interactions are cutoff

In this version GROMACS always uses a cutoff radius for the Lennard-Jones interactions[23] and sometimes for the Coulomb interactions as well. Due to the minimum-image convention (only one image of each particle in the periodic boundary conditions is considered for a pair interaction), the cutoff range can not exceed half the box size. That is still pretty big for large systems, and trouble is only expected for systems containing charged particles. But then truly bad things can happen, like accumulation of charges at the cutoff boundary or very wrong energies! For such systems you should consider using one of the implemented long-range electrostatic algorithms, such as particle-mesh Ewald.[24]

6 - Boundary conditions are unnatural

Since system size is small (even 10,000 particles is small), a cluster of particles will have a lot of unwanted boundary with its environment (vacuum). This we must avoid if we wish to simulate a bulk system. So we use periodic boundary conditions, to avoid real phase boundaries. But liquids are not crystals, so something unnatural remains. This item is mentioned last because it is the least of the evils. For large systems the errors are small, but for small systems with a lot of internal spatial correlation, the periodic boundaries may enhance internal correlation. In that case, beware and test the influence of system size. This is especially important when using lattice sums for long-range electrostatics, since these are known to sometimes introduce extra ordering.

1.2.9 Minimization methods

The potential energy of a molecule is directly correlated to its geometry and to its chemical characteristics. The minimization

[23] J Lennard-Jones, *Proceedings of the Physical Society*, **1931**, 43, 461.
[24] T Darden, D York, L Pedersen, *J Chem Phys*, **1993**, 98, 10089.

process consists in iterative mathematical operations in order to optimize the structural geometry and reach the coordinates set corresponding to the energy minima.

Three problems are present in the potential energy function analysis:

- The choice of the initial direction of optimization, critical because the space is multidimensional;
- The research of the minimum number of steps to reach the nearest minima. It is important to remember that after every coordinate variation the force field has to be re-applied for the potential energy calculation;
- The choice of the mathematical method for the determination of the reaching of the minima (converging criteria).

Traditional minimization methods are able to search only for the nearest minima. Critical factor in this sense is the starting conformation: the only way to find the absolute minimum is to make a conformational search to obtain one or several starting conformation(s) for the minimization.

The methods used for these processes can be divided in two categories on the basis of the type of algorithm used:

- Non derivative methods: the most used is the simplex method,[24] based on a very light mathematical algorithm in terms of calculation complexity. It is scarcely efficient, its application is restricted to cases where the potential energy surface is extremely complex. This method acts on every atom until the forces are under a certain value.
- Derivative methods: as already seen, one of the fundamental requirement for a function describing a force field is being continuous and differentiable in every point. In fact, from the analysis of the first and second derivatives of these mathematical functions we can have information about the topology of the potential energy surface. In this topology three main approaches exist:

Steepest descent, Conjugated gradient and Truncated Newton.[25]

The Steepest descendent method searches the minimum going to the direction of the maximum slope on the potential energy surface. This method is called *line searching* and it acts by changing the direction of search always perpendicularly. This is not the best minimization algorithm and is not accurately convergent, but it can be used taking into account that the result will not be very close to the minimum. It can be defined as an approaching algorithm.

The Conjugated gradient is an evolution of the previous one, it uses the line searching method as before for the pathway optimization, but in this case every step is stored to avoid that the same pathway is covered for a second time. This process is more expensive because the pathway chosen at each step follows the analysis of the previous steps. The increase in efficiency justifies the increase of computational cost.

The truncated Newton method uses the gradient for the direction identification and a curvature function (second derivative). This method is used when the minimum is near the starting point or when the function is almost harmonic, otherwise divergences are possible (going far away from the minimum).

A geometrical and energetic criteria are taken into account when reaching the minimum. The gradient of these criteria is analysed, when the gradient is zero the minimum is reached.

For the theoretical achievement of the converge criteria, a high number of steps is necessary, for this reason a value close to zero is set.

[25] W Press, B Flannery, S Teukolski, W Vetterling, *The Art of Scientific Computing*, **1988**,Cambridge University Press, Cambridge, UK.

1.2.10 Homology alignment

Homology alignment is a bioinformatic technique that determines the correspondence of two or more protein chains. The main assumption is that the two chains are related.

There are different data banks and tools for the similarity search of the sequences (BLAST, FASTA), and other for the pairwise alignment (LFASTA, WISE, SIM) and multiple alignment (ClustalW, MAP).

The instrument used for this research work is MOE-align,[26] which can align several protein sequences at the same time (multiple sequence alignment tool). The software is also able to use information coming from the primary structure (secondary structure prevision) to perform multiple sequences alignment. The information based on the structure can be also used when the structural information are not available for all the protein chains; therefore MOE-align can work with mixed sets of information about sequences and structures. The alignment can be optimized using constrains in the structure alignment and using a manual repositioning of residues.

MOE-align is the modified version of the original alignment methodology introduced into molecular biology by Needleman and Wunsch.[27] In this method the alignments are computed optimizing the base scoring function based on the similarity of the residues (obtained applying an aminoacid substitution matrix to the residue aligned couple) and gap penalties. Gap penalties are set in order to introduce and extend gaps in one sequence respecting the other one. The optimized final value is defined as alignment score. MOE includes a matrix that derives directly from the family of the aligned proteins, e.g. a matrix that derives from

[26] Molecular Operating Environment versione **2006.08**, Chemical Computing Group, Montreal, Canada.

[27] S B Needleman, C D Wunsch, *J Mol Biol,* **1970**, 48, 443.

the application of an evolution model to the matrix set correlated. Moe-align is able to calculate alternated matrices of similarity using the Needleman-Wunsch procedure as well as gap-penalties for specific position. For example, MOE-align can adapt similarity matrices and gap penalties using the secondary structure prevision or the real secondary structure.[28]

Since the Needleman-Wunsch procedure allows the alignment of an arbitrary number of sequences, the calculations become computationally expensive when the number of chains is higher than four. For this reason multiple align protocols act redoing the pairwise alignment of the groups of already aligned chains. Nevertheless, the theoretical difficulty comes from the scoring of the new gaps. MOE-align perform the multiple alignment in four steps:

1. Pairwise initial building: the starting evaluation of the alignment can be calculated in two different modes:

- Progressive: align the chains 1 and 2, then align to the result the chain 3 and so on until all the chains are aligned
- Sensible to the chain order; more expensive in computational terms

2. Round-robin realignment: the initial alignment evaluation (or the evaluation of the present alignment) is improved with a series of single round-robin alignments, where every chain is extracted from the global alignment and realigned to the chains left.

3. Random iterative refinement: the results of the initial alignment and of the round-robin can be sensible to the order in which the chains are processed. To reduce this dependency, the series of alignment can be calculated dividing randomly the chains in two groups and aligned independently. If the new alignment has a better score it is accepted, otherwise no.

4. Structure based realignment: the chains that contain the information about their α-carbons can be re-aligned taking into

[28] W Kabsch, C Sander, *Biopolymers*, **1983**, 22, 2577.

account the spatial information of their structure. From the starting evaluation the method then generates a new similarity matrix using the relative coordinates of the C-α which results from a multi-body overlapping. This matrix is used for the realignment of the chains populated by C-α only. This operation is repeated until the Root Mean Square Distance (RMSD) of the superimposition is no more improving. At this point the chains are reintroduced as indivisible units not computed yet and the steps from 1 to 3 are repeated.

1.2.11 Homology modelling

To perform the homology modelling in this work, the MOE-homology tool was used. The tool use a comparative modelling procedure to build complete models with all atoms of the sequence based on one or more structure template.

MOE-homology consist of three steps:

1. Initial specification of the partial geometry: an initial partial geometry for the sequence is specified. The MOE-homology copy the geometry of the regions from one or more template chains. The residue identity is preserved between the template and the model and all its coordinates are copied, or just the backbone coordinates are used.

2. Building of intermediate model: the independent models of the target protein structure are build up using the Boltzmann-weighted randomized modelling procedure used by Levitt,[29] combined with a specialized logic for the correct treatment of the residues different from the template.[30] Everyone of those intermediate models is evaluated by the residue packing quality function which is sensible to the exposition rank of the non polar side chains and

[29] M Levitt, *J Mol Biol*, **1992**, 226, 507.

[30] T Fechteler, U Dengler, D Schomberg, *J Mol Biol,* **1995**, 253, 114.

to the opportunity to hydrogen bond formation. First the list of molecular data are collected in order to model the missing atoms.

- Indels list: list of backbone fragment that allow the inclusions in the target sequence;
- Sidechain list: list of the alternative conformations of the side chain belonging to residues with non modelled atoms:
- Outgap list: list of backbone fragments to model residues which can need more exceptions of N- and C- terminal in the template chain.

The Indel data are collected searching backbone segments through the high resolution chains of the protein data bank (PDB),[31] which are well overlapping with the ending residues at the end of the adding area. During this segment searching the back shifting of the indel regions is possible if no one segment agree with the RMSD criteria.[29] The data on the side chain are built starting from a large rotamers collection, generated with a systematic grouping of the high resolution PDB data.

After the data collecttion, a number of independent models is created. First the loops in random order are modelled. For every loop the contact energy function analyse the list of candidates collected in the segment collecting step, taking in account all the atoms already modelled and all the user specified one because are members of the modelling environment (e.g. ligand bonded to the template).

Those energies are used for choosing the candidates generated by the Boltzmann-weighted procedure, which the coordinates are copied into the model. Every missing atom is modelled using the same procedures. The sidechain atoms of the residues which the coordinates are copied from the template are modelled first, followed by the sidechain loops. The outgaps and their sidechains are the last to be modelled;

[31] The UniProt consortium, *Nucleic Acids Res,* **2006**, 36, 190.

3. Building of the final model: The coordinates of the final model are generated as the average of the coordinates of the modelled atoms and as the coordinates of the model with the best *peacking quality function*.

The final procedure is to check the *Protein Report* to verify the presence of *outliers* (parameters that are outside allowable values). Bonds length, angles, dihedral angles and atoms contacts are the major parameters reported in the *Protein Report*. In case of *outlier*, restrained local minimizations can be applied for corrections. Ω angles, chirality of α carbons, X angles bond agles of the main chains, and bond length constitute another set of important parameter for the model quality evaluation. They are compared to data of statistical maps from *Protein Data Bank*.[32] Non-bonding forces between heavy atoms are tested in relation with the sum of their Van deer Waals interactions. Two type of *outlier* are not defined by limit values: Ψ and Φ angles, which could be simply placed in "not allowed" regions of the Ramachandran plot and they must be corrected during the model refinement.

1.2.12 Docking simulations

All the procedures named molecular docking include all the simulations where molecules approach each other, aimed to the study of their interactions. This technique is mainly based on the analysis of the electrostatic and steric interactions of the involved species. Docking allows, for instance, the placement of a substrate into an enzymatic active site. Therefore this type of analysis is able to provide very relevant information for the identification of the most important structural elements that permit an optimal

[32] R Laskowski, A Moss, S Thornton, *J Mol Biol*, **1993**, 231, 1049.

interaction concerning either the receptor and the ligandl. This information, for instance, is useful for engineering new ligands with better affinity or for the creation of plausible hypothesis for action mechanisms.

Docking applications concern different research fields:

- Interactions between macromolecular receptor and ligand with low molecular weight (i.e. enzyme-substrate).
- Interactions between macromolecular receptor and macromolecular ligand (i.e. protein-protein, DNA-protein, DNA-DNA).
- Interaction between low molecular weight receptor and low molecular weight ligand (i.e. inclusions).

Fundamental requirement for this technique is the knowledge of the three-dimensional structure of the target molecule, which has to be highly defined to assure high quality results. Docking usually takes place searching the best interaction between a rigid macromolecular target (usually a protein) and a mobile and flexible ligand of limited size.

Docking can be applied on the whole macromolecular target, but the research of bonding is usually focused on a smaller and specific area named *site*.

A lot of different docking software have been developed like DOCK, FlexX, Combi-DOCK, and they use different algorithms for the scoring of the poses they calculate. In this work the software MOE was used.

The docking analysis consists in several steps:

Conformational analysis: to search different ligand conformations.

Placement: different poses of a single ligand conformation are generated. This application uses different placement techniques, everyone of those with different properties. These methods are deterministics and it is normal that *outputs* change from a calculation to another.

The available methodologies are:

- *Alpha Triangle:* it is the standard method. Positions are generated by overlapping the coordinates of the ligand atoms with the coordinates of the receptorial site represented by alpha spheres. Alpha spheres are virtual spheres, which are generated inside receptorial pockets. They does not contain any atoms inside and they are big enough to allow the ligand placement at bond distance from the protein atoms. A binding site is described as a set of spheres in contact with each other. At every iteration a ligand conformation is randomly chosen, and the algorithm tries to overlap the ligand atoms with a randomly chosen set of alpha spheres.
- *Triangle Matcher:* this method generates conformations aligning the coordinates of the ligand atoms with the coordinates of the alpha sphere in a systematic mode.
- *Alpha PMI:* it is a method that generates positions aligning significantly ligand conformations with randomly chosen alpha spheres. This is the best method for narrow receptorial pockets, it is a fast method and the research space is more strictly defined.

1. Pharmacophore filtering: it is possible to limit the freedom of conformations generation forcing determined interactions in order to satisfy an arbitrary condition to respect an eventually parmacophoric group. This set can be used as a docking filter: all the results that does not satisfy this condition will be delete.

2. scoring: every pose generated with the placement methodology is subjected to a *score* (scoring) which identifies the most favourite poses. Typically, the scoring functions highlight hydrophobic contacts, ionic contacts and hydrogen bonds.

Available methodologies are:

- *London dG Scoring:* it is the function that calculates the free energy of binding of the ligand in that pose. The functional form is the sum of the following terms:

$$\Delta G = c + E_{flex} + \sum_{h-bonds} c_{HB}\, f_{HB} + \sum_{m-lig} c_M\, f_M + \sum_{atoms\ i} \Delta D_i$$

Where *c* represents average gain/lose of rotational and translational entropy; E_{flex} is the energy due to the lost flexibility of the ligand (calculated by the topology); f_{hb} measures the geometrical imperfections of the hydrogen bonds and it can assumes the values [0,1]; c_{HB} is the energy of an ideal hydrogen bond; f_M measures the imperfections of the coordination bonds and it can assume the values [0,1]; c_M is the energy of an ideal coordination bond; D_i is the desolvation energy of the atom *i*. The difference of the desolvation energies is calculated by the formula:

$$\Delta D_i = c_i R_i^3 \left\{ \iiint_{u \notin A \cup B} |\, u \,|^{-6}\; du - \iiint_{u \notin B} |\, u \,|^{-6}\; du \right\}$$

Where *A* and *B* are the volumes of the protein and/or of the ligand with the atom *i* belonging to the volume *B*; R_i is the solvation diameter of the atom *i* (taken as parameter OPLS-AA Var der Waals σ plus 0.5 Å); and c_i is the desolvation coefficient of the atom *i*. Coefficients {c, c_{HB}, c_M, c_i} are obtained from~400 x-ray crystal structures of protein-ligand complexes with available experimental pKi. Atoms are categorized in~12 atom type for the assignment of the c_i coefficients. Integrals are rounded using the generalized formulas of the Born integral.

- *Affinity dG Scoring:* it is the function that calculates the enthalpy at the bound free energy using the linear function:

$$G = C_{hb}\, f_{hb} + C_{ion}\, f_{ion} + C_{mlig}\, f_{mlig} + C_{hh}\, f_{hh} + C_{hp}\, f_{hp} + C_{aa}\, f_{aa}$$

Where *f* defines the fraction of the specific atomic contacts and the *C* coefficients measure the contribution of the terms of the affinity evaluation. The individual terms are:

I. *hb*: interaction between a couple of donor-acceptor of hydrogen bonds; it is supposed that two hydroxides group interact in the most favourable mode.
II. *ion*: ionic interactions; the Coulombian term is used for the evaluation of interactions between charged groups, which can contribute to rise or decrease the bond affinity.
III. *mlig*: coordination bond; interactions between Nitrogen/Sulphur and transition metals are calculated as coordination bonds.
IV. *hh*: hydrophobic interactions; i.e. the interactions among aliphatic carbons. These interactions are usually favourable.
V. *hp*: interactions between hydrophobic and polar atoms; these interactions are usually unfavourable.
VI. *aa*: interactions between any atoms; these interactions are usually weak but favourable.

- *Depth HB Scoring:* this scoring is a linear combination between two terms. The first term measures how much deep is posed the ligand into the active site and it is the sum of all the atomic scores of the ligand. The score is roundly the fraction of the volume of a sphere of 5 Å radius, this sphere is centred on the atom and the volume is the volume occupied by the receptor atoms. The second term measures the effects of hydrogen bonds. Score of +2 is assigned if the site is occupied by the favourable atom. Otherwise, if it is occupied by any other ligand atom, the assigned score is -1. Concerning donor and acceptor sp3, all the favourable atoms within 3.5 Å contribute with a score of +1; while all the others contribute with a -1 score. Metals on receptors are treated like acceptor but with a triplicate effect.

1.3 GRID

The Molecular Discovery programs[33] are used to predict specific non-covalent interactions between a molecule with known three-dimensional structure (*target*) and a small chemical group (*probe*), mimicking various chemical functionalities. The employed probes can be characterized by different nature and include, among the others, water, methyl group, amino nitrogen, carbonyl and hydroxyl. The procedure builds a three-dimensional grid all around the target molecule and it calculates the interaction energy between the target and the probe at every grid knot. The calculation output is a three-dimensional matrix of interaction energies named Molecular Interaction Field (MIF). Energies are computed basing on attractive and repulsive forces, their combination leads to simple functions for energy minima visualization, which corresponds to a favourable non-bond organization of atoms and molecules. This is expressed by the Lennard-Jones function:[23]

$$E_{lj} = A / d^{12} - B / d^{6}$$

Where d is the distance between the non-bonded atom couple, which potential energy of Lennard-Jones E_{lj} is described by the empirical parameters A and B.[33]

A huge number of target can be studied, including enzymes, nucleic acids, polysaccharides, glycoproteins, peptides, membranes, crystals, drugs and a lot of other organic compounds.

The MIF can be used in many different ways, the most obvious one is its visualisation as isopotential energetic surface. It can also be used as molecular descriptor in structure-activity correlation studies.

[33] P Goodford, *J Med Chem,* **1985**, 28, 849.

1.4 Protein Data Bank

As stated above, it is important to know the three-dimensional structure of molecules under examination. These structures are a fundamental starting point for most computational studies. Their definition plays a key role for the success of the undertaken research. Therefore, structures solved by x-ray crystallography as well as structures solved by other techniques like 2D/3D-NMR and homology modelling are generally used.
The Protein Data Bank (PDB),[34] that collects three-dimensional structures coming from x-ray or NMR studies is an extremely source of data for molecular modelling. PDB is supported by the *Research Collaboratory for Structural Bioinformatics* (RCSB). It is a no-profit consortium which is aimed to improve the comprehension of biological systems by the study of the three-dimensional structures of macromolecules. Actually 56635 structures are freely available.

1.5 QSAR

QSAR methods (*Quantitative structure-activity relationship*) try to figure out, for a compound set, a correlation between chemical-physics properties and a generally defined activity, which can be any measurable properties of the compound under study. Generally a QSAR equation is a linear equation:

$$Activity = Cost + (c_1 P_1) + (c_2 P_2) + ...(c_n P_n)$$

Where the parameters P_i are calculated for each molecule of the

[34] H Berman, J Westbrook, Z Feng, G Gilliland, T Bhat, H Weissig, I Shindyalov, P Bourne, *Nucleic Acids Res,* **2000**, 28, 235.

set by a computer, and the coefficients c_i are calculated by a correlation between the parameters variations and the variations of activity.

The correlation between structure and activity needs a mathematical expression of molecular structure, that is usually called molecular descriptor.

A lot of parameters can be used as a molecular descriptor: for instance, the *Hansch* approach[35] (defining the beginning of QSAR paradigm) uses electronic and structural characteristics of the molecule, described by the σ parameter.

In every QSAR the parameters choice is the first and most important step. The success is strictly correlated with the use of appropriate molecular descriptors. Only if the chosen parameters and the activities are closely related a model able to predict the activity is feasible. The QSAR techniques depend on the assumption that each compound of the series interacts with the target molecule in the same way. Since the activity depends from the affinity of the ligand for the receptor, which is a function of the ligand structure, QSAR can be used for receptor binding studies, which is the standard QSAR application.

The main problem of QSAR models is that they are not *ab-initio* methods: a series of experimental values are necessary to build the model. Therefore the model is trained to predict a certain type of structure-activity relationship and they are not able to predict properties of compounds with no reference to the structural variability which is represented by compounds of the training set. In other words is not possible to predict the activity of a molecule totally different from the molecules of the training set. Moreover, if a model is built with high activity molecules it will not be precise in the prediction of low activity molecules and viceversa.

Essential requirements in order to obtain a good QSAR model are

[35] H Kubinyi, R Mannhold, P Krogsgaard-Larser, H Timmerman, *QSAR: Hansch Analysisand Related Approaches*, **1993**, VCH, Weinheim.

the choice of an appropriate data set and a good description of the chemical properties of the molecules under examination.

The 3D-QSAR techniques (based on three dimensional descriptors) have the advantage of considering a more refined model of compounds compared with bidimensional methods. Information related to the spatial component of the molecule allows to obtain more realistic systems and therefore to extract more accurate molecular description. This generally improves the predictivity of the outcoming models.

1.6 Chemometrics-Multivariate analysis

Chemometrics is a science based on mathematical and statistical methods for the resolution of multivariate chemicals problems. It can be defined as the application of mathematics, statistics and graphical methods to chemistry in order to maximize the extraction of information from data. The term multivariate analysis indicates an approach that considers more than one variable at the same time. Multivariate analysis are used in 3D-QSAR studies in order to summarize all the informations which are contained in the variables matrix. Important and most used techniques to perform multivariate analysis are PCA (principal component analysis)[36] and PLS (partial least square or projection to latent structures).[37]

1.6.1 Principal components analysis (PCA)

In almost any 3D-QSAR, especially the ones based on the GRID method, the number of molecular descriptors is very high. The

[36] R N Carey, S Wold, J O Westgard, *Anal Chem*, **1975**, 47, 1824.

[37] A Hoskuldsson, *J Chemom*, **1988**, 2, 211.

PCA method reduces this number of original variables by combining them in a series of latent variables called Principal Components. By doing so, the method preserves the intrinsic variability of the original variables in terms of physical-chemical information they contain. Moreover, this method is mostly useful to understand the differences among the studied compounds and estimate the quality of the produced molecular descriptors. It is in fact important to verify that the selected description of the involved compounds can be able to discriminate between two clearly distinct ones. For a GRID analysis the PCA method can be useful to identify variables that contain similar information from those that contribute differently to the description of the compounds.

Two interesting characteristics of the Principal Components are that they are orthogonal to each other, and there is no correlation between the information contained in each of them and the second one is that they are extracted in and order of importance, meaning that the first PC contains more information that the second one and so on.

These characteristics allow to overcome the general limitations present in the multiple linear regression method in which it is important that the variables are independent one form another and that the number of objects exceeds the number of variables.

In the PCA method, the user can choose how many PC he wants to extract. Nevertheless this number may not be big since generally the first five PCs contain around 90% of the model's variance.

1.6.2 Projection on latent structures (PLS)

PLS are methods that are used to generate regression models. In the regression models, the correlation between the molecular descriptors and the experimentally obtained biological activity is achieved by the extraction of the Principal Components in the presence of the activity data. In other words, the extracted PCs are not only the ones responsible for the explanation of the maximal variability of the system, but also the ones that show the best correlation to the experimental data. Therefore the PCs extracted in this procedure are somewhat different to the ones of the PCA method in a way that they have to meet the necessity to maximize the correlation between the molecular descriptors and the activity values of the compounds in the data set.

This multivariate statistical method is the mostly used one in the QSAR studies. Nevertheless, the predictivity of the model is clearly attributed to the existence of a tight correlation between the calculated and the measured properties.

1.6.3 Model validation

Once the model has been obtained, it has to be evaluated in terms of its quality and predictivity. This is usually done in two ways: via an internal (or cross) validation or an external validation.

The cross-validation in based on reduced models that do not contain the entire data set which are then used to predict the properties of the excluded objects. The predicted properties are confronted to the experimentally obtained ones, and the goodness of the prediction is evaluated by different indicators such as the r^2 (correlation coefficient), the SDEP (standard deviation of error prediction) and the q^2 (prediction correlation coefficient).

Another type of validation is the external validation in which the complete 3D-QSAR model is used for the prediction of the activity of one or more new compounds (a so called *test set*) , not included in the initial data set. These new compounds are built and their activity is experimentally measured and compared to the activity predicted by the model.

CHAPTER 2

AIM

The use of enzymatic reactions to catalyse chemical transformations in commercial processes, in competition with conventional chemical catalysis, is becoming increasingly convenient and affordable to many industries in recent years. Biocatalytic reactions are generally more energy efficient, have lower cost, and produce less hazardous waste than chemical catalytic reactions. Biocatalysts are used in many sectors, including the food, textile, pharmaceutical, chemical and energy industries.

There is still a technological gap, basically due to the need to understand many of the mechanisms at the basis of biocatalysed reactions, that limits the diffusion of industrial application of biocatalysis. The most efficient strategy to face this limitations is probably the synergy between dry and wet labs' investigations. It is clear that the development of new computational methods, specifically thought to be complementary to the experimental activity, represent a greatly interesting field.

The work of this thesis is aimed at the development of computational methods for simulating enzymes under operational conditions simulating chemical systems in as “real” (realistic) as possible operating environments.

The work was initially focused on a well known enzyme class, the lipases. The development of computational methodologies started with the analyses of the activation mechanisms of these enzymes. A number of different lipases, extremely eterogeneous from the structural point of view, but very similar from the mechanism of action and general characteristics, was taken into account. First the physical-chemical properties of the enzymes surfaces were investigated with the GRID methods analysis in order to find common features. Afterwards lipases activation was studied by applying different types of molecular dynamics (MD) simulations. The opportunity to investigate the conformational possibilities of enzymes in chemically defined environments makes MD simulations a suitable technique in order to understand common activation mechanisms. Classical MD approaches were applied together with other strategies, such as steered MD and coarse grained force field based MD, to tackle different aspects of lipase application to synthetic processes.

In particular, the potentiality of the MARTINI force field was exploited. This particular coarse grained force field was applied to investigations of lipases orientation at the oil/water interface as well as to the study of lipase stability.

While MD demonstrated its potential in the investigation of solvent dependent enzyme behaviour and to the study of the dynamics of activation/inactivation, more complex problems, like the study of enzyme-substrate interactions, require different computational approaches to be efficiently investigated. The study of lipase enantio-selectivity was studied by the application of a hybrid method based on the combination of both MD simulations and 3D-àQSAR approach, based on chemometric analysis.

Finally, the promising results got with lipases pushed the need to verify the general applicability of the concepts. For this reason, the same approaches were used during the investigation of a completely different enzyme, the Alkanesulfonate monooxygenase. This enzyme was chosen because it catalyses very attractive reactions from an industrial point of view. Moreover enzymatic redox reactions are not well diffused in industrial applications, basically because of the lack of knowledge on monooxygenase and other related enzyme classes.

Computational studies object of this thesis should be able to describe how the enzyme can be affected by the surrounding environment and to predict properties like enzyme stability, conformational changes and substrate selectivity.

The idea is providing answers to the industrial and academic requirements in terms of information concerning not only the isolated biocatalyst, but especially focusing on conditions of its real utilization. Obviously to be really complementary (and useful) to the experimental practice, computational tools must be competitive in terms of time and costs, therefore major attention have been put on the control of computational cost, exploring the simplification of the models on the basis of the various research goals.

CHAPTER 3

LIPASES:
Physical-Chemical Characteristics and Activation Mechanisms

3.1 Introduction

Lipases constitute an important group of biotechnologically valuable enzymes, mainly because of the versatility of their properties and ease of mass production. They are triacylglycerol ester hydrolases (EC 3.1.1.3) that catalyze the hydrolysis of long-chain acylglycerols. Lipases are widely diversified in their enzymatic properties and substrate specificity, making them very attractive for industrial applications. In the industrial segment, lipases and cellulases are anticipated to post the best gains. It is expected that in the next few years lipases will benefit from their versatility and continued penetration into the detergent and cosmetics markets. Cellulases, which share lipases' versatility, will continue to be used to emulate the stone-washing of denim while making substantial gains in the pulp and paper industry as bleaching and lignin-removal agents. Lipases and cellulases, like most specialty and industrial enzymes, will increasingly be produced via recombinant DNA technology.[1]

In fine chemistry, lipases are valued biocatalysts because they act under mild conditions, are highly stable in organic solvents, show broad substrate specificity, and usually show high regio- and/or stereoselectivity in catalysis. The usefulness of bacterial lipase in commerce and research stems from its physiological and physical properties.[2]

Bacterial lipases are generally more stable than animal or plant lipases. They are active under ambient conditions reducing the energy cost required for high temperature and pressures processes, avoiding at the same time the instability of temperature labile reactants and products. In the industrial application, lipases share the general advantages of biocatalysis over traditional synthetic processes of the reduction of side products, milder conditions,

[1] K E Jeager, B W Dijsktra, M T Reetz, *Annu Rev Microbiol*, **1999**, 53, 315.

[2] E A Snellman, E R Sullivan, R R Colwell, *FEBS J*, **2002**, 269, 5771.

reduction of wastes, offering cost-effectiveness over with traditional downstream processing. Moreover, their remarkable stability in organic solvents represents a plus for the industrial applicability.[3]

The hydrolytic activity of most lipases, but not esterases, is enhanced hugely upon contact with a lipid–water interface,[4, 5] a phenomenon known as interfacial activation.[6, 7]

Three-dimensional structures of lipases coming from a wide variety of sources[8] help to understand this property. Thus, a partial explanation of interfacial activation comes from the presence of an amphiphilic flexible lid,[9,10] a protein domain switching from a so-called closed conformation (or inactive state) coving the active-site entrance and an open conformation (or active state) allowing full access to the inner part of the pocket to substrates. Though this evidence might demonstrate a very simple activation mechanism, the structural basis of lipase interfacial activation, i.e. the distinction between lipases and esterases, is intrinsically much more complex, as it can be deduced from the fact that not all lipases with a lid domain exhibit this behavior[11] and, conversely, there are lipases without a lid, that show interfacial activation.[12] For these reasons, lipases can be defined pragmatically as esterases that act on long-chain acylglycerols.

[3] F Hasan, A A Shah, A Hameed, *Enz Microb Technol*, **2006**, 39, 235.

[4] H L Brockman, J H law, F J Kézdy, *J Biol Chem*, **1973**, 248, 4965.

[5] L Sarda, P Desnuelle, *Biochem Biophys Acta*, **1958**, 30, 513.

[6] P Desnuelle, L Sarda, G Alihaud, *Biochem Biophys Acta*, **1958**, 37, 570.

[7] R Verger, *Methods Enzymol*, **1980**, 64, 340.

[8] J Pleiss, M Fischer, M Peiker, C Thiele, R D Schmid, *J Mol Catal sect B,* **2000**, 10, 491.

[9] J D Schrag, M Cygler, *Methods Enzymol*, **1997**, 284, 85.

[10] K E Jaeger, M T Reetz, *Trends Biotechnol*, **1998**, 16, 396.

[11] M Nardini, B W Dijsktra, *Curr Opin Struct Biol*, **1999**, 9, 732.

[12] J C Chen, L J Miercke, J Krucinski, J R Starr, G Saenz, X Wang, *Biochemistry*, **1998**, 37, 5107.

Lipases are widely used for industrial purposes. They are efficient stereoselective catalysts in the kinetic resolution of a wide variety of chiral compounds[13] and are useful in transesterification, synthesis of esters and peptides, and resolution of racemic mixtures to produce various optically active compounds. Several organochemical and crystallographic studies have provided some insight into their enantioselectivity.[14] On the basis of these studies, a general rule for the enantiopreference towards the production of a secondary alcohol, and the positioning of the scissile fatty acid chain and ester bond has been proposed.[15]

Lipases are, in general, highly variable in size and the sequence similarity among them is limited to short spans located around the active-site residues. However, the three-dimensional structures of lipases, in their cores, share a common fold motif, known as an α/β hydrolase fold.[16] This α/β hydrolase fold has been identified in many other distantly or closely-related enzymes.

The general α/β hydrolase fold (Figure 3.1) consists of eight central, mostly parallel β sheet strands of which the second strand is antiparallel. The parallel β3 to β8 strands are connected by α helices, packing on either side of central β sheet. The β sheet has a left-handed superhelical twist such that the surface of the sheet covers about half a cylinder and the first and last strands cross each other at an angle of around 90°. The curvature of the β sheet may differ significantly among the various enzymes, and also, the spatial positions of topologically equivalent α helices may vary considerably. They differ substantially in length and architecture, in agreement with the large substrate diversity of these enzyme.

[13] E Santaniello, P Ferraboschi, P Grisenti, A Manzocchi, *Chem Rev*, **1992**, 92, 1071.

[14] R J Kazlauskas, A N E Weissfloch, A T Rappaport, L A Cuccia, *J Org Chem*, **1991**, 56, 2656.

[15] M Cygler, A H Gupta, *J Am Chem Soc*, **1994**, 116, 3180.

[16] D L Ollis, A Goldman, *Protein Eng*, **1992**, 5 197.

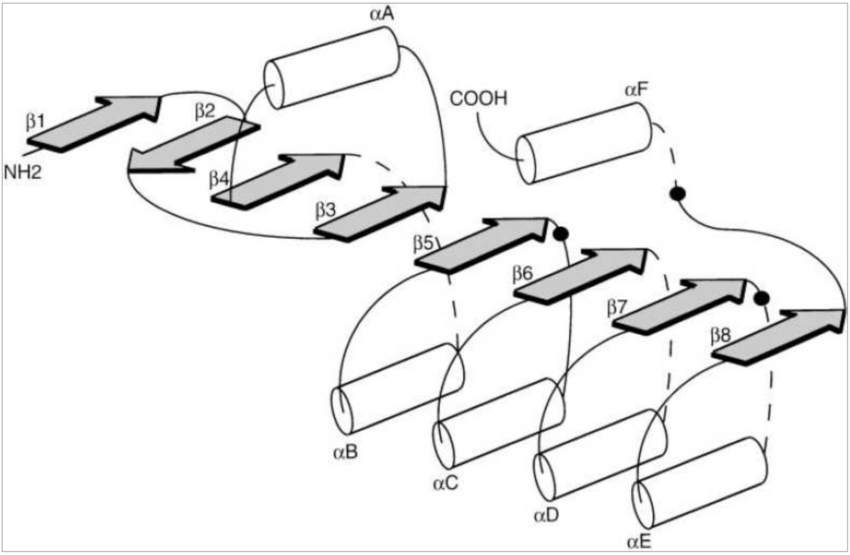


Figure 3.1: Canonical α/β fold. α Helices are indicated by cylinders and β strands are indicated by shaded harrows. The topological position of the active site residues is shown by solid circles; the nucleophile is the residue after β strand 5, the Asp/Glu residue is after β strand 7, and the histidine residue is in the loop between β8 and αF.

Lipases are hydrolases acting on the hydrolysis of ester bonds connecting fatty acids and glycerol. Their active site consists of a Ser-His-Asp/Glu catalytic triad. This catalytic triad is similar to that observed in serine proteases, and therefore lipases catalysis is thought to proceed along a similar pathway. Hydrolysis takes place in two steps (Figure 3.2). It starts with an attack by the oxygen atom of the hydroxyl group of the nucleophilic serine on the activated carbonyl carbon of the susceptible lipid ester bond (Figure 3.2). A transient tetrahedral intermediate is formed, which is characterized by a negative charge on the carbonyl oxygen atom of the scissile ester bond and four atoms bonded to the carbonyl carbon atom arranged as a tetrahedron (Figure 3.2). The intermediate is stabilized by the helix macrodipole of helix C (Figure 3.1), and hydrogen bonds between the negatively charged carbonyl oxygen atom (the “oxyanion”) and at least two main-chain NH of OH groups (the “oxyanion hole”). One of the NH groups is from the residue just behind the nucleophilic serine; the other one is usually from the residue at the end of strand β3.[17] The nucleophilicity of the attacking serine is enhanced by the catalytic histidine, to which a proton from the serine hydroxyl group is transferred. This proton transfer is facilitated by the presence of the catalytic acid, which precisely orients the imidazole ring of the histidine and partly neutralizes the charge that develops on it. Subsequently, the proton is donated to the ester oxygen of the susceptible bond, which is cleaved. At this stage the acid component of the substrate is esterified to the nucleophilic serine (the “covalent intermediate”), whereas the alcohol component diffuses away (Figure 3.2). The next stage is the deacylation step, in which a water molecule hydrolyzes the covalent intermediate. The active-site histidine activates this water molecule by drawing a proton from it. The resulting OH^- ion attacks the carbonyl carbon atom of the acyl group covalently attached to the serine (Figure 3.2). Again, a transient negatively charged tetrahedral intermediate is formed, which is stabilized by interactions with the oxyanion hole. The histidine donates a proton to the oxygen atom of the active serine residue, which then releases the acyl component. After diffusion of the acyl product the enzyme is ready for another round of catalysis.[18, 19]

[17] R J Kazlauskas, *Trends Biotechnol*, **1994**, 12, 464.

[18] K H G Verschueren, F Seljée, H J Rozeboom, K H Kalk, B W Dijkastra, *Nature*, **1993**, 363, 693.

[19] A M Brozozowski, L Thim, *Nature*, **1991**, 351, 491.

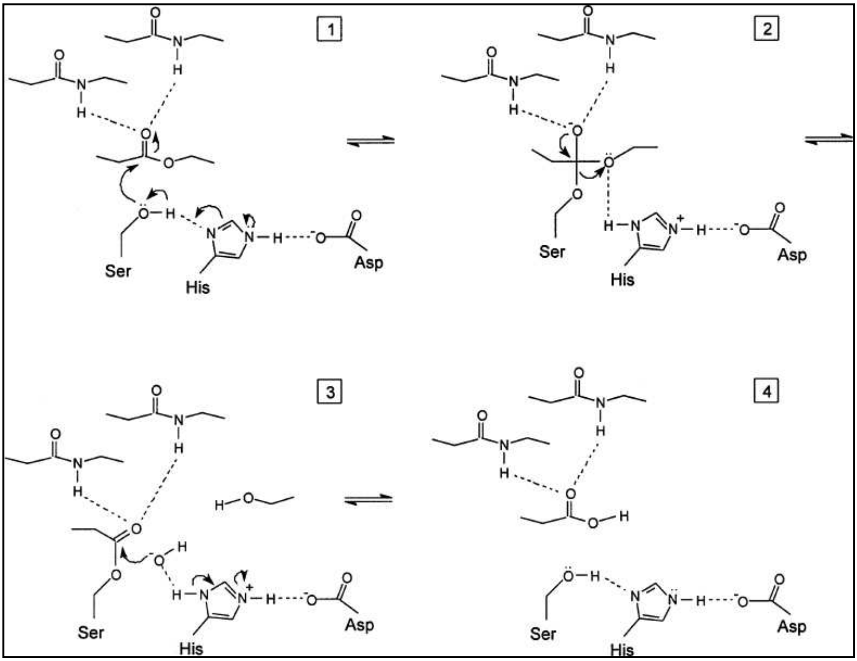


Figure 3.2: Reaction mechanism of lipases. [1] Binding of lipid, activation of nucleophilic serine residue by neighboring histidine and nucleophilic attack of the substrate's carbonyl carbon atom by Ser O^{-}. [2] Transient tetrahedral intermediate, with O^{-} stabilized by interactions with two peptide NH groups. The histidine donates a proton to the leaving alcohol component of the substrate. [3] The covalent intermediate ("acyl enzyme"), in which the acid component of the substrate is esterified to the enzyme's serine residue. The incoming water molecule is activated by the neighboring histidine residue, and the resulting hydroxyl ion performs a nucleophilic attack on the carbonyl carbon atom of the covalent intermediate. [4] The histidine residue donates a proton to the oxygen atom of the active serine residue, the ester bond between serine and acyl component is broken, and the acyl product is released.

3.2 Choice of the lipases for the study

For a deeper investigation of enzymatic activation phenomena of lipases we chose to study an heterogeneous group of members of this enzyme class. This way it is possible to obtain different data related to lipases from different taxa and genus, with the aim to have a sample as homogeneous and representative as possible.

The enzymes object of this study are listed in the table below (Table 1.1). For the most of them the crystal structure was available, at least for one of the two conformations; for three of them (*Candida rugosa* lipase; *Humicola lanuginosa* lipase; *Rhizomucor miehei* lipase) the two conformations were both available; for two of them (*Pseudomonas fluorescens* lipase; *Rhizopus oryzae* lipase) it was not possible to find any crystal structure, therefore homology modeling procedures were necessary.

LIPASE	TAXON
Bacillus subtilis lipase	Bacteria
Burkholderia cepacia (*Pseudomonas cepacia*) lipase	Bacteria
Candida antarctica lipase B (CaLB)	Yeast
Candida rugosa lipase	Yeast
Geotrichum candidum (*Botrytis geotricha*; *Torula geotricha*) lipase	Fungus
Humicola lanuginosa (*Termomyces lanuginosus; Monotospora lanuginosa; Sepedonium lanuginosum*) lipase	Fungus
Pseudomonas aeruginosa lipase	Bacteria
Pseudomonas fluorescens lipase	Bacteria
Rhizomucor miehei (*Mucor miehei*) lipase	Fungus
Rhizopus niveus lipase	Fungus
Rhizopus oryzae lipase	Fungus

Table 1.1: List of used lipases and their taxa; in black the lipases with only one conformation available, in blue lipases with both conformations available, in red lipase structures obtained by homology modelling.

3.2.1 *Bacillus subtilis* lipase

Several structures of *Bacillus subtilis* extracellular lipase (BsL) are deposited in the PDB. Among them, 1ISP is the most accurate with a resolution of 1.3 Å. This structure was used in this thesis for all the studies concerning *Bacillus subtilis* lipase.
BsL is encoded by the lipA gene,[20] has a molecular weight of 19,4 kDa, which is exceptionally low for a member of the bacterial lipase family; the range is generally 30-75 kDa. BsL is stable even

[20] V Dartois, A Baulard, K Schanck, C Colson, *Biochim Biophys Acta*, **1992**, 1131, 253.

under highly alkaline conditions (pH 12) and has optimal activity at pH 10; it is therefore regarded as an alka-liphilic lipase.[21] The activities of ordinary lipases are known to be enhanced greatly in the presence of their substrate lipid micelle,[22] implying that lipases act on their substrates at the lipid-water interface (so-called “interfacial activation”). The enzymatic activity of BsL, however, does not depend on the formation of the substrate micelle, indicating that BsL possesses no inter-facial activation, and takes place even at a low concentration of the substrate. Because of these unique characteristics, BsL is thought to be widely applicable to industrial uses.

The three-dimensional structures of BsL variants have been determined by X-ray crystallography (Figure 3.3).[23]

[21] E Lesuisse, K Schanck, C Colson, *Eur J Biochem*, **1993**, 216, 155.
[22] L Sarda, P Desnuelle, *Biochim Biophys Acta*, **1958**, 30, 513.
[23] Z S Derewenda, *Adv Protein Chem*, **1994**, 45, 1.

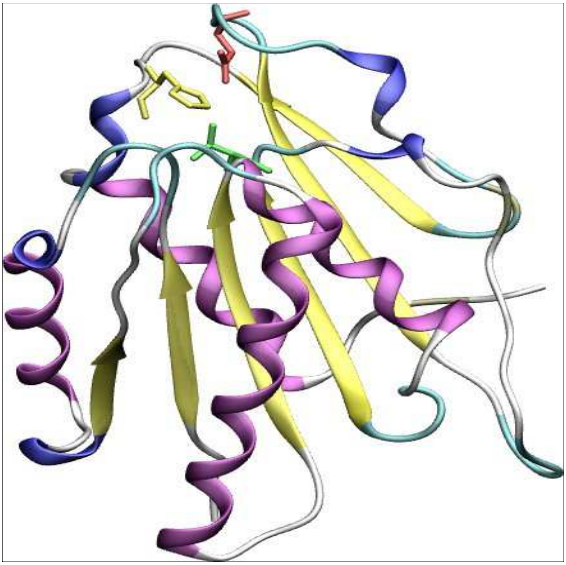


Figure 3.3: Structure of *Bacillus subtilis* lipase (PDB 1ISP) in new cartoon representation, coloured by secondary structure; the catalytic triad is highlighted in licorice mode: Ser in green, His in yellow and Asp in red.

They share a common topology named the α/β-hydrolase fold[24] consisting of a six to eight-stranded parallel β-sheets. The β-sheet is connected through α-helices, which are generally located surrounding the β-sheet. The active site of lipase is constructed in the C-terminal portion of the β-sheet and consists of Ser, His and Asp (the catalytic triad). In ordinary lipases, the active site is covered by a hydrophobic `lid' consisting of one or two α-helices,[25, 26] in this case there is no lid, and this can be the

[24] D L Ollis, E Cheah, M Cygler, B Dijkstra, F Frolow, S M Franken, M Harel, S J Remington, I Silman, J Schrag, J L Sussman, K H G Verschueren, A Goldman, *Protein Eng*, **1992**, 5, 197.

[25] A M Brzozowski, U Derewenda, Z S Derewenda, G G Dodson, D M Lawson, J P Turkenburg, F Bjorkling, B Huge-Jensen, S A Patkar, L Thim, *Nature*, **1991**, 351, 491.

[26] U Derewenda, A M Brzozowski, D M Lawson, Z S Derewenda,

motivation of the absence of interfacial activation.[27]

3.2.2 *Burkolderia cepacia* lipase

The lipase from *Burkolderia cepacia* (*Pseudomonas cepacia,* PcL) represents a widely applied biocatalyst for highly enantioselective resolution of chiral secondary alcohols. Its stereopreference is determined predominantly by the substrate structure, while stereoselectivity depends on atomic details of interactions between substrate and lipase.[28]

Several structure of this enzyme, in the open conformation only, are available from PDB. 1YS1 was chosen to be used in the study because of its highest resolution (1.10 Å).

The structure of PcL (Figure 3.4) is made up by 320 aminoacid residues (33 kDa) and shows an highly open conformation, typical for most bacterial lipases, which is likely to represent the active state of the enzyme at an oil–water interface.

Biochemistry, **1992**, 31, 1532.

[27] K Kosei, K Hidemasa, S Mamoru, O Satoru, T Sakae, *Acta Cryst*, **2002**, D58, 1168.

[28] T Schulz, J Pleiss, R D Schmid, *Protein Sci*, **2000**, 9, 1053.

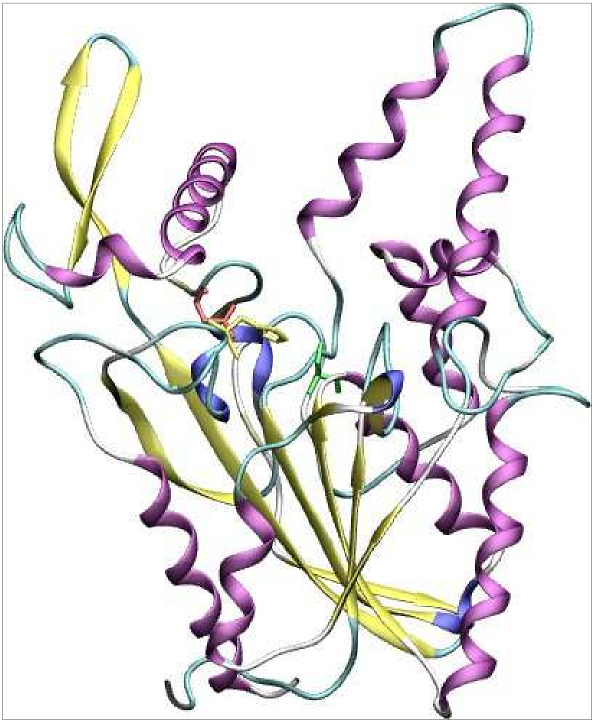


Figure 3.4: Structure of *Bulkolderia cepacia* lipase (PDB 1YS1) open conformation in new cartoon mode coloured by secondary structure; the catalytic triad is highlighted in licorice mode: Ser in green, His in yellow and Asp in red.

PcL is a globular enzyme with approximate dimensions of 30 Å × 40 Å × 50 Å; the comparison of its structure with the general hydrolase fold points out an additional strand, lined up with the sixth strand, but in the opposite direction.[29]

The active site Ser (Ser-His-Asp represent the catalytic triad) lies at the bottom of a cleft in the protein and is probably fully exposed to the solvent when the enzyme is in solution. The

[29] K K Kyeong, K S Hyun, H S Dong, Y H Kwang, W S Se, *Structure*, **1997**, 5, 173.

entrance of the cleft has an ovoid shape and it is 10 Å × 25 Å across and about 15 Å deep.[30]

3.2.3 *Candida antarctica* lipase B

Candida antarctica lipase B (CaLB) is an efficient catalyst for hydrolysis in water and esterification in organic solvents.[31] It is used in many industrial applications because of its high enantioselectivity, wide range of substrates, thermal stability, and stability in organic solvents.[32] Different structures of CaLB are available from PDB, 1TCA is the most accurate one with a resolution of 1.55 Å; this was the structure used in all the investigation about this enzyme.

CaLB (Figure 3.5) is a lipase with 33 kDa molecular weight and belongs to the α/β hydrolase fold family with a conserved catalytic triad consisting of Ser, His, and Asp.[33, 34]

[30] J D Schrag, Y Li1, M Cygler1, D Lang, T Burgdorf, H J Hecht, R Schmid, D Schomburg, T J Rydel, J D Oliver, L C Strickland, C M Dunaway, S B Larson, J Day, A McPherson, *Structure*, **1997**, 5, 187.

[31] E M Anderson, K M Larsson, O Kirk, *Biocatal. Biotrnsform*, **1998**, 16, 181.

[32] D Rotticci, J C Rotticci-Mulder, S Denman, T Norin, K Hult, *ChemBioChem*, **2001**, 2, 766.

[33] J Uppenberg, M T Hansen, S Patkar, T A Jones, *Structure*, **1994**, 2, 293.

[34] J Uppenberg, N Ohrner, M Norin, K Hult, G J Kleywegt, S Patkar, V Waagen, T Anthonsen, T A Jones, *Biochemistry*, **1995**, 34, 16838.

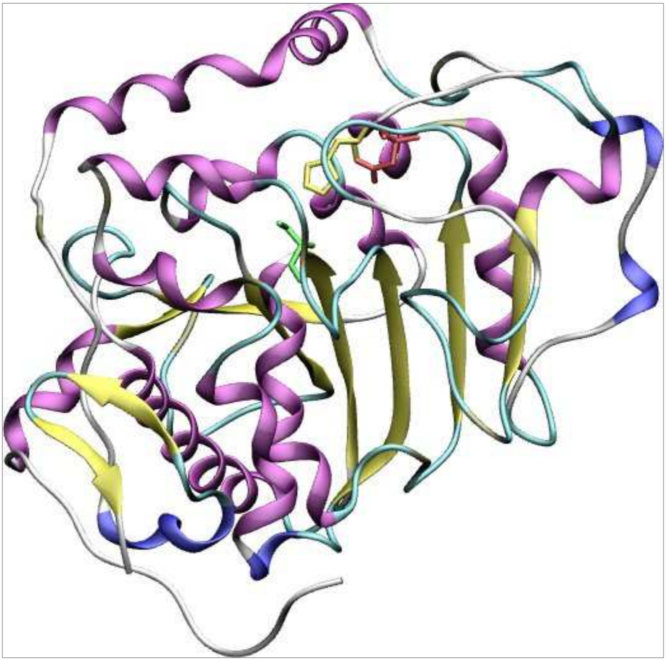


Figure 3.5: Structure of *Candida antarctica* lipase B (PDB 1TCA) in new cartoon mode coloured by secondary structure; the catalytic triad is highlighted in licorice mode: Ser in green, His in yellow and Asp in red.

The binding pocket for the substrates consists of an acyl-binding pocket, a large and a medium binding pocket for the small and large moiety of secondary alcohols, respectively. In contrast to most lipases, CaLB has very small lid which is not big enough to cover the entrance to the active site and therefore the enzyme shows no interfacial activation.[35, 36]

[35] M Martinelle, M Holmquist, K Hult, *Biochim Biophys Acta*, **1995**, 1258, 272.
[36] P Trodler, J Pleiss, *BMC Struct Biol*, **2008**, 8, 9.

3.2.4 *Candida rugosa* lipase

Lipase from *Candida rugosa* (CrL) is a versatile biocatalyst which catalyzes hydrolysis, alcoholysis, esterification and transesterification of triacylglycerols and other hydrophobic esters. It is widely applied in a variety of biotechnological applications as diverse as production of carbohydrate esters of fatty acids, stereoselective synthesis of pharmaceuticals and a multitude of applications in food and flavour production.[37] Structures of this enzyme in its open and closed conformation are available from PDB. Structures 1CRL and 1GZ7 representing open and closed conformation respectively, were used in the study.

Candida rugosa expresses a mixture of lipase isoforms which differ in substrate specificities. Each gene codes for a 534 amino acid residue polypeptide chain, with molecular masses of around 60 kDa (Figure 3.6).[38]

[37] A Padney, S Benjamin, C R Soccol, P Nigam, N Krieger, V T Soccol, *Biotechnol Appl Biochem*, **1999**, 29, 119.

[38] J M Mancheno, M A Pernas, M J Martinez, B Ochoa, M L Rua, J A Hermoso, *J Mol Biol*, **2003**, 332, 1059.

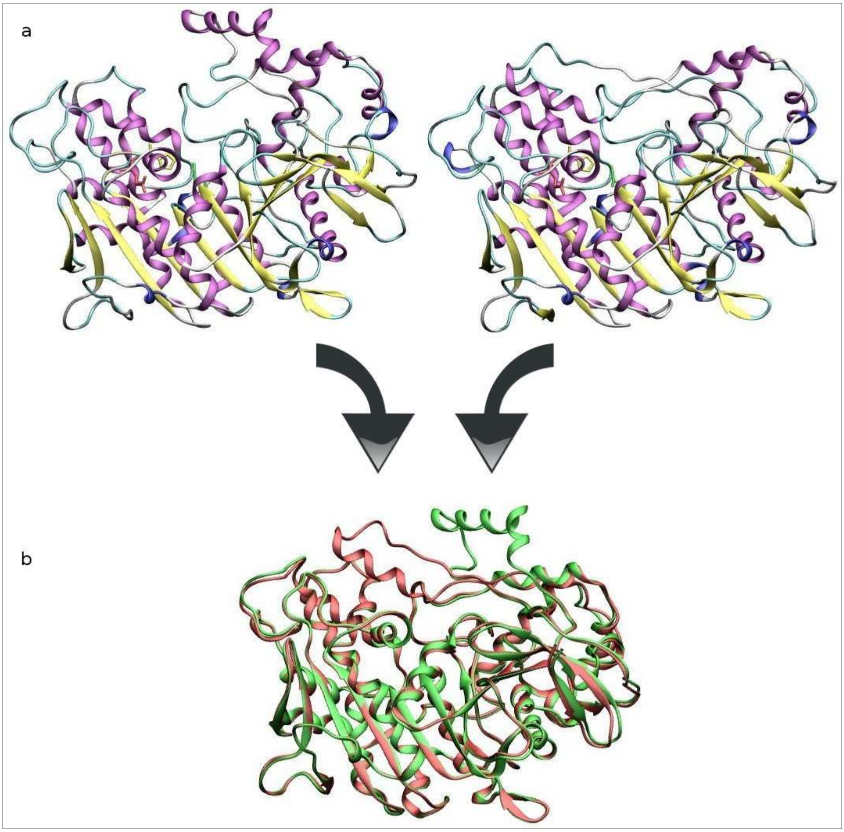


Figure 3.6: Structure of *Candida rugosa* lipase a- in open (on the left) and closed (on the right) conformation (PDB 1CRL and 1GZ7 respectively) in new cartoon mode coloured by secondary structure; b- the two conformation overlapped, the open one in green and the closed one in red.

Like other microbial lipases, CrL is a member of the α/β hydrolase fold family. A mobile element covers the catalytic site in the inactive form of the lipase. In the open, active form the lid moves away and makes the binding site accessible to the substrate.[39] Lid movement is clearly showed in figure 3.6.

[39] J Schmitt, S Brocca, R D Schmid, J Pleiss, *Protein Eng*, **2002**, 15, 595.

3.2.5 *Geotrichum candinum* lipase

Lipases produced by the fungus *Geotrichum candidum* belong to the class of big lipases (approximately 60 kDa) with significant amino acid similarity to many esterases.[40] Early reports regarding the substrate specificity of *G. candidum* lipase (GcL) indicated its preference for long chain fatty acids.[41]
Only one structure of this enzyme, in its closed conformation, is available from PDB with the code 1THG (Figure 3.7).

[40] M Cygler, J D Schrag, J L Sussman, M Harel, I Silman, M K Gentry, B P Doctor, *Protein Sci*, **1993**, 2, 366.
[41] R G Jensen, J Sampugna, J G Guinn, D L Carpenter, T A Marks, *J Am Chem SOC*, **1965**, 42, 1029.

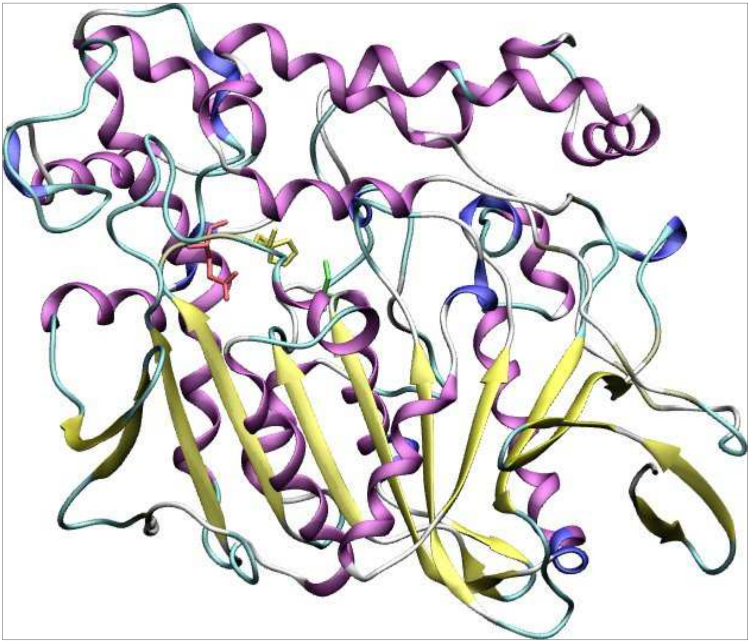


Figure 3.7: Structure of *Geotricum candidum* lipase (PDB 1THG) closed conformation in new cartoon mode coloured by secondary structure; the catalytic triad is highlighted in licorice mode: Ser in green, His in yellow and Asp in red.

GcL is another member of α/β hydrolase fold family, it has a big mobile lid, that can cover the active site in the closed inactive conformation; the catalytic triad is represented by Ser-His-Glu.[42]

3.2.6 *Humicola lanuginosa* lipase

Structures of *Humicola lanuginosa* lipase (HlL) in open and closed conformations are available from PDB; structures with codes 1DTE and 1TIB for open and closed conformation

[42] J D Schrag, M Cygler, *J Mol Biol*, **1993**, 239, 575.

respectively are the most accurate ones and were used for all the calculations about this protein.

HlL is an enzyme of 30 kDa and 219 amino acids[43] which consists of a single, roughly spherical domain containing a central eight-stranded, predominately parallel β-pleated sheet and five interconnecting α-helices, compacted to a volume of approx. $9{,}7x10^3$ $Å^3$. The active site of HlL is composed of a Ser-His-Asp catalytic triad[44] (Figure 3.8).

[43] A M Brzozowski, H Savage, C S Verma, J P Turkenburg, D M Lawson, A Svendsen, S Patkar, *Biochemistry*, **2000**, 39, 15071.

[44] K Zhu, A Jutila, E K J Tuominen, S A Patkar, A Svendsen, P K J Kinnunen, *Biochim Biophys Acta*, **2001**, 1547, 329.

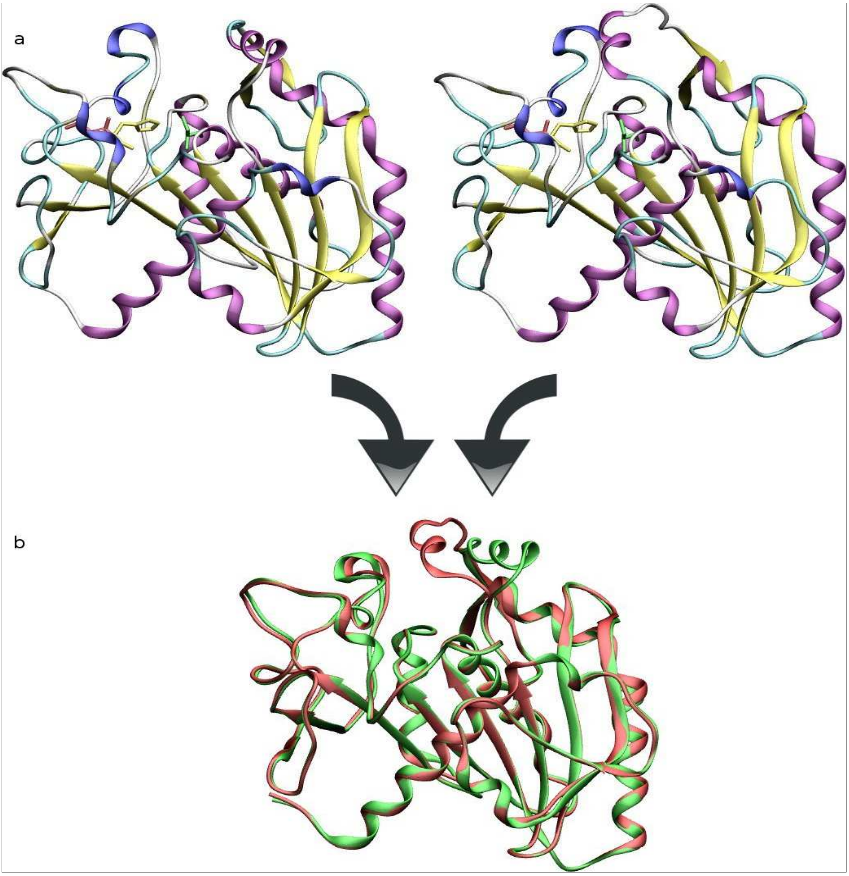


Figure 3.8: Structure of *Humicola lanuginosa* lipase a- in open (on the left) and closed (on the right) conformation (PDB 1DTE and 1TIB respectively) in new cartoon mode coloured by secondary structure; b- the two conformation overlapped, the open one in green and the closed one in red.

The coparison of the two crystal structures shows that the main difference by the two enzyme conformations is represented by the flexible lid domain.

3.2.7 *Pseudomonas aeruginosa* lipase

The structure of *Pseudomonas aeruginosa* lipase (PaL) in open conformation is available from PDB with code 1EX9. This is the only one available for this lipase.

PaL has a nearly globular shape with approximate dimensions of 35x40x50 Å. Its structure consists of a "core" domain, showing the typical features of the α/β hydrolase fold topology,[45] and a "cap" domain, with four α-helices that shape the active site cleft (Figure 3.9).

[45] P Heikinheimo, A Goldman, C Jeffries, D L Ollis, *Structure*, **1999**, 7, 141.

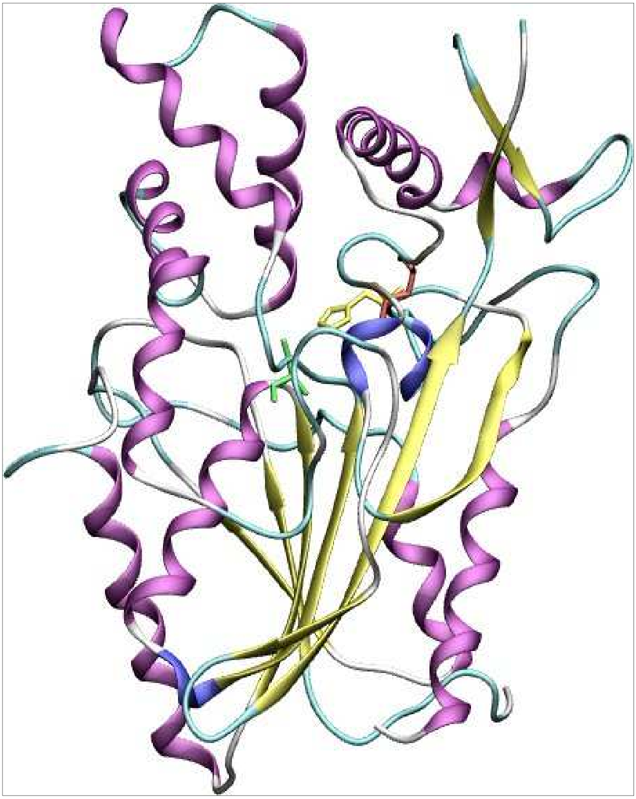


Figure 3.9: Structure of *Pseudomonas aeruginosa* lipase (PDB 1EX9) open conformation in new cartoon mode coloured by secondary structure; the catalytic triad is highlighted in licorice mode: Ser in green, His in yellow and Asp in red.

PaL structure is similar to the lipase structures from Burkolderia glumae, Burkolderia cepacia, and Chromobacterium viscosum, which show 42% amino acid sequence identity to PaL.[46, 47] The structural similarity is mainly localized in the core domain, where

[46] D Lang, B Hofmann, L Haalck, H J Hecht, F Spener, R D Schmid, D Schomburg, *J Mol Biol*, **1996**, 259, 704.

[47] D A Lang, M L M Mannesse, G H De Haas, H M Verheij, B W Dijkstra, *Eur J Biochem*, **1998**, 254, 333.

the secondary structure elements have an almost equal structural distribution .[48]

3.2.8 *Rhizomucor miehei* lipase

Rhizomucor miehei lipase (RmL) is a single chain protein consisting of 269 aminoacids with a total molecular weight of 29 kDa and an isoelectric point of 3.5. RmL is probably one of the most widely used fungal lipase.[49] The stereochemistry of the oxyanion hole of RmL is not clear. It has been proposed, on the basis of X-ray crystallography studies, that the oxyanion hole of RmL can exist only in the open, active conformation.[50]

Structure of this enzyme in its open and closed conformation are available from PDB; 4TGL for the open and 3TGL for the closed conformation respectively are the most accurate structures for this protein and were used for the studies (Figure 3.10).

[48] M Nardini, D A Lang, K Liebeton, K E Jaeger, B W Djkstra, *J Biol Chem*, **2000**, 275, 31219.

[49] B Folmer, K Holmberg, M Svensson, *Langmuir*, **1997**, 13, 5864.

[50] M Norin, F Haeffner, A Achour, T Norin, K Hult, *Protein Sci*, **1994**, 3, 1493.

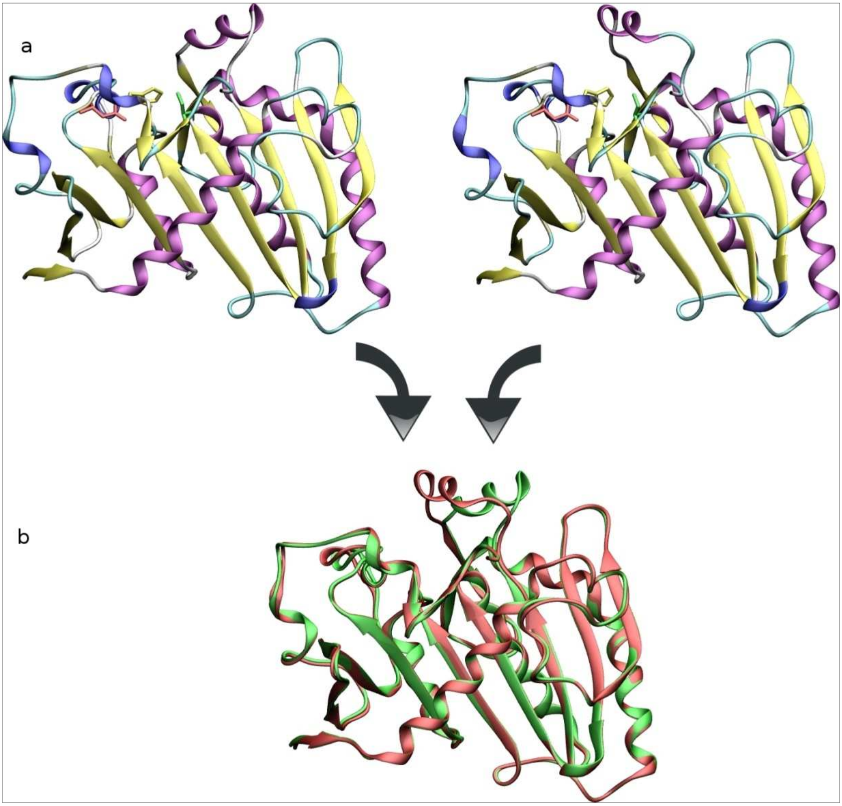


Figure 3.10: Structure of *Rhizomucor miehei* lipase a- in open (on the left) and closed (on the right) conformation (PDB 4TGL and 3TGL respectively) in new cartoon mode coloured by secondary structure; b- the two conformation overlapped, the open one in green and the closed one in red.

Once again the main difference between the two enzyme conformation is well showed in figure 3.10 and is represented by the lid domain.

3.2.9 *Rhizopus niveus* lipase

The structure of *Rhizopul niveus* lipase (RnL) in its open conformation is available from PDB with code 1LGY and it is the only one available.

The mature form of the enzyme (lipase II) is 269 aminoacids long. Lipase II come from lipase I form, which contains two polypeptide chains combined through non-covalent interaction. The structure of Lipase II (Figure 3.11) shows a typical α/β hydrolase fold containing the so-called nucleophilic elbow (a conserved lipase domain between strand β5 and helix α4 where is usually located the catalytis Ser, this domain is located deep within the core of classical lipase structure showed in figure 3.1). The catalytic center of this enzyme is analogous to those of other neutral lipases and serine proteases.[51]

[51] M Kono, J Funatsu, B Mikami, W Kugimiya, Y Morita, *J Biochem*, **1996**, 120, 505.

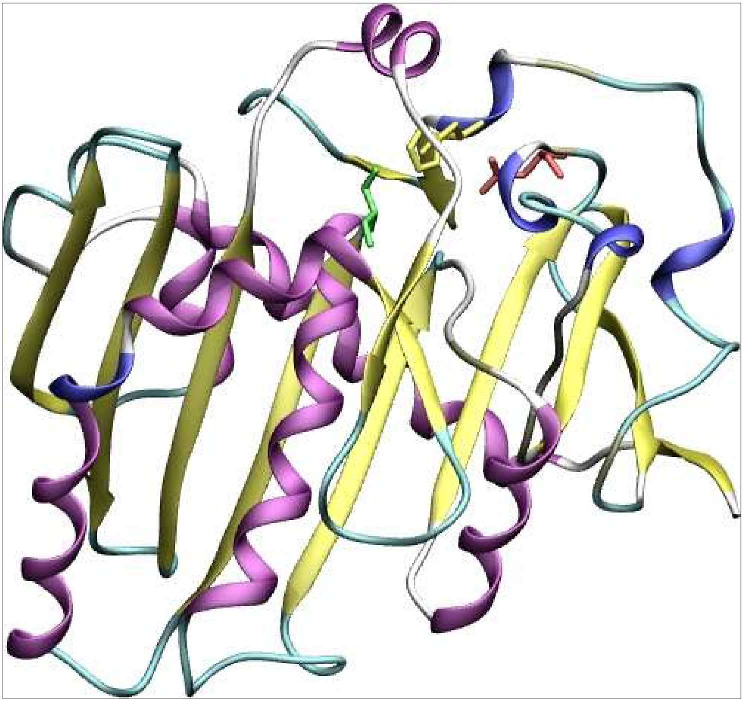


Figure 3.11: Structure of *Rhizopus niveus* lipase (PDB 1LGY) open conformation in new cartoon mode coloured by secondary structure; the catalytic triad is highlighted in licorice mode: Ser in green, His in yellow and Asp in red.

3.3 Homology modelling

In several cases only the aminoacidic sequence of an enzyme is available. In these cases three-dimensional structure can be calculated by homology modelling. The first step of this procedure is the homology alignment; which is a bioinformatic technique that determines the correspondence between protein sequences. Therefore, if the protein of unknown structure has a strong sequence correspondence (> 70%) with a protein which structure is known, it will be possible to build the homology structure. The known structure is used as a template to build up the structure of the other enzyme. Afterwards, refinement processes are needed in order to relax the structure and achieve a structure of acceptable quality.

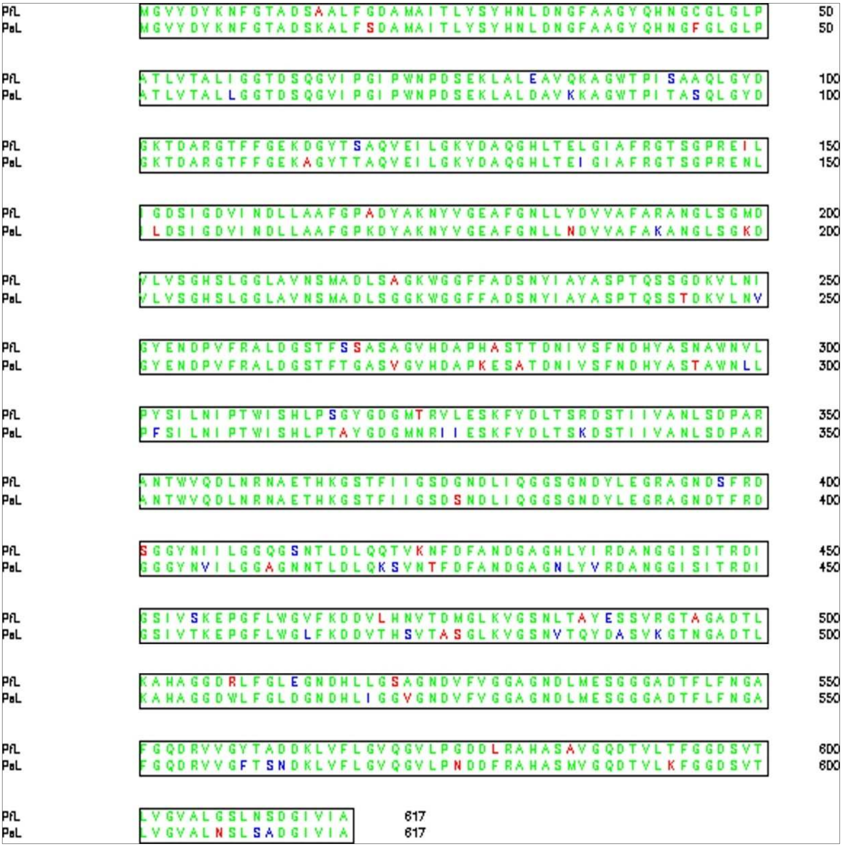
3.3.1 *Pseudomonas fluorescens* lipase

The aminoacidic sequence of *Pseudomonas fluorescens* lipase (PfL) is available on UniProtKB[52] database with the code Q0PM63_PSEFL and its length is 617 aminoacids. A homology search of the aminoacid sequence was performed on SRS@EBI[53] website. PfL is 88 % homologous to *Pseudomonas sp.* MIS30 lipase (PsL) as it can be seen in Figure 3.12. The structure of PsL is available on PDB with the code 2Z8X. Starting from the structure of PsL, used as template, a homology model of the structure of PfL was generated. The sequences of the two enzymes were first aligned with the MOE align tool. Then the residues of the catalytic triad of PsL (Ser28, His30, Glu77) were constrained to the corresponding residues of PfL in order to keep the spatial geometry of the catalytic machinery. Ten homology models were generated, the structure with the highest score was selected for the next steps. The generated structure was minimized and analysed with the protein report tool which takes into account the allowed geometrical parameters of the residues. The PfL catalytic triad is Ser 207, His 313, Asp 255. The model refinement was performed by total and local energy minimisation and local molecular dynamics simulations in order to achieve acceptable protein report

[52] The UniProt Consortium, *Nucleic Acids Res*, **2006**, 36, 190.
[53] N Harte, V Silventoinen, E Quevillon, S Robinson, K Kallio, X Fustero, P Patel, P Jokinen, R Lopez, *Nucleic Acid Res*, **2004**, 18, W3.

parameters for all the aminoacids.

The final structure (figure 3.14) was evaluated by means of Ramachandran plot (figure 3.13). Only eight outliers are present and they are generally close to allowed regions of the plot; the Ramachandran high score indicates the general quality of the model, in fact 87% of the residues fall in the core region.



```
PfL   MGVYDYKNFGTADSAALFGDAMAITLYSYHNLDNGFAAGYQHNGCGLGLP   50
PsL   MGVYDYKNFGTADSKALFSDAMAITLYSYHNLDNGFAAGYQHNGFGLGLP   50

PfL   ATLVTALIGGTDSQGVIPGIPWNPDSEKLALEAVQKAGWTPISAAQLGYD   100
PsL   ATLVTALLGGTDSQGVIPGIPWNPDSEKLALDAVKKAGWTPITASQLGYD   100

PfL   GKTDARGTFFGEKDGYTSAQVEILGKYDAQGHLTELGIAFRGTSGPREIL   150
PsL   GKTDARGTFFGEKAGYTTAQVEILGKYDAQGHLTEIGIAFRGTSGPRENL   150

PfL   IGDSIGDVINDLLAAFGPADYAKNYVGEAFGNLLYDVVAFARANGLSGMD   200
PsL   ILDSIGDVINDLLAAFGPKDYAKNYVGEAFGNLLNDVVAFAKANGLSGKD   200

PfL   VLVSGHSLGGLAVNSMADLSAGKWGGFFADSNYIAYASPTQSSGDKVLNI   250
PsL   VLVSGHSLGGLAVNSMADLSGGKWGGFFADSNYIAYASPTQSSTDKVLNV   250

PfL   GYENDPVFRALDGSTFSSASAGVHDAPHASTTDNIVSFNDHYASNAWNVL   300
PsL   GYENDPVFRALDGSTFTGASVGVHDAPKESATDNIVSFNDHYASTAWNLL   300

PfL   PYSILNIPTWISHLPSGYGDGMTRVLESKFYDLTSRDSTIIVANLSDPAR   350
PsL   PFSILNIPTWISHLPTAYGDGMNRIIESKFYDLTSKDSTIIVANLSDPAR   350

PfL   ANTWVQDLNRNAETHKGSTFIIGSDGNDLIQGGGSGNDYLEGRAGNDSFRD   400
PsL   ANTWVQDLNRNAETHKGSTFIIGSDSNDLIQGGGSGNDYLEGRAGNDTFRD   400

PfL   SGGYNIILGGQGSNTLDLQQTVKNFDFANDGAGHLYIRDANGGISITRDI   450
PsL   GGGYNVILGGAGNNTLDLQKSVNTFDFANDGAGNLYVRDANGGISITRDI   450

PfL   GSIVSKEPGFLWGVFKDDVLHNVTDMGLKVGSNLTAYESSVRGTAGADTL   500
PsL   GSIVTKEPGFLWGLFKDDVTHSVTASGLKVGSNVTQYDASVKGTNGADTL   500

PfL   KAHAGGDRLFGLEGNDHLLGSAGNDVFVGGAGNDLMESGGGADTFLFNGA   550
PsL   KAHAGGDWLFGLDGNDHLIGGVGNDVFVGGAGNDLMESGGGADTFLFNGA   550

PfL   FGQDRVVGYTADDKLVFLGVQGVLPGDDLRAHASAVGQDTVLTFGGDSVT   600
PsL   FGQDRVVGFTSNDKLVFLGVQGVLPNDDFRAHASMVGQDTVLKFGGDSVT   600

PfL   LVGVALGSLNSDGIVIA   617
PsL   LVGVALNSLSADGIVIA   617
```

Figure 3.12: Sequence alignment of PfL against PsL, in green the identical residues, in blue the similar residues and in red the other residues.

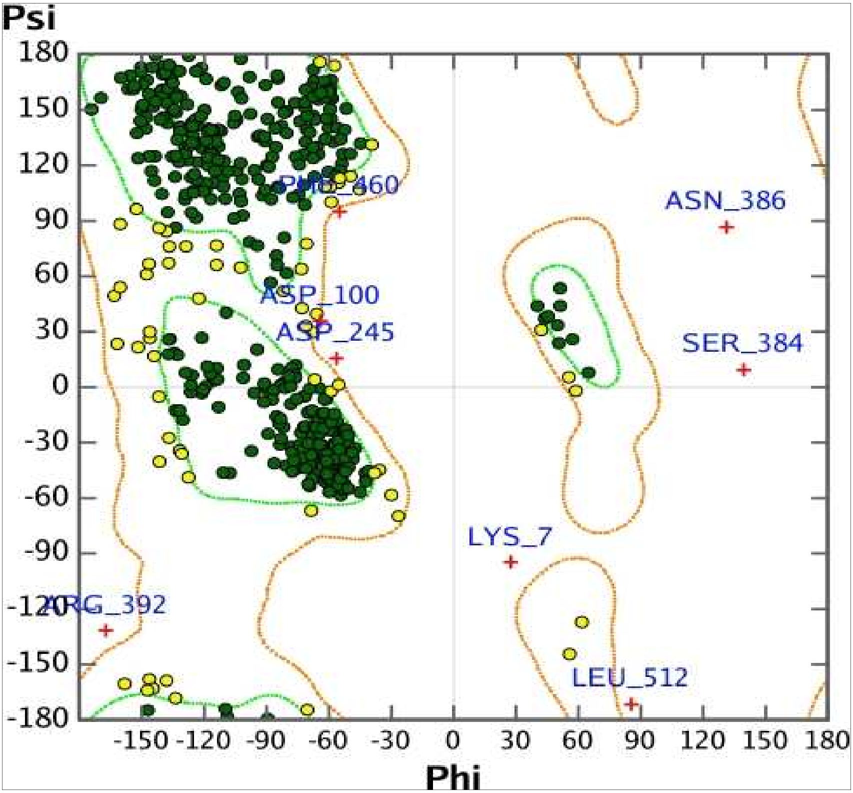


Figure 3.13: Ramachandran plot of the generated structure of PfL; in green circles the residue in the core structure, in yellow circles the allowed residues and in red crosses the outliers.

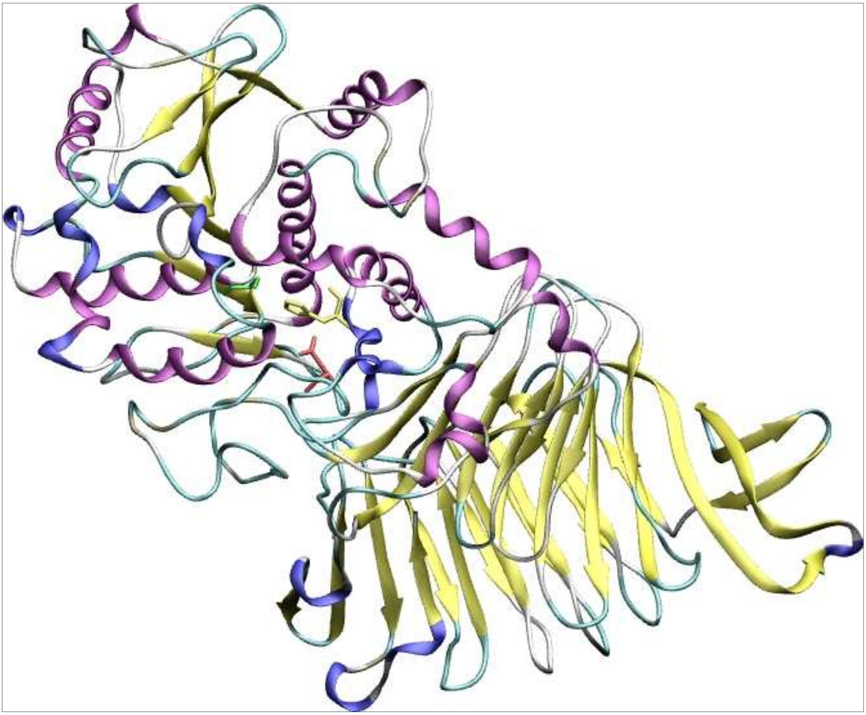


Figure 3.14: Generated structure of *Pseudomonas fluorescens* lipase's open conformation in new cartoon mode coloured by secondary structure; the catalytic triad is highlighted in licorice mode: Ser in green, His in yellow and Asp in red.

The structure of PfL seems to be made by two distinct structural domains. Two thirds of the enzyme is exposed by a well structured domain made by a core of β-sheets connected by random coil parts; while the active site domain shows the typical α/β hydrolase fold.

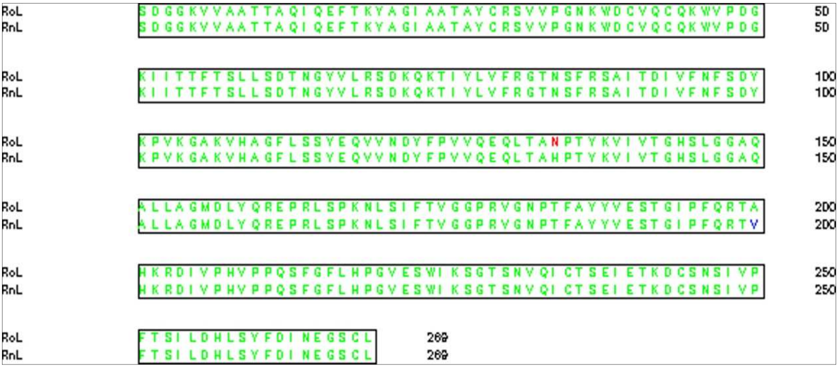
3.3.2 *Rhizopus oryzae* lipase

In the case of *Rhizopus oryzae* lipase (RoL), the generation of the homology model was performed with a different strategy. The primary structure of RoL is available in the UniProtKB[52] database with the code Q2QFX1_RHIOR. A homology search of the aminoacid sequence was performed on SRS@EBI[53] website. RoL is closely related to *Rhizopus niveus* lipase but it is interesting to note that RoL can be present in a pre-mature enzymatic form. The mature forms of these two enzymes are 99% homologous in their primary structure, just a two residues difference is found between these two enzyme sequences (Figure 3.15). This differences are His134 of RnL which is replaced by an Asn in RoL and Val200 of RnL which is replaced by an Ala in RoL.

In this case the three dimensional structure of RoL was generated by a simple *in silico* mutation of these different residues performed on the available structure of RnL.

The PDB structure 1GLY was mutated using the mutagenesis PyMol[54] tool and then minimized in MOE using AMBER 99 force field. Therefore the generated RoL structure was analysed in order to assure the reliability of the model. The Ramachandran plot (Figure 3.15) shows that just four residues are considered as outliers, but they are close to the allowed regions. None of these outliers are mutated residues. The high score of the Ramachandran plot indicates the general quality of the generated structure, in fact 92% of the residues are in the core region.

[54] PyMol 0.99, *DeLano Scientific,* Palo Alto, CA, USA.



```
RoL     SDGGKVVAATTAQIQEFTKYAGIAATAYCRSVVPGNKWDCVQCQKWVPDG     50
RnL     SDGGKVVAATTAQIQEFTKYAGIAATAYCRSVVPGNKWDCVQCQKWVPDG     50

RoL     KIITTFTSLLSDTNGYVLRSDKQKTIYLVFRGTNSFRSAITDIVFNFSDY    100
RnL     KIITTFTSLLSDTNGYVLRSDKQKTIYLVFRGTNSFRSAITDIVFNFSDY    100

RoL     KPVKGAKVHAGFLSSYEQVVNDYFPVVQEQLTANPTYKVIVTGHSLGGAQ    150
RnL     KPVKGAKVHAGFLSSYEQVVNDYFPVVQEQLTAHPTYKVIVTGHSLGGAQ    150

RoL     ALLAGMDLYQREPRLSPKNLSIFTVGGPRVGNPTFAYYVESTGIPFQRTA    200
RnL     ALLAGMDLYQREPRLSPKNLSIFTVGGPRVGNPTFAYYVESTGIPFQRTV    200

RoL     HKRDIVPHVPPQSFGFLHPGVESWIKSGTSNVQICTSEIETKDCSNSIVP    250
RnL     HKRDIVPHVPPQSFGFLHPGVESWIKSGTSNVQICTSEIETKDCSNSIVP    250

RoL     FTSILDHLSYFDINEGSCL     269
RnL     FTSILDHLSYFDINEGSCL     269
```

Figure 3.15: Sequence alignment of RoL against RnL, in green the identical residues, in blue the similar residues and in red the other residues.

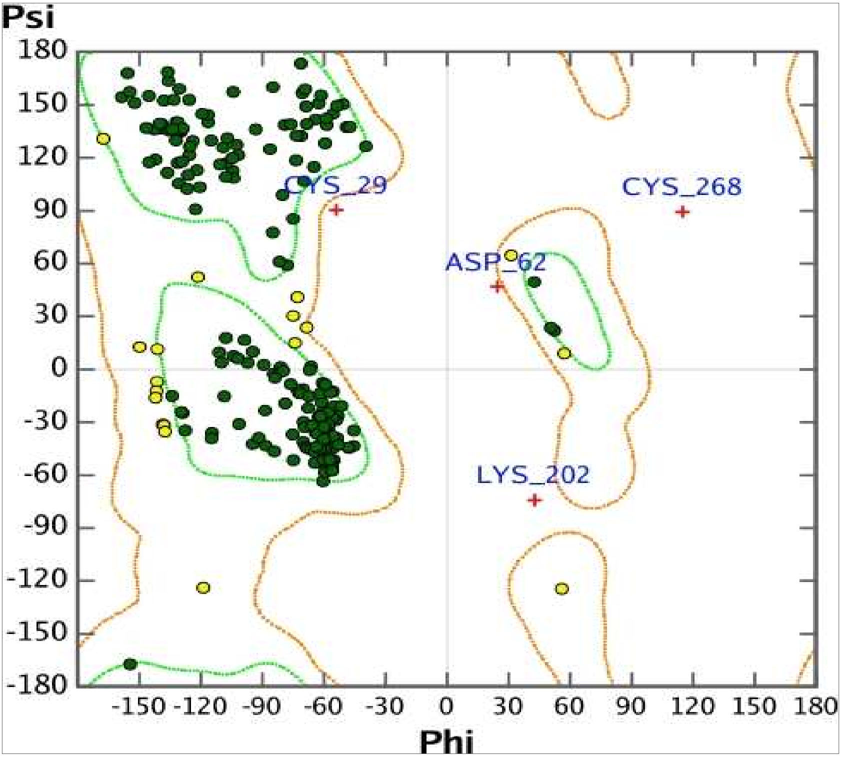


Figure 3.16: Ramachandran plot of the generated structure of RoL; in green circles the residue in the core structure, in yellow circles the allowed residues and in red crosses the outliers.

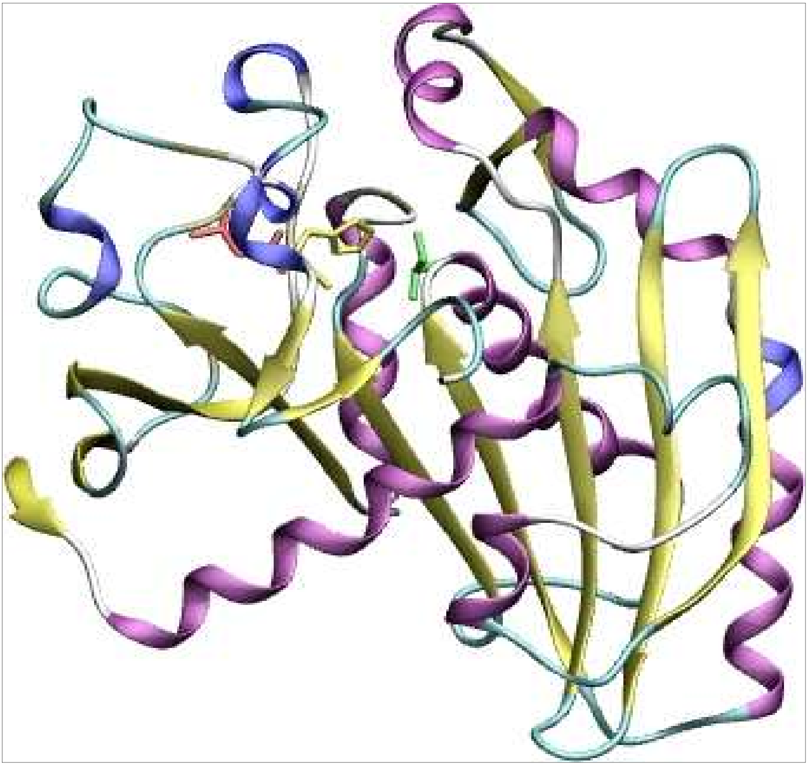


Figure 3.17: Generated structure of *Rhizopus oryzae* lipase open conformation in new cartoon mode coloured by secondary structure; the catalytic triad is highlighted in licorice mode: Ser in green, His in yellow and Asp in red.

The structure of RoL (Figure 3.17) is substantially identical to RnL and presents the classical α/β hydrolase fold.

3.4 Surface analysis

A detailed surface analysis was performed on the lipase structures in order to map the distribution of the hydrophilic and hydrophobic zones of the enzymes' surfaces.

The study of the physical-chemical properties of the enzymatic surface can be very useful because the information can be used for rationalize the experimental work. A significant number of enzyme properties is correlated to the characteristics of its surface and the most of them affects enormously the experimental activity. The dependency of enzyme action on polarity of medium, or the immobilization on solid supports of different chemical nature are two representative examples of the role that can be played by the surface analysis in the experimental practice. Another example can be the analysis and prediction of enzyme stability, as recently reported by our group.[55]

As described in the previous paragraphs, lipases are generally characterised by the tendency to be governed by interfacial activation. Although some differences do exist, common tracts on their structural organization group them in a single big ensemble of enzymes, displaying common behaviours, despite often major structural differences among them emerge very clearly.

The analysis of the protein surface can be made by many different computational strategies. The generation of the Molecular Interaction Fields (MIFs) represents one of the most powerful. A MIF is a tridimensional map of the interaction between a given molecule and a chemical probe, mimicking a given interaction capability (i.e. hydrogen bonding donor, acceptor, dipolar interaction, etc.). The software GRID (version 22) was used to measure the non-covalent interactions between the target protein structure and two different probes:

- WATER, for the simulation of the properties of a water

[55] P Braiuca, A Buthe, C Ebert, P Linda, L Gardossi, *J Biotechnol*, **2007**, 2, 214.

molecule able to accept and donate recognize hydrogen bonds;

- DRY, to recognize non-polar areas on the enzyme surface and describe hydrophobic interactions.

The lipases object of this study were submitted to this procedure in order to investigate and compare their surface characteristics.

As far as *Bacillus subtilis* lipase is concerned, the surface analysis of the structure shows a big hydrophobic area placed in the correspondence of the active site of the enzyme, whereas the rest of the structure is prevalently hydrophilic (Figure 3.18).

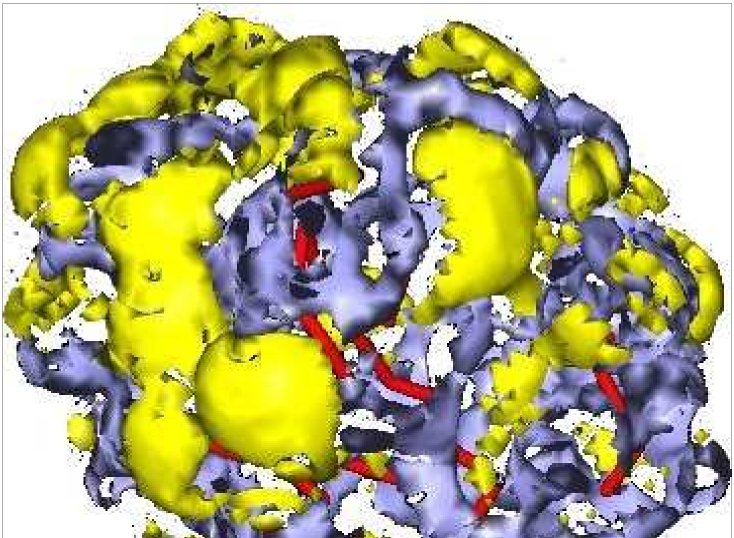


Figure 3.18: Surface analysis of *Bacillus subtilis* lipase (PDB 1ISP) in yellow the hydrophobic areas (DRY probe) of the enzyme surface and in blue the hydrophilic ones (WATER probe).

The surface analysis of the structure of *Pseudomonas cepacia* lipase shows a similar behaviour; the structure is very distinctly polarized and a large hydrophobic area is located just above the zone of the big active site of the protein (Figure 3.19). This picture indicates the probable behaviour of the enzyme at the interface, the active site will be oriented to the non-polar solvent.

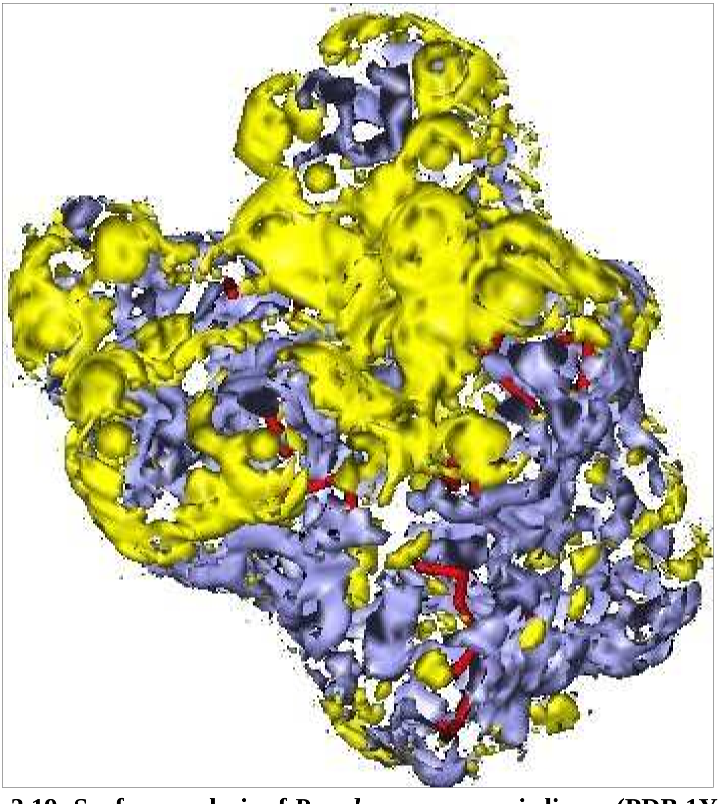


Figure 3.19: Surface analysis of *Pseudomonas capacia* lipase (PDB 1YS1) in yellow the hydrophobic areas (DRY probe) of the enzyme surface and in blue the hydrophilic ones (WATER probe).

The surface characteristics of *Candida antarctica* lipase B (CaLB) are also somewhat related with the previous analysis (Figure 3.20). In this case hydrophobic zones are still concentrated in the active site area but are also spread on the rest of the enzyme surface.

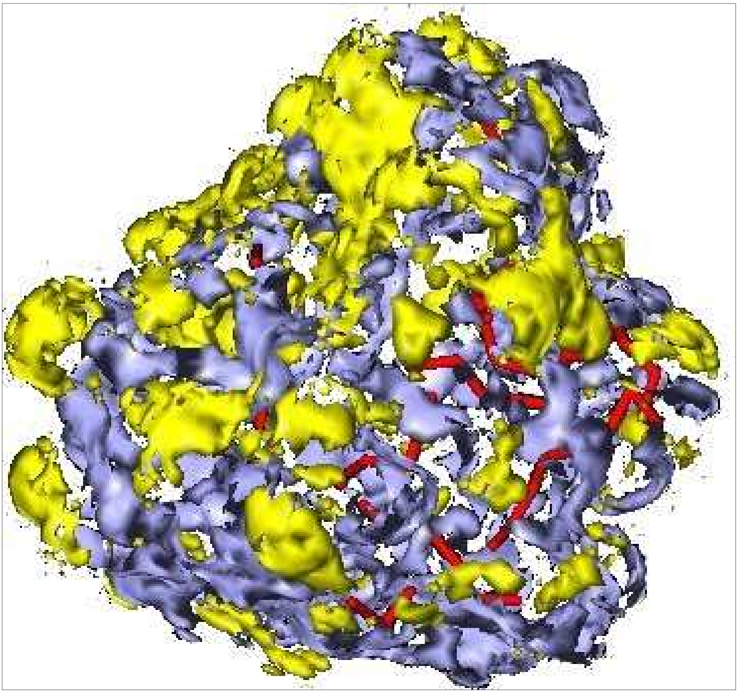


Figure 3.20: Surface analysis of *Candida antarctica* lipase B (PDB 1TCA) in yellow the hydrophobic areas (DRY probe) of the enzyme surface and in blue the hydrophilic ones (WATER probe).

In the case of *Candida rugosa* lipase the analysis was performed on both structures: open and closed. The results of the surface analysis of the structure in the open conformation are similar to the results described before. On the other hand, the surface analysis of the structure in the closed conformation is different, the enzyme is less hydrophobic than in the open conformation (Figure 3.21). This is due to the lid movement that covers the active site of the protein and its hydrophobicity.

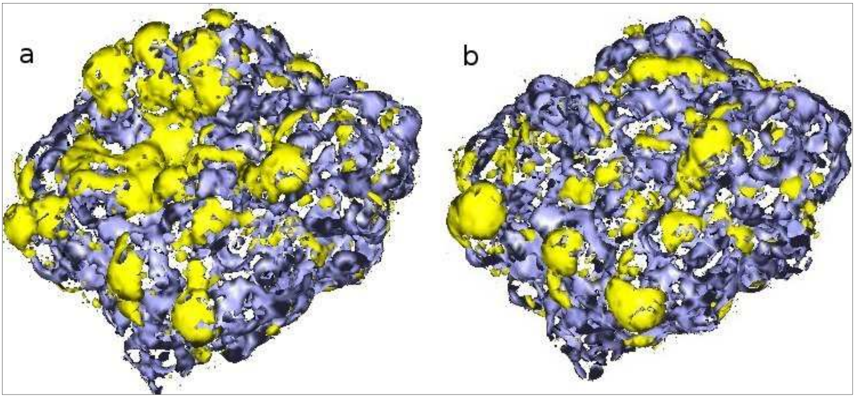


Figure 3.21: Surface analysis of *Candida rugosa* lipase a- in the open conformation (PDB 1CRL); b- in the closed conformation (PDB 1GZ7); in yellow the hydrophobic areas (DRY probe) of the enzyme surface and in blue the hydrophilic ones (WATER probe).

Concerning *Geotrichum candidum* lipase, the analysed structure was the one in the closed conformation. For this reason the hydrophobic regions are, as expected, quite small and mostly located near the active site area (Figure 3.22).

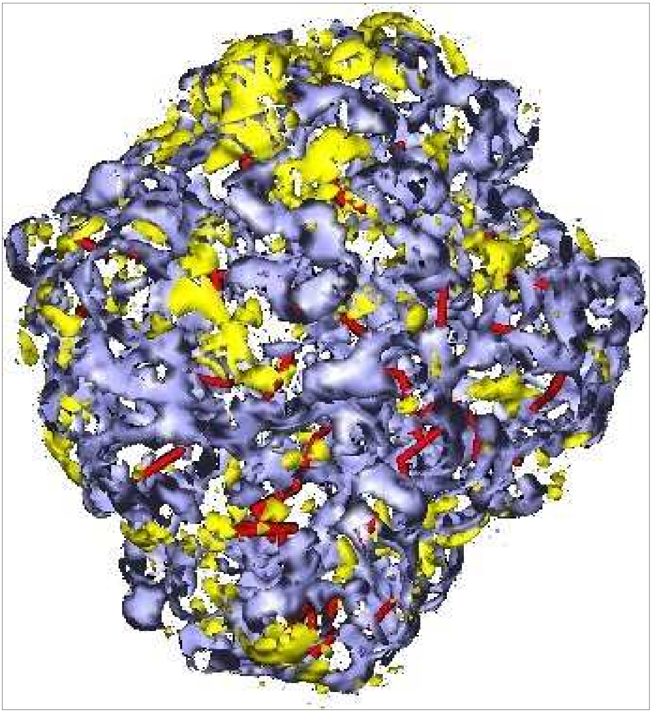


Figure 3.22: Surface analysis of *Geotricum candidum* lipase (PDB 1THG) in yellow the hydrophobic areas (DRY probe) of the enzyme surface and in blue the hydrophilic ones (WATER probe).

For *Humicola lanuginosa* lipase, the surface analysis of the structure in the open conformation shows a big hydrophobic area in correspondence of the active site. On the other hand, the surface analysis of the structure in the closed conformation is different, the enzyme is less hydrophobic than in the open conformation (Figure 3.23). This is due to the lid movement that covers the active site of the protein and its hydrophobicity, similarly to what observed in the case of *Candida rugosa* lipase.

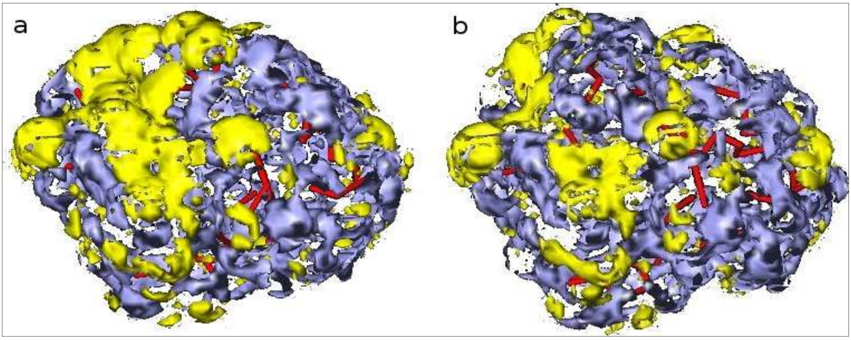


Figure 3.23: Surface analysis of *Humicola lanuginosa* lipase a- in the open conformation (PDB 1DTE); b- in the closed conformation (PDB 1TIB); in yellow the hydrophobic areas (DRY probe) of the enzyme surface and in blue the hydrophilic ones (WATER probe).

The surface analyses allowed to point out a common feature for the studied lipases; all of them display a neatly polarized surface. The active site of this enzymes is a big hydrophobic area, whereas the rest of the enzyme surface is usually balanced in small zones of hydrophobic and hydrophilic character. The balance of the non catalytic part of the enzyme depends on the intrinsic characteristics of the different lipases and on their natural phylogenetic evolution. The polarization phenomenon can also explain how they act on water/oil interfaces, positioning the catalytic part of the enzyme into the non-polar phase. The lid movement is able to influence the hydrophobicity of the enzyme surface covering the active site. Even if the active site is covered by the lid some hydrophobic regions are always present near the covered part, and they probably have the role of driving the right movement and the positioning of the enzyme at the interface, as well as affecting the initial part of the interfacial activation process affecting in this way the lid opening movement.

3.5 Molecular dynamics simulations

Molecular dynamics (MD) is a technique that explores the conformational possibilities of the system during time. The opportunity to simulate and therefore to study complex phenomena with high accuracy makes MD simulations a suitable technique for investigating lipases' activation processes.

Different variants of the basic molecular dynamics procedure can be performed in order to perturb the system equilibrium, to change the accuracy and consequently reduce the required simulation time. The idea of MD is simulating a system to observe a phenomenon dependent on the time evolution of the system. The nature of the phenomenon defines the probability of its happening in the time of the simulation. In other words the probability of observe a given phenomenon is a function of its intrinsic nature and of the total time simulated. A very probable event will happen many times in a single short trajectory, a rare event can not happen at all in a very long one.

The idea of non equilibrium MD (or steered MD) is forcing an alteration of the simulated chemical system, that spontaneously tends to a thermodynamic equilibrium, to accelerate the desired event, or to make its occurrence more probable. In principle the concept is very simple and it is based on the alteration of one or more parameters during the simulation (pressure gradient, forces gradient, etc.).

Another obvious strategy to observe a slow or improbable phenomenon is increasing the simulation time. Although letting the system to evolve spontaneously for the necessary time is much more rigorous than forcing non-equilibrium by applying forces to the system, the computational cost is usually too big to pursue this route. The need to reduce the computational cost can only be satisfied by reducing the accuracy of the simulation. Many efforts have been spent on this concept and nowadays it is possible to put simplification in the definition of the chemical system at the cost of a loss of accuracy, often tolerable if it does not affect severely the phenomenon of interest.

Accuracy alterations can be achieved principally changing the force field type. Various types of force fields can be employed during an MD simulations, and they are classified on the basis of different simulation targets they describe best (proteins, organic molecules, DNA, etc.) or on the basis of their accuracy (fine grain, coarse grain).

Each MD type and each force field has advantages and disadvantages and the operator should be able to choose the right simulation conditions on the basis of the system that has to be simulated and on the basis of the aim that has to be achieved.

In figure 3.24 an example of the parametrization of a molecule of propanol using different force fields is reported.

- fine grained force fields, such as OPLS-AA,[56] typically all atom force fields, where molecules are actually represented by all their atoms;
- united atom force fields, such as GROMOS,[57] represent only the polar hydrogens, other hydrogens are not explicitly simulated but they are considered together with the heavy atoms they are bonded to (i.e. CH_3);
- coarse grained force fields, such as MARTINI,[58] represent molecules with molecular building blocks, where every simulated sphere can represent more than one chemical group (i.e. CH_4-CH_3-CH_3-).

The use of united atoms or coarse grained force fields is possible when the object of the study is a phenomenon not related, or not heavily affected by the simplification at the basis of the force field

[56] W L Jorgensen, J Triado Rives, *J Am Chem Soc*, **1988**, 110, 1657.

[57] W R P Scott, P H Huenenberger, I G Tironi, A E Mark, S R Billeter, J Fennen, A E Torda, T Huber, P Krueger, W F van Gunsteren, *J Phys Chem A*, **1999**, 103, 3596.

[58] S J Marrink, H J Risselada, S Yefimov, D P Tieleman, A H de Vries, *J Phys Chem B*, **2007**, 111, 7812.

(i.e. non polar hydrogen atoms in the united atoms representation). The effect of simplifications must be taken into account in the analysis step of the work, in order to assess the quality of the model.

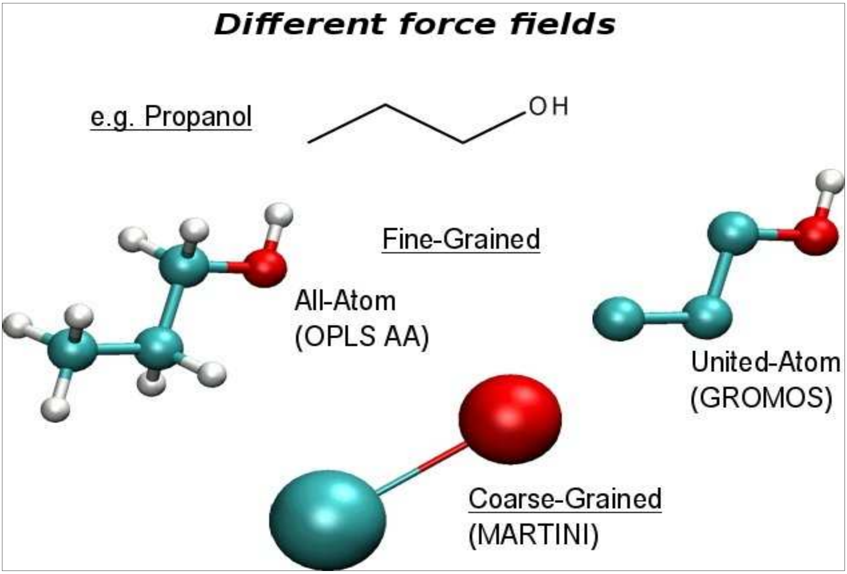


Figure 3.24: different force fields parametrization of a molecule of propanol; hydrogens in white spheres, carbons in cyan spheres and oxygen in red spheres.

3.5.1 Classical MD simulations

Lipases’ activation phenomena are governed by the lid movement. This kind of event has a probabilistic nature and in a simulation protocol its observation can require several nanoseconds.

Lipase activations, at molecular level, have not been deeply investigated yet, just few information on some lipases are available from literature.[59, 60] In order to investigate the nature of the lid movements and the environmental factors acting of it classical MD simulations were performed.

The lid's closure movement is exactly the opposite of its opening movement but from a practical point of view it is easier to simulate the deactivation process. This is due to the fact that the lid closure (enzyme inactivation) happens in polar media, such as water, which is the easiest case of solvation and parametrization for an MD run. All the systems were parametrized using the GROMOS force field[57] and all the MD simulations were performed starting from enzymes in their open conformation in water environment.

3.5.1.1 Classical MD on *Pseudomonas cepacia* lipase

PcL is is characterised by a huge lid domain formed by 30 aminoacids: residues from Gly116 to Leu149. For this reason the simulation of its conformational change is particularly interesting. Similarly to a common door, the lid hinges are constituted by hydrogen bonds on its ends. The first one is fixed by hydrogen bonds formed between His114 and Ser271, and between Gly 116 and Ser271. On the other hand, the second hinge is due to hydrogen bonds between Thr150 and Ala24, and Ser152 and Asp21. This lid has an arch shape with its first half (from Gly116 to Pro131) more rigid and stabilised by 2 hydrogen bonds, between Ser117 and Leu167 and between Asp121 and Thr169. The second half of the arch (from Thr132 to Leu149) is stabilised by just one hydrogen bond between Asp144 and Ala160 (Figure 3.25).

[59] P Trodler, R D Shmid, J Pleiss, *BMC Struct Biol*, **2009**, 9, 38.

[60] S L Cherukuvada, A S N Seshasayee, K Raghunathan, S Anishetty, G Pennathur, *PLoS Comput Biol*, **2005**, 28, 182.

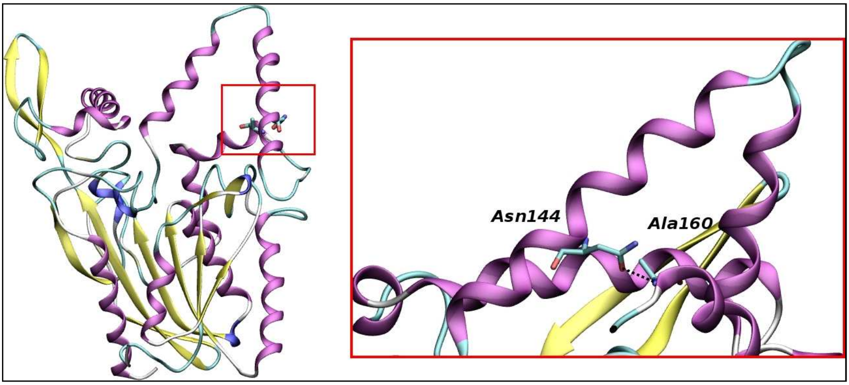


Figure 3.25: Hydrogen bond that stabilize the second half of the lid of PcL.

The simulation was performed starting from the open conformation of the enzyme. PcL (PDB 1YS1) was put in the centre of an 80 X 80 X 80 Å^3 cubic system and solvated with water. The system was first minimised and equilibrated with a 500 ps of molecular dynamic simulation in NPT conditions, during this equilibration step the enzyme was keep restrained in its position.

Afterwards the restrain on the enzyme was removed. MD simulation was performed for 10 ns in NPT conditions. The system was then minimised and equilibrated with 500 ps of MD simulation in NPT conditions. The trajectory was analysed measuring the minimum distance between two residues located on the opposite side of the active site cleft, namely Ala141 and Ala247. The calculated structure was overlapped with the starting structure (Figure 3.26).

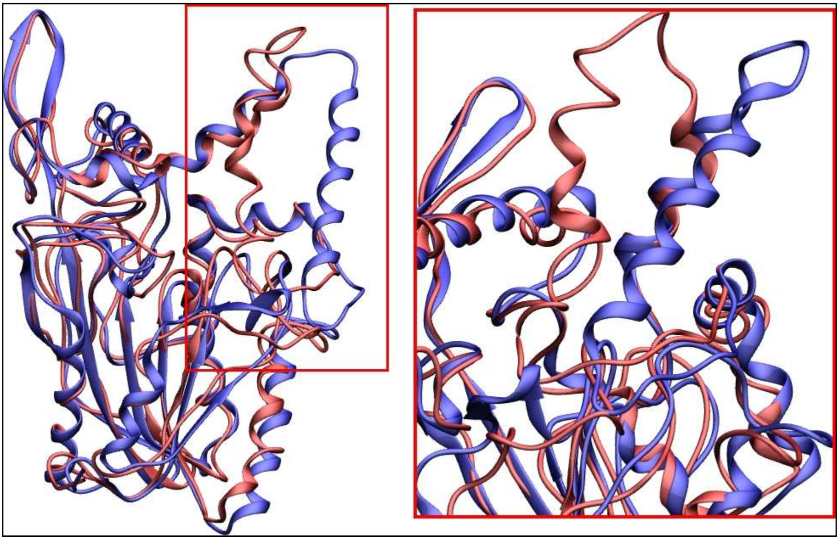


Figure 3.26: Superimposition of the crystal PcL (PDB 1YS1) and the structure resulting afret classical MD simulation. In the red box the lid is highlited.

It is clearly observable a lid position variation. In order to verify the achievement the correct deactivation movement a GRID analysis of the calculated structure was performed (Figure 3.27).

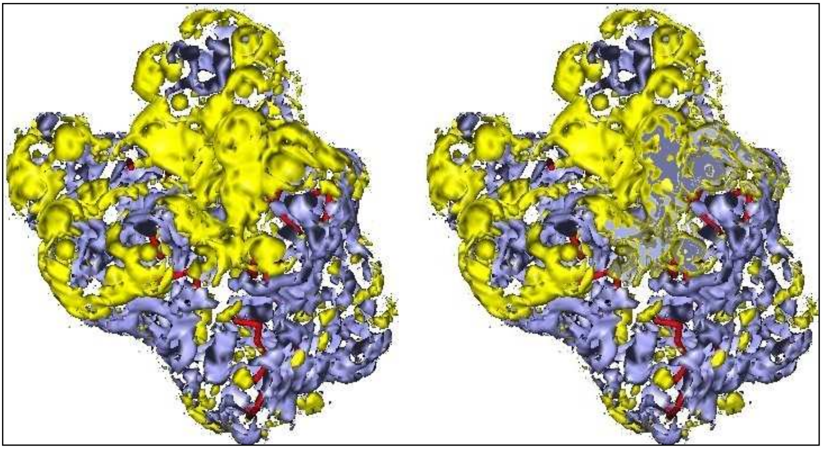


Figure 3.27: Comparison between the GRID analysis of the crystal structure of PcL (PDB 1YS1) and its calculated structure after classical MD simulation. In yellow the hydrophobic areas and in blue the hydrophilic ones.

Lid domain movement was not enough pronounced to cover the hydrophobicity of the active site, being the polarity of the active site area substantially unchanged. The simulation was then not able to reproduce a complete inactivation of the enzyme, despite the suitable simulation time and the repeated calculations. This result demonstrates that these kinds of phenomena are not easy to simulate and the lid of this enzyme is very strongly stabilised in its open conformation, even if put in a polar solvent. Considering once again the probabilistic nature of the lid movement, new attempts to simulate the deactivation phenomenon should be achieved performing simulations longer than 10 ns. Nevertheless simulations of more than 10 ns become extremely expensive in terms of computational time and this was against the thesis aims.

A different strategy, able to keep the necessary precision on classical MD approach, but with a significantly reduced computational cost had to be pursued. The choice fell to steered MD simulations and will be described in paragraph 3.5.2.

Recently a work of Pleiss and co-workers demonstrated that the complete closure of PcL lid requires at least 20 ns of a classical MD trajectory, confirming what observed.[59]

3.5.1.2 Classical MD on *Pseudomonas aeruginosa* lipase

Structural comparison of PaL and PcL points out a striking similarity. Nevertheless a major difference emerges very clearly. It is given by one single domain, which is structured by two antiparallel β sheets (Figure 3.28). This domain is located at the opposite side of the active side entrance with respect to the lid and it might have a role in the activation mechanism. Classical MD simulations were applied to the study of this aspect.

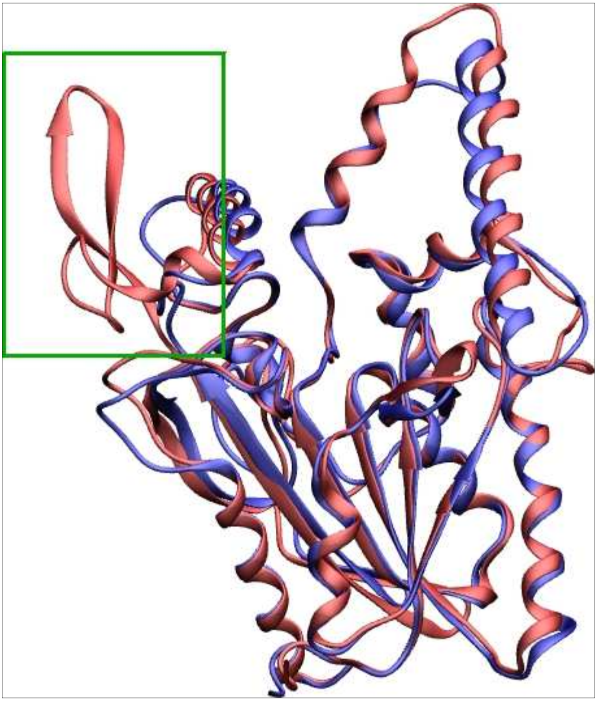


Figure 3.28: superimposition of PaL (PDB 1EX9) in blue and PcL (PDB 1YS1) in red. The main difference between the two structures is highlighted in the green box.

Like in the case of PcL the lid is the biggest among lipases, arch shaped with hinges at the ends. The first half hinge is constituted by the hydrogen bond between Gly111 and Ser236; this domain part is stabilised in its open conformations by hydrogen bonds between Ser112 and Leu162, and between Asp116 and Ser164. The second half of the lid has a hinge formed by the hydrogen bond between Ser146 and Asp20, and it is stabilised in its open conformation by hydrogen bonds between Ser143 and Gly148 and interactions that Asn136 establishes with Gln153 and Ser155 (Figure 3.29).

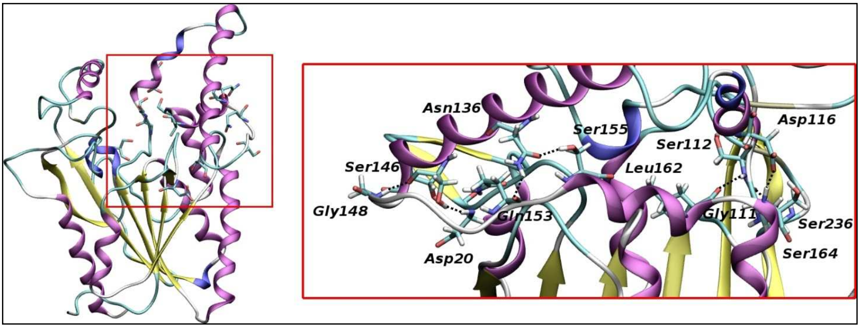


Figure 3.29: Hydrogen bonds that stabilize the lid of PaL.

The simulation was performed starting from the open conformation of the enzyme which is also the only crystal structure available for this enzyme (PDB 1EX9). PaL was put in the centre of an 80 X 80 X 80 Å^3 cubic system and solvated with water. The system was minimised and equilibrated with a 500 ps of molecular dynamic simulation in NPT conditions, during this equilibration step the enzyme was keep restrained in its position. Afterwards the restrain on the enzyme was removed and the system was subjected to 10 ns MD simulation in NPT conditions. The system was then minimised and equilibrated with 500 ps of MD simulation in NPT conditions.

In this case, the data analysis shows a consistent reduction of the distance between the two edges of the active site cleft. It is possible to verify that lid closure was achieved within 6 ns.

The GRID analysis confirms that the conformational change is actually an inactivation mechanism (Figure 3.30).

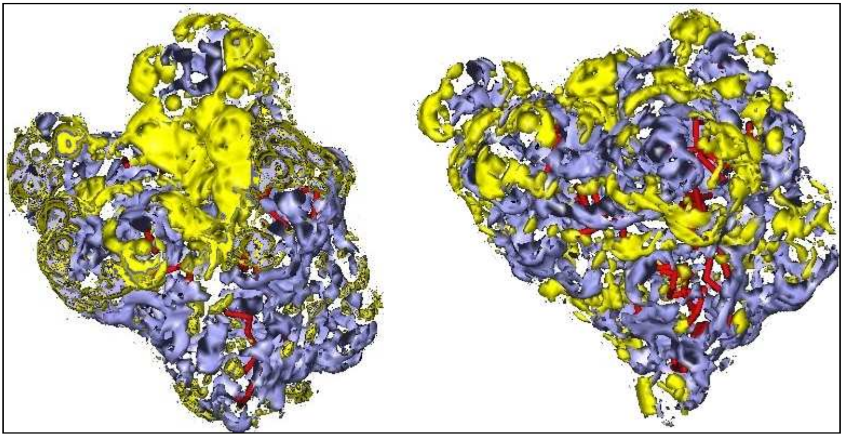


Figure 3.30: Comparison between the GRID analysis of the crystal structure of PaL (PDB 1EX9) and its calculated structure after classical MD simulation. In yellow the hydrophobic areas and in blue the hydrophilic ones.

In fact the active site hydrophobicity was appreciably reduced at the end of MD trajectory, confirming the quality of the simulated movement. A superimposition of the starting PaL structure with its calculated closed conformation was performed (Figure 3.31).

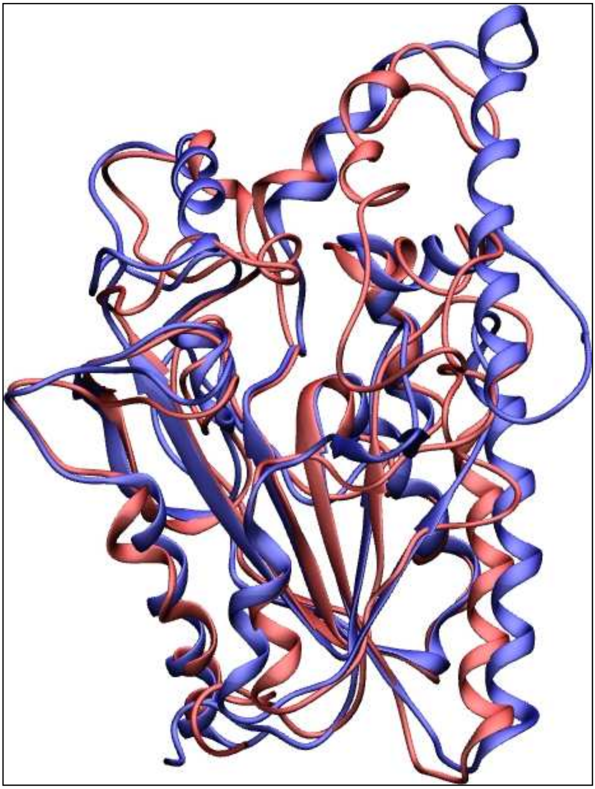


Figure 3.31: Superimposition of the crystal structure of PaL (PDB 1EX9) in blue, with the calculated closed structure of PaL, in red.

Surprisingly the superimposition shows that the conformational change is due to the movement of two protein domains: the already mentioned lid and another domain (from Leu208 to Thr221) located on the opposite side of the active site cleft, just next to the zone that differs from PcL. This concerted mechanism appears unique among lipases.

This “cooperative” lid was further analysed. It also stabilised by a network of hydrogen bonds, similarly to the “real” lid domain. The bonds are formed between Asp209 and Val258 and between Asp212 and Thr205 (Figure 3.32).

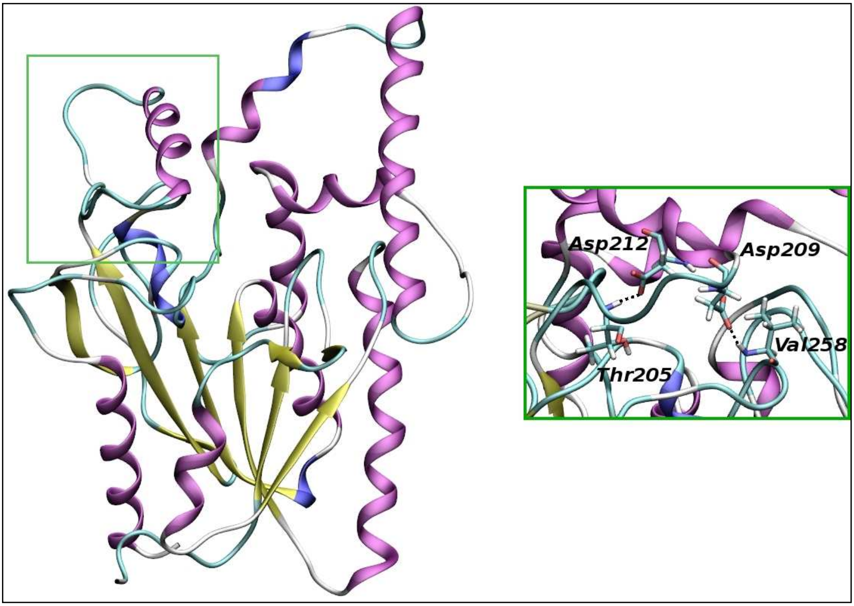


Figure 3.32: Hydrogen bonds that stabilize the cooperative lid of PaL.

The two lids are stabilised in their closed conformations by the establishment of hydrogen bonds between each other (Figure 3.33), particularly between Leu138 and Phe214 and Gly139 and Pro210.

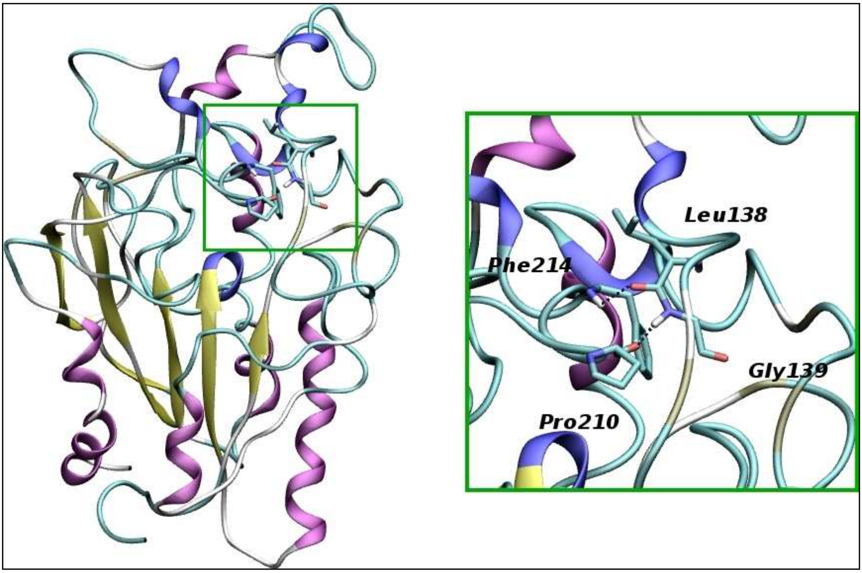


Figure 3.33: Hydrogen bonds that stabilise the calculated closed conformation of PaL.

Cooperative lids movement of PaL was compared with the incomplete movement of PcL. This analysis found some common behaviour among the two lipase movements. In both cases the lid movement, the “main” lid in the case of PaL, follows the movement of another little domain (from Asp21 to Glu28 in PcL and from Asp20 to Tyr27 in PaL). This small domain in both cases is constituted by 2 β sheets. These sheets flex toward the core of the enzyme generating the necessary space for the lid movement.

On the other hand, the two antiparallel β sheets (D-domain) that structurally distinguish PcL from PaL stabilise the potential second lid of PcL with supplementary hydrogen bonds (Figure 3.34).

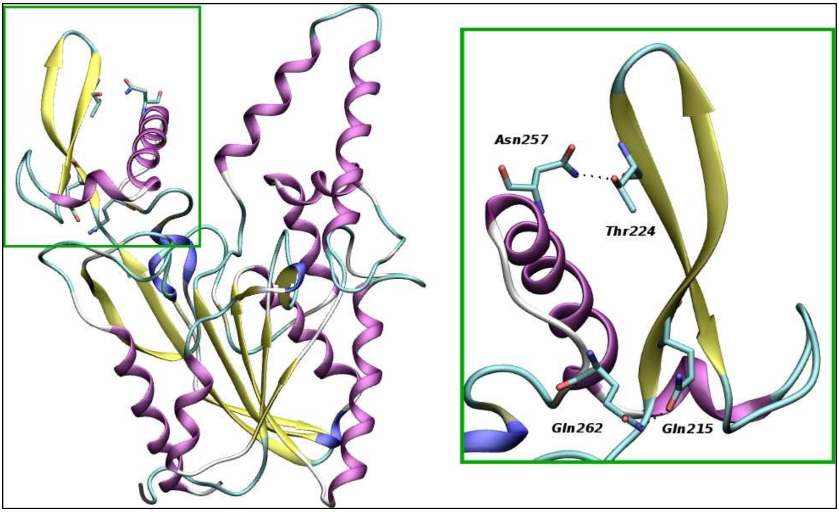


Figure 3.34: Hidrogen bonds that stabilise the putative second lid of PcL.

The high number of weak bonds acting on this cooperative lid in PcL stabilizes significantly its conformation, subsequently, despite the strong homology, the lid activation of PcL is entrusted by a single lid domain only. The stabilising effect of the D-domain explains the different activation time required by the two enzymes. The different evolution of these clearly phylogenetically related lipases is nevertheless fascinating and could be topic of further studies in the future.

3.5.1.3 Classical MD on *Humicola lanuginosa* lipase

Classical MD simulations have been applied on bacterial lipases also, representing a different taxa in the lipase family. The HlL was chosen for this purpose. Crystal structures of this enzyme are available in both open and closed conformations as mentioned at the beginning of this chapter. These structures should allow to verify the quality of the simulated conformational changes. As briefly mentioned in paragraph 3.2.6, comparing the open structure of the protein with the closed one it is evident that the main difference is represented by the position of the lid domain only (Figure 3.8), without any other significant difference.

Differently to PcL and PaL, the lid is much smaller, composed by residues from Gly82 to Ile90, but only residues from Ser83 to Glu87 show a remarkable difference in terms of spatial position in the open and close conformations. The lid is stabilized in its open conformation by several hydrogen bonds between different residues of the lid and different aminoacids of the nearest domain just behind it. The two most important hydrogen bonds are between Ser85 and Asp62 and between Asn88 and Asp62. The rest of the lid is stabilized by seven other hydrogen bonds, making this part of the domain basically fixed (Figure 3.35). This is particularly relevant, since not the entire domain is mobile, mainly because of these seven H-bond interactions.

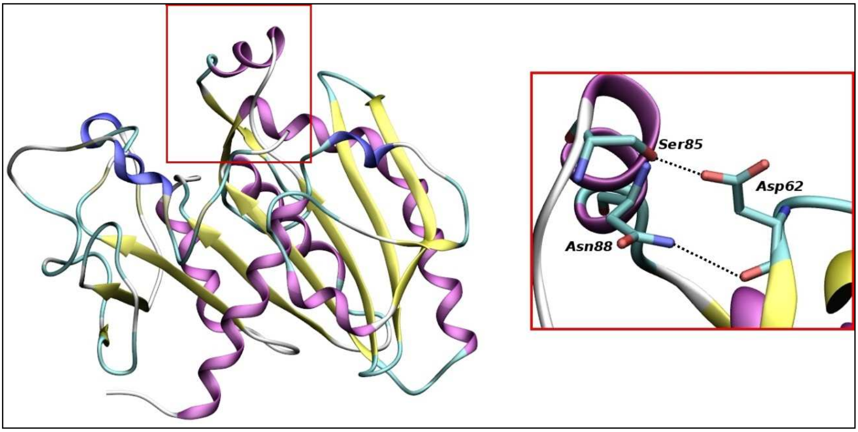


Figure 3.35: hydrogen bonds that stabilize the mobile part of the lid of HlL.

HlL (PDB 1DTE) was put, like in the other cases, in the centre of a 80 X 80 X 80 Å^3 cubic system and explicitly solvated with water. Afterwards the system was minimised and equilibrated with a 500 ps of molecular dynamic simulation in NPT conditions. During this equilibration step the enzyme was kept restrained in its position.

After equilibration the restrain on the enzyme was removed. MD simulation was calculated for 10 ns in NPT conditions. The system was then minimised and equilibrated by 500 ps of MD simulation in NPT conditions.

The results of the trajectory analysis show that the lid movement is reproducible and the final achieved conformation demonstrated to be reliable, since during the last equilibration phase it remains stable in the closed position. The trajectory of the simulation was used to investigate the minimum distance between Ser85 and Asp254, which showed that the lid closure takes about 3 ns, and the total lid movement is about 10 nm.

The analysis of the calculated closed conformation shows that the lid is stabilized in its position by the establishment of new hydrogens bonds (Figure 3.36) between Ser83 and His258 and Arg84 that can establish interaction with Gly266 or with Cys268.

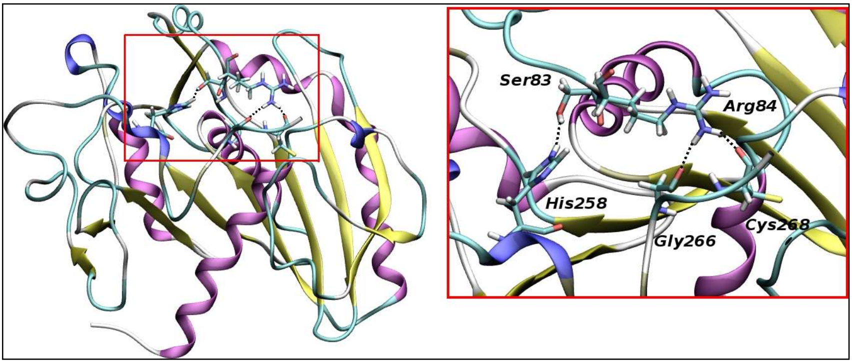


Figure 3.36: Hydrogen bonds that stabilize HlL in its closed conformations.

The calculated closed conformation was superposed with the crystal one (Figure 3.37) in order to prove the quality of the simulation result.

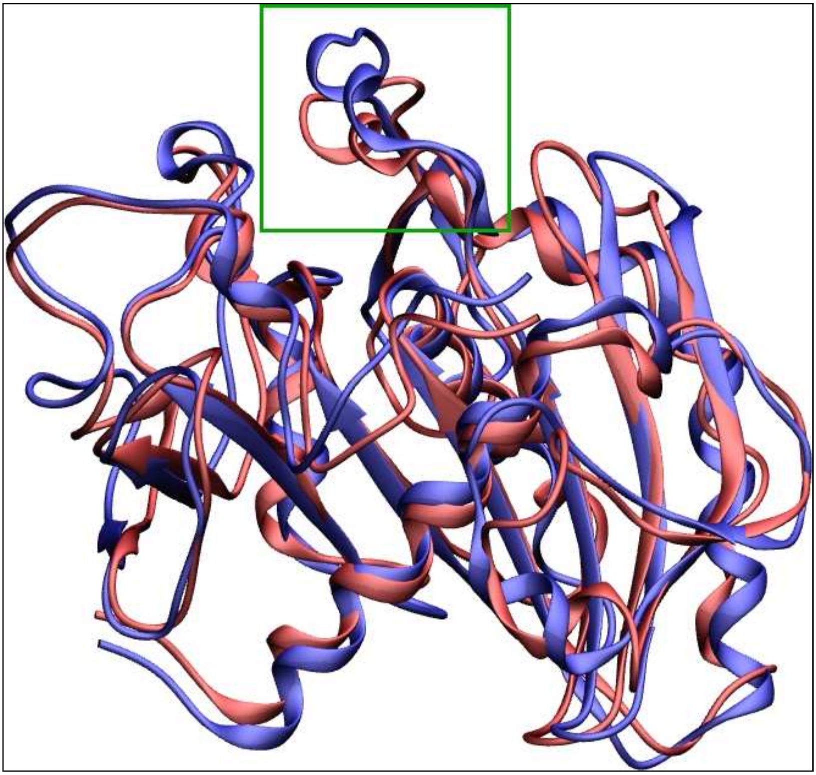


Figure 3.37: Superimposition of the crystal structure of HlL (PDB 1TIB) and the calculated one by classical MD simulation; in the green box the lid domain.

The comparison proves the quality of the simulation; some differences are observable especially in the lid domain area. These differences are expected since lid is the most mobile part of the protein and they are still small enough to prove the quality of the simulation.

The equilibrium of the MD trajectory is clearly drifted towards the closed conformation of the enzyme, cinfirming the tendency of inactivation in water environment. At he end of the trajectory the lid covers the hydrophobicity of the active site, increasing the polar surface of the enzyme, as demonstrated by surface analysis described above. Since the simulation is calculated in water, this is the driving force of the deactivation process. The loss in hydrophobicity of the surface represents an energy gain for the enzyme immersed in a polar environment, while the open conformation, remark by a more hydrophobic situation, is energetically favored in hydrophobic environment or at the interface.

3.5.2 Steered MD simulations

The classical MD application to lipase inactivation proved to be able to reproduce what it is possible to indirectly derive experimentally. Nevertheless, in one case it was not possible to observe the complete inactivation process because of the computational cost needed and in the successful cases, still at least 10 ns were necessary to be simulated. Ten ns of lipases classical MD simulation in explicit solvent condition can take several days of calculation time, even if performed on modern computers. This makes application of classical MD scheme not particularly intriguing and definitely too far from a high throughput application. Therefore, in order to reduce the computational cost of such studies a different scheme was designed. Non-equilibrium MD can accelerate the observation of lid movements by constantly perturbing the equilibrium of the simulated system. A force vector was put on the lid to accelerate the lid movement and increasing the probability of observing the phenomenon.

The attention was focused particularly on the force calibration and direction, since the aim was to accelerate the activation/deactivation movement but not to influence it too much, nor to induce undesired structural alterations on the protein structure. The lid’s closure movement is exactly the opposite of its opening movement but from a practical point of view it is easier to simulate the deactivation process. This is due to the fact that the lid closure (enzyme inactivation) happens in polar media, such as water, which is the easiest case of solvation and parameterization for an MD run. All the systems were parameterized using the GROMOS force field[57] and all the steered MD simulations were performed starting from enzymes in their open conformation in water environment, applying a fine tuned force to observe lid closure mechanism.

3.5.2.1 Steered MD on *Pseudomonas cepacia* lipase

Classical MD simulations on PcL did not permit to observe inactivation in water. It was obvious to test the approach on this enzyme first. New simulations on this enzyme were performed with a different protocol in order to observe the desired phenomenon in a reduced simulation time.

The PcL's lid was described in detail in paragraph 3.5.1.1, but the description is reported again here below, for making the analysis easier to follow.

This lid is similar to a common door with two well distinguishable hinges, the lid hinges are constituted by hydrogen bonds on its ends. The first one is fixed by hydrogen bonds formed between His114 and Ser271, and between Gly 116 and Ser271. On the other hand, the second hinge is due to hydrogen bonds between Thr150 and Ala24, and Ser152 and Asp21. This lid has an arch shape with its first half (from Gly116 to Pro131) more rigid and stabilised by 2 hydrogen bonds, between Ser117 and Leu167 and between Asp121 and Thr169. The second half of the arch (from Thr132 to Leu149) is stabilised by just one hydrogen bond between Asp144 and Ala160 (Figure 3.25).

During this steered MD simulation a force of 0.3 nm/ps^2 was applied on the lid region including residues from Ile148 to Leu134. Concerning the force direction, it was calculated starting from the Cα of Ala141 to the Cα of Ala247 (Figure 3.38).

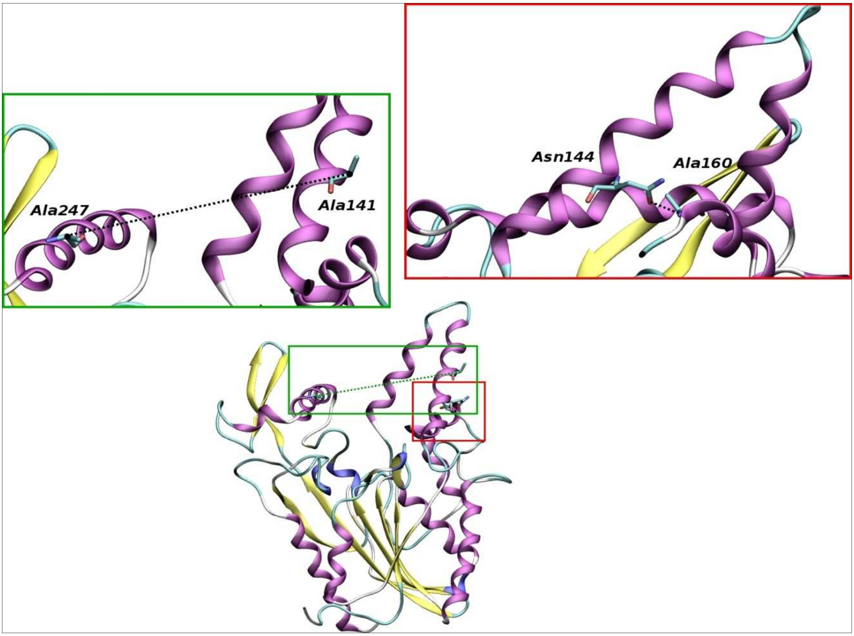


Figure 3.38: Vector force applied on PcL; on the bottom the secondary structure of the enzyme in its open conformation (PDB 1YS1), in the red box representation of the hydrogen bonds that stabilize the lid; in the green box representation of the vector orientation.

The simulation was performed starting from the open conformation of the enzyme. PcL (PDB 1YS1) was put in the centre of an 80 X 80 X 80 $Å^3$ cubic system and solvated with water. The system was first minimised and equilibrated with a 500 ps of molecular dynamic simulation in NPT conditions, during this equilibration step the enzyme was keep restrained in its position.

Afterwards the force was applied and the restrain on the enzyme was removed. A small region opposite to the lid, from Ala1 to Tyr4, was freezed in its Cartesian position during the simulation to avoid protein rotation caused by the force. MD simulation was performed for 1.5 ns in NPT conditions. At the end of this trajectory, the force was removed and also the freezing block. The system was then minimised and equilibrated with 500 ps of MD simulation in NPT conditions.

Despite the simulation has been calculated several times, the steered MD simulation on this enzyme was not successful. It was not possible to observe a correct lid closure mechanism. Different forces, changing in strength and direction were applied, but the only result was a structural distortion.

The impossibility to successfully simulate the deactivation mechanism of PcL defines this lipase as an outlier if compared to the rest of the family members. As already discussed in paragraph 3.5.1.2 its structural similarity to PaL can be the source of an explanation. In fact, PaL is substantially identical but its activation/deactivation is a concerted conformational change of two different domains, which make it apparently easier and faster. PcL at the contrary, besides having a big lid, apparently significantly flexible, probably follow a much more complex machinery.

3.5.2.2 Steered MD on *Humicola lanuginosa* lipase

The same steered MD protocol was applied on HlL also. The availability of crystal structure of both active and inactive state of the enzyme made the validation of the simulation possible. The starting point of the study was the evaluation of the point of force application, before proceeding with the simulation. Lid description analysis was performed again in order to clarify the vector force application. The lid is stabilized in its open conformation by several hydrogen bonds between different residues of the lid and different aminoacids of the nearest domain just behind it. The two most important hydrogen bonds are between Ser85 and Asp62 and between Asn88 and Asp62. The rest of the lid domain is stabilized by seven other hydrogen bonds, making this part of the domain basically fixed. This is particularly relevant, since not the entire domain is mobile, mainly because of these seven H-bond interactions (Figure 3.35).

For these reasons, the vector was put on the flexible lid region between Ser85 and Asn88 (Figure 3.40). The vector force had to be intense enough to accelerate the movement of the mobile part of the lid but also weak enough to assure a correct and not distorted movement. A trial and error strategy was applied, validating the result after each run by comparison with available crystal structures. After several tests an acceptable force was identified in 0.3 nm/ps^2.

The force vector direction was oriented from Ser85 toward Asp254 (considering the coordinates of the Cα of the residues) which lie exactly on the opposite sides of the active site entrance.

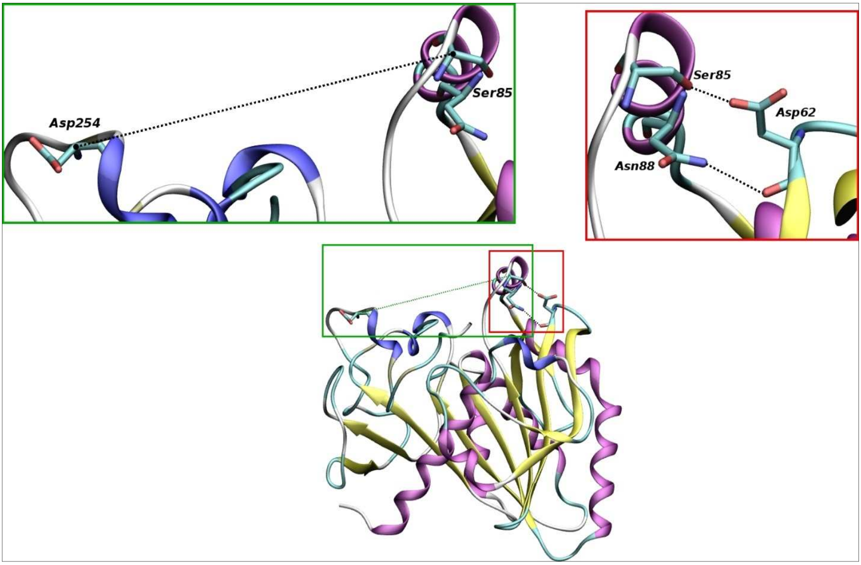


Figure 3.40: Vector force applied on HlL; on the bottom the secondary structure of the enzyme in its open conformation (PDB 1DTE), in the red box representation of the hydrogen bonds that stabilize the lid mobile part; in the green box representation of the vector orientation.

The simulation was performed starting from the open conformation of the enzyme (PDB 1DTE). HlL was put in the centre of a 80 X 80 X 80 Å^3 cubic system and explicitly solvated with water. Afterwards the system was minimised and equilibrated with a 500 ps of molecular dynamic simulation in NPT conditions. During this equilibration step the enzyme was kept restrained in its position.

After equilibration, the force was applied and the restrain on the enzyme was removed. Cartesian coordinates of a small region opposite to the lid, comprising Glu1 and Val2, were kept constant during the simulation, to avoid rotation of the whole protein in reaction to the application of the force vector. MD simulation was calculated for 2 ns in NPT conditions.

At the end of this trajectory, the force and the Cartesian constraint were removed. The system was then minimised and equilibrated by 500 ps of MD simulation in NPT conditions.

The results of the analysis show that the lid movement is reproducible and the final achieved conformation demonstrated to be reliable, since during the last equilibration phase it remains stable in the closed position. The trajectory of the steered MD simulation was used to investigate the minimum distance between Ser85 and Asp254, which showed that the lid closure takes about 750 ps with the application of this specific force. For the rest of the simulation time the lid position is quite stable even if the vector is still acting on it (Figure 3.41). This behaviour proves that the force applied was strong enough to just accelerating the closure process while not causing structural distortions, or other artefacts on the system. In this case the extension of closure movement, regarding its mobile part, was of about 9 Å (distance from the initial 19 Å to about 10 Å). During the last 250 ps of the simulation the measured Ser85-Asp254 distance was slightly increasing, despite the force. The final equilibration step stabilized it at around 11 Å.

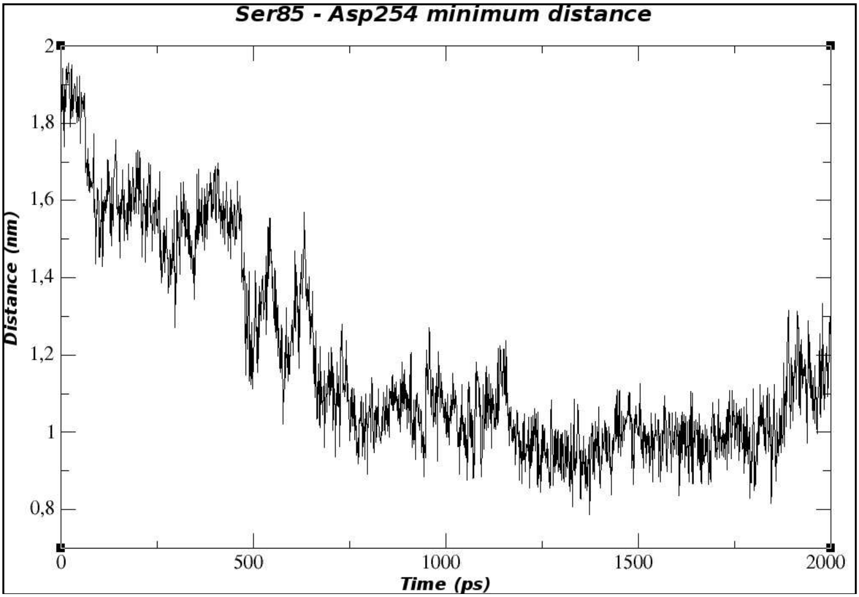


Figure 3.41: Steered MD simulation of HlL analysis; minimum distance between Ser85 and Asp254.

The calculated structure was superimposed with the PDB 1TIB structure (HlL in its closed conformation) in order to compare the results obtained from the simulation with the data coming from the crystal structure (Figure 3.42).

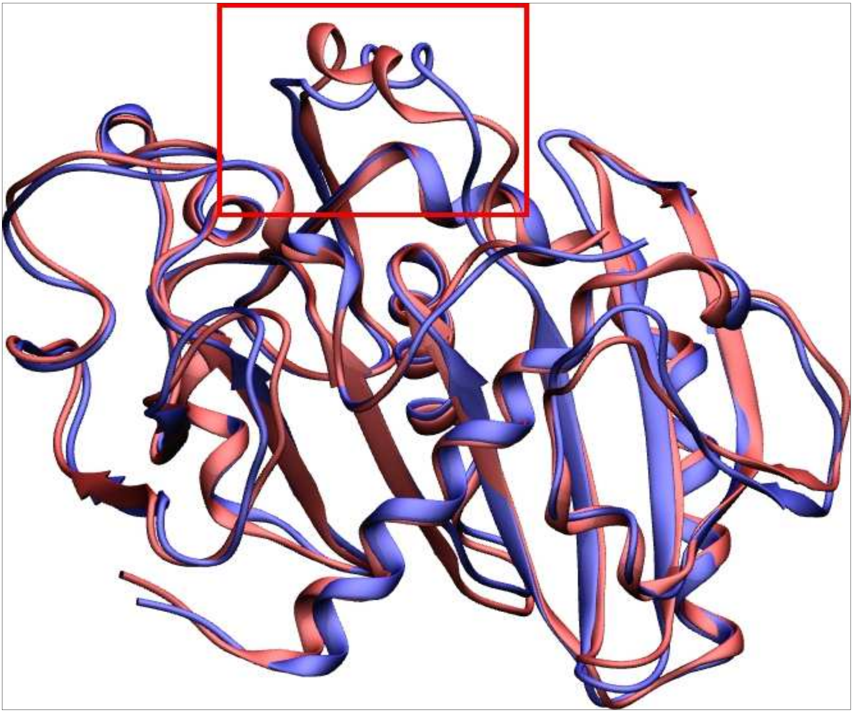


Figure 3.42: Superimposition of the calculated structure of HlL, in blue, with the crystal structure of the same enzyme (PDB 1TIB), in red; the lid is highlighted in the red box.

The comparison between the two structures demonstrates the good quality of the calculated structure. The two structures are quite similar. Some negligible differences are however visible in the lid domain where major differences could be expected as it was the most stressed part of the protein during the simulation. A quantitative comparison can be achieved by the Root Mean Square Deviation (RMSD) calculations which compute the differences in terms of spatial position among two structures. The RMSD for the crystal HlL against the calculated one was 1.32 Å. The detailed RMSD for each residue (Figure 3.43) shows that the most different domain between the two structures was represented by the residues constituting the lid domain, whereas the rest of the protein is almost identical. These minimal differences are expected because even if the applied force was accurately tuned a little deformation of the stressed region is not avoidable. However these differences cannot be considered as significant artefacts they are not evidence of bad quality of the simulation. Nevertheless particular attention has to be used during the interpretation of steered MD simulations results.

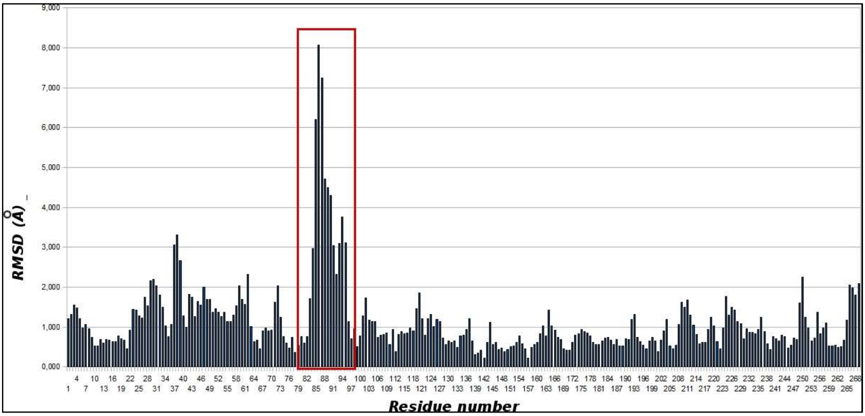


Figure 3.43: RMSD of HlL, The RMSD for each residues shows that the two proteins have almost the same structure, except for the lid domain, in the red box, where there are some differences.

Looking more closely at the calculated conformation it is clear that the lid is stabilized in its position by the generation of new weak interactions, one between Ser83 and His 258, and another between Arg84 and both Gly266 and Cys268 (Figure3.44).

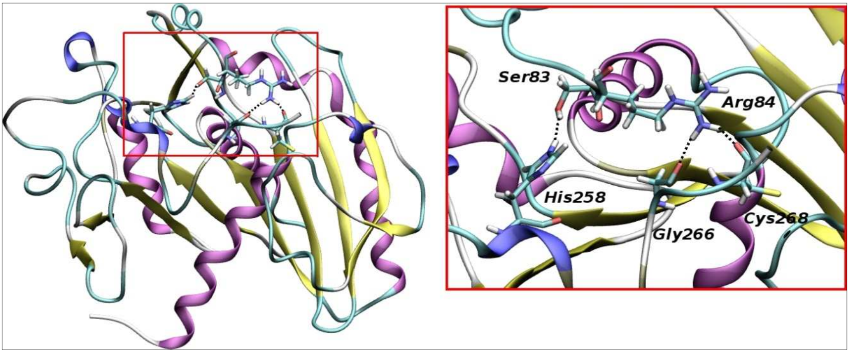


Figure 3.44: Calculated structure of HlL in its closed conformations, weak interactions that stabilised the lid in its closed form are highlighted in the red box.

Comparing the results obtained with the classical MD simulations described before, no significant differences emerged. These two simulations converge substantially to the same results.
The steered MD simulation produces an efficient lid movement with a significant gain in terms of computational time.

3.5.2.2 Steered MD on *Candida rugosa* lipase

Steered MD was used to study CrL, applying the same strategy used for HlL described above. Again, structures of CrL are available in both open and closed conformations from PDB (1CRL and 1GZ7 respectively). CrL has a big lid domain formed by residues from Glu66 to Ser94. Comparing open and closed crystal structure there is no evidence of lid sub-domains characterised by different mobility, as seen in the previous case. The lid is stabilized all along its length by three hydrogen bonds formed with aminoacids placed in the nearest domain just behind it. These three hydrogen bonds are due to the interaction between Glu67 and Gly295, Lys75 and Asn292, Gln83 and Glu287 (Figure 3.45). A force of 0.3 nm/ps^2 was put on the lid region from Glu71

to Ala89; this region comprises two of the three lid stabilising hydrogen bonds.
The force direction calculation was performed considering the Cα coordinates of Val86, which is one of the lid residues, and Ile453, which is a residue positioned on the opposite site of the active site cleft (Figure 3.45).

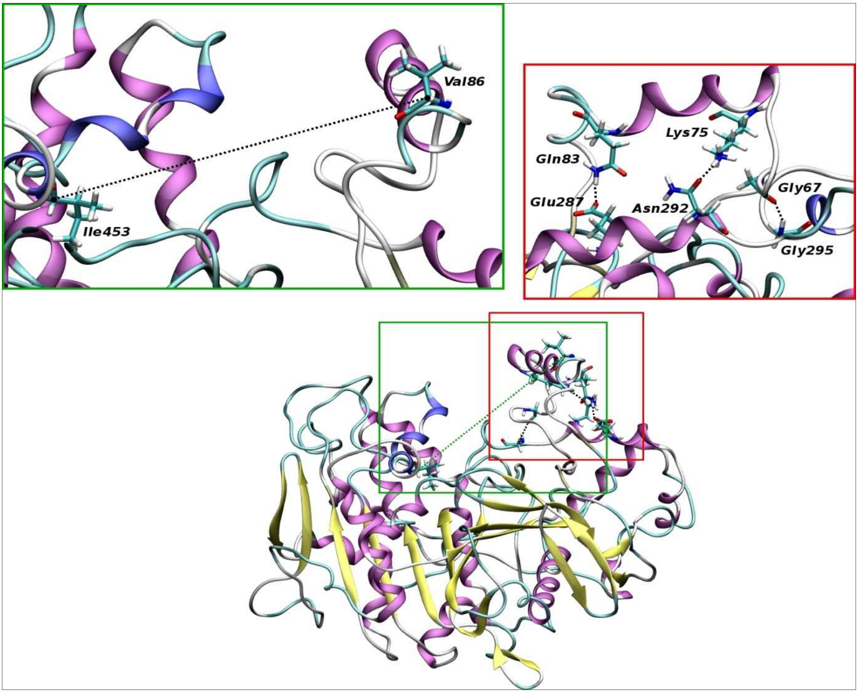


Figure 3.45: Vector force applied on CrL; on the bottom the secondary structure of the enzyme in its open conformation (PDB 1CRL), in the red box representation of the hydrogen bonds that stabilize the lid, aminoacids in licorice mode; in the green box representation of the vector orientation, aminoacids in licorice mode.

Following the same strategy applied in the case of HlL, the simulation was performed starting from the open conformation of the enzyme (PDB 1CLR). CrL was put in the centre of an 80 X 80

X 80 Å^3 cubic system and solvated with water. The system was then minimised and equilibrated with a 500 ps of molecular dynamic simulation in NPT conditions, with the enzyme kept restrained in its initial position.

Afterwards the force was applied and the restrain on the enzyme was removed. A small region opposite to the lid, from Ser94 to Asn500, was freezed in its initial cartesian position during the simulation, to avoid protein rotation as a reaction to the application of the force. MD simulation was performed for 1.5 ns in NPT conditions. At the end of this trajectory, the force and the Cartesian constraint were removed. The system was then minimised and equilibrated with 500 ps of MD simulation in NPT conditions.

The analysis of the results shows that the lid is stable in its position after the last equilibration step and remains in its closed conformation. The trajectory of the MD simulation was investigated measuring the minimum distance between Val86 and Ile453 (figure 3.46). The lid closure movement takes about 150 ps under the action of the force vector. After the closure is completed, it remains stably in the closed conformation for the rest of the simulation, even if the applied force is still acting on it. As previously stated, this behaviour proves that the force applied was just accelerating the closure process and was not causing destabilization of the whole structure. In this case the extension of closure lid movement was about 12 Å, significantly bigger than the cases inspected above.

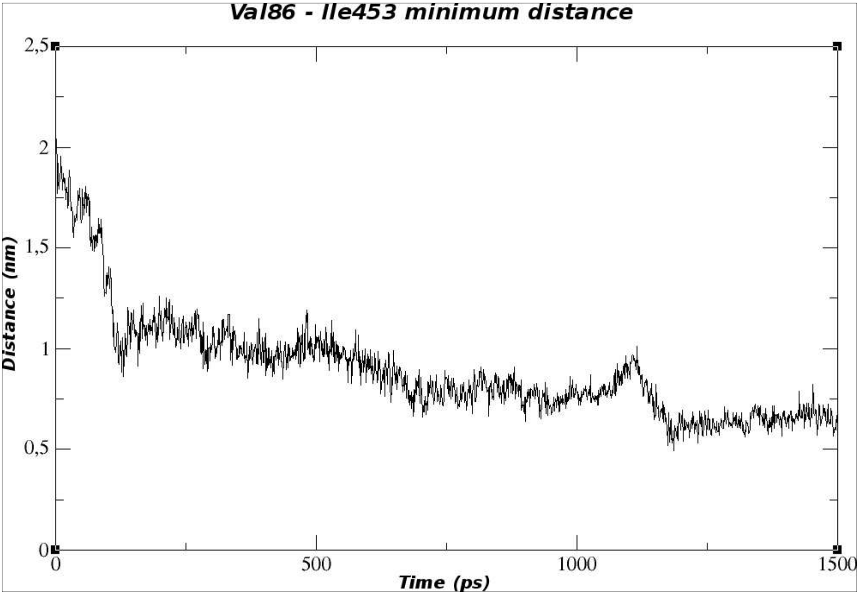


Figure 3.46: Steered MD simulation of CrL analysis; minimum distance between Val86 and Ile453.

The calculated structure was superimposed with the PDB 1GZ7 structure (CrL in its closed conformation) in order to compare the results obtained from the simulation with the data coming from the crystal structure (Figure 3.47).

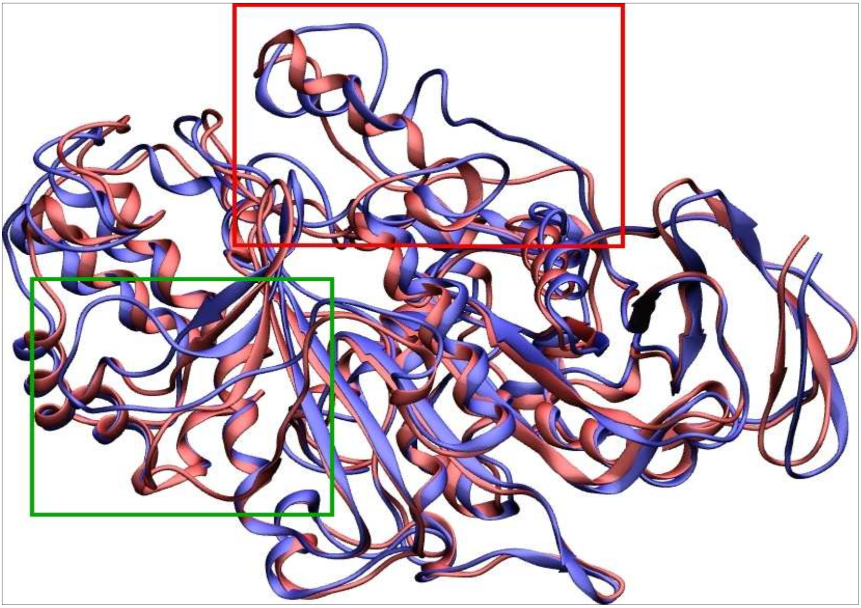


Figure 3.47: Superimposition of the calculated structure of CrL, in blue, with the crystal structure of the same enzyme (PDB 1GZ7), in red; lid is highlighted in the red box; friezed domain in the green box.

The comparison between the two structures demonstrates the high quality of the simulation. Some differences in the lid domain are actually present, but they are due to the structural stress caused by force application. Nevertheless the calculated lid conformation was still acceptably structured. There were some appreciable differences also in a loop on the surface of the enzyme opposite to the lid that corresponds to the freezed part during the simulation. This loop seems to be very mobile and the difference is caused by the freezing procedure. RMSD calculations for the crystal CrL against the calculated structure resulted into a value of 2.74 Å. The detailed RMSD for each residue (Figure 3.48) shows that the major differences between the two structures were represented by the lid domain and by the freezed domain which has the highest RMSD. The total RMDS value is surely affected by the hyperflexibility of the freezed domain since the rest of the two structures were almost identical.

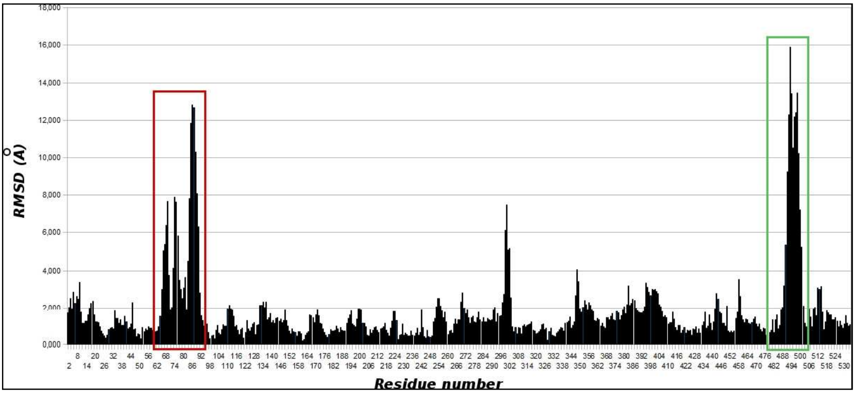


Figure 3.48: RMSD of CrL, the RMSD for each residues shows that the two proteins have almost the same structure, except for the lid domain, in the red box, and the friezed domain in the green box.

3.5.3 MD simulations using the MARTINI force field

MARTINI is a coarse grain (CG) force field. The use of CG models in a variety of simulation techniques has proven to be a valuable tool to reduce the time and the length scale of the studied systems. A large diversity of CG approaches is available; they range from qualitative, solvent-free models, through more realistic models with explicit water, to models including chemical specificity.[61, 62] MARTINI force field, has also been developed in close connection with atomistic models; however its philosophy is different. Instead of focusing on an accurate reproduction of structural details at a particular state for a specific system, MARTINI aims for a broader range of applications without the

[61] M Venturoli, M M Sperotto, M Kranenburg, B Smit, *Phys Lett*, **2006**, 437, 1.
[62] M Mùller, K Kastov, M Schick, *Phys Lett*, **2006**, 434, 113.

need to re-parametrize the model each time. It was designed by an extensive calibration of the chemical building blocks that constitute the force field, using thermodynamic data as a reference, in particular oil/water partitioning coefficients. The same concept has also been applied by the widely employed united atoms GROMOS force field.[57] The use of a consistent strategy for the development of compatible CG and atomic level force fields is of additional importance for its use in multiscale applications.[63]

Another important feature of this CG force field is determined by its simple molecule parameterization which makes MD simulations performed with MARTINI extremely fast in terms of computational time. This time gain is due to the less accuracy of the force field definition. Compared with fine grain force fields, MARTINI needs just a few numbers of particles to define the same system as briefly explained in paragraph 3.5.

The application of CG forcefield to simulate a conformational process governed by a few intramolecular weak interactions is extremely challenging. The reason for that is the dramatic simplification of the atomistic description in the CG forcefield. Nevertheless MARTINI demonstrated to be particularly successful in simulating molecular events influenced by the surrounding environment. Moreover the reduction of computational cost would make the application of this procedure particularly appealing as a support to the experimental practice.

3.5.3.1 MARTINI MD simulation on *Humicola lanuginosa* lipase

Considering the above mentioned limitations that the application of this simulation intrinsically possesses, it was necessary to focus

[63] J W Chu, G S Ayton, S Izvekov, G A Voth, *Mol Phys*, **2007**, 105, 167.

the attention to the simplest possible case study. The HlL enzyme was selected since it resulted as the most successful in the other MD approaches. HlL was defined in the force field using specific tools. The enzyme was put in the centre of a cubic space of the same dimension of the previous cases (521000 Å^3). Afterwards the system was completely filled with water (the weight of one water particle in MARTINI is 72 Da because it represents four real water molecules). After a minimization step the protein was restrained in its position and the system was equilibrated with 1 ns of molecular dynamic simulation. The restrain on the protein was then removed, and subsequently to a new minimization step, the system was subjected to a 100 ns MD simulation.

The simulation trajectory was analyzed once again measuring the minimum distance between Ser85 and Asp254 (Figure 3.49). The analysis of this trajectory generated by this type of simulation is more complex in respect to other MD protocols because of the intrinsic properties of the force field which generate simulations with a high noise. This noise is due to the force field derived system simplifications, the simulated particles are less defined and consequently less controlled than particles of fine grain force field definitions. In other words, the number of weak interactions truly taken into account is way lower than the ones calculated by fine grained force fields therefore the vibration of the system particles results much more intense. In fact comparing the same simulation performed in GROMOS and in MARTINI the average RMSD of the whole protein during the dynamics is about five times higher for the MARTINI system definition.

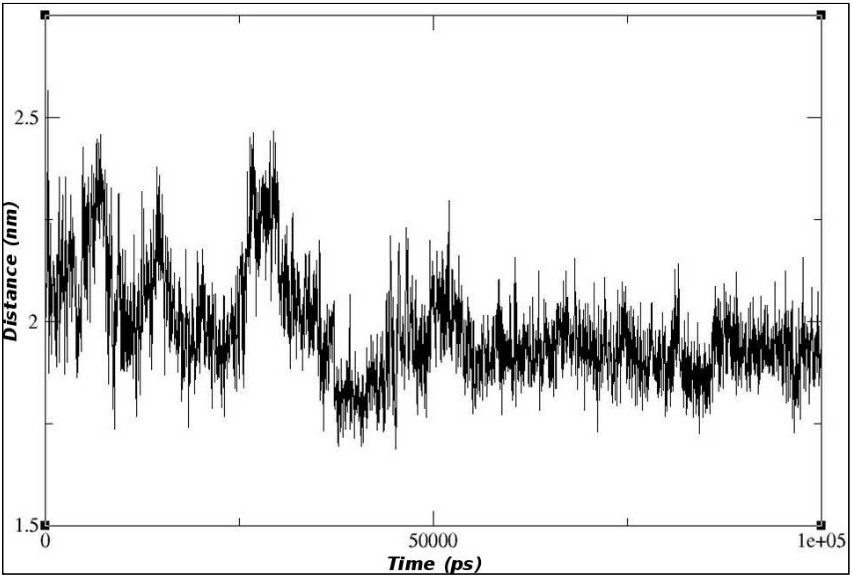


Figure 3.49: Minimum distance between Ser85 and Asp254 during the MARTINI MD simulation.

The interpretation of the analysis shows that a movement of about 0.5 nm seems to appear after 30 ns of simulation, but this is not in agreement with previous simulations, steered and classical MD, during which the lid movement is completed within 1 ns and 7 ns respectively. The noise level made any other consideration impossible.

The complete analysis of the results led to the conclusion that MARTINI lacks the necessary accuracy to be applied in the study of such a fine conformational process.

3.5.4 Other MD simulations using the MARTINI force field

Understanding the factors that can influence lipase activation is as important as gaining knowledge on the lid movements. It is obvious that important information for a full comprehension of the lipase nature can be obtained by investigating its behaviour in its natural operating conditions.

The operating environment can influence not only the activation phenomena but also the lipase localization in the medium. As described above, lipases are polarized enzymes and this feature is responsible for their orientation and localization, especially if they are put in heterogeneous systems. A correct enzyme orientation and localization is an important variable that is necessary to take into account. For instance, polymers in the catalysis environment can influence the orientation of the enzyme's active site towards the solvent or towards the solid phase. At water/oil interface a lipase will locate presumably in the middle of the two phases with the active site oriented towards the non polar phase. It is perfectly understandable that if the enzyme is too strongly attracted towards the polymer, the active site will be less accessible to soluble substrates or, on the other hand, if the catalytic machinery is completely immerse in a non polar phase it cannot act on molecules solvated in the aqueous phase.

In silico simulations of lipase orientation are significantly complicated from an operational point of view because of the time scale for the phenomenon observation is much longer than the one needed for observing conformational changes, but they do not require the same precision needed for the studies of the lid movements. For these reasons, the application of MARTINI force field appeared particularly suitable.

3.5.4.1 MARTINI interface simulation

The generation of the interface environment started with the definition of a cubic space big enough to include both the interface solution and the enzyme that has to be simulated. Usually for a simulation of a single, medium weight lipase the cubic space was set to 80 Å X 80 Å X 80 Å. Afterwards the enzyme was put in the system in a random position. Initially the cube was filled with all water molecules. The water in MARTINI is parameterized as a single sphere representing four water molecules (weight of 72 Da). This cube filled of water molecules was then minimised and equilibrated by a 500 ps of MD simulation, during which the enzyme structure was restrained in its original Cartesian position.

The hydrophobic phase was simulated afterwards; octane was chosen for the purpose. Octane molecules were added replacing water molecules in the cube and since this solvent is parameterized as two beads, two water spheres were replaced by one octane molecule. Octane was added in the quantity necessary to replace two thirds of the water particles in order to have enough space for the enzyme in each phase and to have an equimolar water/octane solution.

After the substitution (the octane molecules were randomly added in the box) the solvent box needed to be equilibrated for creating the interface. After this step, the cube was re-minimised and subjected to MD simulation for 1 ns, an equilibration step during which the lipase was restrained in its position. At this point the interface was ready.

Starting from this point the enzyme was free to move and to orient itself at the interface. Different simulations of 10 ns each were performed changing the starting position of the enzyme in order to assure that the same enzyme orientation was achieved starting from different situations. These type of studies were performed on two enzymes: CaLB and PcL. These proteins were selected because PcL is one of the investigated lipases with the highest hydrophobic surface while CaLB represents the opposite situation. In each trajectory, for both enzymes, the simulation proved that the interface was stable and enzymes were able to orient their active site toward the hydrophobic phase represented by octane. The correct orientation time depends on the starting position of the enzyme but in all the cases it takes place in the first 3 ns.

After the orientation phenomena, in all the simulations, enzymes started to rotate on its axis randomly in clockwise or contraclockwise direction and changing frequently the rotation direction (figure 3.50). This observation was particularly intriguing, since it seems like the enzymes were “trying to found something to eat (to catalyse). This behaviour is highly reproducible and it might have entropic reasons concerning diffusion of reactants and products in and out of the active site.

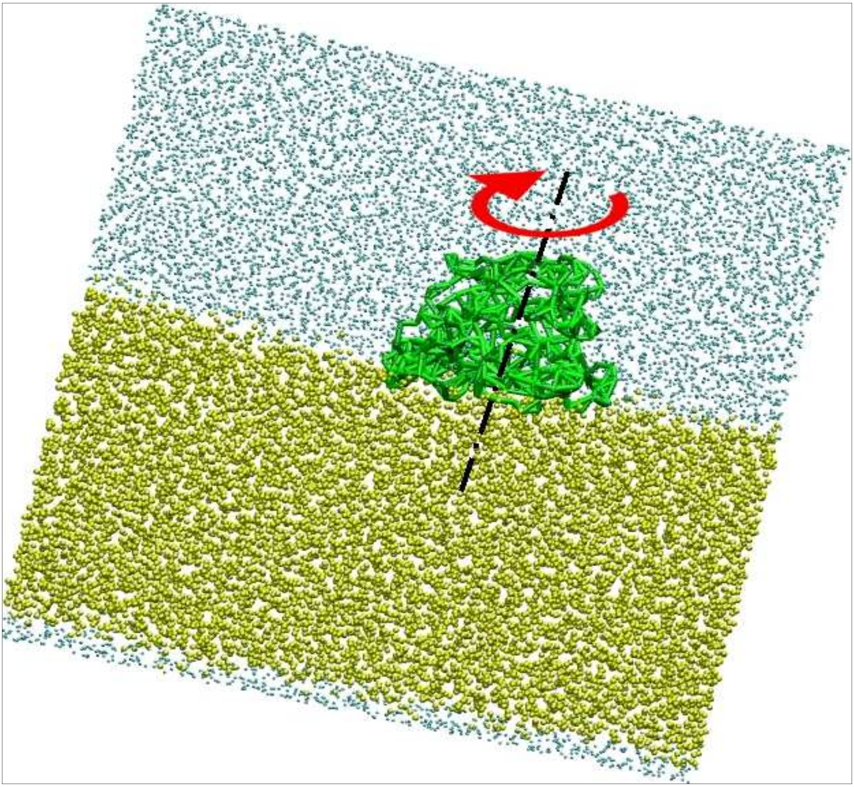


Figure 3.50: Enzyme oriented at the interface; in blue the water phase, in yellow the octane phase, in green CaLB with the active site toward the hydrophobic phase, the black line represents the enzyme axis, the red arrow indicates the enzyme rotation.

These simulations demonstrate that the MARTINI force field is perfectly able to simulate enzyme orientation phenomena, building blocks which is parameterised with are perfectly able to take into account differences in terms of hydrophobicity. These features of the force field can be used in various applications which are based on non bonding interactions. Next step for taking advantages from these types of simulation is the parameterization of more complex systems, with different solvents, several enzymes and functionalized polymers. The parameterization of these polymers is particularly interesting because it can allow to make predictions in terms of orientation of enzymes during an immobilisation procedure or in terms of proteins separation during a chromatography purification. The loss of accuracy of these CG force field is not compromising information concerning intermolecular interactions and it takes strong advantages in terms of calculation speed.

3.5.4.2 MARTINI polymer simulation

Nevertheless immobilization procedure, performed in order to enhance the protein stability, should be rationally programmed since the enzyme orientation is a key factor of to achieve the best enzymatic performances. An immobilized enzyme can truly improve its stability, but when the enzyme is bonded to a solid phase with a wrong orientation the accessibility of the active site can be compromised. The idea was to evaluate the possible enzyme orientation during an immobilization procedure. The simulation of big systems with important intermolecular interaction was performed once again with the employment of MARTINI force field. The first step of this procedure was the polymer definition. A commercial polymer, based on methacrylic units was selected for this simulation challenge (Figure 3.51).

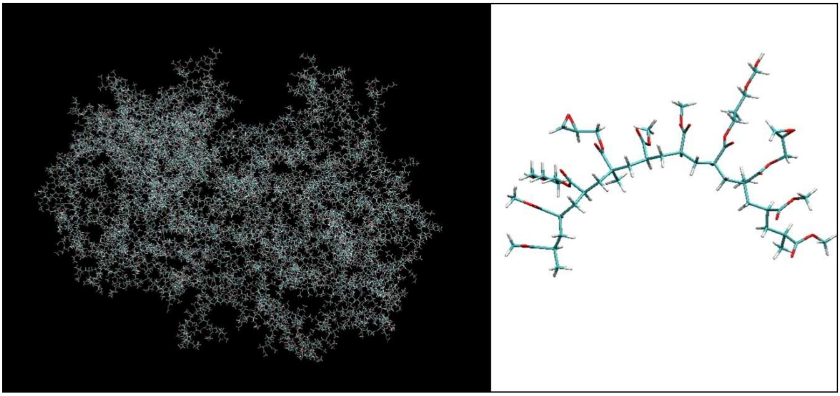


Figure 3.51: Simulated polymer. On the left an example of polymer sphere; on the right polymer single repeated unit polymer.

The correct polymer simulation in MARTINI starts with the polymer definition. This process is significantly complex because the polymer structure represents an unusual chemical specie for this force field, although MARTINI is in theory perfectly suitable for its simulation. In order to select the correct building block and to faithfully reproduce the right polymer behavior, a repeatable polymer unit of a commercial immobilization polymer was first defined in the GROMOS force field and simulated for 5 ns in vacuum.

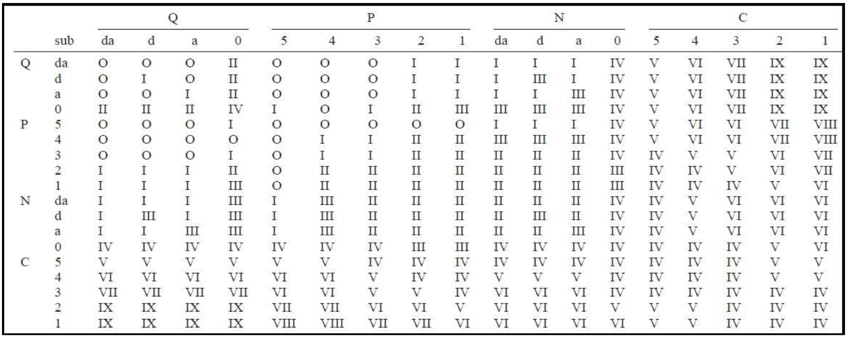
Also the polymer GROMOS definition was not an automatic procedure, but in this case some tools such as the Dundee PRODRG2 server[64] helps during this delicate parameterization.

MARTINI particles are defined focusing particularly attention to the non-bounded interactions defined by energy non-bounded interactions calculated by means of Lennard-Jones potential energy function.[65] There are just four particle types definition in the MARTINI force field considering the polarity, each particle type is differentiated is other four or five sub classes concerning

[64] A W Shuettelkopf, D M F van Aalten, *Acta Crystallog*, **2004**, D60, 1355.

[65] J E Jones, *Proc R Soc Lon A*, **1924**, 106, 463.

the possibilities to establish hydrogen bonds (Table 1.2).



		Q				P					N				C				
	sub	da	d	a	0	5	4	3	2	1	da	d	a	0	5	4	3	2	1
Q	da	O	O	O	II	O	O	O	I	I	I	I	I	IV	V	VI	VII	IX	IX
	d	O	I	O	II	O	O	O	I	I	I	III	I	IV	V	VI	VII	IX	IX
	a	O	O	I	II	O	O	O	I	I	I	I	III	IV	V	VI	VII	IX	IX
	0	II	II	II	IV	I	O	I	II	III	III	III	III	IV	V	VI	VII	IX	IX
P	5	O	O	O	I	O	O	O	O	O	I	I	I	IV	V	VI	VI	VII	VIII
	4	O	O	O	O	O	I	I	II	II	III	III	III	IV	V	VI	VI	VII	VIII
	3	O	O	O	I	O	I	I	II	II	II	II	II	IV	IV	V	V	VI	VII
	2	I	I	I	II	O	II	II	II	II	II	II	II	III	IV	IV	V	VI	VII
	1	I	I	I	III	O	II	II	II	II	II	II	II	III	IV	IV	IV	V	VI
N	da	I	I	I	III	I	III	II	II	II	II	II	II	IV	IV	V	VI	VI	VI
	d	I	III	I	III	I	III	II	II	II	II	III	II	IV	IV	V	VI	VI	VI
	a	I	I	III	III	I	III	II	II	II	II	II	III	IV	IV	V	VI	VI	VI
	0	IV	IV	IV	IV	IV	IV	IV	III	III	IV	IV	IV	IV	IV	IV	IV	V	VI
C	5	V	V	V	V	V	V	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV	V	V
	4	VI	VI	VI	VI	VI	VI	V	IV	IV	V	V	V	IV	IV	IV	IV	V	V
	3	VII	VII	VII	VII	VI	VI	V	V	IV	VI	VI	VI	IV	IV	IV	IV	IV	IV
	2	IX	IX	IX	IX	VII	VII	VI	VI	V	VI	VI	VI	V	V	V	IV	IV	IV
	1	IX	IX	IX	IX	VIII	VIII	VII	VII	VI	VI	VI	VI	VI	V	V	IV	IV	IV

Table 1.2: Level of interaction indicates the well depth in the LJ potential: O, *E*= 5.6 kJ/mol; I, E=5.0 kJ/mol; II, E=4.5 kJ/mol; III, E=4.0 kJ/mol; IV, _) 3.5 kJ/mol; V, *E*=3.1 kJ/mol; VI, *E*=2.7 kJ/mol; VII, *E*=2.3 kJ/mol; VIII, *E*=2.0 kJ/mol; IX, *E*=2.0 kJ/mol. The LJ parameter *δ E*= 0.47 nm for all interacion levels except level IX for which *δ E*=0.62 nm. Four different CG sites are considered: charged (Q), polar (P), nonpolar (N), and apolar (C). Subscripts are used to further distinguish groups with different chemical nature: 0, no hydrogen-bonding capabilities are present; d, groups acting as hydrogen bond donor; a, groups acting as hydrogen bond acceptor; da, groups with both donor and acceptor options; 1-5, indicating increasing polar affinity.

The results of the GROMOS simulation were used as a template for the MARTINI definition. Each particle characteristics and each correlation with the other particle of the molecule has to be accurately regulated. MARTINI particle selection was based on the chemical characteristics of the molecule moiety defined by the MARTINI particle. Afterwards each bond distance, angle and dihedral was deeply investigated by mapping the molecule behavior during the reference GROMOS trajectories in order to reproduce the same behavior in the MARTINI definition.

This step of the work was particularly tedious because each one of these parameter has to be defined with a try and error approach for each bond, angle or dihedral. The verification of the correct parameterization were achieved by 5 ns of MARTINI MD simulation and the behavior of the molecule was compared with the reference force field simulation.

The repeatable polymer unit only was correctly and completely parameterized in the MARTINI force field because of the long definition time required, the polymer definition is still under development (Figure 3.52).

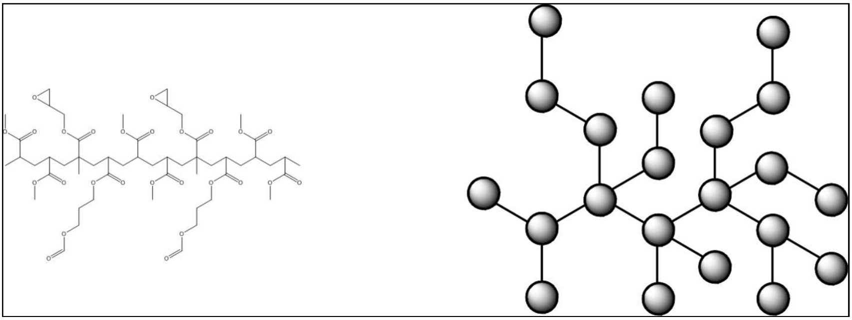


Figure 3.52: Polymer definition; polymer chemical structure on the left; on the right MARITNI force field definition.

3.7 Conclusions

Grid method was successfully applied in the study of chemical-physical properties and demonstrates the role of the lid movement which is involved in lipase activation process. Concepts comprehended by MIFs applications lead to deeper investigation by molecular dynamics simulations. MD simulations results prove to be suitable to reproduce enzymes behaviours. Steered MD simulations were successfully applied to simulate accelerate lid's movement. This technique reduces the computational time necessary to observe enzyme activation. Difficulties in this application are due to the delicate compromise between the event acceleration and the necessity to do not influence the nature of the process and to avoid structural distortions. Moreover the direction of the acceleration has to be carefully selected to avoid effects in contrast with the natural lid dynamics.

Classical MD simulations are able to simulate the same phenomena anyhow and they must be considered the standard reference for those kinds of investigations. They proved to be useful in the study of lipases activation/deactivation, at least when the mechanism can be observed within a reasonable time frame. On the other hand, the computational power requested by this simulation type is at least ten times larger than in the case of steered MD simulation. Moreover, an accurate tuning of force vector to be applied on the lid, with the aim of reducing the energy barrier of the conformational change initiation, ensure an absolute quality of conformations and dynamic, perfectly comparable to the classical MD scheme. It must be taken into account, however, that the delicacy of the parameters settings requested by steered MD simulations requires time for testing to optimize vector force regulations.

Concerning lipases, different activation/deactivations mechanisms were highlighted by simulation results. Huge lid bacterial lipases demonstrate complex lid movement, where more than one protein domain are usually involved during the activation process, while eukaryotic lipases, with a considerable higher molecular weight, usually have a small lid domain acting with more simple movements governed just by a few hydrogen bonds. These behaviours can lead to a new lipases classification based on lid movement complexities.

Finally, MARTINI force field was successfully applied to more complex simulations, namely in interface simulations. This promising force field shows all its potential where several weak bonds are involved like in division coefficient problems or enzymes orientation. After an initial time consuming part necessary for system parameterization, simulations based on this force field result very fast. The loss of accuracy due to the intrinsic characteristics of the force field is balanced by a huge gain in terms of computational time. Nevertheless, MARTINI demonstratd to be inappropriate to simulate phenomena finely governed by few weak interactions, such as activation/inactivation mechanisms. Applications of MARTINI force field are strongly suggested where the simulation of the problem does not require high accuracy like intermolecular events investigations while for intermolecular studies a fine grain force field like GROMOS represents a better choice.

3.8 Experimental section

The protein structures used for this study were retrieved from the Protein Data Bank. Initial structures were pretreated in pymol by removing the crystallographic water molecules and sugar molecules eventually present (usually N-acetylglucosamine) in the pdb file. Hydrogen atoms were added in dependence on the force field characteristics.

Homology modelling

Aminoacid sequences of the enzymes generated by homology modelling were taken from UniProtKB database. Homology search was performed on SRS@EBI server by using BlastP algorithm.[66] Chain alignment were performed with the align algorithm of the MOE program by using the Blosum 30 amino acid substitution matrix with a tree-based method.[67]
The construction of the three-dimensional models was carried out with the MOE homology modelling module calculating 10

[66] S F Altschul, W Gish, W Miller, E W Myers, D J Lipman, *J Mol Biol*, **1990**, 215,403.

[67] S Henikoff, J G Henikoff, *Proc Nat Acad Sci U S A*, **1992**, 89, 10915.

intermediate models for each enzyme. The obtained structures were ranked by the structure quality Z score of the MOE and one model for each enzyme was chosen on the basis of the Z score. The quality of the generated models was assessed by Ramachandran plot.

Final structures were parametrised in AMBER94 force field[68] and refined with the program MOE by energy minimisations, using both steepest descendent and conjugated gradient algorithms, and molecular dynamics simulations in NVT conditions.

Surface analysis

Matrices of interaction energies (MIFs) were calculated by using the GRID methods with WATER and DRY probes. For each structure the calculation was performed simulating a grid with 0.5 Å knots distance. The calculated MIFs were visualized with the program Gview setting -4.00 kcal for WATER interactions and -0.25 kcal for DRY interactions.

Classical MD simulations

All the steered molecular dynamics simulations were performed using the software GROMACS with the GROMOS96 53a6 force field.

Protein structures were implemented in the force field in gro file format by using the automatic tool of the GROMACS software which also add the necessary hydrogens. Proteins were solvated with explicit water in virtual cubic boxes of 512 nm^3. All the

[68] W D Cornell, P Cieplak, C I Bayly, I R Gould, K M Merz, D M Ferguson, D C Spellmeyer, T Fox, J W Caldwell, P A Kollman. *J Am Chem Soc*. **1998**, 117, 5179.

dynamics were performed in a NPT environment simulating the temperature of 300 K and keeping the pressure constant (Berendsen-thermostat and pressure),[69] cut-off for electrostatic interaction was setted at 1.4 nm and the limit for the van der Waals interactions setted at 1.4 nm. Only for the minimization procedures the PME algorithm[70] (Particle Mesh Ewald and not a simple cut-off) was used for the calculation of the electrostatic interactions setting the limit at 1.0 nm and using both steepest descendent and conjugated gradient algorithms and performing a steepest descendent step every 100 conjugated gradient step. Minimization were performed until the maximum force was smaller than 10 kJ/mol·nm or at least for 1000 steps. Systems were minimized every molecular dynamics calculation.

Molecular dynamics analysis were performed with GROMACS tools; distances measurements were calculated using g_mindist which computes the minimum distance between two residues and using g_rms which computes the RMSD between two structures; the results were visualized using Grace software.

Steered MD simulations

All the steered molecular dynamics simulations were performed using the software GROMACS with the GROMOS96 53a6 force field.

Protein structures were implemented in the force field in gro file format by using the automatic tool of the GROMACS software which also add the necessary hydrogens. Proteins were solvated with explicit water in virtual cubic boxes of 512 nm^3. All the dynamics were performed in a NPT environment simulating the temperature of 300 K and keeping the pressure constant (Berendsen-thermostat and pressure),[69] cut-off for electrostatic

[69] H J C Berendsen, J P M Postma, A DiNola, J R Haak, *J Chem Phys*, **1984**,

interaction was setted at 1.4 nm and the limit for the van der Waals interactions setted at 1.4 nm. Only for the minimization procedures the PME algorithm[70] (Particle Mesh Ewald and not a simple cut-off) was used for the calculation of the electrostatic interactions setting the limit at 1.0 nm and using both steepest descendent and conjugated gradient algorithms and performing a steepest descendent step every 100 conjugated gradient step. Minimization were performed until the maximum force was smaller than 10 kJ/mol·nm or at least for 1000 steps. Systems were minimized every molecular dynamics calculation. Forces intensity were selected by a try and error approach. Directions were calculated considering the Cα coordinates of the two residues for each enzyme.

Molecular dynamics analysis were performed with GROMACS tools; distances measurements were calculated using g_mindist which computes the minimum distance between two residues and using g_rms which computes the RMSD between two structures; the results were visualized using Grace software.

MD using the MARTINI force field

Molecular dynamics simulations were performed using the software GROMACS with the MARTINI force field.

Protein structures were implemented in the force field in gro file format by using the necessary scripts available on the MARTINI web site[71] and the DSSP program[72] for the necessary secondary structure definition. Proteins were solvated with explicit water in virtual cubic boxes of 512 nm^3. All the dynamics were performed in a NPT environment simulating the temperature of 300 K and

81, 3684.

[70] P Ewald, *Ann Phys*, **1921**, 369, 253.

[71] http://md.chem.rug.nl/cgmartini/index.php/home.

[72] K Wolfgang, C Sander, *Biopolymers*, **1983**, 22, 2577.

keeping the pressure constant (Berendsen-thermostat and pressure),[69] cut-off for electrostatic interaction was setted at 1.4 nm, the limit for the van der Waals interactions setted at 1.4 nm and the time step for integration set to 4 fs (usually this value is set to 2 fs). Minimization procedures were performed using cut-off for electrostatic interactions setting the limit at 1.4 nm and using steepest descendent algorithm. Minimization were performed until the maximum force was smaller than 10 kJ/mol·nm or at least for 1000 steps. Systems were minimized every molecular dynamics calculation.

CHAPTER 4

SOLVENT STABILITY

4.1 Introduction

The possibilities to perform enzymatic transformations in pure (neat) organic solvent is an accepted and applied property of several enzyme class.[1] Moreover the use of organic solvents in biocatalysis applications is a fundamental requirement that can be due to different factors like substrate solubility as well as the reaction type.[2]

Usually enzymes show a more stable behaviour in pure organic solvent than in water-organic mixtures.[3] In these kind of mixtures water acts as a molecular lubricant while in neat solvent enzymes result very rigid.[4] Therefore is plausible that in organic media enzymes have the naturally trend to unfold but not the necessary flexibility to do it.[5]

The stability of an enzyme is affected by many factors, such as temperature, pH, oxidative stress, the solvent, binding of metal ions or co-factors, and the presence of surfactants. The effect of organic solvents is important since the presence of such solvents is often essential when applying enzymes for the production of fine chemicals.[6]

It is often assumed that enzymes with improved thermal stability also become more resistant to other denaturing factors. However, this correlation is not absolute, especially not when it comes to denaturation processes which do not, or to a minor extent, depend

[1] A M Klibanov, *Nature*, **2001**, 409, 241.

[2] V G H Eijsink, S Gaseidnes, T V Borchert, B van den Burg, *Biomolecular Engineering*, **2005**, 22, 21.

[3] S Lapanje, *Physicochemical Aspects of Protein Denaturation*, **1978**, Wiley, New York.

[4] J A Rupley, G Careri, *Adv Protein Chem*, **1991**, 41, 37.

[5] A J Straathof, S Panke, A Schmid, *Curr Opin Biotechnol*, **2002**, 13, 548.

[6] V G H Eijsink, S Gaseidnes, T V Borchert, B van den Burg, *Biomolecular Engineering*, **2005**, 22, 21.

on folding stability.[7]

There is still a lack of knowledge concerning the factors that can influence enzyme stability, especially in aqueous-organic mixtures. Nevertheless there is a clear evidence about the need of a predicting tool, or at least of a set of rules to speed up the process design when dealing with aqueous-solvent, pure solvent, or multiphasic systems. The application of hybrid molecular modelling and pure experimental studies could actually shed light on this matter.

From the computational point of view MD simulations can represent a suitable tool for investigation of enzyme stability, although the simulation of a complex chemical system can push the computational cost too further for a really applicable tool. Nevertheless, as deeply described in chapter 3, coarse grained force field based MD, more precisely MARTINI force field based MD, demonstrated to be valuable in the simulation of lipases in bi-phasic systems, while reducing significantly the calculation time.

It must be underlined that enzyme stability is a complex phenomenon. Instability can be well represented by the loss of tertiary structure, but unfolding usually happens in many tens of nanosecond, also in the most disadvantageous conditions, therefore it is necessary to increase significantly the simulation time. MARTINI has a strong parameterization and testing in case of proteins and it is reasonably cheap in terms of computational resources. For that reasons it seems an ideal candidate for the development of a computational scheme for predicting enzyme stability.

[7] S D’Amico, J C Marx, C Gerday, G Feller, *J Biol Chem*, **2003**, 278, 7891.

4.2 General strategy

Stability of lipases to alcohol is of particular interest. Being hydrolases, the reaction they catalyse usually produces alcohol, or uses alcohol to synthesize an ester. A remarkable number of industrial applications, especially in the food industry and in the biodiesel production, would benefit from a rational approach to the improvement of lipase stability to alcohol. Experimental measurements can be time consuming and significantly costly, therefore a support tool for predicting solvent stability would be of great utility for the experimental practice.

This is why the stability of CaLB in water/propanol mixtures was taken as a case study.

Enzyme rigidity, as explained in the introduction above, is supposed to be strictly related with enzyme stability. Structural flexibility is a feature that can be easily simulated and measured by analysing the outcome of extended MD simulations. The rigidity of a given enzyme can be measured simulating, for a sufficiently long time, the protein in operational conditions and subsequently calculating the Root Mean Square Deviation (RMSD) of the structure. This is a qualitative parameter, obviously affected by the force field parameterization. Nevertheless, as soon as a set of simulations is calculated under the same protocol and using the same force field, the structural RMSD of the protein can represent a powerful score of enzyme stability. For that reason, performing simulations in different aqueous-organic mixtures and extracting RMSD for each condition, makes possible to obtain estimations of the relative enzyme stability in the different conditions, as soon as the force field has an appropriate parameterization for the system of interest.

In order to transform these qualitative values into true quantitative estimations, RMSD data were correlated with the residual activity of the enzyme measured after experimental incubation in the same conditions. Regression between the two sets of data enables the calculations of enzymes stability models, that can find practical application in the set up of biocatalysed synthetic processes.

4.3 RMSD calculations

Calculations of RMSD were performed analysing the outcome of each MD simulation performed in different conditions. The extensive simulations of different aqueous-organic mixtures are the typical cases where the MARTINI force field[8] represents a first choice. MARTINI force field guarantees minimum computational cost and it is particularly suitable for simulating different solvents and solutions.

The simulation space was defined as a 512 nm^3 cube which has in its geometrical centre the protein already defined into the MARTINI force field. The remaining volume was filled with all water molecules. It is important to remember that water in MARTINI is parametrized as a single sphere which represents four real water molecules, for this reason a MARTINI water sphere weight is 72 Da. After a minimisation step the system was minimised with a 1 ns MD simulation during which the enzyme was restrained in its position.

The organic fraction was simulated using propanol. Propanol molecules were added replacing water molecules in the cube; this solvent is parametrized as one spheres also, thus four water molecules were replaced by one propanol molecule. Propanol was added in different molar fractions depending on the desired aqueous-organic mixture concentration taking into account. The replaced water fraction was calculated considering the nature of water sphere: for example a solution formed by 10 water beads

[8] S J Marrink, H J Risselada, S Yefimov, D P Tieleman, A H de Vries, *J Phys Chem B*, **2007**, 111, 7812-7824.

and 10 propanol beads represents a solution of 20 % of propanol because each water sphere simulate four water molecules. After the substitution (the propanol molecules were randomly positioned in the box) the solvent needs to be equilibrated for the homogeneous solution generation. Therefore the cube was re-minimised and subjected to MD simulation for 1 ns. During this equilibration step the lipase was restrained again in its position. At this point the solution is ready.

Afterwards, the restrain on the enzyme was removed and each solution system was subjected to extended MD simulation.

4.4 CaLB RMSD calculation

A first stability model was generated on the enzyme CaLB. 65 to 100 ns MD simulations of CaLB in different water/propanol solutions were performed and global structural RMSD was measured during the trajectory. The data (Figure 4.1) show that in all the simulations RMSD value reaches a plateau after approximately 40-50 ns and this value is strongly dependent on the concentration of propanol. The simulations reproduced remarkably well the expected decrease of stability with the increase of propanol concentration in the mixture. For instance in 50% propanol the enzyme reaches a 2nm RMSD, increased to 3nm in 100% propanol. Obviously those values are indices of real protein denaturation. At the contrary, the RMSD value of the 100% water simulation is about 0.7 nm, perfectly compatible with a completely and correctly folded structure.

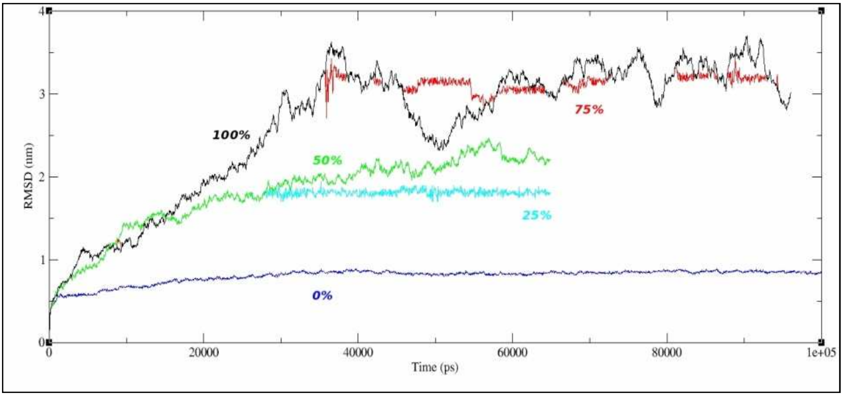


Figure 4.1: RMSD calculation of CaLB in different water/propanol simulations; the percentage indicates the propanol quantity.

4.5 Stability model

Commercial preparation of native lipase (Lipozyme CaLB-L, purchased from Novozyme) was previously subjected to a dialysis procedure in order to eliminate all the undesired compounds that can potentially interfere with activity assay such as preserver or stabilizer; the final preparation was solved in Kpi buffer 0.1 M pH 7.0. Afterwards the dialyzed protein solution was characterized in terms of protein content and specific activity; more in detail the protein solution concentration was 13.11 mg/mL with a specific activity of 385.28 U/mg.

Different solution of water/propanol mixtures were prepared in order to study the enzyme stability. Each stability test was performed in a final volume of 5 mL of the desired solvent mixture using 65.55 µg of enzyme at constant temperature of 30 °C. Enzyme activity assay were performed at different time for each solvent mixture tested.

A general qualitative agreement is evident comparing the experimental measurement with the results of the MD simulations.

The experimental data (Figure 4.2) show that the residual activity decreases with the increase of the propanol concentration. In pure water the protein retains more than 80 % of its activity after 1500 minutes, while residual activity decreases to 20 % for the 100 % propanol condition.

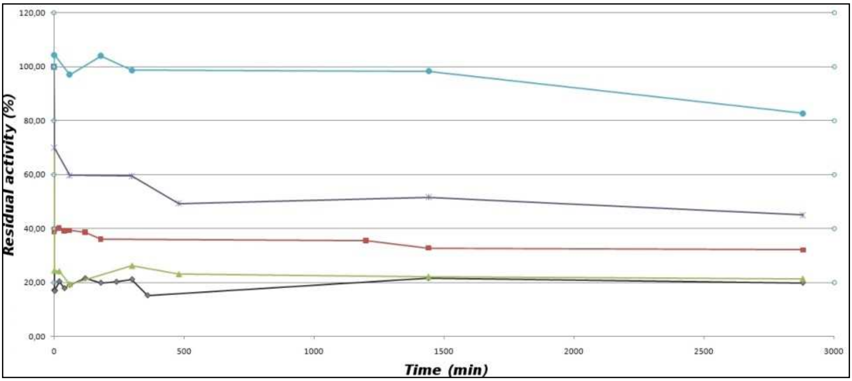


Figure 4.2: Experimental CaLB stability in different water/propanol mixtures, residual activity expressed as percentage. In blue roundes 0 % propanol, in violet crossess 25 % propanol, in red squares 50 % propanol, in green triangles 75 % propanol and in charcoal rhombus 100 % propanol.

The regression of the average RMSD achieved after 40 ns of simulation and the residual activity after 1500 minutes allowed to find out a linear correlation (with a correlation coefficient R^2 of 0.96) between the two sets of data. The outcoming linear equation represents a simple yet useful stability model (Figure 4.3) for CaLB. The impressive correlation between the experimental data and the simulations demonstrates the effectiveness of the computational approach for predicting such a complex mechanism.

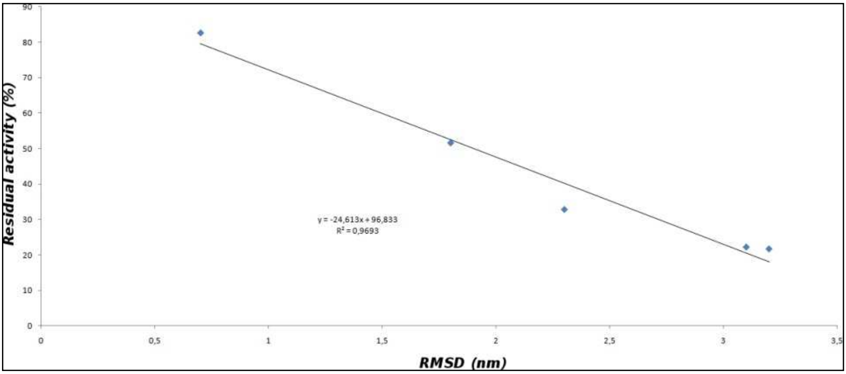


Figure 4.3: CalB stability model, on top left the straight line equation with its relative R^2 value.

4.6 Conclusions

These models prove that the RMSD can be used as a parameter to evaluate the enzyme stability in different solvent conditions. These stability models can be easily generated for any lipase with just two information needed: the protein structure and the possibility to experimentally measure residual activity in different conditions.

With the availability of these kind of models it will be possible to select in advance the suitable solvent mixture in order to have a sufficiently high enzyme activity and perhaps to meet industrial requirements.

The scheme demonstrated its potential in the case of CaLB, but it will be applied to other enzymes in the prospected work, to assess the general applicability of the concept.

Another possible step in this direction will be the identification of structural parameters that can be directly used as descriptors of enzyme rigidity, but for that purpose several steps in the comprehension of structural mechanisms of protein folding still appear necessary.

4.7 Experimental section

MD simulations

All the molecular dynamics simulations were performed using the software GROMACS with the MARTINI force field.

Protein structures were implemented in the force field in gro file format by using the necessary scripts available on the MARTINI web site[9] and the DSSP program[10] for the necessary secondary structure definition. Proteins were solvated with explicit water in virtual cubic boxes of 512 nm^3. All the dynamics were performed in a NPT environment simulating the temperature of 300 K and keeping the pressure constant (Berendsen-thermostat and pressure),[11] cut-off for electrostatic interaction was setted at 1.4 nm, the limit for the van der Waals interactions setted at 1.4 nm and the time step for integration set to 4 fs (usually this value is set to 2 fs). Minimization procedures were performed using cut-off for electrostatic interactions setting the limit at 1.4 nm and using steepest descendent algorithm. Minimization were performed until the maximum force was smaller than 10 kJ/mol·nm or at least for 1000 steps. Systems were minimized before every molecular dynamics calculation. RMSD were calculated with the GROMACS tool g_rms.

[9] http://md.chem.rug.nl/cgmartini/index.php/home.

[10] K Wolfgang, C Sander, *Biopolymers*, **1983**, 22, 2577.

[11] H J C Berendsen, J P M Postma, A DiNola, J R Haak, *J Chem Phys*, **1984**, 81, 3684.

Dialysis

2 mL of commercial enzyme preparation Lipozyme CaLB-L purchased from novozyme, was diluted in Kpi buffer 0.01 M pH 7.0 to a final volume of 10 mL. The preparation was then put in a dialysis membrane (14 kDa pore size), and dialyzed versus Kpi buffer 0.01 M pH 7.0. The procedure was performed for 24 hours with renewer of the washing buffer every 8 hours.

Protein determination

The protein content of Lypozyme CaLB-L was determined by using bicinchoninic acid kit
(SIGMA) - Pierce method, using BSA as standard protein.

Activity assay

The assay is based on the hydrolysis of glycelyl tributyrate into butyric acid. 30 mL of an emulsified solution 0.17 M of glyceryl tributyrate was added with 50 µL of dialyzed enzyme solution at constant temperature of 30 °C. The butyric acid produced by the enzymatic hydrolysis was measured by reaction with NaOH 0.1 M. The reaction was followed during time. One activity unit corresponds to the amount of enzyme that hydrolyses 1 µmol of glyceryl tributyrate in one minute at 30 °C.

CHAPTER 5

STATISTICAL ANALYSIS

Enantioselectivity Prediction

5.1 Introduction

Different molecular dynamics strategies were described in the previous chapters. Important information concerning potential MD applications on the study of enzymes activation and stability were already gained. Although in many cases the goals of reaching a quantitative estimation of the desired properties and of keeping the computational cost as low as possible have been achieved, there are other cases where MD still shows strong limitations. The investigation of catalytic properties, such as enzyme selectivity, are certainly one of those cases. In the study of enzyme-substrate interactions MD surely represents the reference for conformational analysis, but it still is unsatisfactory in the generation of robust quantitative predictions of enzyme (enantio)selectivity, even when applying very detailed simulation schemes, such as free energy perturbation. Very complex approaches can give practical indications, but they are too computationally expensive and their outcome cannot usually be considered more that a qualitative evaluation of enzyme kinetics. For that reasons the need for completely different approaches is emerging.

In the present chapter an original protocol based on the concept of 3D-QSAR (three-dimensional Quantitative Structure-Activity Relationships) is described. The idea comprises the combination of molecular modelling techniques, molecular descriptors calculation and statistical regression to a set of available experimental data by means of multivariate statistics.

Once again *Candida antarctica* lipase B (CaLB) was taken as a case study. Despite its extensive application and huge number of publications based on it, its peculiar enantioselectivity is still very hard to be predicted quantitatively.[1] This makes it the perfect

[1] R J Kazlauskas, A N E Weissfloch, A T Rappaport, L A Cuccia, *J Org Chem*, **1991**, 56, 2656.

candidate for the development of the concept.

The CaLB active site is located in a deep and relatively small cavity if compared to the total size of the enzyme. The binding site of the enzyme has a funnel shape and it contributes to drive the substrate in the active site assuring the correct orientation. Like all the lipases, CaLB presents a catalytic site constituted by the aminoacidic triad Ser105-His224-Asp187. The so called catalytic machinery is completed by the hoxyanion site (Thr40, Gln106) which makes possible tetrahedral intermediate (TI) formation. The spatial geometry of the hoxyanion site makes possible the formation of three hydrogen bonds with the carbonilic oxygen of the substrates, which is negatively charged in the TI. The CaLB lid corresponds to a little α helix formed by few amino acids. Therefore it can just partially limit the active site access, as a matter of fact CaLB does not show interfacial activation.[2] CaLB enantioselectivity also depends on its particular active site, which is composed by two distinct subsite: the acylic one and the alcholic/aminic one (Figure 5.1).

[2] M Skiot, L De Maria, L Chatterijee, A Svendsen, S A Patkar, P R Ostergaard, J Brask, *ChemBioChem*, **2009**, 10, 520.

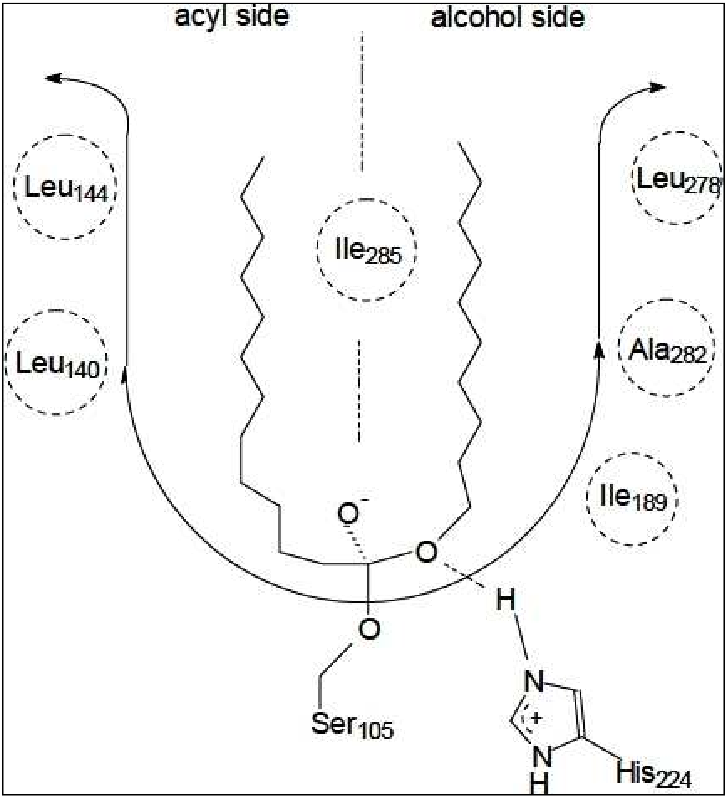


Figure 5.1: CaLB schematic division of the active site in the two distinct subsite.

These two subsites receive different moieties of the substrate. Moreover a stereospecific pocket (Thr42, Ser 47, Trp104) which is involved in the enantiomer discrimination does also exist (Figure 5.2).[1]

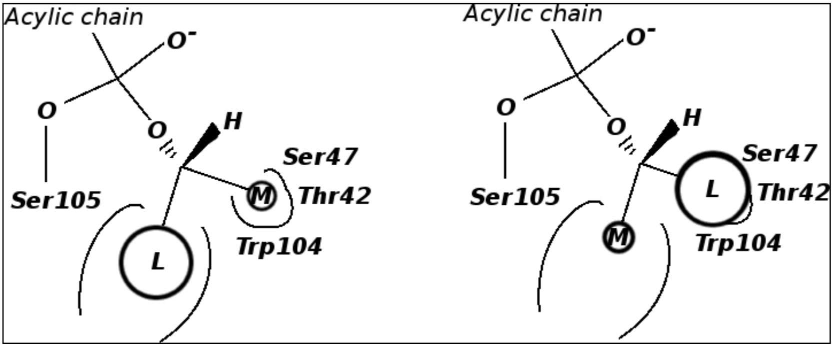


Figure 5.2: Schematic CaLB active site with the representation of the stereospecific pocket, and different enantiomer orientation during the catalysis, L is the large substituent and M the medium one.

5.2 General Strategy

Is generally accepted that the enantioselectivity depends on the interactions that a given molecule establishes with enzyme’s active site.[1] Therefore, a promising approach is represented by the calculation and comparison of the interactions established by a set of different substrates.

The idea is taking a number of substrates, calculating their conformations in the enzyme active site, discard the enzyme and focus the analysis on the comparison of the substrates. The use of a set of experimental measurements and multivariate analysis for the regression of the structural differences among the compound of the set will finally allow the generation of a mathematical model for the prediction of the desired enzyme kinetic property.

In this perspective a molecular descriptor able to identify the interaction capabilities of the compounds is necessary. MIF (Molecular Interaction Field), obtained by GRID analysis, contains an entire set of information based on the interaction possibilities of a molecule. If MIFs of two molecules differ in a particular space region it should be possible in theory to identify this difference as the source of different kinetics by enzyme action. Moreover this should guide to focus the study of enzyme-substrate interactions in specific areas, shedding light on the molecular basis of substrate (enantio)recognition. Of course, spatial orientation and conformation of compounds affect heavily the corresponding MIF. Therefore it is of basic importance to pay the necessary attention to conformational analysis and spatial alignment of the molecules in the data set. It has been demonstrated that MIF based 3D-QSAR models can be extremely predictive in the case of penicillin amidase.[3]

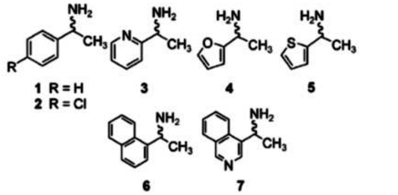
When developing a 3D-QSAR model, the first step is always choosing the data set. The set of compounds used as a reference will affect enormously the final outcome. The training set was chosen combining seven racemic amines and twelve racemic alcohols and their corresponding experimentally measured values of enantiometic ratio (E) in the CaLB catalysed synthesis reaction.[4, 5, 6] (Table 5.1).

[3] P Braiuca, A Buthe, C Ebert, P Linda, L Gardossi, *Adv Synth Catal*, **2006**, 348, 773.

[4] L E Iglesias, V M Sanchez, F Rebolledo, V Gotor, *Tetrahedron:Asimmetry,* **1997**, 8, 2675.

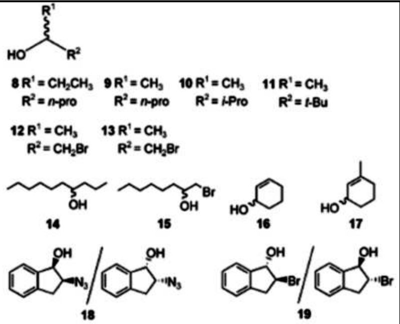
[5] K A Skupinska, E J McEchern, I R Baird, R T Skerlj, G J Bridger, *J Org Chem*, **2003**, 68, 3546.

[6] J Ottosson, L Fransson, K Hult, *Protein Sci*, **2002**, 11, 1462.



Amine	Donor	E /F.r.[a]
1	Ethyl acetate	110 (*R*)
2	Methyl methoxyacetate	232 (*R*)
3	Ethyl acetate	66 (*R*)
4	Ethyl acetate	>100 (*R*)
5	Ethyl acetate	32 (*R*)
6	Ethyl acetate	24 (*R*)
7	Ethyl acetate	120 (*R*)

[a] F.r. = fast reacting enantiomer.



Alcohol	Donor	E /F.r.[a]
8	Octanoyl acetate	340 (*R*)
9	Octanoyl acetate	8 (*R*)
10	Octanoyl acetate	760 (*R*)
11	Octanoyl acetate	430 (*R*)
12	Octanoyl acetate	100 (*R*)
13	Octanoyl acetate	370 (*R*)
14	Octanoyl acetate	10 (*R*)
15	Octanoyl acetate	7 (*S*)
16	Octanoyl acetate	1.6 (*R*)
17	Octanoyl acetate	62 (*R*)
18	Octanoyl acetate	1.3 (*S,R*)
19	Octanoyl acetate	90 (*S,S*)

[a] F.r. = fast reacting enantiomer.

Table 5.1: Data set of the resolution of amines on the left and alcohols on the right.

The choice fell on these reactions among the numerous examples reported in the literature, because they meet the necessary requirements in terms of molecular diversity and E values range and homogeneity of distribution, crucial for the generation of a consistent 3D-QSAR model.[7] The distribution of enantiomeric ratio values throughout the data set is well balanced and structural diversity of nucleophiles is significant. Some difficult cases are included, such as nucleophiles bearing halogen substitution in the medium-sized chain, which are not resolved by CaLB because of polarity effects.[7]

In a general 3D-QSAR model there is the biunivocal correspondence of each compound of the data set with a given

[7] P Braiuca, L Knapic, V Ferrario, C Ebert, L Gardossi, *Adv Synth Catal*, **2009**, 351, 1293.

"activity". When facing the prediction of enantioselectivity, the "activity" is represented by E values that are properties of a couple of compound, not of each single molecule. Therefore it was not possible to apply the same scheme used in the case of PGA.[3]

The novelty of the concept was the generation of a novel class of MIF-based molecular descriptors, melting the information of two enantiomers into a single molecular entity,making possible the biunivocal association with a E value. The accomplishment of this task was achieved with the calculation of a second generation MIF, the Differential MIF (DMIF). It is basically the mathematical difference between the MIFs of the two enantiomers of each couple, calculated by a simple matrix subtraction. The DMIF represents a sort of molecular hybrid, a virtual molecule comprising all the information of each enantiomer couple. In particular it is very useful because it amplifies by definition the structural differences of the two enantiomers.

In fact, if a given variable of the MIF carries the same information for both enantiomers (the same value of interaction energy for the two enantiomers), the corresponding DMIF value will be zero. At the contrary if the variable has a different value for the two enantiomer, the generated DMIF variable will be as big as the difference between the original MIFs values. Therefore the DMIF assumes null values in the zones where the structures two enantiomer are identical and it assumes high values where there are significant differences. This is very important because big structural differences indicate different interactions with the enzyme active site.

5.3 Active conformers calculations

The structures of the compounds of the data set were manually generated and minimized in the AMBER 99 force field[8] using the software MOE.[9]

The first and the most delicate step of this investigation involved the calculation and the assessment of the tetrahedral intermediates for each acylation reaction by molecular modelling techniques.

For this purpose, the corresponding esters and amides were docked into the active site of the lipase and the best conformers were chosen on the basis of the results of the docking algorithm scoring function (London dG) as well as by evaluating the geometric compatibility with the initiation of the catalytic mechanism. Different criteria were taken into account during the structural compatibility assessment: I) the correct orientation of the acylic and nucleophilic portion of the conformer inside the hydrophilic/hydrophobic pocket of the active site; II) the distance of the catalytic Ser105 from the carbonyl carbon of the substrate, which must be compatible with the nucleophilic attack; III) the correct orientation of the carbonyl oxygen toward the Thr40 and Gln106 that constitute the oxyanion hole. The tetrahedral intermediates were then simulated by forming a covalent bond between the hydroxy group of the catalytic serine (Ser105) and the carbonyl carbon of the acylated substrate, thus resulting in the corresponding oxyanions.

After the formation of the tetrahedral intermediates the obtained systems were minimised and each enzyme-substrate complex was subjected to 300 ps of molecular dynamic simulation using the software MOE (NVT conditions at the temperature of 300 K and implicit water solvation), in which only amino acid residues within a 10 Å radius sphere from the catalytic serine (Ser105)

[8] J Wang, P Cieplak, P A Kollman, *J Comp Chem*, **2000**, 21, 1049.

[9] MOE **2006.08**, *Chemcomp*, Montreal, Canada.

were allowed to move. The rest of the protein was kept constrained. The simulations generated energy stable complexes within the first few tens of ps of the simulations.

The simulation during time of enzyme-substrate complexes allows a complex evaluation of the interaction that occurs during substrate stabilization in the active site. Moreover, it was possible to investigate the space and the conformational freedom of the substrates and of the enzyme either. The comparison between the enantiomeric couples allows to identify the main structural basis of CaLB enantioselectivity.[7] For instance, in the case of amides with high E values, important structural differences were observed between the TIs of the fast- and the slow-reacting enantiomers. As shown in Figure 5.3 for substrate **1**, the TI of the fast-reacting enantiomer is embraced inside the hydrophilic pocket on the right hand portion of the active site (the so-called alcoholic sub-site).

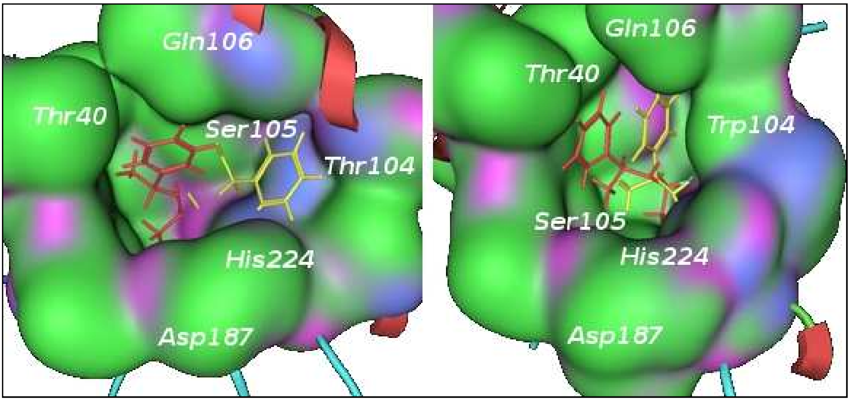


Figure 5.3: Initial (red) and final (yellow) conformation of the slow-reacting (S), on the left, and the fast-reacting (R) nantiomer, on the right, of substrate 1.

On the other hand, the TI of the slow-reacting enantiomer remains at the outer region of the active site which makes the second nucleophilic attack unfeasible. Another evident discriminating factor is illustrated in 5.4, which represents the outcome of the

MD based conformational search of the two enantiomers of substrate **8**.

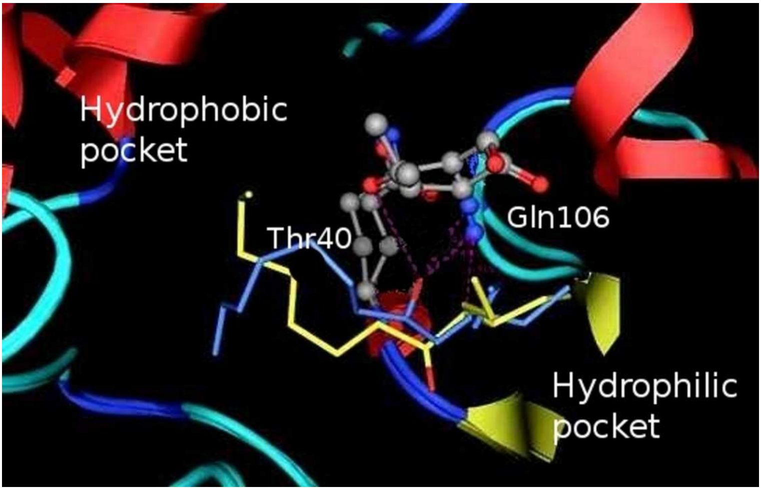


Figure 5.4: Energy minima conformations of the fast-reacting (blue) and the slow-reacting (yellow) enantiomers for the acylation of compound 8 obtained by MD simulations. The different orientation of the oxyanions (in red) is clearly visible: only the fast-reacting enantiomer is stabilized through the formation of hydrogen bonds (dashed lines) with Thr 40 and Gln 106.

In the case of the slow-reacting enantiomer, the minimum energy conformer is not able to perfectly place the oxyanion in the oxyanionic hole, with a consequent energy destabilization as compared to the fast-reacting enantiomer where stabilizing hydrogen bonds take place between the oxyanion and the Thr40 and Gln106 residues of the oxyanion hole. In this case the MD causes the evolution of the slow-reacting enantiomer towards an unproductive conformation, as defined by the criteria used for the docking scoring. This means that the initiation of the reaction for that enantiomer is unfavorable and consequently the E values is very high. Although this leads to the comparison of productive and unproductive conformations in the QSAR, this dramatic conformational difference is certainly correlated to the high E, so that both the productive and unproductive conformers must be included in the model.[7]

5.4 DMIFs calculation

The outcome of every MD simulation was carefully analysed and conformers with the lowest potential energy (as calculated by MD algorithm for the whole unconstrained part of the system, therefore within the active site region) were selected as the best simulations of the different TIs and they were used for the construction of the 3D-QSAR model. The enzyme substrate complexes were superimposed by overlaying the catalytic triad and the oxyanion hole of all selected configurations. This was necessary because after the MD simulations the Cartesian coordinates of the systems were perturbed, although the conformational changes of the active site residues were always negligible. The protein structures were discarded after the removal of the covalent bond between the substrates and the catalytic serine, while the overall geometry of the substrate conformers was kept unaltered, to generate a so-called “supermolecule”, which consisted in all the 38 active conformers (19 enantiomeric couples), both R and S. GRID analysis was then performed by setting the dimensions of the grid to contain all the conformers and each of them was analyzed separately (14 Å x 24 Å x 21 Å, knots every 1 Å). In order to take into account the most important non-covalent interactions, two probes with diverse physico-chemical properties were used in the calculation of the molecular interaction fields, namely the WATER and the DRY probe. The WATER probe describes and quantifies the dipolar interactions and the hydrogen bond formation, whereas the DRY probe considers all the hydrophobic interactions.[10] The DMIF calculation was performed in a matrix differential procedure where each variable of the MIF of the slow-reacting enantiomer was mathematically subtracted from the corresponding variable of the MIF of the fast-reacting enantiomer (Figure 5.5).

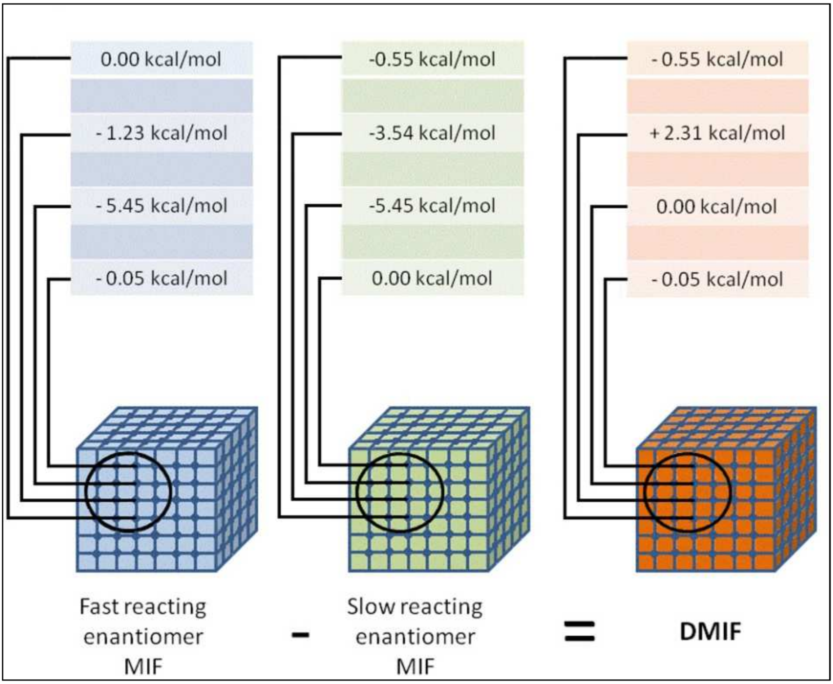


Figure 5.5: The procedure used for the calculation of the DMIFs taking as example the case of interaction energies between the water probe and the two enantiomers of compound 16. The mathematical difference between matrices was calculated to generate a single “differential matrix”.

It must be noted that the redundancy of the information contained in the calculated MIFs was cut by operating a “zeroing values

[10] J Pleiss, Enzymes in Lipid Modification, *Wiley-VCH Verlag, Weinheim*, **2005**, 85.

pretreatment": all the positive/unfavourable interaction energies were set to zero, because every cluster of positive variables (nodes of the MIFs grid) has a corresponding cluster of negative variables that contains information that is identical from the statistical point of view. This procedure led to the quantitative evaluation of the differences in interactions between the two enantiomers and both polar and hydrophobic regions of the active site. Consequently, as explained before, DMIFs present null values in the correspondence of areas where the enantiomers establish identical interactions with the active site, whereas high absolute values indicate that the enantiomers establish different interaction patterns with the enzyme.[7]

5.5 Generation of the mathematical model

The energy values contained in the "differential matrices" of the DMIFs were statistically analyzed to generate PLS[11] models able to correlate the quantitative differences between the two enantiomers with the experimental E values. The three-dimensional DMIFs are unfolded to form the so called bi-dimensional X-matrix, where each row corresponds to an enantiomeric couple and each column to a MIF grid node, matching a specific three-dimensional position. Each column of the X-matrix (containing the values of the DMIFs) is an X variable and the enantiomeric ratio E is the Y variable (or the "dependent" variable). As a matter of fact, molecular interaction fields describe the steric and electrostatic properties of substrates by sampling the interaction energies at all predefined gridpoints. The multitude of gridpoints and, therefore, the quantity of variables present in a DMIF can be extremely high even in the case of small molecules. Moreover, some of these variables are

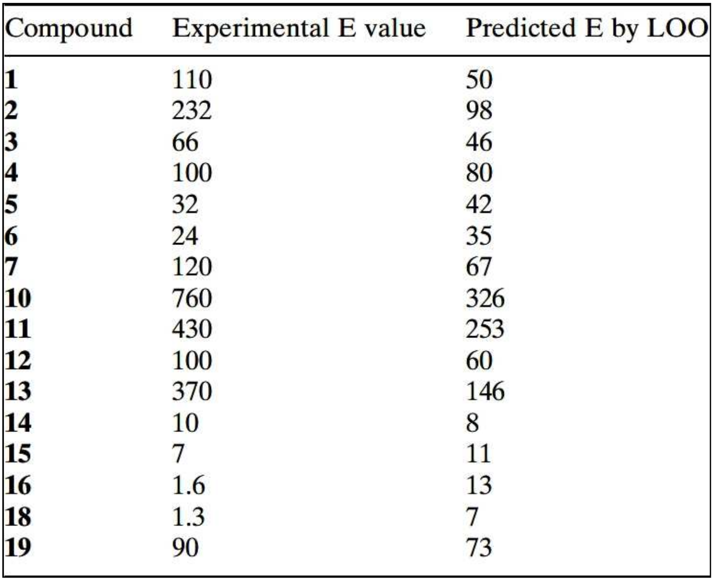
[11] J Ottosson, L Fransson, K Hult, *Protein Sci*, **2002**, 11, 1462-1471.

more informative than others. Although the procedure of DMIF calculation halved the number of objects, the number of independent X variables was still unvaried and it amounted to 24,950 for each couple of enantiomers. Therefore, the first stage in the statistical analysis was the choice of the most important variables and the discarding of the insignificant and redundant ones. For this purpose the GOLPE[12] program was used. GOLPE is a software package largely used for the construction, the validation and the interpretation of 3D-QSAR models. It is particularly adequate for models with large numbers of variables since it has a variety of tools for their selection. Once the DMIFs were calculated, all those variables having very low absolute values were discarded due to their negligible contribution to the quantification of the differences in enzyme-enantiomer interactions. Then, variables with a standard deviation close to zero were discarded as well because of their small variation through all of the DMIFs, that makes them useless in discriminating the objects in the data set. A last action was performed on the remaining active variables by using the “block unscaled weights” algorithm that attributes different weights to all blocks of variables giving them the same initial importance in the model without modifying the variable scale. This latter step was necessary because the polar interaction energies are significantly higher in absolute value than the hydrophobic interaction energies, therefore the statistical analysis would overestimate their importance in the model. Finally, the standard GOLPE procedure was applied on these pre-treated data, by employing the D-optimal pre-selection and the FFD variable selection algorithm which conserved only 568 active variables.[13] The multivariate statistical analysis was performed on 16 of the initial 19 compounds, by performing the PLS regression and five principal components

[12] GOLPE **4.5**, *Multivariate Infometric Analysis srl*, Perugia, Italy.

[13] S Raza, L Fransson, K Hult, *Protein Sci*, **2001**, 10, 329-338.

were calculated. Three compounds, fulfilling the requirement of having small, medium and high E values, were randomly chosen and excluded from the training set. The predictivity of the model was evaluated by means of the leave-one-out (LOO) cross-validation method as well as by performing an external validation using the GOLPE PLS external prediction on the three compounds not included in the training set (Table 5.2).



Compound	Experimental E value	Predicted E by LOO
1	110	50
2	232	98
3	66	46
4	100	80
5	32	42
6	24	35
7	120	67
10	760	326
11	430	253
12	100	60
13	370	146
14	10	8
15	7	11
16	1.6	13
18	1.3	7
19	90	73

Table 5.2: Comparison between the measured experimental E values and the values calculated by the model in the LOO (leave-one-out) cross-validation procedure applied on the training set.

The predictive correlation coefficient (q^2) provides the quantitative evaluation of the consistency of the model. The best q^2 value for the model is 0.76 on the third principal component and 99 percent of the variance of the model is explained by the first two principal components (expressed by the correlation coefficient r^2). Although the mathematical model was constructed on the basis of an experimental data set with a broad distribution of E values, the algorithm proved to be quite predictive and robust as illustrated in (Figure 5.6).

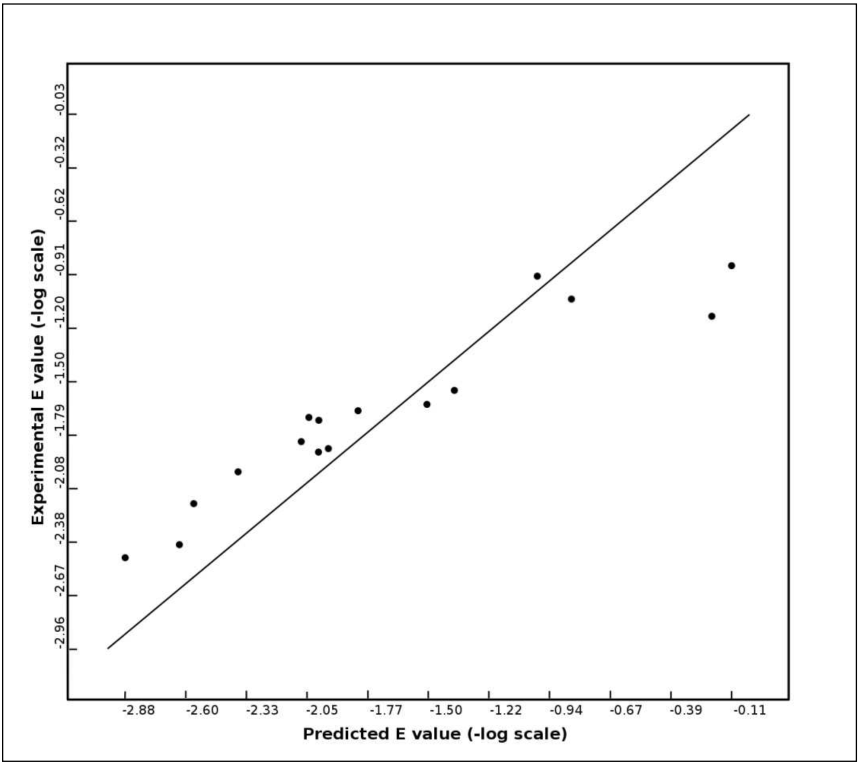
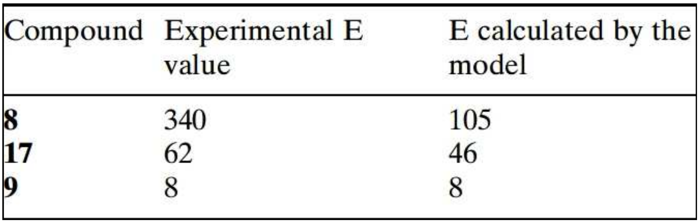


Figure 5.6: Predictivity of the model in terms of experimental versus predicted E values.

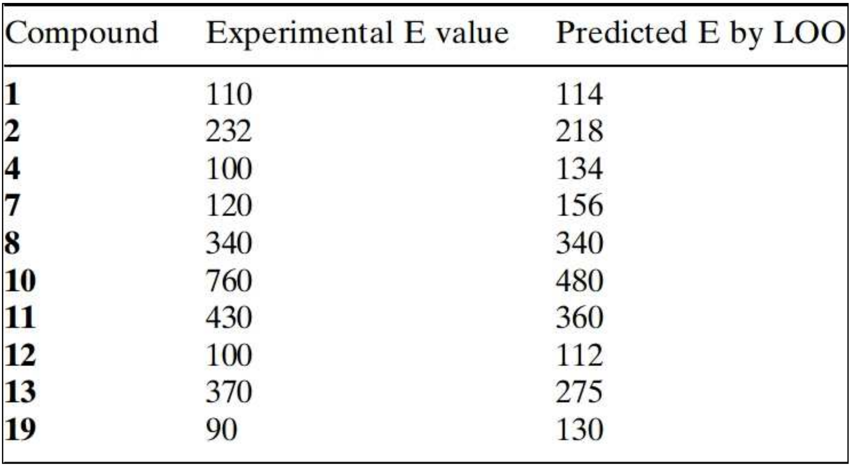
The worst predictions are represented by compounds **16** and **18** that are, however, characterized by extremely low E values (1.6 and 1.3. respectively). Because of these two compounds the model appears to be more predictive towards compounds having higher E values. It is an intrinsic property of any QSAR model to be more robust for the compounds in the middle of the activity range, simply because this zone is usually more populated. Nevertheless, in all cases the model is able to identify correctly the fast-reacting enantiomer and, more importantly, to recognize those couples of enantiomers characterized by poor enantiodiscrimination (for substrates **16** and **18** the calculated E values are < 15 in both cases). The external validation was performed on three additional compounds (**8**, **17**, **9**) not originally included in the training set that were chosen due to their low, intermediate and high E values, respectively. For every compound the complete procedure was repeated, as described above, in order to generate the molecular descriptors (DMIFs). Their E values were then predicted by applying the generated 3D-QSAR model. As it can be seen from Table 5.3, although the model predicts with good precision the ability of the enzyme to enantiodiscriminate within a couple of enantiomers (predictivity expressed as q^2= 0.78), in the case of compound **8** the E value is underestimated.



Compound	Experimental E value	E calculated by the model
8	340	105
17	62	46
9	8	8

Table 5.3: External validation of the calculated PLS model (q2=0.78) towards compounds not included in the training set.

This underestimation was observed also for compounds **2**, **10**, **11**, **13** and it might suggest that the variables crucial for structural discrimination for substrates having low E values are different as compared to variables describing substrates with high E values. In other words, the 3D-QSAR model is trained on the basis of a pattern of interactions which are actually different as compared to those taking place in the case of substrates with high E values, and this might limit the predictivity of the PLS model. It should be noted that in the case of compounds **2** and **13** the underestimation given by the model could be ascribed to the presence of the halogen substituent, whose polar character might be measured with insufficient precision by the WATER probe. Even though this probe can adequately estimate polar interactions that are not correlated to hydrogen bonding, the halogen atoms might not be described comprehensively by the force field parameterization of the WATER probe. Therefore, a second model, specifically trained for the prediction of high E values, was calculated in order to refine the quantitative predictivity of E values for those enantiomers that are efficiently enantiodiscriminated by the enzyme. The new data set was constructed by setting the value of E= 50 as a threshold since E values lower than 50 correspond to enantiomers poorly enantiodiscriminated (examples are compounds **3**, **5**, **6**, **15** in Table 5.4). Indeed, the predictivity of this second model improved (q^2= 0.88) and, as expected, the same model was less efficient in predicting the E values for copules of enantiomers that are poorly enantiodiscriminated.



Compound	Experimental E value	Predicted E by LOO
1	110	114
2	232	218
4	100	134
7	120	156
8	340	340
10	760	480
11	430	360
12	100	112
13	370	275
19	90	130

Table 5.4: Data set used for the calculation of the second model and for its validation in terms of predicted E values by the LOO (leave-one-out) cross-validation procedure (q2=0.88).

This second "specialized" model is based on a larger number of variables as compared to the first general model (618 instead of 568 of the general model) and its increased predictivity suggests that the variables involved in the two models are substantially different, not only quantitatively. It must be noted that each variable corresponds to a specific grid point, therefore to a specific Cartesian coordinate in the active site of the enzyme. To understand the differences between the two models in deeper detail, each single variable was analysed and its position in the space refolded. The two models share nearly 25% of the variables (125 variables), while they differ for the rest of them. A detail of the analysed space with the spatial position of included variables is represented in Figure 5.7. It is evident that in the first general model the crucial variables are scattered throughout the active site, whereas in the second "specialized" model the crucial interactions are concentrated in the oxyanion hole and in the alcoholic subsite (central and the right-hand part of the molecule).

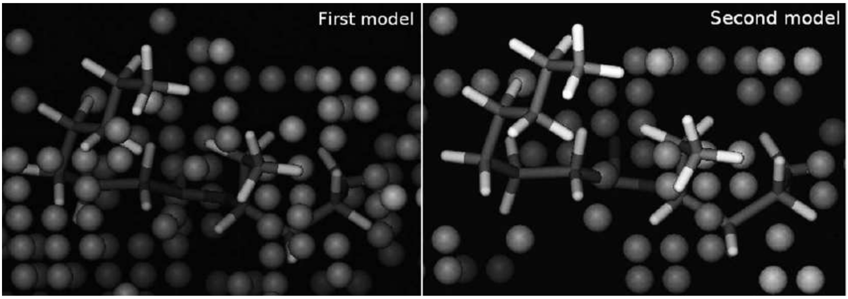


Figure 5.7: Representation of variables utilized for the construction of the first general predictive model (left) and the second "specialized" model (right). Substrate 8 is shown in both images, as an example.

The analysis confirms what emerged from the conformational analysis: for substrates characterized by high E value the slow-reacting enantiomer either cannot place the oxyanion into the oxyanionic hole or it cannot place the alcoholic moiety inside the corresponding subsite. As a rule of thumb, when dealing with the prediction for a new substrate, the first general model should be used to obtain an initial classification of the CaLB enantiodiscriminating potential. If the first model predicts a high E value, the second specialized model should be used for obtaining a more refined quantitative prediction.[7]

5.6 Conclusion

The combination of molecular modelling with multivariate statistics constitutes a powerful tool for predicting and also interpreting the enantioselectivity of biocatalysts. The remarkable flexibility of this "hybrid" computational tool makes it adaptable to the solution of different problems as well as to the investigation of the molecular basis of enantiodiscrimination. By definition, the success of any 3D-QSAR strategy depends strongly on the experimental data set used for the training of the mathematical model. Moreover, the generation of the PLS model heavily relies on the selection of the most informative variables of the whole data set. This statistical procedure is of fundamental importance and constitutes one of the bases of the QSAR paradigm.

5.7 Experimental section

The protein structure used for this study was retrieved from the Protein Data Bank (1TCA). This initial structure was pretreated in MOE[9] by removing the crystallographic water molecules and two molecules of NAG (N-acetylglucosamine) present in the pdb file. Hydrogen atoms were added and their position was optimized with an energy minimization procedure in the Amber99 force field[14] in its MOE implementation. Subsequently a minimization of the side chains was performed keeping the backbone atoms fixed. The substrates were built, minimized and then docked into the active site of CaLB by means of the DOCKING module of MOE. The docking was performed on a 10 Å radius selected area surrounding the active site. The force field used for the docking was MMFF94x,[15] the charges of substrate atoms were calculated at the QM PM3 semi-empirical level, by means of the MOPAC7 program. The initial positions of the substrates were manually set in order to meet the criteria previously reported. For each substrate, the conformation presenting the highest score and fulfilling the structural requirements for the initiation of the enzymatic catalysis, was chosen. Construction of Tetrahedral Intermediates All tetrahedral intermediates were sketched bonding the hydroxy oxygen of the serine residue (Ser 105) and the reactive carbonyl carbon of the substrate. This carbon atom changes to a tetrahedral sp3-hybridized configuration. The partial

[14] J Wang, P Cieplak, P A Kollman, *J Comp Chem*, **2000**, 21, 1049.

[15] T A Halgren, *J Comp Chem*, **1996**, 17, 490.

charges and geometry of this chemical species (the substrate and the serine) were calculated by an ab initio algorithm, based on DFT-TZVB, by Turbomole.[16] In the molecular mechanics calculations, the standard MMFF94x atom types were used for the atoms of the tetrahedral intermediates, while bond lengths, angles and torsions on tetrahedral carbon were constrained to the values obtained by the ab initio optimization.

Molecular Dynamics

The molecular dynamics simulations were performed using the DYNAMICS module of MOE. All the dynamics were performed in an NVT environment simulating the temperature of approximately 300 K. In order to reduce the calculation time, the attention was focused on the relevant part of the system: all the atoms of the substrate and the protein residues within a sphere of 10 _ radius, centered on the catalytic serine (Ser105) were allowed to move, all the rest of the system was kept constrained. An integration time of 2 fs was used and a frame of the trajectory was saved every 10 fs. Each substrate conformation chosen for the construction of the data set for the QSAR analysis, was the one characterized by the lowest potential energy out of all the frames saved in the dynamics database. All enzyme structures chosen by these criteria were superimposed with the database viewer superpose implementation. The active conformers were than extracted.

[16] TURBOMOLE **5**, *Cosmologic*, Leverkusen, Germany.

GRID

The GRID analysis was performed on every constituent of the data set. The chosen dimensions of the cage were 14 Å X 24 Å X 21 Å with NPLA (number of grid planes per Angstrong) set to 2 while the probes used were DRY probe and WATER probe. Once the MIFs have been calculated, all the unfavorable interactions were set to zero. For the DMIFs calculation a specific algorithm was constructed which performs the matrix differential procedure for the subtraction of the two MIFs.

GOLPE

The pretreatment section of GOLPE was used to perform the variable selection. All the variables having an absolute value lower than 0.1 for the water probe and 0.03 for the dry probe were set to zero and those with standard deviation of less than 0.2 for the water probe and 0.06 for the dry probe were discarded. The pretreatment was eventually completed with the block unscaled weight application. Both PLS models with 5 principal components were computed and validated with the LOO (leave-one-out) method. The prediction ability of the general model was tested on a test set by using the PLS predictions module of the GOLPE program.

CHAPTER 6

ALKANESULFONATE MONOOXYGENASE

6.1 Introduction

In the previous chapters several different computational techniques for the study of lipases were reported. The most of the work object of this thesis was based on molecular dynamics and on calculation of molecular descriptors and statistical analysis. Several different simulation schemes and applications have been described, all together representing a powerful set of tools for the study of lipases from many different points of view. Nevertheless, although in principle the approaches should have general applicability, the limitation of application to lipases only cannot assess this important aspect. For that reason the combination of MD for testing the dynamics of enzyme action and 3D-QSAR for constructing quantitative predicting models of enzyme selectivity was applied to a completely different enzyme.

The alcansulfonate monooxigenase is an enzyme with relatively unknown properties, extremely different, both structurally and from the point of view of the mechanism of action. It catalyses an extremely different reaction and it is a cofactor-dependent enzyme.

6.1.1 The alcansulfonate monooxygenase

In bacteria sulfur is mainly assimilated from inorganic sulfate via the cysteine biosynthetic pathway. In nature, where the levels of inorganic sulfate may be low, bacteria have to rely on organosulfur compounds such as sulfate esters, sulfamates and sulfonates as sulfur sources. Under conditions of sulfate starvation, *Escherichia coli* synthesizes the TauABCD and SsuEADCB proteins, which cover the full range of uptake and desulfonation activities for growth with taurine and a variety of aliphatic sulfonates as sole sources of sulphur.[1] Sulfite liberated inside the cells from taurine and alkanesulfonates serves as sulfur for cell growth.

SsuD is an $FMNH_2$-dependent monooxygenase, which catalyzes the oxygenolytic cleavage of the C-S bond of 1-alkanesulfonates by monooxygenation leading to the release of the corresponding aldehyde and sulfite. Catalysis is absolutely dependent upon oxygen and reduced FMN, the latter of which is provided by the associated NAD(P)H:FMN oxidoreductase SsuE.[2] SsuD or SsuD-like enzymes showing high sequence identity to *E. coli* SsuD, and whose function in sulfur-scavenging from varying ranges of organosulfur compounds was assessed are found in *Bacillus subtilis*,[3] *Pseudomonas* strains[4] and *Corynebacterium glutamicum*.[5] Elucidating the detailed catalytic mechanism of SsuD is of interest as it will allow shedding light on the unique role in C-S bond cleavage of this among microorganisms widespread enzyme.

The 3-dimensional structure of SsuD was solved showing a TIM-barrel fold as structural core and the location of the active site was proposed at the end of the β–barrel based on biochemical observations.[6] This fold, which is adopted by many enzymes for flavin binding[7] has been found in all members of the luciferase family whose structure have been sold so far. Luciferase LuxAB,[19]

[1] J R van der Ploeg, E Eichhorn, T Leisinger, *Review Arch Microbiol*, **2001**, 176, 1.

[2] E Eichhorn, J R van der Ploeg, T Leisinger, *J Biol Chem*, **1999**, 274, 26639.

[3] J R van der Ploeg, N J Cummings, T Leisinger I F Connerton, *Microbiology*, **1998**, 144, 2555.

[4] A Kahnert, P J Vermeij, C Wietek, P James, T Leisinger, *J Bacteriol*, **2000**, 182, 2869.

[5] D J Koch, C Ruckert, D A Rey, A Mix, A Puhler, J Kalinowski, *Appl Environ Microbiol*, **2005**, 71, 6104.

[6] E Eichhorn, C A Davey, D F Sargent, T Leisinger, T J Richmond, *J Mol Biol*, **2002**, 324, 457.

[7] G K Farber, G A Petsko, *Trends Biochem Sci*, **1990**, 15, 228.

long-chain alkane monooxygenase LadA,[8] F_{420}-dependent N^5,N^{10} methylenetetrahydromethanopterin reductase Mer,[9] and F_{420}-dependent secondary alcohol dehydrogenase Adf from methanogenic *Archaea*.[10] All use reduced flavin or the flavin analogue F_{420} as a cosubstrate and show high structural similarity. The striking homology between SsuD, luciferase and LadA translates into highly conserved amino acid residues among these proteins where the SsuD active site has been proposed. SsuD's Cys54, His228, Arg297 and Tyr331 find their counterparts in luciferase at Cys106, His44, Arg291 and Tyr110, where His44 mutations produced inactive luciferase.[11] The reactive thiol of Cys106 was observed in the crystal structure of luciferase-bound FMN to project directly towards position C(4a) of the isoalloxazine ring, the site of flavin oxygenation.[12] They also find their counterparts in LadA at Cys14, His311 and Tyr63, where activity was completely abolished when mutating His311 and Tyr63[8] as well as in Cys14 mutants of which were unable to produce the dimeric assembly required for enzymatic activity. SsuD enzymes involved in sulfur scavenging from organosulfur sources are synthesized under sulfate-starvation conditions; they show a very low content in sulfur-containing amino acids. Cys54 of *E. coli* SsuD is remarkably conserved, suggesting that this amino acid may play an important role in catalysis of alkanesulfonate desulfonation. Labeling of Cys54 in SsuD led to inactive enzyme, but the role of this residue was not elucidated.[6]

[8] L Li, X Liu, W Yang, F Xu, W Wang, L Feng, M Bartlam, L Wang, Z Rao, *J Mol Biol*, **2008**, 376, 453.

[9] S Shima, E Warkentin, W Grabarse, M Sordel, M Wicke, R K Thauer, U Ermler, *J Mol Biol*, **2000**, 300, 935.

[10] S W Aufhammer, E Warkentin, H Berk, S Shima, R K Thauer, U Ermler, *Structure*, **2004**, 12, 361.

[11] A J Fisher, T B Thompson, J B Thoden, T O Baldwin, I Rayment, *J Biol Chem*, **1996**, 271, 21956.

[12] Z T Campbell, A Weichsel, W R Montfort, T O Baldwin, Biochemistry, 2009, 48, 6085.

Very recently Carpenter *et al.*[13] established by means of mutagenesis studies that Cys54 may be directly or indirectly involved in stabilizing the C(4a)-hydroperoxyflavin intermediate formed during catalysis through hydrogen-bonding interactions.

6.2 Enzyme structure

The initial characterization of SsuD did not address the question of the enzyme's reaction mechanism. Also, interactions between the enzyme and its substrates were not elucidated when the crystal structure of the enzyme was solved: no crystals of enzyme-susbstrate complexes were obtained. However, many studies have been conducted with this enzyme by means of biochemical, mutagenesis and spectrometry techniques to get insight into enzyme-substrate interactions.

The work presented here aims at shedding light on the SsuD interactions with cofactor and substrates, its most significant structural elements, either static or dynamic, determining its substrate specificity at molecular level. Moreover, the information gathered by molecular docking, molecular dynamics and analysis of molecular interaction fields were used for building a QSAR model[14] for the quantitative prediction of SsuD substrate selectivity.

The lack of knowledge about the enzyme action at molecular level required a structural comparison with other members of the same enzyme family. The presence of preserved aminoacids in the same spatial positions among the active sites of the structures means that these residues can play an important role in the enzyme's activity and selectivity.

[13] R A Carpenter, X Zhan, H R Ellis, *Biochim Biophys Acta*, **2010**, 1804, 97.

[14] P Braiuca, L Boscarol, C Ebert, P Linda, L Gardossi, *Adv Synth Catal*, **2006**, 348, 773.

The structure of SsuD was compared with the structures of highly homologous enzymes not only based on their sequence similarity, but also in terms of tertiary structure.

The crystal structure of SsuD was taken from the Protein Data Bank (PDB)[15] where it is available since July 2002 under the entry code 1m41 and having a resolution of 2.3 Å.

In the asymmetric unit, SsuD appears as a homo-dimer (subunit A and B); each subunit consists of a single domain, an eight-stranded α/β-barrel motif. This fold classifies SsuD as a member of the big $(\beta/\alpha)_8$ TIM-barrel family (Figure 1).[6]

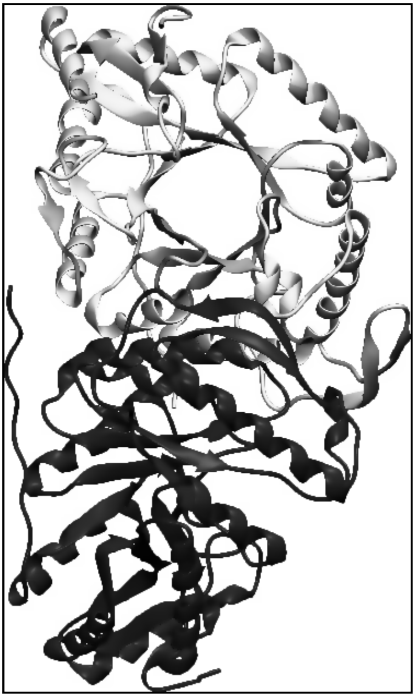


Figure 6.1: Secondary structure of SsuD; subunit A in silver and subunit B in charcoal.

[15] H Berman, J Westbrook, Z Feng, G Gilliland, T Bhat, H Weissig, I Shindyalov, P Bourne, *Nucleic Acids Res,* **2000**, 28, 235.

SsuD is an $FMNH_2$-dependent enzyme, but in the 1m41 PDB structure there is no co-crystallised reduced FMN cofactor. The following members of the $(\beta/\alpha)_8$ TIM-barrel structures were used for the comparison:

- Bacterial luciferase (PDB 1luc);
- Long-chain alkane monooxygenase (LadA) co-crystallised with FMN (PDB 3b9o);
- Methylenetetrahydromethanopterin reductase (Mer) co-crystallised with a different cofactor, namely F_{420} (PDB 1z69).

The first two enzymes, bacterial luciferase and LadA, are classified as FMN/$FMNH_2$ monooxygenases and are members of the same family along with SsuD. Therefore they presumably share a similar catalytic mechanism, since they also share a homologous aminoacidic organization within the active site.[11]

In particular, the structures 1luc, 3b9o and 1m41 showed after superimposition three highly conserved residues: a His, a Trp and an Arg (His 11, Trp 196 and Arg 127 fur SsuD; His 44, Trp 194 and Arg 107 for bacterial luciferase; and His 311, Trp 348 and Arg 157 for LadA). It was clear that the spatial position of the Arg residue is highly conserved, whereas more divergences can be seen in the cases of His and Trp (Figure 2).

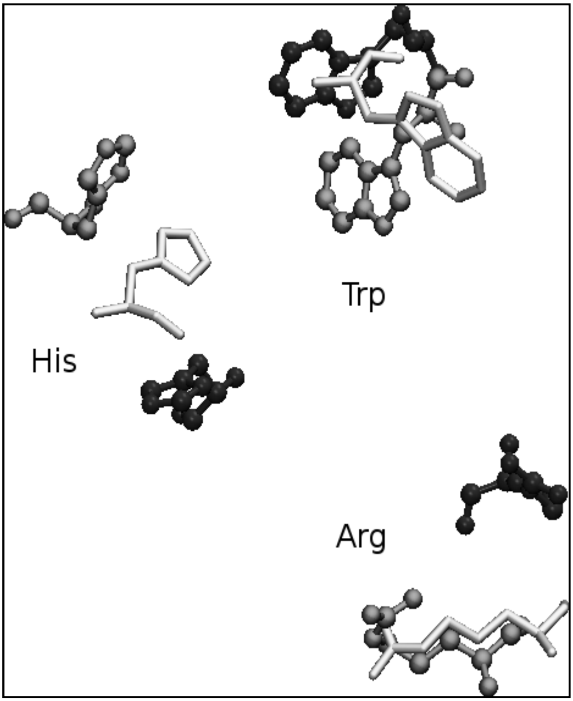


Figure 6.2: Superimposition of the most preserved residues, His, Trp and Arg in the active site area among the compared structures: grey balls and sticks for SsuD, charcoal balls and sticks for bacterial luciferase and silver sticks for LadA.

The alkalinity of the Arg residue plays an important role in the stabilization of the phosphate group of the FMN cofactor. As a matter of fact, Mer lacks this residue, probably due to a different and bulkier cofactor, which steric hindrance is an obstacle for the presence of such a residue in this position.

A Trp residue is located in the middle of the active sites of all three enzymes and creates a hydrophobic pocket to accommodate the substrates. Moreover, observing the Trp positions in LadA and in Mer, it appears evident that its indolic ring stabilizes the cofactor by interacting with its isoalloxazine tricycle ring.

Concerning the role of SsuD's His a valuable explanation was found in a previous publication on Bacterial luciferase, whose His 44 was proven to be the crucial aminoacid[11] for the catalytic activity, responsible for the substrate stabilization in the active site. Moreover, its distance from $FMNH_2$ seems to be sufficient for the accommodation of substrates.

The spatial variability of this His residue among the superimposed enzymes is due to the different substrate specificity spectrum of the enzymes.

6.3 Surface analysis

In order to map the physical chemical properties, a detailed GRID[16] analysis of the enzyme surface was performed. Three different probe types were used: WATER probe which describes and quantifies the dipolar interactions and the hydrogen bond formation, DRY probe which describes all the hydrophobic interactions and OS, sulphone/sulphoxide probe, necessary for the specific substrate affinity of SsuD.

Looking at the enzyme surface, the prevalence of hydrophilic regions is in agreement with the cytosolic origin of the protein; nevertheless small hydrophobic regions are present and spread all over the surface. A bigger hydrophobic area is located at the entrance of the active site, which might have the function of facilitating the substrate entrance in the catalytic site and of reducing the entropic penalty caused by the expulsion of water generated during catalysis (Figure 3).

[16] P G Goodford, *J Med Chem*, **1985**, 28, 849.

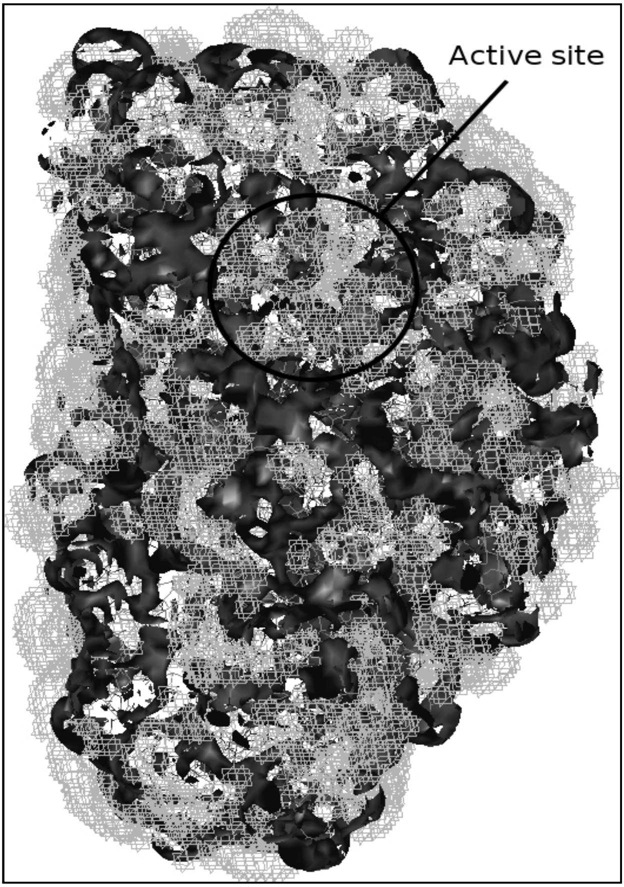


Figure 6.3: Isopotential surfaces calculated by the GRID analysis of SsuD, the hydrophilic areas are displayed in grey crosses (-0,7 kcal/mol) while the hydrophobic ones are visualized in charcoal solid surface (-2,5 kcal/mol).

The molecular interaction fields generated by GRID in the inner part of the active site are characterised by the presence of a big hydrophobic zone, which is mostly due to Trp 196. This result indicates the existence of significant and widespread area, characterised by a neat hydrophobicity, mainly due to Trp, that has a major role in cofactor stabilization (Figure 4).

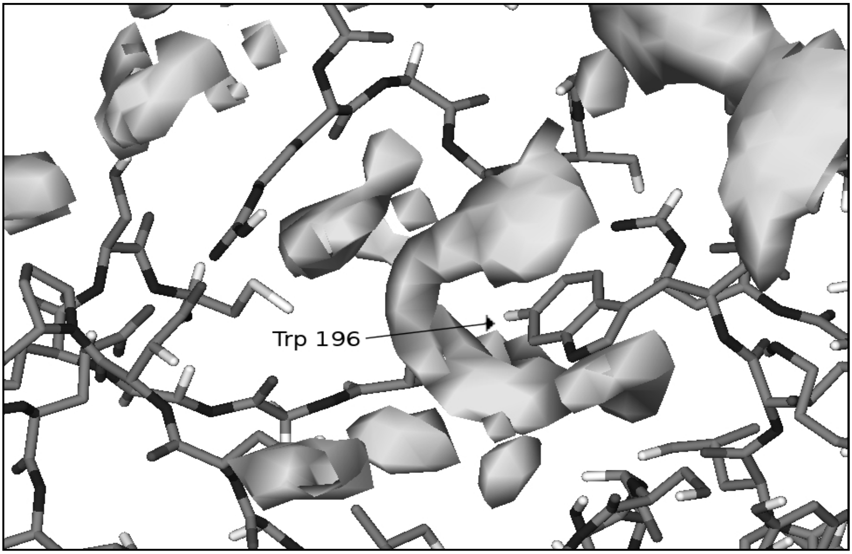


Figure 6.4: An insight of the active site of SsuD. The hydrophobic zone (grey surface) around Trp 196 is clearly visible.

The substrates of alkanesulfonate monooxygenase (SsuD) are aliphatic sulphonic acids, the probe OS was therefore used in order to simulate the HSO_3^- group of the substrate. Alkaline residues show high affinity for this probe and since the active site contains a few of them, an unambiguous binding position for the substrate was not recognisable with this procedure (Figure 5). Besides this finding does not give a conclusive evidence for formulating a hypothesis about the substrate binding, it still represents valuable information.

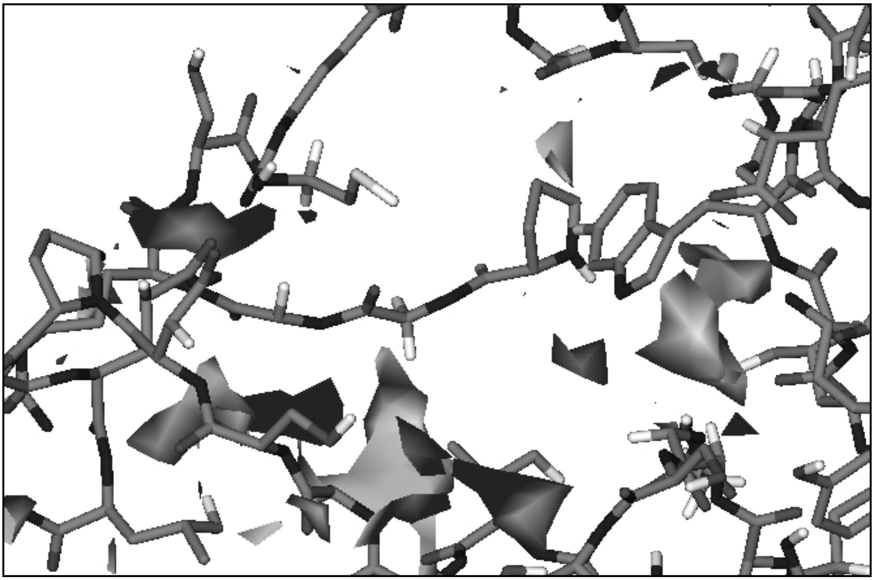


Figure 6.5: The same perspective of the active site of SsuD as in figure 4. Areas with affinity for the OS probe are highlighted by the gray surfaces.

6.4 Cofactor interaction

The following step of the SsuD analysis was the calculation of the interaction with the cofactor. This was made by docking[17] and molecular dynamics[18] of the enzyme-substrate complex.

As previously stated, SsuD is a homo-dimer in the asymmetric unit, and in its crystal structure two active sites are present, one in each subunit. A detailed analysis pointed out that they show a different conformation: the active site of subunit A seems to be more closed than the active site of subunit B.

[17] T Lengauer, M Rarey M, *Curr Opin Struc Boil*, **1996**, 6, 402.

[18] B J Alder, T E Wainwright, *J Chem Phys*, **1959**, 31, 459.

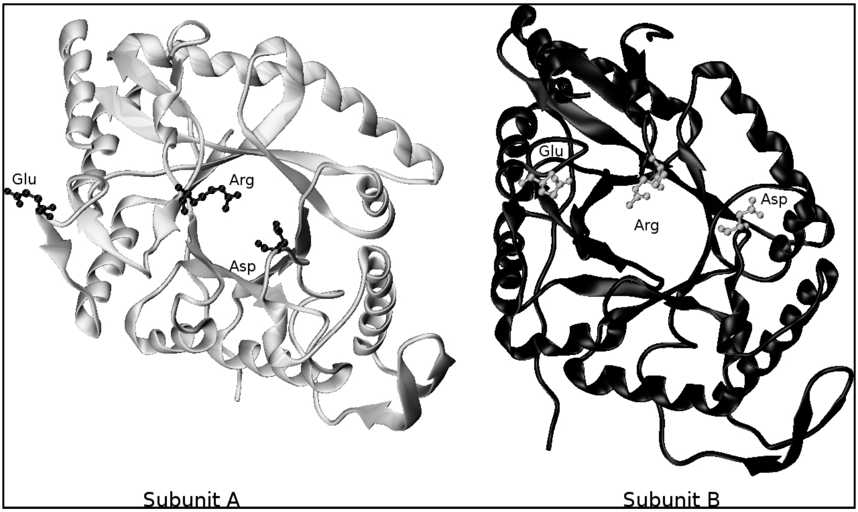


Figure 6.6: Active site of SsuD, closed conformation for Subunit A on the left and open for Subunit B on the right. The residue Asp interacts with the residue Arg in the closed conformation of Subunit A and with the residue Glu in the open conformation of Subunit B.

The conformational differences are mainly due to the interactions of three aminoacids: Arg 297, Asp 111 and Glu 21. Arg 297 can establish ionic interactions with both acidic residues (Asp 111 and Glu 21), switching the active site conformation from closed to open state and vice-versa. As a matter of fact, these two different states suggest the existence of a dynamic switching mechanism ruled by Arg 297 conformational changes (Figure 6). Moreover, the enzyme-cofactor interaction may be affected by this switch.

The cofactor was docked in both active sites (subunit A and B). A large number of diverse poses were generated and a relevant difference between the two active sites (being in two different states) emerged. Therefore an unambiguous pose for the cofactor was not determinable with the sole docking simulation. By looking at the scoring function (London dG)[19] as well as at the

[19] MOE **2006.08**, *Chemcomp*, Montreal, Canada.

potential energies of the different conformations, it was clear that the overall interactions between FMN and the active site of subunit A (closed conformation) were more favourable than the interactions with the open form of subunit B.

A further scoring criterion was necessary for screening the calculated cofactor conformations. The three-dimensional coordinates of LadA co-crystallized FMN was used for this purpose. This way it was possible to univocally select one conformation satisfying the energy based scoring function and the structural requirements derived from the superimposition with LadA. Nevertheless, the obtained structure had a high potential energy due to its unrefined position, an energy minimisation was therefore necessary.

6.5 Molecular dynamic simulation of active site conformational changes

The enzyme-cofactor interactions were further studied by subjecting the enzyme-cofactor complex to ns-scale molecular dynamics simulations.

The enzyme-cofactor minimized complex, with $FMNH_2$ present in the active site of subunit A only, was explicitly solvated in water and re-minimized using the PME[20] algorithm implemented in GROMACS[21] with GROMOS-96 53a6[22] force field. The generated system was used as input for 20 ns molecular dynamics simulation.

[20] U Essmann, L Perera, M L Berkowitz, T Darden, H Lee, L G Pedersen, *J Chem Phys*, **1995**, 103, 8577.

[21] H J C Berendsen, D van der Spoel, R van Drunen, *Comp Phys Comm*, **1995**, 91, 43.

[22] C Oostenbrink, A Villa, A E Mark, W F van Gunsteren, *J Comput Chem*, **2004**, 25, 1656.

The study of the active site conformational change was performed by analysing the dynamics trajectories focusing on the movement of the Arg 297 residue. The initial distance between Arg 297 and Asp 111 is about 7 Å; this distance decreases during the simulation and reaches equilibrium at the distance of 2 Å in approximately 10 ns, after which it remains stable for the rest of the time.

This movement is due to the presence of $FMNH_2$ that perturbs the system, which takes approximately 10 ns to reach a new equilibrium state. The ionic interactions between Arg 297 and Asp 111 prevent the expulsion of the cofactor; while the distance between Arg 297 and Glu 21 is variable during the entire period of the simulation fluctuating between 12 and 3 Å (Figure 7).

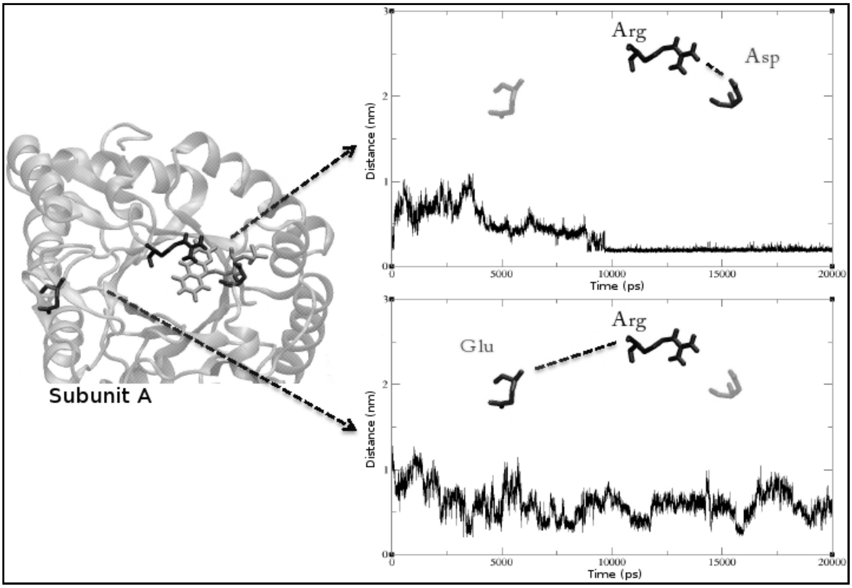


Figure 6.7: Evaluation of the minimal distances between Arg 297, Asp 111 and Glu 21 during the 20 ns molecular dynamics simulation of Subunit A.

The results of the simulation of subunit B show an opposite phenomenon: the distance between Arg 297 and Asp 111 is not stable and fluctuates between 5 and 12 Å, on the contrary the distance between Arg 297 and Glu 21 is stable at 2 Å during the entire simulation (Figure 8).

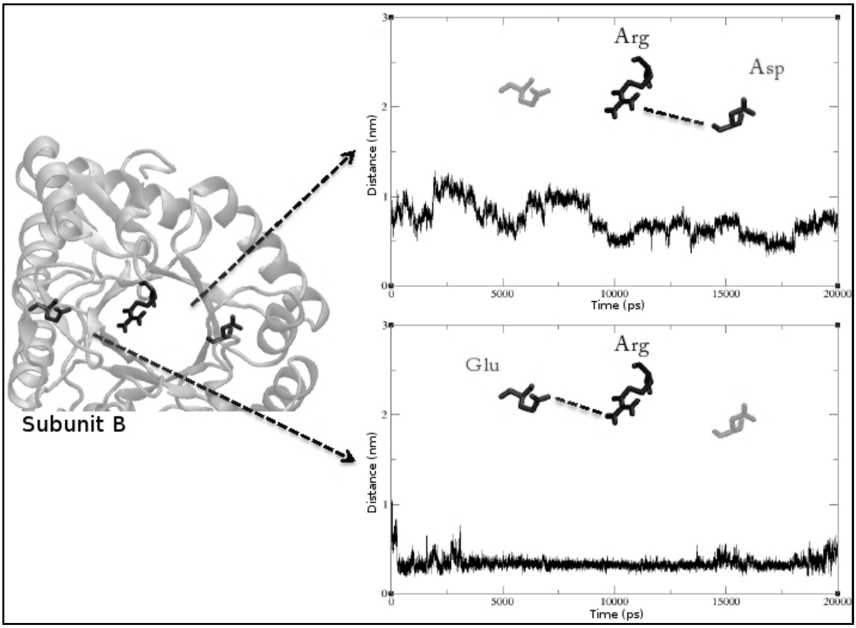


Figure 6.8: Evaluation of the distances between Arg 297, Asp 111 and Glu 21 during the 20 ns molecular dynamics simulation of Subunit B.

This simulation showed that the presence of the cofactor influences the progressive stabilization towards the closed conformation of the active site, while the empty site remains stable in the open state.

After this simulation the cofactor was removed from subunit A and the system was submitted to another MD simulation, to verify whether the hypothesized dynamic switch does exist. The mechanism was not observed after 20 ns, but increasing the simulation time to 100 ns the dynamic switch of Arg 297 was noticeable. In this simulation after approximately 25 ns Arg 297 breaks its interaction with Asp 111 and establishes new hydrogen bonds with Glu 21, the distance between these two residues gradually changes from the initial 15 Å to a final distance of 3 Å (Figure 9).

These simulations confirm that Arg 297, Asp 111 and Glu 21 are important residues for the enzymatic activity because they are responsible for cofactor entrapment. A dynamic conformational change of the active site, switching from a closed to an open state, due to Arg 297 was also demonstrated. The presence of $FMNH_2$ undoubtedly stabilizes the active site in the closed conformation, while on the other hand the empty active site is energetically more stable in its open conformation and therefore accessible for cofactor accommodation. Nevertheless the opening-closing mechanism is necessary for substrate access and cofactor regeneration.

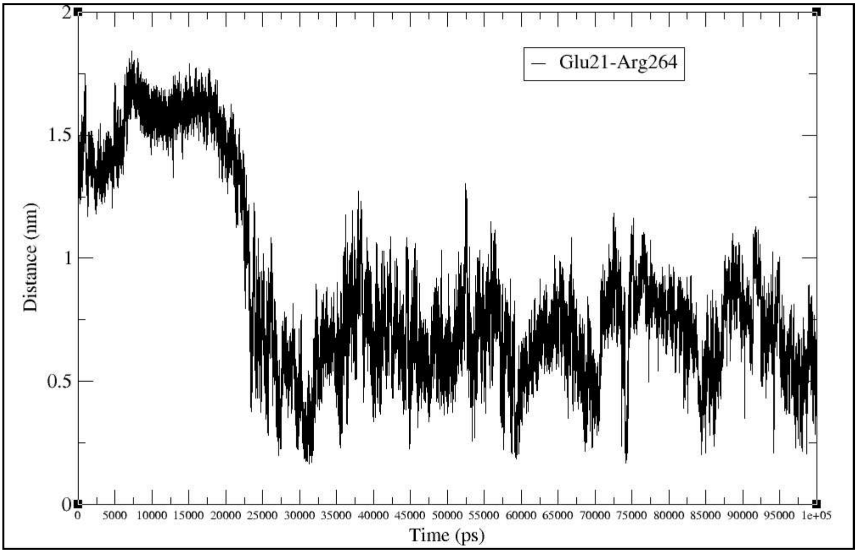


Figure 6.9: Evaluation of the minimal distance between Arg264 and Glu21 during 100 ns MD simulation of subunit A after cofactor removal.

6.6 Substrate interactions and selectivity

After having obtained the enzyme-cofactor complex and having understood its major structural features and dynamic behaviour, it was possible to study the interaction with substrates. SsuD catalyses the oxidation of aliphatic sulphonates and though its specificity seems to be pretty narrow towards long linear alkyl compounds, it can accept a significantly diverse set of compounds. A set of substrates was taken from the literature,[2] molecular docking and molecular dynamics were used for the calculation of their complexes with the enzyme and a subsequent quantitative structure activity relationship (QSAR) approach was applied to the system for generating a predictive model of the substrate specificity.

The substrates used in this part of the study and their corresponding experimentally measured k_{cat}/K_M are reported in table 1. The data set meets the necessary requirements in terms of structural diversity, k_{cat}/K_M range and homogeneity of distribution, which are crucial for the generation of a consistent QSAR model.

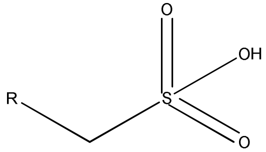
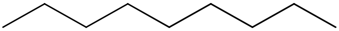
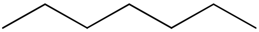
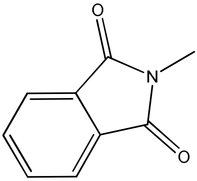
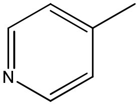
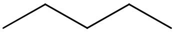
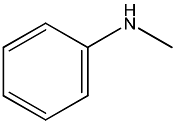
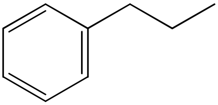
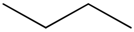
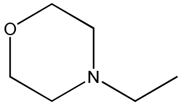
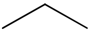
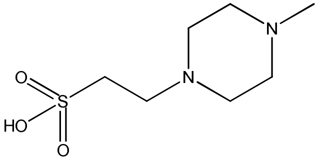
Compound	 R-CH_2-SO_3H	Experimental[2] k_{cat}/K_M (min^{-1}µM^{-1})
1		6,7
2		6,1
3		6,0
4		4,6
5		4,0
6		3,2
7		2,7
8		1,8
9		1,1
10		0,6
11		0,4

Table 6.1: Data set of the 3D-QSAR model.

The starting point for the determination of the substrates' active conformers was the equilibrated enzyme-cofactor complex structure after 20 ns dynamics simulation. The substrates were docked into the same active site and the criteria for the screening of the generated poses was based on the molecular interaction fields calculated by GRID, especially based on OS probe and described above.

Despite several different interaction regions were pointed out by the OS probe, this information was of outstanding importance for the docking scoring step.

The scoring of the docked conformations of the substrates was based on three criteria:

i) Position of the sulphonic group in the areas determined by OS probe molecular interaction field and correct orientation towards the alkaline side chain of His 11;

ii) The geometric compatibility with a possible catalytic mechanism, namely the distance between the sulphonic group and the $FMNH_2$-bonded oxygen molecule involved in the catalysis;

iii) The docking algorithm energy based scoring function (London dG).[19]

Following this scheme one single conformation for each substrate was chosen and the system was subjected to energy minimization and molecular dynamics for a more detailed conformational analysis. In particular, each system was solvated with a water shell and energy minimized using PME[20] algorithm by GROMACS[21] and GROMOS-96 53a6[22] force field. Subsequently a 300 ps molecular dynamics simulation for each enzyme-substrate complex was carried out.

The outcome of every MD simulation was carefully analysed and the substrate conformers with the lowest potential energy were selected and subsequently used for the construction of the QSAR model. The GRID independent descriptors (GRIND)[23] implemented in the software ALMOND[24] were used, since this procedure is alignment independent and permits to avoid errors coming from superimposition of substrates, common obstacle in 3D-QSAR.

GRIND starts from the molecular interaction fields (MIFs), obtained with the GRID approach, pointing out regions where the molecule can produce energetically favourable interactions with its environment, simulated by the probes. Then the method transforms these MIFs in a relatively small number of variables, namely an ensemble of vectors coupling points of the MIFs (representing favourable interactions) at different distances, having the highest possible energy. The vector can join points belonging to the same MIF, or to different MIFs. This way the variables are grouped in blocks generally called correlograms and specifically auto-correlograms when the vectors join points within a single MIF, cross-correlogram when they join points belonging to different MIFs. In other words, the GRIND descriptors are auto- and cross-correlation vectors that join the MIF's points with the highest energy products.

Four probes with different physico-chemical properties were used in the calculation of the molecular interaction fields, namely the WATER, the DRY, the Ca++ and the TIP probe. The WATER probe describes and quantifies the dipolar interactions and the hydrogen bond formation, the DRY probe considers all the hydrophobic interactions, the Ca++ probe takes into account the interactions with charged groups, for instance the sulphonic group, and the TIP probe generates a MIF that is strictly dependent on the shape of the molecule. Those four probes generated totally 10 correlograms, four auto-correlograms and six

[23] M Pastor, G Cruciani, I McLay, S Pickett, S Clementi, *J Med Chem*, **2000**, 43, 3233.

[24] ALMOND **3.3.0**, *Molecular Discovery Ltd*, Perugia, Italy.

cross-correlograms (all possible couples), corresponding to 600 independent variables (vectors).

The energy values contained in the matrices of the GRIND descriptors were statistically analysed to generate PLS[25] models able to correlate the molecular descriptors with the experimental k_{cat}/K_M values.

The first model included all 600 variables, but its predictivity was low. In order to improve the consistence of the generated model, a variable selection process was necessary. All the vector blocks containing insufficient information in terms of substrate discrimination (vector blocks with standard deviation close to zero) were discarded. In this operation all the vectors coming from the Ca++ probe were deleted; probably because the Ca++ probe generates high interaction values only in correspondence of the sulphonic group which is present in every object of the data set.

Then we applied the FFD variable selection algorithm[26] which conserved only 136 active variables. The final PLS analysis was performed only on 10 of the initial 11 compounds; one compound (**5**) with average k_{cat}/K_M value was excluded from the training set due to its outlier behaviour. Five principal components were calculated.

The model was validated by means of the leave-one-out (LOO) cross-validation procedure (Table 2). The predictive correlation coefficient (q^2) which provides the quantitative evaluation of the consistency of the model was as high as 0.719 on the third principal component whereas 71 percent of the variance of the model was explained by the first three principal components and showed an r^2 of 0.976 on the third PC.

The predictivity of the model was satisfactory, showing a good quality of prediction especially for the molecules lying in the

[25] F Ildiko, J Friedman, *Technometrics*, **1993**, 35, 109.

[26] G Cruciani, S Clementi, M Baroni, Theory Methods and Applications, *H Ed. ESCOM: Leiden*, **1993**, 551.

medium to high part of the k_{cat}/K_M range (see Table 2 and Figure 10). The performances of the model slightly decreases for the molecules having low k_{cat}/K_M, simply because that zone of the plot is less populated. This means that in principle the predictivity of the model could be even higher than measured by the LOO procedure.

Compound	**Experimental[2] k_{cat}/K_M ($min^{-1}\mu M^{-1}$)**	**Predicted k_{cat}/K_M by LOO ($min^{-1}\mu M^{-1}$)**
1	6,7	4,9
2	6,1	6,3
3	6,0	3,9
4	4,6	3,5
6	3,2	3,6
7	2,7	3,2
8	1,8	1,4
9	1,1	0,6
10	0,6	1,3
11	0,4	1,2

Table 6.2: Comparison of the measured experimental k_{cat}/K_M values and the k_{cat}/K_M predicted by LOO cross-validation procedure applied on the data set.

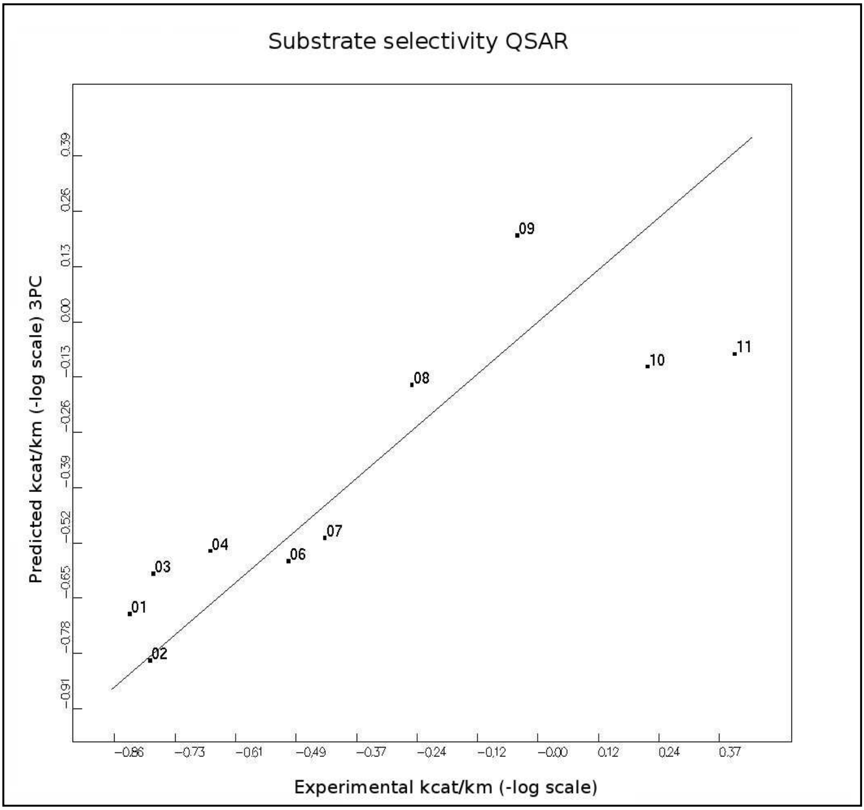


Figure 6.10: Predictivity of the model in terms of experimental versus predicted k_{cat}/K_{M}; values expressed in p (-Log) scale.

The interpretation of a QSAR model based on GRIND descriptors can sometimes be a difficult task. The advantage of having alignment free mathematical description of the data set causes the loss of the direct connection between molecule's chemical characteristics and their absolute position in the Cartesian space, making the comparison of the different molecules of the data set particularly difficult. Nevertheless it is still possible to get important information from the analysis of the model's variables having the highest statistical weight; this is being highly correlated with the activity of interest. The analysis of the correlograms and the PLS weights profile plot is of great utility in this perspective.

Statistics point out that among all the original variable blocks, only those coming from DRY and TIP probes MIFs were informative.

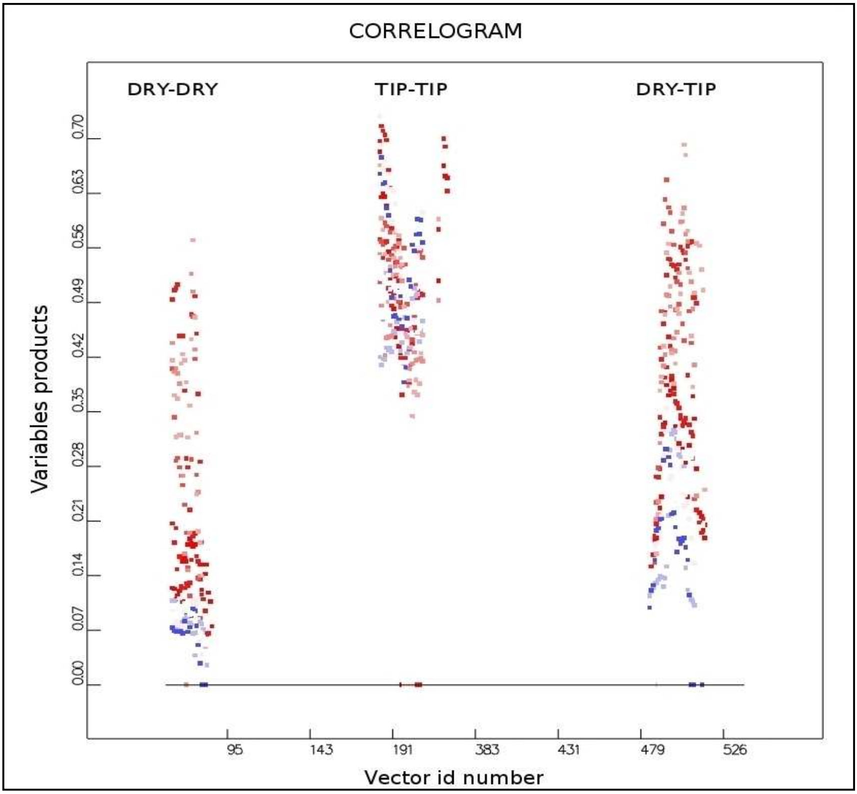


Figure 6.11: Auto- and cross-correlograms of the probes DRY and TIP; each single point corresponds to a vector correlated with a measured activity, high activity values in red, low in blue.

The correlograms show that the GRIND descriptors are able to separate the active compounds from the inactive ones (Figure 11).

The DRY-DRY auto-correlogram generates vectors with small module (the product of MIFs original variables it connects) for low activity (blue points) and vectors with big module for highly active molecules (red points). The same trend is preserved also for the DRY-TIP cross-correlogram. The TIP-TIP auto-correlogram discriminates among the activities in a different, yet similar way. The discrimination is 'horizontal', with the active compounds having a broader ensemble of vectors whereas the inactive compounds show narrower correlograms. In other words, the substrates having a high k_{cat}/K_M generate longer vectors than the compounds characterized by low k_{cat}/K_M. This is associated with the length and shape of the molecules. TIP probe describing long linear alkyl chains usually generates two strong interaction zones at the ends of the chain, while with globular or complex shaped structures it generates several different zones around the molecule. In this latter case the average distances among these zones are significantly shorter, thus generating an ensemble of shorter vectors.

The comparison of two molecules, lying on the opposite side of the activity graph (Figure 10), is useful to understand how a chemical meaning can be extracted from the statistical model. Highlighting the vectors having the major contribution to the PLS in both the DRY and TIP MIFs demonstrates how the distribution of hydrophobicity and the shape of the substrates in their active conformation are the key elements for interpretation. Comparing for instance molecule **1** and **9** (Figure 12), it appears clearly that extended linear conformations, presenting significant hydrophobicity in their central parts are necessary for displaying high k_{cat}/K_M. Compound **1** is substantially linear and hydrophobic, the vectors corresponding to the highest PLS weights are extremely similar and they univocally connect the two ends of the molecule. The DRY auto-correlogram is made by a complex network of relatively short vectors, connecting the extensive central hydrophobic zone of the compound. On the other hand, the shape of compound **9** shows a slightly globular character, affecting significantly the ensemble of vectors of the TIP auto-correlograms, as it can be seen in figure 11. The hydrophobic character is still present, but it is unhomogeneously distributed and the molecule is remarkably shorter than compound **1**, altering the network of vectors of the DRY auto-correlogram.

The TIP representation is correlated with the shape of the active site cleft, which is long and narrow, and the DRY representation is correlated to the interactions that the substrate establishes with the isoalloxazine ring of the cofactor.

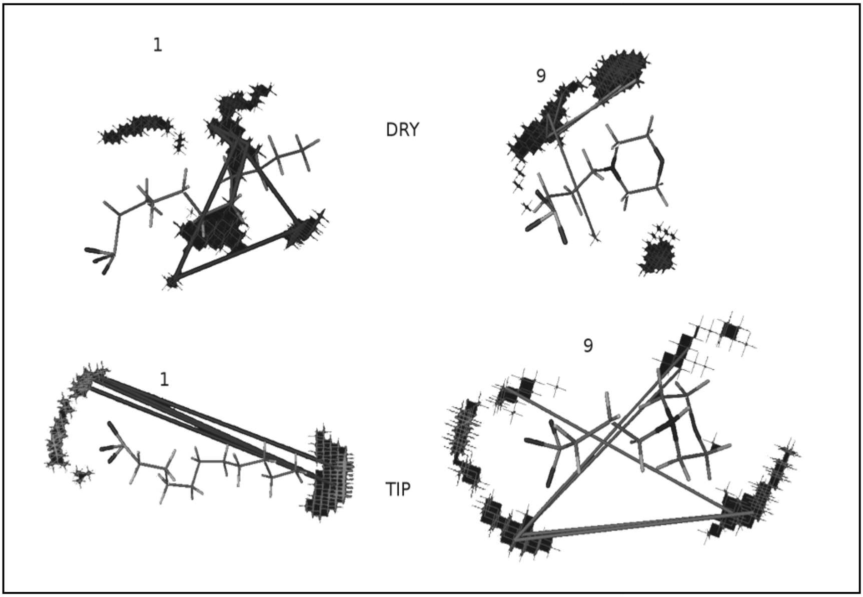


Figure 6.12: ALMOND vectors for the auto-correlograms of the DRY probe on top and for the TIP probe on bottom; compound 1 on the left, compound 9 on the right.

Besides being an instrument for the prediction of alkanesulfonate monooxygenase substrate specificity, the model offers a guideline describing the physico-chemical characteristics necessary for being accepted by this enzyme. In this context it can be used as a reference for reaction engineering.

6.7 Conclusions

Molecular simulation methods have been applied to the study of SsuD action at molecular level, shedding light to many aspects of enzyme-substrate interaction mechanisms. The enzyme showed a two-states dynamic mechanism, switching between an open and a closed conformation of its active site. The presence of $FMNH_2$ cofactor stabilizes the closed state, while the empty active site showed to be energetically more stable in the open state. This mechanism has never been observed before and its function is clearly the regulation of substrates access and cofactor regeneration.

The role of some residues, already pointed out as relevant for the enzyme’s catalytic activity, has been confirmed and clarified. For instance, Arg 297, Asp 111 and Glu 2 govern the dynamical switch between the open and closed conformations, while His 44 and Trp 196, on the other hand, are crucial for stabilizing the sulphonic group of the substrates and the cofactor respectively.

The conformational analysis of a set of substrates, of known k_{cat}/K_M, performed by MD simulations was used for the generation of a 3D-QSAR model for the prediction of SsuD substrate specificity. The Grid INdependent Descriptors (GRIND) and Partial Least Squares (PLS) have been used for building up the model, which demonstrated to be statistically robust and predictive. The interpretation of the 3D-QSAR model helped in pointing out the most relevant structural characteristics correlated with enzyme specificity. The distribution of hydrophobicity and the shape of the substrates in their active conformation showed to be essential in this aspect. Substrates whose active conformation is elongated, with a neat and defined polar part, corresponding to the sulphonic group, linked to a long and narrow hydrophobic structure are those showing the highest k_{cat}/K_M. On the contrary, the globular character of the hydrophobic part of the substrate is inversely proportional to enzyme substrate specificity.

The model can be used to predict SsuD selectivity towards new compounds, thus representing a helpful tool for the support in developing synthetic processes involving this enzyme.

The results achieved during the studies on Alkanesulfonate monoxygenase prove that the computational methodologies previously employed on lipases can be successfully applied on other enzymatic systems with equal success. Molecular dynamics simulations demonstrated once again their high efficiency in conformational sampling, while hybrid approaches obtained by the synergic application of MD simulations, molecular descriptors calculation and multivariate statistics lead to the generation of very powerful quantitative predictive models.

6.8 Experimental section

The protein structures used for this study were retrieved from the Protein Data Bank (Id:1m41 for SsuD, 1luc for bacterial luciferase, 3b9o for LadA and 1z69 for Mer). These initial structures were pretreated in MOE by removing the crystallographic water molecules present in the pdb file. Hydrogen atoms were added and their positions were optimized with an energy minimization procedure in the Amber99 force field in its MOE implementation. Subsequently a minimization of the side chains was performed keeping the backbone atoms fixed.

Superimpositions of protein structures was performed using the MOE Superpose tool fixing the spatial position of the most preserved residues (His, Trp, Arg) in all the superimposed structures. The alignment was performed using the blosum30[27] matrix.

GRID

The GRID analysis was performed on the SsuD protein, choosing a cage big enough to include the whole protein. The grid nodes were set every 0.5 Å. The probes used for the calculation of the molecular interaction fields were DRY (hydrophobic probe), WATER (H_2O probe) and OS (sulphonic acid mimic probe).

Docking

The Docking procedure was performed using the docking module of MOE.

Concerning $FMNH_2$, the procedure was performed on a 10 Å radius selected area surrounding the active site, taken from the coordinates of active His, Trp and Arg residues. The force field used for the docking was MMFF94x, the partial charges of the atoms were calculated at the PM3 semi-empirical level, by means of the MOPAC7 program and the different poses were placed with the alpha PMI method and scored by means of the London dG scoring function. The final $FMNH_2$ pose was finally located by evaluation of the scoring function (London dG) and the pose similarity with the FMN co-crystallized in the PDB structure of LadA (PDB 3b9o).

The docking simulations of the substrates were performed in the same way considering a 12 Å radius selected area surrounding the FMN; the initial positions of the substrates were manually set in order to meet the three different criteria: i) sulphonic group

[27] S Henicoff, J F Henicoff, *PNAS*, **1992**, 89, 10915.

located in a OS probe interacting area; ii) sulphonic group distance from $FMNH_2$; iii) London dG scoring function. The substrate molecules were inserted by building each molecule with the MOE builder tool and subsequently minimized. For each substrate, the conformation presenting the highest score and fulfilling the structural requirements for the initiation of the enzymatic catalysis was chosen.

Molecular dynamics

The molecular dynamics simulations were performed using the software GROMACS with the GROMOS-96 53a6 force field. The SsuD pdb crystal structure was implemented in the force field in gro file format by using the automatic tool of the GROMACS software which also add the necessary hydrogens. The protein was solvated with explicit water in a virtual box of 1331 nm^3. All the dynamics were performed in a NPT environment simulating the temperature of 300 K and keeping the pressure constant (Berendsen-thermostat and pressure), cut-off for electrostatic interaction was setted at 1.4 nm and the limit for the van der Waals interactions setted at 1.4 nm. Only for the minimization procedures the PME algorithm (Particle Mesh Ewald and not a simple cut-off) was used for the calculation of the electrostatic interactions setting the limit at 1.0 nm. The $FMNH_2$ and the substrate molecules were parametrised in the GROMOS-96 53a6 force field by using the Dundee PRODRG2 server[40] and manually refined in order to meet the correct force field definition. The reduced FMN and the substrate molecules were manually added in the gro file taking the coordinate from the docking results. The system was previously minimized with 1000 step of steep descendent algorithm before every molecular dynamics calculation.

Molecular dynamics analysis were performed with GROMACS tools; distances measurements were calculated using g_mindist and computing the minimum distance between two residues; the results were visualized using Grace software.

Each substrate conformation chosen for the construction of the data set for the QSAR analysis was the one characterized by the lowest potential energy out of all the frames saved in the dynamics trajectories.

Almond

Four different probes were used for the Almond vectors generation, namely DRY (hydrophobic probe), WATER (H_2O probe), Ca++ (charged probe) and TIP (shape probe). These probes generated in total ten correlograms (four auto-correlograms and six cross-correlograms) corresponding to 600 variables (vectors).

All the vectors generated from the data coming from the Ca++ probe were discarded (standard deviation close to zero).

Variables selection was performed by using the FFD algorithm keeping the uncertain variables, 136 variables were active after this operation. One compound (**5**) with average k_{cat}/K_M value was excluded from the training set (outlier behaviour).

PLS models with 5 principal components were computed and validated by LOO (leave-one-out) method. QSAR substrate predictivity model is expressed in terms of experimental versus predicted (on three principal component) k_{cat}/K_M with values expressed in p (-Log) scale.

CHAPTER 7

CONCLUSIONS

Great advantages of biocatalytic processes in relation with classical chemical approaches have been universally accepted during the last few years. More and more biocatalytic applications are now commonly employed in industrial processes.

A comprehensive knowledge about the enzyme of interest is a key factor for the improvement and expansion of biocatalysis processes, nevertheless several aspects of enzymatic behaviour, as well as new enzyme activities, have not been completely investigated yet.

The classical experimental trial and error approach needs to be reviewed in order to maximise the information gained from any experimental attempt and rationalize at molecular level the enzyme behaviour.

New strategies based on molecular modelling techniques have been developed during the last ten years. Molecular dynamics simulations (MD) are able to investigate enzyme behaviours at molecular level. The principal limitation of this computational methodology is its inefficiency for simulations regarding complex phenomena. In these cases a different statistical approach is more recommended and it can be used to extract relevant information from any experimental test and create a predictive model.

All these techniques represent the bases for the development of new hybrid approaches for enzyme and biocatalyticaly based procedures.

This idea was applied to one of the most important and used enzyme classes, the lipases. Several lipases were investigated looking at their crystal structure and studying their surface properties in order to find some common features during the lipase activation process. This process depends on the movement of a protein domain called lid which covers the active site when the enzyme is in its inactive conformation and undergoes a conformational change, exposing the active site and making it accessible for the substrate in the enzyme’s active conformation. This phenomenon concerning the lid movements of several lipases was investigated by performing molecular dynamics simulations. Different molecular dynamics simulation techniques were used, starting from classical MD simulations, steered MD simulations which reduce the computational time required to simulate the desired event and finally coarse grained MD simulations with the MARTINI force field. Classical MD simulations demonstrated to be able to simulate correctly the process of lipase activation. Steered MD simulations effectively reduced the computational time needed for the study of these phenomena. Nevertheless, results obtained by steered MD simulations need to be carefully analysed because of the slight movement distortion caused by the method itself.

Different concepts were acquired during the work of this thesis, not only from the methodological point of view. Concerning lipases, it was observed that small bacterial lipases are usually characterised by a huge lid which is affected by a complex movement that involves more than one protein domain. On the other hand, other eukaryotic lipases with a consistent molecular weight usually have a small lid and its movement is governed just by the breaking and the formation of few hydrogen bonds. These lid movement features can be used for a new type of lipase classification. Moreover the force that drives these conformational changes has to be found in the media characteristics because in polar environments the protein hinders the hydrophobicity of its active site in order to minimise the unfavourable interactions that would be established between the solvent and the protein active site.

MARTINI demonstrated to lack the necessary accuracy to simulate a fine event like the lid movement. On the other hand MARTINI was successfully applied in the simulation of enzyme orientation at the interface and more generally it proved to be suitable for the simulation of big system in order to study enzyme orientation.

The intrinsic characteristics of MARTINI make it suitable for the estimation of the average molecular vibrations of an enzyme in different water-solvent mixtures which were performed in order to study enzyme stability in different conditions. The combination of these computed vibrations with experimental data of enzyme residual activity after incubation in the same water-solvent mixtures that were parametrised in silico led to the generation of a predictive model. This was performed for the *Candida antarctica* lipase B.

Another important enzyme characteristic given by the enzyme selectivity was investigated during this work. In order to quantitatively describe this feature a statistical approach based on 3D-QSAR was performed on the same enzyme. This consisted in the development of a new class of molecular descriptors, namely differential Molecular Interaction Field. The intrinsic properties of these newly developed descriptors is the ability to describe two objects simultaneously and can therefore be applied as the descriptors for the characterization of the enantioselectivity of an enzyme.

To test the general applicability of the computational approaches developed for the study of lipases, they were finally tested on a completely different and relatively unknown enzyme in order to prove the universality of these methods. MD simulations to study the dynamics of enzyme activation/inactivation, as well as 3D-QSAR approach to study the substrate selectivity were successfully applied to Alkanesulfonate monooxygenase.

This enzyme is different from the other studied enzymes in terms of its classification, mechanism of action as well as the necessity of a cofactor involved in the catalytic process. The application of the previously established techniques proved to be adequate in this case study as well. Different interesting notions about the enzyme properties were highlighted during the study. It was seen that the enzyme undergoes a conformational change in the presence/absence of the cofactor given by a motion of a particular structural domain. With the aid of molecular docking and MD based approaches the correct substrate collocation inside the active site was comprehended. Nevertheless, for a deeper understanding of the enzyme substrate recognition, at molecular level, a 3D-QSAR methodology was applied. The constructed model proved to be consistent for the description of the desired properties as well as predictive for the enzyme selectivity estimation.

ACKNOWLEDGMENTS

ACKNOWLEDGENTS

Fisrt of all I want to thank my tutor Prof.ssa Lucia Gardossi for her total support and confidence. She always believed in me and a lot of time she tolerated me.

Thanks to Dr. Paolo Braiuca for his patience especially during my thesis writing and for all the things he taught me: some good and also a lot of bad things, with his behavior he made me stronger than before.

Thanks to Prof. Cinthia Ebert and Prof. Paolo Linda for their kind words of encouragement.

Special thanks to Prof. Siewert-Jan Marrink, when I was visiting his research group he welcomed me very warmly and he is always available.

Many thanks to Prof. Francesco Molinari, my supervisor when I was just an undergraduate student, the person who suggested me to plunge myself into this PhD adventure.

I can not forget my parents and the rest of my family for their indestructible faith in my possibilities.

Special thanks to Lorena, she helped me a lot in every situation.

Thanks to Patty, Ale, Sara, Loris, laughing Steve, Debrah, Lince istintiva, Paoletta and Science Student for their attentions and their gaiety at work.

The Buon Pier, Matej and his staff, as well as Michele, Tommaso and the buon Pablo for their help and for all the nights that we had fun together, the drinks and their psychological help.

MetJuti for his wonderful sgnape.and kaiPiroska.

Of course the Capperi Marshall Vittorio, we had together a very crazy period wet by a lot of drinks.

Pato and Tux thanks a lot for their support and for the crazy and philosophical speeches we used to have.

Thanks to all the Groningen people of the MD group, the NMR group, as well as the others visiting students Jorge and Byron, my period there was very beautiful also thanks to their friendship.

Thanks to my old volleyball team friends: marca tre sul tabellino!!!

Special and very great thanks to my own town friends: Carra, Marco, Simo, Pippo, Massimino seven beauties, Tony, We Wescio, Ini, Zanzi, and all the others.

Last but not least thank you very much to my Bella Pollastrella…

I love her but I would like to hate her.

I want also to thank all the people that I forgot in these too short acknowledgments.