XXIV CICLO DEL
DOTTORATO DI RICERCA IN
SCIENZE E TECNOLOGIE CHIMICHE E FARMACEUTICHE

“COMPUTATIONAL METHODS FOR RATIONAL SCREENING AND ENGINEERING OF ENZYME PROPERTIES”
(Settore scientifico-disciplinare CHIM/06)

DOTTORANDO
Lorena Knapic

RESPONSABILE DEL DOTTORATO DI RICERCA
Chiar.mo Prof. Enzo Alessio
Università degli Studi di Trieste

RELATORE
Chiar.mo Prof. Cynthia Ebert
Università degli Studi di Trieste

ANNO ACCADEMICO 2010/2011
This PhD project was part of the EU-FP7 “IRENE“ project and it has received funding from the European Community's Seventh Framework Programme under the FP7-KBBE-2008-2B grant agreement n° 227279

ESTECO S.r.l is greatly acknowledged for the support
Table of contents

Abstract III

Chapter 1 Introduction to computational biocatalysis 1
1.1 Biocatalysis: how and why 1
1.2 Enzyme engineering and overcoming application limitations 8
1.3 Computational approaches 15
1.4 Aim of the thesis 22

Chapter 2 Differential RMSD as a novel descriptor for rational redesign of natural enzymes in case of bulky substrates 25
2.1 Summary 25
2.2 Introduction 27
  2.2.1 Lipase catalyzed synthesis of polyesters 27
  2.2.2 Chiral polymers by lipase catalysis 33
  2.2.3 Molecular dynamics and RMSD calculation 37
  2.2.4 Aim of the work 39
2.3 Results and discussion 40
  2.3.1 In silico simulation of enzyme-substrate complexes 40
  2.3.2 Glycolide, DD-lactide and LL-lactide effects on CALB structure 41
  2.3.3 Differential RMSD calculation 48
  2.3.4 Caprolactone as a model substrate 52
  2.3.5 In vitro production of mutants and experimental validation 56
2.4 Experimental section 59
2.5 Conclusions 60

Chapter 3 Converting a hydroxynitril lyase into an esterase; a computational study 62
3.1 Summary 62
3.2 Introduction 64
  3.2.1 Alpha/Beta-Hydrolase fold enzymes 64
3.2.2 Reaction promiscuity and evolutionary implications 66
3.2.3 Hydroxynitril lyases and esterases 70
3.2.4 GRID-PCA and its application in rational enzyme engineering 75
3.2.5 Aim of the work 78
3.3 Results and discussion 79
3.3.1 Data set selection, sequence and structural similarity 80
3.3.2 GRID-PCA analysis and differences identification 85
3.3.3 Rational identification of hot-spot for conversion of HNL into Esterase and *in silico* mutant generation 90
3.4 Experimental section 96
3.5 Conclusions 98

**Chapter 4 Development of an automatic tool for *in silico* enzyme engineering** 100
4.1 Summary 100
4.2 Introduction and aim of the work 101
4.2.1 The case study 103
4.3 Results and discussion 108
4.3.1 The variables 110
4.3.2 The generation of the objects 110
4.3.3 Simulations 112
4.3.4 Scoring and objectives 115
4.3.5 Workflow description: automation and evolution 124
4.3.5.1 3D-QSAR scoring 124
4.3.5.2 Hydrogen bond scoring 127
4.4 Experimental section 137
4.5 Conclusions 144

**Chapter 6 Conclusions and future perspectives** 146
Abstract

The major advantage offered by computational methods over experimental approaches in biocatalysis is the possibility of describing the phenomena at a molecular level. The rational description and understanding of the molecular basis of enzyme's catalysis and properties that can be derived is useful in many ways. It can be applied, for example, to rational enzyme engineering, process engineering, and enzyme immobilization. Many successful examples of rationally improved enzyme properties can be found in the literature. Nonetheless, existing in silico methodologies are far from being conclusive or able to solve any given problem. On the contrary, there are many challenges laying ahead and waiting to be solved. Some of these challenges require high computational power and long simulation times and their solutions are therefore not affordable on a routine basis. Others have not yet been solved or the so-far proposed solutions are not satisfying enough.

In the present work, state of the art computational techniques were applied to several current research topics in biocatalysis such as substrate promiscuity, reaction promiscuity and high throughput mutant generation and screening. The studied subjects are of great interest to industrial biocatalysis nowadays and can find large application for rational redesign of inefficient biocatalysts and fast substrate engineering and screening. The overall work can be divided into three principal areas i.e. understanding catalytic mechanisms, description of enzyme-substrate interactions and integration of available computational methods for the development of a novel automatized tool for enzyme engineering. In each of these three areas, the goal has been to test the existing methodologies as well as the development of new descriptors and ready to use strategies.

Thus in Chapter 2, the work is focused on redesigning Candida antarctica lipase B to increase its catalytic efficiency in polymerization reactions. In this specific application the substrate structure is particularly big. The enzyme-substrate complexes clearly indicated considerable steric clashes and a novel class of molecular descriptors was developed to quantify the undesired effect and identify appropriate mutation hot spots. Mutants were produced in vitro and characterized; higher reaction rates were obtained.
Chapter 3 deals with the introduction of novel catalytic functionalities into existing enzyme scaffolds, a phenomenon known as reaction promiscuity. Such modifications are generally useful when wild-type enzymes are insufficiently robust to resist harsh application conditions or the host enzyme is suitable to give significantly broader/different substrate specificity. In Chapter 3 the focus has been on the description of the active sites of two classes of enzymes, namely esterases and hydroxynitril lyases mostly based on the comparison in terms of steric and electrostatic properties.

Chapter 4 describes a step by step development of a strategy that includes all the necessary steps for the in silico design, generation and screening of mutants. The aim was the design of a single software instrument to perform the whole multi-step route (relying on many different softwares) for the simulation of enzyme engineering. The resulting tool is an integration of different modelling softwares into one single interface in which each of the included softwares performs a specific task. The resulting major advantage is the mergence of information deriving from other, separate studies. The flexibility of the generated tool is huge and can therefore be tuned to adapt to any given enzyme and design need, provided that appropriate software for the necessary simulations and an appropriate numerical scoring method of the outcoming enzyme mutant model are available. In this chapter, an redesign of the the wild-type C. antarctica lipase B is taken as the case study, with the aim of introducing amidase activity.
1. Introduction to computational biocatalysis

1.1. Biocatalysis: how and why

Biocatalysis, as the word says, is the process in which a chemical reaction is catalysed by a natural (bio) catalyst, *i.e.* an enzyme. The enzyme can be used in various forms, from pure, isolated enzymes to enzymes still residing in an intact living cell. Historically, biocatalytic reactions have been used for a huge amount of years, prevalently for making food in processes such as the production of cheese, beer, leaven, wine and vinegar. The oldest example of chemical transformation known to humans is brewing, with records belonging to the Sumerians, 6000 years ago.

With time, many different industries learned to use enzymes to optimize and improve their products, such as the leather and flax industry. However, it was only with the development of fermentation technologies that the utilization of pure enzyme preparations became possible. The real boost of biocatalysis is though not older than three decades, a period in which the developments of enzymology happened along with that of biotechnology, and in particular with that of recombinant DNA techniques. In this last period, there has been a substantial increase in the application of biocatalysis to produce fine chemicals.

Table 1.1 provides an overview of historic events in biocatalysis and biotechnology. It demonstrates that biotechnology is an old science, or even an old art. The big events and driving forces for biocatalysis in the 20th century were twofold: first, the idea of catalysis as transition-state complementarity in 1944 and second, the development of molecular biology after 1978 to allow the design of enzymes and their production vehicles.
Table 1.1. Historical events in enzyme catalysis and biotechnology.

<table>
<thead>
<tr>
<th>Year(s)</th>
<th>Who?</th>
<th>Where?</th>
<th>What?</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC</td>
<td>Unknown</td>
<td>Old World</td>
<td>chymosin from calf and sheep stomach utilized for production of cheese</td>
</tr>
<tr>
<td>1783–1836</td>
<td>Spallanzani</td>
<td></td>
<td>verifies in vitro “digestion” of meat in stomach juice: factor called “pepsin”</td>
</tr>
<tr>
<td>1876</td>
<td>Kühne</td>
<td>Berlin Agricultural College, Germany</td>
<td>term “enzyme” for catalysts not bound to living cells (“unorganized ferments”)</td>
</tr>
<tr>
<td>1877</td>
<td>Eduard Buchner (Nobel prize 1907)</td>
<td>Berlin Agricultural College, Germany</td>
<td>1st alcoholic respiration with cell-free extract: vital force, <em>vis vitalis</em>, does not exist</td>
</tr>
<tr>
<td>1893</td>
<td>Wilhelm Ostwald</td>
<td>Leipzig Univ., Germany</td>
<td>definition of term “catalyst” (Nobel prize 1909)</td>
</tr>
<tr>
<td>1894</td>
<td>Emil Fischer</td>
<td>Berlin Univ., Germany</td>
<td>“lock-and-key” concept (Nobel prize 1902)</td>
</tr>
<tr>
<td>1903</td>
<td>Henry D. Dakin</td>
<td>London, UK</td>
<td>1st enantioselective synthesis, with oxynitrilase</td>
</tr>
<tr>
<td>1908</td>
<td>Otto Röhm</td>
<td>Darmstadt, Germany</td>
<td>Patent for enzymatic treatment of leather (with trypsin)</td>
</tr>
<tr>
<td>1913–1915</td>
<td>Röhm Company</td>
<td>Darmstadt, Germany</td>
<td>1st laundry detergent with enzyme (pancreatin): “Burnus”</td>
</tr>
<tr>
<td>1926</td>
<td>James B. Sumner</td>
<td>Cornell Univ., Ithaca, NY, USA</td>
<td>1st enzyme crystallized: urease from jack beans (Nobel prize 1946)</td>
</tr>
<tr>
<td>1936</td>
<td>Ernst Sym</td>
<td>Caltech, Pasadena, CA, USA</td>
<td>lipase reaction in organic solvent</td>
</tr>
<tr>
<td>1944</td>
<td>Linus Pauling</td>
<td>Caltech, Pasadena, CA, USA</td>
<td>1st attempt to explain catalysis as transition-state complementarity</td>
</tr>
<tr>
<td>1950</td>
<td>Pehr Edman</td>
<td>Univ. of Lund, Sweden</td>
<td>protein degradation developed</td>
</tr>
<tr>
<td>1951</td>
<td>Frederick Sanger and Hans Tuppy</td>
<td>Univ. of Cambridge, UK</td>
<td>sequence determination of insulin β-chain: each protein is characterized by a sequence (Nobel prize 1978)</td>
</tr>
<tr>
<td>1960</td>
<td></td>
<td>Novo ( Bagsvaerd, Denmark)</td>
<td>large-scale protease production from <em>Bacillus licheniformis</em> in submerged culture</td>
</tr>
<tr>
<td>1963</td>
<td>Stanford Moore and William Stein</td>
<td>Rockefeller Univ., New York, USA</td>
<td>amino acid sequence of lysozyme and ribonuclease elucidated (Nobel prize 1972)</td>
</tr>
<tr>
<td>1978</td>
<td>Stanley Cohen and Herbert Boyer</td>
<td>Stanford, CA, USA</td>
<td>method of recombination of DNA developed</td>
</tr>
<tr>
<td>1985</td>
<td>Michael Smith</td>
<td>Univ. of British Columbia, Canada</td>
<td>site-directed gene mutagenesis to change enzyme sequence (Nobel prize 1993)</td>
</tr>
<tr>
<td>1988</td>
<td>Kary B. Mullis</td>
<td>Cetus Corp., CA, USA</td>
<td>invention of PCR (Nobel prize 1993)</td>
</tr>
<tr>
<td>2000</td>
<td>Celera Genomics</td>
<td>Gaithersburg, MD, USA</td>
<td>sequencing of human genome announced (3 billion basepairs)</td>
</tr>
</tbody>
</table>
Introduction

Nowadays, biocatalytic transformations show a wide variety of applications and are used in virtually every industrial sector that counts on some kind of chemical reactions. The three principal areas of application are:

1. Production and transformation of compounds, mainly in the chemical and pharmaceutical industry
2. Analytics and diagnostics, mainly in medicine
3. Environmental protection and bioremediation

whereas, the most important industries that apply these areas are pharmaceutical, food, fine chemicals, basic chemicals, material chemistry, pulp and paper, agriculture, medicine, energy production and mining industries.

The interdisciplinarity required by the complexity of biocatalytic applications evolved with time as well. Nevertheless, different disciplines perceive differently biocatalysis and biotechnology. Thus chemistry and chemists emphasize a molecularly-oriented perspective dominated by compounds and transformations, chemical engineering and chemical engineers favor a process-oriented perspective of reactions and processes, whereas biology and its practitioners contribute a systems-oriented perspective of description at the organism level as well as in their view of evolution.

Different parts of each of the three disciplines are needed for the successful practice of biocatalysis: biochemistry and organic chemistry from chemistry; molecular biology, enzymology, and protein (bio)chemistry from biology; and catalysis, transport phenomena, and reaction engineering from chemical engineering are indispensable. Both biotechnology and biocatalysis are interdisciplinary areas; as most practitioners tend to hail from one of the three major contributing disciplines, hardly anybody has an equally strong command of all the sub-disciplines of biocatalysis.

This central role of biocatalysis between multiple disciplines and its outstanding applications in different industries is schematically shown in Figure 1.1.
This breakthrough of accepting biocatalysis as a good alternative to classical organic chemistry has not been straightforward, and is actually still in growth. There have been several critics focused on the drawbacks of enzymes that are somewhat still considered to be important. The most disputed critics and the suggested answers are listed below:

- **enzymes only feature limited substrate specificity**- this is true for enzymes that evolved to convert small molecules such as hydrogen peroxide or urea and for very large enzymes acting as multi-enzyme complexes such as fatty acid synthase complex. Other, ordinary-sized enzymes working on medium-sized substrates in most cases feature broad substrate specificity.
- **there is only limited availability of enzymes**- until recently this was indeed a problem. However, recombinant DNA technology, discovered in 1978 by Cohen and Boyer, over the next 20 years allowed enzymes to be produced much more efficiently, in higher purity, and more inexpensively, so that today a multitude of enzymes are commercially available from suppliers.
• **only a limited number of enzymes exist**- this could be an actual drawback, depending on the point of view. Nature's variety of enzyme is surely big enough to cover every conceivable organic chemical reaction. It is the human knowledge and ability to use them that is limited. Moreover, there are enzymes sought for many more reactions than there are enzymes available. New enzymes are constantly being discovered as well as engineered through rational and random mutagenesis.

• **protein catalyst stability is limited**- one of the major drawbacks of enzymes. They commonly require temperatures around ambient to perform (15–50°C), pH values around neutral (pH 5–9) and aqueous media. Under almost any condition, native proteins, with their Gibbs free enthalpy of stability of just a few kilojoules per mole, are never far away from instability.

• **enzyme reactions are saddled with limited space–time yield**- the notion that biocatalysts are slow catalysts is false. Optimized syntheses not only produce very good selectivities or total turnover numbers but also satisfactory to excellent space–time yields. The question of high volumetric productivity is coupled to the solubility of substrates. High space–time yields have been demonstrated to be correlated with high solubility of substrates.

• **enzymes require complicated co-substrates such as cofactors**- requirements for cofactors constitute a technological challenge but one that has been met successfully and so, should not be regarded as impeding for the use of biocatalysts in processing. There have been examples demonstrating the feasibility of industrial processing with cofactor-requiring enzymes such as the scale-up of l-aspartate decarboxylation to l-alanine with the help of PLP-requiring l-aspartate decarboxylase. Moreover, cofactors are no longer the dominating cost component, as was believed until recently.

There are some actual current drawbacks of today's biocatalysis. It is still a rather new technology that obviously due to its young nature brings along some essential unsolved drawbacks:

1. Biocatalysts are often not sufficiently stable in the desired media
2. Too few biocatalysts exist for the desired reactions from available substrates to targeted products
3. Development cycles are too long for new and improved biocatalysts

ad 1) Biocatalysts are often not sufficiently stable in the desired media. As mentioned above, this still is an essential drawback of biocatalysts. As even conformational changes of less than a few Ångströms can cause a precipitous decline in activity, retention of activity is a stringent criterion for the integrity of a protein molecule. Enzymes deactivate under a range of conditions such as extremes of temperature or pH value, physical forces such as cavitation by pumps and aqueous–organic or gas-liquid interfaces, or specific covalent interactions.

ad 2) Too few biocatalysts exist for the desired reactions from available substrates to targeted products. This argument can be approached from two vantage points. There are biocatalysts for almost any reaction. However, most biocatalysts (there are now more than 4000 known) are either not well characterized, or proprietary, or at least not commercially available. The situation, though, improves steadily: just fifteen years ago, only about a dozen enzymes were available commercially, whereas nowadays the number has increased more than tenfold. Further rapid progress is to be expected.

ad 3) Development cycles are too long for new and improved biocatalysts. The average duration of the development cycles is around 15 years. One reason for such timelines is the as-yet incomplete knowledge base of biotechnology and biocatalysis. With an improved knowledge base stemming from intensive research efforts, development times will certainly be decreased. Shortening the development cycle time for biocatalysts is therefore a topic of active research. Most of the work of this thesis can be regarded as a contribution to this problem. Finally, compared to other types of catalysis, enzymes feature advantages and disadvantages shown in Table 1.2.
Table 1.2. Advantages and disadvantages of biocatalysts and enzymes.

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very high enantioselectivity</td>
<td>Often low specific activity</td>
</tr>
<tr>
<td>Very high regioselectivity</td>
<td>Instability at extreme T and pH</td>
</tr>
<tr>
<td>Transformation under mild conditions</td>
<td>Availability for selected reactions only</td>
</tr>
<tr>
<td>Solvent often water</td>
<td>Long development times for new enzymes</td>
</tr>
</tbody>
</table>

As far as the advantages of using enzymes for catalyzing chemical reactions are concerned, there are many features that make biocatalysis appealing. Moreover, biocatalysis sometimes offers options and solutions that are not available through any other technology; more often it offers one solution among several others. The choice is then evaluated through several criteria such as yield of product, selectivity, productivity, (bio)catalyst stability and space-time yield. Biocatalysis is certainly the solution allowing to have the following characteristics:

- **environmentally benign manufacturing** - both economically and environmentally healthy processes
- **sustainable development** - non-renewables are minimized as much as possible
- **green chemistry** - chemical products and processes that reduce or eliminate the use and generation of hazardous substances

Biocatalysis combines the goals of all three topics above. Biocatalysts, as well as many of the raw materials, especially those for fermentations, are themselves completely renewable and for the most part do not pose any harm to humans or animals. Through the avoidance of high temperatures and pressures and of large consumptions of metals and organic solvents, the generation of waste per unit of product is drastically reduced.
Finally, the most important property of a catalyst for the application in a process is not its activity but rather its selectivity, followed by its stability, which is just activity integrated over time. In comparison with other catalysts, enzymes often feature superior selectivity, especially regio- and enantioselectivity. Enzymes are destined for selective synthesis of molecules with several similar functional groups or chiral centers. A growing emphasis is laid on the synthesis of enantiomerically pure compounds (EPCs). The interest in EPCs shown by all areas of the life science industries (e.g., pharmaceuticals, food and agriculture), derives from the challenge to develop structurally optimized inhibitors, almost always containing chiral centers. These, biocatalytically obtained EPCs are often advanced pharmaceutical intermediates (APIs) that are of huge importance for drug synthesis.

1.2. Enzyme engineering and overcoming application limitations

Enzymes are produced by living organisms and, as such, adapted to function in conditions that are similar to those the parent organism has evolved for. This can on the one hand be a positive trait, allowing biocatalysed processes to be environmentally friendly, but on the other hand brings along many of the drawbacks typically associated with enzyme use. As previously mentioned, the major problems are the low stability at high temperatures and extreme pH values as well as the unfolding of the protein in some non-aqueous media. This, in some cases, unfavorable features can be overcome or at least alleviated by redesigning a wild-type enzyme by means of genetic engineering, i.e. producing mutants.

Apart from the mentioned problems, which are usually associated to the overall fold and superficial characteristics of the enzyme, another typical example for resorting to enzyme engineering are cases when the active site needs to be modified for some reason, e.g. to increase the enzyme selectivity, increase or change enantioselectivity or increase the reaction rate. All this examples can be the limiting factors for applying a biocatalyst in industrial processes and therefore mutations are an important tool for modifying the natural enzymes, making them suited for a desired scope.

Enzyme engineering, as an example of protein engineering, results in the targeted mutation of one or more specific nucleotides in a gene, which then codes for an altered protein compared to
the wild-type. When starting any protein engineering project, two questions must be asked, “what template should I start with” and “what tools should I use to reach the goal”? The answer to these two questions depends upon the following criteria:

- the depth of accuracy of available sequence and structure information: crystal structures can be accurate up to <1Å allowing the positioning of water, often critical to the reaction mechanism, while many data-driven concepts depend on the availability of as many as possible sequences with the maximum possible level of identity

- the expected involvement of amino acid residues close or distant to the active site: changes in activity or selectivity often involve residues close to/in the active site, whereas improvements of stability in most cases require targeting residues away from the active site

- the library screening capacity: no longer is a large library size the driving attribute for protein development. Nowadays, data-driven protein engineering strives for the smallest possible library of mutants.

The most common engineering goals are thermostability, selectivity and stereoselectivity. There are basically three general strategies used to optimize natural enzymes: purely rational redesign, semi rational (data-driven) design and random mutagenesis and directed evolution. Usually the highest rate of success is hit by a combination of these methods.7

The first solution is when protein engineering results in rational rearrangement or mutation of specific protein sequences on the DNA level. The rational redesign of an enzyme is a knowledge-guided process. Therefore, this approach is only used if detailed structural information on the parental enzyme is available. Ideally, a crystal structure of the enzyme in complex with its substrate, reaction product or inhibitor is available, in conjunction with a detailed understanding of the reaction mechanism. Under these circumstances, crucial amino acid residues in the active site can be targeted by site-directed mutagenesis.

Secondly, changes can be introduced through directed evolution, which mimics the process of Darwinian evolution. In the initial step, genetic diversity is generated by creating a collection of mutants, which are then subjected to a screening (or selection) process to isolate the variants with beneficial traits. This process of diversification and selection is repeated until enzymes with the desired properties are obtained.
If the active site of an enzyme is known, but detailed understanding of the contribution of single amino acids in unknown, semi rational redesign can be a successful strategy. In this case, several or all residues that form the active site are varied by saturation mutagenesis. Depending on how many mutants can be screened, mutagenesis of the second shell residues neighboring the active site can also be considered, but the risk of encountering the bottleneck of screening too many variants is high. To alleviate this problem, several cleaver procedures have been developed to reduce the library size by increasing the probability of beneficial mutants in a given library.

Computational de novo design is probably the most remarkable kind of engineering given its ability to alter an enzyme's active site and endow it with a completely new catalytic function. The first example of de novo design dates to 2008 when Jiang et al. developed retro-aldol enzymes able to break carbon-carbon bonds of non natural substrates.

There are several ways to approach an enzyme engineering problem and recent advances in the development of both experimental and computational enzyme engineering tools enabled a number of further successes in the development of biocatalysts ready for large scale applications. Key tool are first, the targeting of mutant libraries, leading to far smaller but more useful libraries than in the past, second, the combination of structural, mechanistic and sequence-based knowledge often based on prior successful cases, and third, the advent of structurally based algorithms allowing the design of novel functions.

A description of individual tools follows:

**Iterative saturation mutagenesis (ISM):** has become known as an essential tool in designing novel biocatalysts. The saturation of individual residues has been proven to create novel functionality, enantioselectivity and thermostability. The incremental accumulation of single site mutations likely mirrors the evolutionary path of nature. Nevertheless, this stepwise procedure is not the most direct method and will likely miss any synergistic effects from neighboring residues.

**Combinatorial active site saturation testing (CASTing):** another powerful tool for developing more active, thermostable, or selective mutants. CASTing's usefulness lies in its ability to test the synergistic effects of saturating small portions of the protein's active site. However, CASTing libraries can easily exceed screening capacity because of the exponential growth of the screening
requirement of each additional mutation. CASTing is best applied when activity toward the desired substrate already exists at some level, otherwise low levels of activity may be missed depending on the precision of the screening assay. Despite their low activity, these transitional mutants are essential in the iterative process to a successful catalyst. The temperature value (B-FIT) approach, a variation of CASTing, targets amino acid positions based on high crystallographic B-factors indicating high flexibility associated with positions in the crystal structure, with the goal of improved thermostability.

**Restricted libraries:** one of the simplest ways to improve screening efficiency is the appropriate application of degenerate codons, which can be used to limit mutation of target sites to a desirable group of amino acids. Selection of the allowed mutation can be restricted based on rational design or broad mutation to efficient codons, such as NDT. This method removes the redundancy of analogous codons seen in conventional codon degeneracy (20 amino acids/32 codons). Further restrictions of degenerate codons can be used to reduce library size even further.\(^\text{13}\)

**Structure-guided consensus:** combines sequence based and structural data and employs a set of phylogenetically diverse but functionally proven proteins. The resulting alignment of sequences often has a low level of sequence identity (40-80%).\(^\text{14}\) From this alignment mutational points of interest are limited to those positions that match the desired consensus cut-off. Further construction of the remaining residues is achieved via structural criteria, such as keeping a distance from the active site, introducing helix stabilizers and not braking salt bridge formation.

**SCHEMA:** a computational algorithm which estimates the amount of disruption caused upon DNA recombination. Libraries of chimeras are scored (E = average disruption) according to the number of residue-residue contacts which broken when compared to the parent proteins. Screening can then be focused to those chimeras with less structural disruption (low E). This algorithm ultimately results in an enhanced probability of functional chimeras with relatively low identity to the parent sequences, effectively generating a family of diverse, folded and functional proteins.\(^\text{15}\) Additionally, many nonfunctional chimeras with low E have regained activity by low error-rate random mutagenesis.\(^\text{16}\)
**ProSAR:** developed by Fox *et al.*, a strategy for improving recombination based directed evolution by applying statistical analysis of PROtein Sequence Activity Relationships (ProSARs). After initially screening a diverse group of mutants for activity towards the desired function, this data is applied to the ProSAR model as a training set. The ProSAR model sorts mutations based on their predicted effect on activity. Once beneficial mutations are identified, they are fixed for subsequent rounds. After each iteration, the resulting mutants showed constant improvement over the parental protein.

**Rosetta:** currently the most rational of the computational techniques. Rosetta applies the quantum mechanical and molecular mechanical design of the active site around a reaction's transition state to create a novel biocatalyst from an existing template structure. This method has been shown to create previously nonexistent Kemp eliminase and Diels-Alderase enzymes. Despite exhibiting relatively low catalytic activity, they achieve some level of base activity to be improved upon. As protein sequence-function knowledge advances to incorporate further interactions, purely rational techniques such as Rosetta will create increasingly robust enzymes.

Three historically successful examples of enzyme engineering are described below, whereas a selection of more recent achieved results is shown in Table 1.3.

- The serine protease *subtilisin* is used in bulk amounts in laundry detergents for its broad specificity and high alkaline stability. The oxidative bleaching force of hydrogen peroxide, another laundry detergent component, deactivates native subtilisin by oxidizing Met222, immediately adjacent to the active-site Ser221, to sulfoxide or sulfone. Site-specific mutagenesis of Met222 into alanine (M222A) or serine (M222S) resulted in remarkably stabilized enzyme, expressed as residual activity in 2 M H$_2$O$_2$ solution after 1 h.

- Structurally and sequentially related NAD(H)-dependent lactate and malate dehydrogenase, LDH and MDH, have been revealed to contain potential amino acid targets for site-specific mutagenesis that will change the substrate binding pocket without affecting their identical catalytic mechanisms. Exchange of a crucial amino acid within
the substrate binding pocket (Q102R) resulted in a highly specific catalyst for the new substrate oxaloacetate.\textsuperscript{20}

- Using \textit{subtilisin BPN}' six individual amino acid substitutions (N218S, G169A, Y217K, M50F, Q206C, N76D) at separate positions in the tertiary structure were found to increase the stability of the enzyme at elevated temperatures (65°C) and extreme alkalinity (pH 12).\textsuperscript{21} Under these denaturing conditions, the rate of deactivation of the six-fold variant was 300 times slower than that of the wild-type subtilisin BPN'. An additional disulfide bond linkage through site-directed mutagenesis between residues Cys61 and Cys98 of subtilisin E resulted in increased thermostability.\textsuperscript{22} A single amino acid change, Ser236Cys, in subtilisin E resulted in the forming of an intermolecular disulfide bridge between two subtilisin E molecules and also in enhanced thermostability at 60°C.\textsuperscript{23}
Table 1.3. Selected results of protein engineering in biocatalysis.

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Engineering goal</th>
<th>Tools</th>
<th>Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATA-117 D-amino acid transferase</td>
<td>Overall enzyme redesign, substrate specificity,</td>
<td>Targeted restricted codon library around</td>
<td>200g/L substrate converted with &gt;99.5% ee, 92% assay yield obtained in ~24h</td>
</tr>
<tr>
<td></td>
<td>thermostability, organostability</td>
<td>active site, ProSAR, directed evolution</td>
<td></td>
</tr>
<tr>
<td>Halohydrin dehalogenase</td>
<td>Activity</td>
<td>ProSAR</td>
<td>Volumetric productivity: 1000g L⁻¹ d⁻¹ (4000 fold improvement)</td>
</tr>
<tr>
<td>Ketoreductase</td>
<td>Organostability, enantioselectivity</td>
<td>ProSAR</td>
<td>&gt;100g/L substrate converted with &gt;99% ee (initially 63%) &gt;99% assay yield obtained in ~24h</td>
</tr>
<tr>
<td>Glucose dehydrogenase</td>
<td>Thermostability, organostability</td>
<td>Structure-guided consensus</td>
<td>T₁/₂ &gt; 3.5 days at 65°C (10^6 fold improvement)</td>
</tr>
<tr>
<td>Ketoreductase/glucose dehydrogenase</td>
<td>Activity and stability</td>
<td>Directed evolution, DNA shuffling</td>
<td>160 g/L substrate converted with &gt;99.9 ee, 95% isolated yield obtained in ~8h</td>
</tr>
<tr>
<td>Kemp eliminase</td>
<td>Novel catalysis</td>
<td>De novo design via Rosetta, directed evolution</td>
<td>kₜₐₜ= 1.375 s⁻¹, kₜₐₜ/Kₐ= 2590 M⁻¹ s⁻¹</td>
</tr>
<tr>
<td>Nucleoside kinase</td>
<td>Altered substrate specificity</td>
<td>Rosetta modelling</td>
<td>Relative specificity (kₜₐₜ/Kₐ)ddT/(kₜₐₜ/Kₐ)T=8.5 (8500 fold improvement)</td>
</tr>
<tr>
<td>Cytochrome P450 (P450-BM)</td>
<td>Regioselectivity</td>
<td>Site saturation mutagenesis based on docking</td>
<td>Indirubin/indigo ratio=9 (51 fold improvement)</td>
</tr>
<tr>
<td>Epoxide hydrolase</td>
<td>Enantioselectivity</td>
<td>Iterative CASTing</td>
<td>E=115 (25 fold improvement)</td>
</tr>
<tr>
<td>Cellobiohydrolase class II</td>
<td>Improved thermostability</td>
<td>SCHEMA</td>
<td>T₁/₂ ~3200 min at 63°C (34 fold improvement)</td>
</tr>
<tr>
<td>Diels-Alderase</td>
<td>Novel catalysis</td>
<td>De novo design with Rosetta</td>
<td>Kₜₐₜ= 0.036 s⁻¹, kₜₐₜ/Kₐ= 0.455 M⁻¹ s⁻¹</td>
</tr>
<tr>
<td>Cellobiohydrolase class I</td>
<td>Improved thermostability</td>
<td>SCHEMA</td>
<td>T₅₀ = 65.7 °C increased 3.4°C compared to parent</td>
</tr>
</tbody>
</table>
1.3. Computational approaches

Ever since they were first recognized, the outstanding catalytic properties of enzymes have inspired awe and envy in chemists. Understanding how enzymes “work” (how they achieve great rate accelerations, efficiency and specificity, under typically mild conditions) is a fundamentally important question in biology and biocatalysis and will also have practical benefits. Understanding enzyme catalytic mechanisms at the atomic level could contribute to a range of technological applications such as the design of inhibitors as pharmaceutical leads, prediction of drug metabolism, and design of catalysts for specific transformations. Determination of enzyme mechanisms has proved difficult in many cases, as differentiation between alternative possibilities is often challenging. There is evidence to suggest that several “textbook” mechanisms are incorrect in important details; hen egg-white lysozyme, for example.24

There have been many controversial proposals seeking to explain catalysis in some or all enzymes; arguments have centered on the possible role of ‘low-barrier’ hydrogen bonds,25 so-called “near-attack conformations”,26 enzyme dynamics,27 enzyme quantum tunnelling,28 and entropic effects.29 The applicability of transition state theory to enzyme reactions has also been questioned. The complexity of enzymes, and the difficulty of studying reactions in them, makes these arguments difficult to resolve through experiments alone.

Modelling and simulation have the potential to give detailed, atomic-level analysis of enzyme mechanisms and catalysis, and of questions such as the effects of mutations and of how chemical and conformational changes are coupled in enzymes. A range of computational methods is now available for the modelling of enzymes and their reactions, complementing experimental approaches. Techniques for modelling enzyme mechanisms and conformational changes have advanced significantly, and are now making an increasingly important practical contribution to enzymology, and biotechnology more widely.30

Molecular modelling and simulation will be, and are proving to be, central to resolving enzyme-related debates, both in the interpretation of experimental data and in providing atomic-level analysis of reactions in enzymes, with correlated engineering of reactions. Molecular simulations can probe enzyme mechanisms, and the origins of catalysis, at a level of detail that cannot, at least for the foreseeable future, be achieved experimentally. Modelling can identify probable chemical mechanisms, analyse the effects of mutations and genetic variations, pinpoint the causes of specificity, and derive structure–activity relationships.
Computational enzymology began in the early to mid-1970s. By the early 1990s, the number of computational mechanistic studies of enzymes was relatively small, but recent years have seen an explosion in the number of computational studies of enzymatic reaction mechanisms. There has also been a transformation in the accuracy of the computational methods. It is now possible to achieve an unprecedented level of accuracy in calculations on enzyme-catalysed reactions with combined quantum mechanics/molecular mechanics (QM/MM) methods. High-level quantum chemical methods can now be used to study enzymatic reactions allowing calculations of energy barriers for enzyme-catalysed reactions, in the best cases, of near “chemical accuracy” (1 kcal/mol). Previously, quantitative predictions at this level were only possible for very small molecules. Moreover, carefully parameterized empirical molecular simulation approaches also give excellent agreement with experiments for enzymatic reactions. In the best cases, calculations can give activation energies that agree very well with experiments. Such good agreement indicates that transition state theory provides a good general framework for understanding the rates of enzyme-catalysed reactions.

An enzyme structure from X-ray crystallography is usually the starting point for modelling enzyme-catalysed reactions or exploring enzymatic conformational changes. In some cases a model constructed based on homology to other structures may be sufficiently reliable, though such models should be treated with much more caution. The first step if studying an enzyme-catalysed reaction is to establish its chemical mechanism. This is far from trivial: as noted above, many ‘textbook’ mechanisms are probably wrong. An initial goal is to determine the functions of catalytic residues, which are often not obvious. Even the identities of the important groups may not be certain. Any specific interactions that stabilize transition states or reactive intermediates (tetrahedral intermediates) should also be identified and analysed. Calculations offer several unique advantages: for example, they can analyse transition states directly. Transition states are central to understanding chemical reactivity and catalysis, but experiments cannot directly study them in enzymes because of their extremely short lifetimes, and because of the large size and complexity of enzymes. Calculations can also identify functional groups and interactions involved in catalysis. Several examples have been published of key catalytic interactions that have been identified by modelling. Functionally important interactions of this type may not be well apparent from experimental structures: for example, they may not exist in structures that can be isolated and crystallized. In addition to providing detailed
understanding of the reaction in the enzyme, identifying interactions of this type may assist ligand design: such interactions potentially offer enhanced affinity if they can be exploited in designed ligands (e.g. pharmaceutical lead compounds), because many enzymes show exceptionally high apparent binding affinities for transition states and reaction intermediates.

Protein dynamics are complex and fascinating. As molecular dynamics simulations and experiments (e.g. NMR) have shown, proteins undergo a wide range of internal motions, some of which are central to their biological function. Many enzymes show large conformational changes during their reaction cycles, and the functions and relationship of these changes to the chemical steps in the reaction are questions of considerable current interest. It has been suggested that protein dynamics may contribute to enzyme catalysis, but simulations indicate that the direct effect of protein dynamics in determining the rates of chemical reactions in enzymes is generally relatively small. On the other hand, protein conformational changes (e.g. involved with substrate binding or product release) may be the rate-limiting step for the overall reaction in many enzymes. It is certainly important to consider the possible effects of protein conformational fluctuations and variations when modelling enzyme reactions, i.e. to consider a representative sample of possible conformations, e.g. through molecular dynamics or Monte Carlo simulations.

In practical terms, often the interest relies in the ability to predict the effects of a mutation on activity, or on the specificity of an enzyme for alternative substrates. Overall, understanding an enzyme mechanism, specificity and catalysis involves many different levels of complexity. This presents a variety of different challenges, and different types of modelling or simulation methods are needed to investigate different types of question.

Enzymes are large molecules, thus modelling the reactions that they catalyze is complex and challenging. This can be complicated further by the need to include part of a particular enzyme’s surrounding environment, such as the surrounding solvent, cofactors, other proteins, a lipid membrane, or DNA. There are many practical considerations in simulating such complex systems, such as interpretation of crystal structures, choice of protonation states for ionizable amino acids etc.

An overview of available simulation methods follows.
Quantum chemical calculations on small (active sites) models

Quantum chemical calculations (i.e. methods that calculate molecular electronic structure, for example *ab initio* molecular orbital or density-functional theory calculations) can provide excellent results for reactions of small molecules. The major problem in applying these methods to model enzyme reactions is that such calculations require very large computational resources (which increase significantly with system size), placing severe practical limitations on the size of the system that can be treated. In most enzymes, however, the chemical changes occurring in the reaction are confined to a relatively small region, the active site of the enzyme. One approach to the study of enzyme-catalysed reactions is therefore to study just the relatively small active site region. Studying small models in this way allows the use of powerful (and potentially highly accurate) quantum chemical methods.

Small “cluster” active site models of up to around 100 atoms (sometimes referred to as the “supermolecule” approach), can represent important features of an enzyme reaction, and identify probable mechanisms. The active site model should contain molecules representing the substrate(s) (and cofactors in cases where they are involved), and protein residues involved in the chemical reaction or binding substrate. The interactions that bind substrates into the active site of an enzyme are typically weak interactions, such as hydrogen bonds, electrostatic and van der Waals interactions. Important functional groups (for example catalytic amino acid side chains) are represented by small molecules. For example, imidazole may be used to represent histidine, acetate to represent an aspartate side chain, and so on. The initial positions of these groups would usually be the coordinates extracted from a representative X-ray crystal structure of an enzyme complex, or perhaps from a molecular dynamics simulation of such a (e.g. enzyme-substrate or enzyme-intermediate) complex.

Calculations on active site models can examine interactions between groups at the active site, and can provide useful models of transition states and intermediates. They can also be useful to test the accuracy of different levels of calculations for a reaction. This approach has shown itself to be particularly useful for studying the reaction mechanisms of metalloenzymes, for which reliable, semi-quantitatively accurate calculations have been made feasible by the development of methods based on density functional theory.

To calculate the energy barrier of a reaction in a cluster model, it is necessary to locate the structures (optimize the geometries) of the reactant, transition state, intermediates and products of the reaction. Technically, it can be difficult to optimize the geometry of the model (for example to
find a transition state structure), while at the same time maintain the correct orientations of the group in the protein. Small molecules can also lack important functional groups and careful thought must be given as to which groups to include, balancing between computational feasibility and the desire for a larger, more extensive model. It is possible to include the effects of the protein and the solvent environment approximately using continuum solvation models, but these cannot fully represent the heterogeneous electrostatic environment in an enzyme.

**Combined quantum mechanics/molecular mechanics (QM/MM) methods**

Combined quantum mechanics/molecular mechanics (QM/MM) methods allow the modelling of reactions within enzymes by combining the power and flexibility of a quantum chemical method with the simplicity of molecular mechanics. The fundamental basis behind the QM/MM approach is simple: a small part of the system is treated quantum mechanically, i.e. by an electronic structure method, for example at the *ab initio* or semiempirical molecular orbital, or density-functional theory QM level. The QM treatment allows the electronic rearrangements involved in the breaking and making of chemical bonds to be modelled. The QM region in a study of an enzymatic reaction mechanism is the enzyme active site, i.e. the reacting groups of the enzyme, substrate and any cofactors. The large non-reactive part is described more simply by empirical molecular mechanics. Different types of coupling between the QM and MM regions are possible. For applications to enzymes, which are polar, it is important to include the interactions between the QM and MM regions. Modern molecular mechanics methods give a good description of protein structure and interactions, so can ensure that these are treated accurately. QM/MM calculations can be carried out at *ab initio* or semiempirical molecular orbital, density-functional or approximate density functional levels of QM electronic structure calculation.

The first enzyme to be studied with QM/MM methods was hen egg-white lysozyme, in a seminal study by Warshel and Levitt in the mid 1970s. Interest in QM/MM methods has grown rapidly in recent years. Following many recent developments and applications, it is now clear that QM/MM calculations can provide useful insight into the mechanisms of enzyme-catalysed reactions. Examples include identifying functions of active site residues investigations of mechanistic questions and analysing catalytic contributions such as the roles of conformational effects and transition state stabilization. Modelling and simulations have been crucial in many cases in formulating and testing mechanisms and hypothesis.
Current extensions in the scope of the use of QM/MM methods include free energy perturbation simulations, e.g. to calculate relative binding affinities and in molecular docking and scoring of binding affinities. QM/MM methods provide several advantages over molecular mechanics methods in studies of ligands bound to proteins, including potentially a better physical description of a ligand and avoiding time consuming MM parameterization of a ligand.

One of the main differences between various QM/MM methods is the type of QM/MM coupling employed \textit{i.e.} in how the interactions (if any) between the QM and MM systems are treated. The simplest linking of QM and MM methods involves a straightforward ‘mechanical’ embedding of the QM region in the MM environment, where the interactions between the QM and MM regions are treated purely classically by MM.

\textbf{Molecular mechanics (MM) methods}

Enzyme-substrate complexes typically contain thousands of atoms, perhaps tens of thousands, particularly when modelled using an explicit representation of surrounding solvent. This puts them currently beyond quantum mechanical or even semiempirical methods for modelling reactions. An equally important consideration in reaction modelling is that important structures (such as transition state structures) and preferably entire reaction pathways should be optimized and simulated. The environment of the enzyme should also be considered; it can be an aqueous solution, organic solvent(s), cell membranes and protein or nucleic acid complexes.

Proteins have many conformations given the huge number of rotational bonds present, and a single structure may not be truly representative. Extensive conformational sampling may be needed to generate a representative ensemble of structures. These are all significant challenges for large structures. For conformational sampling, a useful simulation method must be capable of calculating trajectories of at least a few nanoseconds. Molecular mechanics that treat the arrangement of atoms in a molecule as balls connected by springs, basically applying Hook’s law and standard force fields with empirically derived parameters is the right choice in such cases. The force field is based on classical mechanics. It relates the geometry and the potential energy of a molecule by means of an analytical function. Despite the apparent simplistic treatment, this procedure still allows a fairly accurate description of the three-dimensional structures of molecules even with the size of ribosomes. Molecular dynamics simulations with MM methods can be used to generate several structural models for mechanism calculation, to ensure a wide sampling of possible enzyme configurations. If multiple crystal structures of the same enzyme
are available, these may be suitable as different starting models and similarly help to examine the effects of structural variations on the reaction. Therefore molecular mechanics methods are hardly appropriate for simulating reaction pathways. Useful information can be derived, but the strength of these methods relies in the fact that they are relatively cheap in terms of computational power requested and can nowadays be used to simulate tens of nanoseconds long trajectories. In simple enzymes such lengths are not even necessary to sample the structure's conformational freedom, whereas in more complex enzymes that undergo a significant range of conformational changes molecular dynamics simulations can reveal distinct configurations e.g. associated with the binding and the release of the substrate. As stated, with MM treatment the simulated system can be much bigger. It can include the solvent in which an enzyme is experimentally used, other proteins in case of multi-protein complexes, phospholipid bilayers, DNA or RNA molecules etc. It is therefore the best choice for studying superficial characteristics, substrate or environment related conformational changes (induced fitting, activation/inactivation phenomena) and associated interactions, either as the explanation of experimental data or pure theoretical research.
**Introduction**

**Aim of the thesis**

This thesis aims at developing new computational tool for rational *in silico* screening and engineering of enzymes and enzyme properties. In particular, the focus has been given to substrate and reaction promiscuity. In the first case, the substrate recognition at molecular level has been studied and case-specific molecular descriptors have been developed. The use of such descriptors has the objective to quantify and describe in detail the enzyme-substrate recognition pathways and point out possible unwanted interactions. Based on the results, the descriptor can serve as guidelines for rational redesign of wild-type enzymes and production of mutants with broadened substrate specificities.

The reaction promiscuity is a more complex and ambitious task as it involves the implementation of novel catalytic properties in enzymes naturally evolved to perform specific types of reactions. The *in silico* approach necessary to tackle this kind of problematics is therefore inevitably more complex as well. Many different enzymes with different catalytic properties (hydroxynitrile lyases, esterases, lipases, amidases, proteases) have been analysed to gain knowledge about their mechanistic and structural features. For this purpose several well established methods have been used (sequence alignment, structural superposition, GRID-PCA, 3D-QSAR), and in some cases upgraded, and a novel highly tunable tool has been developed. Additionally, mutations have been proposed for the purpose of redesigning the enzymes catalytic machinery and obtaining biocatalysts with the ability to catalyse new reactions.

The topics of this thesis are of high relevance as the developed computational tools can serve to overcome existing bottlenecks that limit the application of biocatalysis in industries. Many existing and known enzymes could be tuned and optimized for a specific purpose once reliable computational tools exist. The work performed during the three years of this doctorate goes beyond the state of the art of modern molecular modelling both in terms of upgrading of the specific computational tool and in terms of integration of methodologies.
Introduction


23
42 A. Fersht, Structure and Mechanism in Protein Science, New York: Freeman; 1999
2. Differential RMSD as a novel descriptor for rational redesign of natural enzymes in case of bulky substrates

2.1. Summary

Polylactides are important members of the polyester family and are usually synthesised via polycondensation of lactic acid monomers or ring opening polymerization of lactides using metal catalysts. Enzymatic catalysis for the formation of polylactides is still in its early stage. No enzyme have yet shown sufficiently high activity in catalysing this type of reactions and industrial production of polylactides still relies on chemical catalysis. A moderate activity of Candida antarctica lipase B (CALB) towards the ring opening polymerization of DD-lactide has been described in the literature reaching 33% of conversion achieved after 3 days under specific conditions. There can be various reasons why the reaction does not show better, satisfying yields. Besides the problems related to the process, one of the motives can be surely due to the large size of the substrates, especially after the first few monomers have reacted. CALB is known to have a rather small active site, as compared to other lipases, which makes it particularly suitable to hydrolyse esters of small to medium sized fatty acids. Therefore, a molecular modelling study was performed to understand the interactions that are established when the enzyme interacts with the bulky substrates. The study was not restricted only to molecular visualization and conceptualization; it also consisted in molecular dynamics simulations and development of a new class of molecular descriptors based on the simulation itself. The descriptors derive from the root mean square deviation (RMSD) values and were designed to be used as a quantitative description of the effect a given substrate has on the enzyme structure. Thus, their application can have different advantages and can be used to elucidate different kinds of phenomena, from mutation hot spot identification to induced fitting description. The efficiency of the descriptors increases if the monitored effect is of large proportions and results in a consistent modification of enzyme's structural features, such as its secondary or tertiary structure.

In the present chapter, these descriptors have been used to identify which CALB's residues are mostly affected by three different substrates during time as a consequence of unfavorable
interactions such as steric clashes or repulsive electrostatic forces. The substrates were built as a tetrahedral intermediate in the enzyme active site and the systems were evolved in a molecular dynamics simulation. The resulting structures were compared with the empty enzyme (no substrate in the active site) by means of differential RMSDs to emphasize the substrate effect. The outcomes were different histogram like plots indicating the residues and the related RMSD in respect to the empty enzyme. These results were used to identify amino acids that are mainly distorted by the substrates and consequently their respective positions in the gene sequence can be used as possible hot spots for in vitro mutagenesis experiments.
2.2. Introduction

2.2.1. Lipase catalysed synthesis of polyesters

Polyesters are important polymeric materials that find large applications in our everyday life, ranging from different type of clothing to pharmaceutical and fine chemistry products, from bottles for carbonated soft drinks and water, to the base for photographic film and recording tape. In addition, most of the biodegradable polymers employed in medical applications belong to the polyester family.

In the past, the term polyester referred to polymers derived essentially from diols and dicarboxylic acids. Earliest reports of polyester resins of this type include those from Berzelius, who documented resins from tartaric acid and glycerol, Berthelot, who produced a resin from glycerol and camphoric acid, and Van Bemmelen, who synthesised glycerides of succinic acid and citric acid. Back in 1901, Watson Smith had already described the reaction product of glycerol and phthalic anhydride. In 1924, Kienle and Hovey began to study the kinetics of polyesterification reactions between glycerol and phthalic anhydride. Carothers’ pioneering studies were based on aliphatic polyesters and culminated in laying the foundations for condensation and step-growth polymerization. Since then, many research groups have investigated this group of polymers, broadening fundamental studies and working towards developing commercial products.

In recent years, environmental concerns have led to a renewed interest in biodegradable polyesters as an alternative to commodity plastics. Environmental pollution caused by production and disposal of petrochemical-derived plastics have led to the pursuit of alternative approaches using environmentally benign processes to synthesise plastics that are engineered to degrade on demand. The advantage that polyesters have over other types of polymers is the presence of scissible ester bonds. Since ester bonds are frequently encountered in nature it is reasonable to assume that at least a subset of the polyester family will be degradable and environmentally friendly.

In nature, various polymers are constantly being produced in living organisms for their normal metabolic needs. These macromolecules, such as polysaccharides, polynucleotides (DNA and
RNA), proteins, or polyesters, are essential to organism survival. Their synthesis involves enzyme catalysed chain-growth polymerization reactions of activated monomers, which are generally formed within the cells by complex metabolic processes. Following nature's example, enzymatic polymerization reactions gained attention and started to be a promising strategy to develop environmentally benign processes for polyester synthesis. A variety of \textit{in vitro} polyester synthesis reactions have been developed in the last couple of decades.

Okumara \textit{et al.}\textsuperscript{9} were the first to attempt the enzyme-catalysed synthesis of oligoesters from a reaction between dicarboxylic acids and diols. Gutman \textit{et al.}\textsuperscript{10} reported the first study on polyester synthesis by enzyme-catalysed polymerization of A-B type monomers. Two independent groups in 1993\textsuperscript{11,12} were the first ones to report enzyme-catalysed ring opening polymerization (ROP).

Among enzymes, lipases proved to be the most efficient for the \textit{in vitro} polyester synthesis. Lipases are ubiquitous enzymes of considerable physiological significance and perform crucial roles in the digestion, transport and processing of dietary lipids in most living organisms. Thus, lipases can be found in diverse sources, such as plants and animals. More abundantly, they are found in bacteria, fungi and yeasts. Lipases are water-soluble enzymes that catalyse the hydrolysis of ester bonds in water-insoluble, lipid substrates such as triglycerides. On the other hand, in low water concentrations ester hydrolysis is reversible and lipases can be used to catalyse ester synthesis and transesterification reactions, opening up the possibility to synthesise polyesters.

Lipases have several characteristics. As a catalyst for polymerization, they are enantioselective, regioselective and chemoselective, and can be used under mild reaction conditions. The catalyst activity is achieved in pure organic solvents, in supercritical carbon dioxide and even in ionic liquids, at a wide range of temperatures. They are non-toxic catalysts and are often renewable. Considering these characteristics, lipase polymerization catalysis has the potential to contribute to the environmental problems and to maintain a sustainable society conducting green polymer chemistry.

Due to these, as well as other advantages, enzymatic polymerization is considered to be a polymer synthesis method with the potential to realize i) precise control of polymer structures, ii) creation of polymers with new structures, iii) a clean process (without by-products), iv) a low-loading process (energy saving), and v) biodegradable properties of many product polymers.\textsuperscript{13} The structural features of polyesters are obviously dependent on the type of monomer used and
the application of lipases in polymer synthesis expands the array of available products and materials given the intrinsic property of substrate selectivity shown by these enzymes. In this respect it is important to underline that the application of lipases in polymer chemistry does not necessarily aim at the achievement of products with high molecular weight, but also at the synthesis of structural oligomers to be exploited as building blocks via chemical polymerization. This opens the possibility to produce innovative materials, inaccessible to classical chemical synthesis.

Nevertheless, monomers for enzymatic polymerization need to be recognized and activated by the enzyme through the generation of the enzyme-monomer complex (acyl-enzyme) for the polymerization to proceed. Therefore, one of the approaches for enzymatic polymerization is the design of an artificial monomer with the structure required by the enzyme.

In the past, several monomers have been specifically designed and combined with a correctly selected enzyme by considering its function. In this way many polymers have been synthesised including polysaccharides, polyesters, polyaromatics and vinyl polymers. Moreover, some of these polymers had not been successfully synthesised via conventional polymerization and the application of lipases enabled their \textit{in vitro} synthesis for the first time. Some examples are cellulose, chitin, xylan, hyaluronan, chondrotin, etc.\textsuperscript{14,15,16}

On the other hand, the enzyme's active site can be modified through enzyme engineering techniques for a better recognition of the substrate. The present chapter describes an innovative, molecular modelling based approach used to understand the substrate effect on the enzyme (\textit{Candida antarctica} lipase B) conformation and explain the low rates of different reactions. This approach allows to identify amino acid residues that could be targets for mutagenesis in order to achieve better substrate recognition.

From the mechanistic point of view, the \textit{in vitro} polyester synthesis can proceed via two major polymerization modes (Figure 2.1):

1) ring-opening polymerization (ROP) of lactones
2) polycondensation between a carboxyl group and an alcohol group.

Both methods have been very extensively studied.
As previously stated, lipase-catalysed ROP was first found in 1993 when Knani et al and Kobayashi et al independently reported on ring opening polymerization of ε-caprolactone and δ-valerolactone (δ-VL), respectively. Since then a variety of lipases from different sources have been screened for the activity of lactone monomers ROP and some examples are *Pseudomonas fluorescens* lipase, *Candida cylindracea* lipase and porcine pancreas lipase which are all active for lactone ROP. *C. antarctica* lipase B was found to be the most effective for this specific polymerization. Even if different substrates have shown a good activity in the lipase catalysed ROP, ε-caprolactone is still the most extensively studied of various lactone monomers.

Ring opening polymerization is an important route for polymer synthesis, since leaving groups which generally can limit monomer conversion or degree of polymerization, are not generated during the reaction. A mechanism for the lipase-catalysed ROP of lactones (cyclic esters) is postulated by considering the principal reaction course involving an acyl-enzyme intermediate. The hydrolysis of esters by these enzymes is generally accepted to proceed via a similar intermediate. The catalytic site of a lipase is known to be formed by the sidechain of a serine
The key step is the reaction of ring opening of a lacton to give an acyl-enzyme intermediate. The initiation is a nucleophilic attack of a nucleophile, such as water or an alcohol, which is present in the reaction mixture, on the acyl carbon of the intermediate. In the propagation stage the acyl-enzyme is nucleophilically attacked by the terminal hydroxyl group of a propagating chain end to produce a one monomer elongated polymer chain, Figure 2.2.

![Postulated mechanism of lipase catalysed ROP of lactones. EM stands for enzyme-monomer complex; Lip stands for lipase.](image)

Ring opening polymerization of non-substituted lactones has been reported for almost all ring sizes, from 4-membered to 17-membered. Additionally, substituted lactones, cyclic carbonates and cyclic diesters have also been demonstrated to be enzymatically polymerized. On the other hand, the condensation polymerization by lipases relies on the fact that most of the isolated lipases can act as catalysts for the reverse reactions of the inherent lipase catalysis, esterifications and transesterifications. Polycondensation involves a much broader substrate range in comparison with the ROP, since not only cyclic monomers can be used. Differently from the
ROP where the employed monomers have the same functionality on both ends and the reaction propagates in a chainwise fashion with a single monomer at time, in the polycondensation there can be different functionalities on monomer ends and the propagation can proceed in both ends of the growing polyester chain which translates into a fast increase in degree of polymerization. This fact results in the ease of using lipases also for the synthesis of tailored and functionalized polyesters which demand is constantly growing in the current society. Utilizing lipase catalysis, functional aliphatic polyesters have been synthesised extensively.

A scheme of the reactions that can be catalysed by these enzymes is shown in Figure 2.3. The reactions are all reversible, and to make the reaction equilibrium shift to the product side and to make the reaction proceed faster, often products like water or an alcohol are removed or reduced in concentrations as the reaction proceeds. All polycondensation reactions basically involve four modes of reaction i.e. esterification, alcoholysis, acidolysis and intermolecular esterification. The former three reactions are examples of transesterification reactions.

![Figure 2.3. Four types of reactions of lipase catalysed polycondensation.](image-url)
2.2.2. Chiral polymers by lipase catalysis

A wide variety of chemical catalysts is nowadays available to polymerize monomers into well-defined polymers and polymer architectures that are applicable in advanced materials. However, synthetic polymers rarely possess well defined stereochemistries in their backbones. This is in sharp contrast with the polymers made by nature where perfect stereocontrol is the norm. An interesting exception is poly(L-lactide), a polyester that is used in a variety of biomedical applications. By simply playing with the stereochemistry of the backbone, properties ranging from a semi-crystalline, high melting polymer (poly(L-lactide)) to an amorphous polymer (poly-meso-lactide) have been achieved.

The chemical synthesis of known chiral polymers mostly starts from optically pure monomers obtained from the chiral pool. The optically pure fermentation product L-lactic acid, for example, is the starting material for the synthesis of poly(L-lactide). However, converting a racemic or achiral monomer quantitatively into a homochiral polymer is less straightforward. This is surprising considering the enormous potential of biocatalysis that has emerged in the past decades to prepare optically active intermediates.

Enzymes are perfectly equipped to convert substrates into products in high enantio-, regio-, or chemoselectivity, a property that is commonly used in industry to prepare optically active fine-chemical intermediates. More specifically, lipases appeared as ideal catalysts as a result of their high enantioselectivity, broad substrate scope and stability. In addition, lipases are powerful catalysts for the preparation of polyesters, polycarbonates and even polyamides.

Lipases belong to the subclass of α/β-hydrolases and their structure and reaction mechanism are well understood. This common enzyme fold is characterized by an α-helix that is connected with a sharp turn, referred to as the nucleophilic elbow, to the middle of an β-sheet array. All lipases possess an identical catalytic triad consisting of an aspartate or glutamate, a histidine and a nucleophilic serine residue. The serine is located at the nucleophilic elbow and is found in the middle of the highly conserved Gly-X1-Ser-X2- Gly sequence in which X1 and X2 are residues that can vary. The histidine residue is spatially located at one side of the serine residue, whereas at the opposite of the serine a negative charge can be stabilized in the so-called oxyanion hole by a series of hydrogen bond interactions.
Although the naturally occurring fatty acid ester substrates of lipases are achiral in nature, lipases can show excellent enantioselectivity in the deacylation step of their reaction mechanism when a chiral center is present in the nucleophile. The chiral nature of the amino acids and the unique three-dimensional spatial organization of the catalytic residues favor one of the nucleophile enantiomers. Especially when the chiral center is close to the nucleophilic site, the degree of enantioselectivity can be high. In these cases, the enantipreference of lipases is well understood and depends on the relative size of the substituents at the chiral center. Generally speaking lipases prefer R-nucleophiles over S-nucleophiles.\(^\text{30}\) The ratio of the reaction rates for the R- and S-enantiomer of a nucleophile (kR/kS) in the deacylation of an achiral acyl-enzyme intermediate is conveniently referred to as the enantiomeric ratio or E-ratio. If the differences in steric demands of the substituents at the chiral center are sufficiently large, high E-ratios can be observed, meaning that only the preferred enantiomer will react. This will result in products that show a high enantiomeric excess (ee). Chiral substrates in which the center of chirality is located more distant from the reactive functional groups can also be recognized by lipases. However, in these cases the preferred enantioselectivity is often lower and more difficult to predict \textit{a priori}.\(^\text{31}\)

The enzymatic polymerization of optically pure DD-, or LL-lactide for the preparation of biocompatible and biodegradable polylactide has received a lot of attention. In 1997 Matsumura et al. reported the first lipase-catalysed preparation of polylactide. The successful polymerization of LL-lactide, DD-lactide and DL-lactide by \textit{Candida cylindracea} lipase, porcine pancreatic lipase and \textit{Pseudomonas cepacia} (PS) lipase in bulk at 80-130°C was reported to result in polymers in a range of \(M_w = 8 – 27\) kg/mol.\(^\text{32}\) A major drawback was the low yields (3-16%) in all experiments. The same group published the copolymerization of LL-lactide, DD-lactide and DL-lactide with trimethylene carbonate in bulk by porcine pancreatic lipase over a temperature range from 80-100°C yielding polymers in a range of \(M_w = 12 – 21\) kg/mol.\(^\text{33}\) Recently PS lipase was used for the synthesis of branched polylactides in bulk at 140°C and for the copolymerization of LL-lactide and DL-lactide with glycolide in bulk at 100-130°C.\(^\text{34}\) In the latter case the authors concluded that the thermal polymerization without the catalyst could not be excluded.

Later on, the copolymerization of L-lactide with glycolide using both free and immobilized \textit{Burkholderia cepacia} lipase (BCL) was investigated. At reaction temperatures of 100–130 °C, copolymers with small amounts of glycolyl repeat units were formed according to careful...
MALDI-TOF-MS analysis. For polymerization at 100°C, Mw and PDI values were higher when using free enzyme (11.7 kg mol, PDI = 1.8) compared to immobilized lipase (8.7 kg mol, PDI = 1.5).

*Candida antarctica* lipase B, which is widely used due to its exceptionally high catalytic activity and versatility for esterification and transesterification reactions, does not seem to be suitable for the enzymatic ring opening polymerization of DL-lactide and LL-lactide. In order to explain the low efficiency of CALB in the polymerization and copolymerization of LL-lactides and DL-lactides, the high enantioselectivity of CALB towards R-configurated secondary alcohols has to be considered. For the lactide monomers with water as initiator, a secondary alcohol end group is formed upon ring opening (Figure 2.4). It can therefore be hypothesized that LL-lactide and DL-lactide are not polymerizable due to the S-configuration of the secondary alcohol which is formed upon the ring opening of the lactide. On the other hand, the secondary alcohol obtained from DD-lactide has R-configuration and is able to react with another enzyme activated monomer and thus to propagate the chain. Polymerization will take place.

**Figure 2.4.** Enzymatic ring opening polymerization of LL-lactide (LLA) and DD-lactide (DLA) by *C. antarctica* lipase B; i) initiation ii) chain propagation step.

In a recent study it has been reported that wild-type *C. antarctica* lipase B has exactly the expected behavior. Hans *et al.* performed the polymerization reactions using Novozyme 435 (CALB immobilized on a macroporous resin) and DD-lactide and LL-lactide as substrates, as shown in Figure 2.4. While no polymer was detected using LL-lactide, 33% conversion was
achieved after 3 days when using the DD-lactide in toluene at 70°C using 12 wt-% enzyme. The proposed mechanism of the reaction is shown in the figure below (fig 2.5).

![Mechanism of the enzymatic ring opening polymerization of DD-lactide](image)

**Figure 2.5.** Mechanism of the enzymatic ring opening polymerization of DD-lactide; the chain growth sequence

To overcome the low yield of the polymerization of DD-lactide, rational enzyme engineering can be used. CALB has already been used for successful synthesis of different polyesters, in bulk or in solution\(^\text{13}\) and has been therefore proven to be a valid biocatalyst for these types of reactions. On the other hand it is also a versatile enzyme when it comes to *in vitro* engineering, since it has already been successfully redesigned for different applications. Examples are the change of enantioselectivity,\(^\text{36}\) introduction of catalytic promiscuity\(^\text{37}\) and increase of thermostability.\(^\text{38}\)
2.2.3 Molecular dynamics and RMSD calculation

Molecular dynamics (MD) simulations, along with a range of complementary computational approaches, have become valuable tools for investigating the basis of protein structure and function.\(^{39}\) As mentioned in the first chapter, MD simulation is a molecular mechanics based computational approach. The term molecular mechanics (MM) refers to the use of simple potential energy functions and can be used to model molecular systems. MM approaches are widely applied in molecular structure refinement, MD simulations, Monte Carlo (MC) simulations and ligand-docking simulations.

Typically, molecular mechanics models consist of spherical atoms connected by springs which represent bonds. Internal forces experienced in the model structure are described using simple mechanical functions. For example, Hooke's law is commonly used to describe bonded interactions while the non-bonded atoms might be treated as inelastic hard spheres or may interact according to a Lennard-Jones potential. Using these simple models, a molecular dynamics simulation numerically solves the Newton's equation of motion, thus allowing structural fluctuations to be observed with respect to time. Molecular mechanics functions are well known and the mathematic equations that describe the potential energies involved in molecular dynamics can easily be found in the literature.\(^{40}\)

Dynamic simulation methods are widely used to obtain information on the time evolution of conformations of proteins and other biological macromolecules and also kinetic and thermodynamic information. Simulations can provide fine detail concerning the motions of individual particles as a function of time. They can be utilized to quantify the properties of a system at a precision and on a time scale that is otherwise inaccessible. Simulation is therefore a valuable tool in extending our understanding of model systems.

Once the conformational search of the system has been performed by e.g. a molecular dynamics simulation, results are usually analysed by performing some kind of molecular fitting. Fitting is a procedure whereby two, or more, conformations of the same (or different) molecules are oriented in space so that particular atoms or functional groups are optimally superimposed upon each other. It is a method largely used in molecular modelling. For example, fitting is an integral part of many conformational search algorithms, particularly those that require each conformation to be compared with those generated previously in order to check for duplicates.

A molecular fitting algorithm requires a numerical measure of the difference between two
structures when they are positioned in space. The objective of the fitting procedure is to find the relative orientations of the molecules in which this function is minimized. The most common measure of the fit between two structures is the root mean square deviation between pairs of atoms, or RMSD:

\[
\text{RMSD} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \delta^2_i}
\]

where \( \delta \) is the distance between \( N \) pairs of equivalent atoms.

Initially, the pairs of relative atoms in the superposed structures are identified and the distances between the pairs of atoms are measured in each of the X, Y and Z axes. As it can be seen from the formula, these values are squared and summed, and a square root value of the total is calculated and divided by the number of atoms. Additionally, the procedure usually checks for automorphism or different orientations of symmetrical groups in the molecule which are treated as zero RMS deviations.

The root mean square deviation implicitly accounts for the differences between the structures it measures, as can be seen from the equation itself. Therefore the RMSD values are not only used to understand the broadness of the sampled conformational space. These values are also a good description of the structural features the simulated system encounters over the simulation time. From this perspective, it can be used for example as a measure of protein folding or unfolding, distortion of a protein structure under the influence of an external force, the structural difference among different protein crystal structures and many more.

In the present study the RMSD was used to compare the structural differences encountered in time of an enzyme with a large substrate covalently bound to its active site, to the same enzyme without the substrate. Both proteins were subjected to a MD simulation under the same condition and for the same period of time. The resulting structures were compared by means of RMSD as already mentioned. For this particular purpose, the RMSD of every amino acid composing the enzyme was calculated separately. The purpose was to easily identify the mostly influenced residues and to quantify the extent of the effect a large substrate can have on the enzyme structure and on the residues forming the active site.
2.2.4. Aim of the work

None of the enzymatic polymerization reactions have as yet been looked at with computational chemistry techniques. This is surprising since the literature of the last five to ten years provides numerous examples of computational simulations of enzyme-catalysed ester and amide formation and cleavage.

In the present work, the polilactide synthesis by Candida antarctica lipase B was the objective of the study. Three different substrates were chosen for the in silico work, namely glycolide, DD-lactide and LL-lactide. Some experimental evidences were already present in the literature about CALB's ability to catalyse ROP starting from DD-lactide and LL-lactide, whereas glycolide was a substrate that had no experimental data published. As mentioned in chapter 2.2.2, the group of Hans et al reported only a poor catalytic activity for the ROP of DD-lactide when using the wild-type enzyme. Therefore, one of the scopes of the work was to model the behaviour of glycolide derived substrate as compared to the other two, better known substrates.

On the other hand, another scope was to understand the structural basis of substrate recognition by the enzyme. Due to the large size of the substrates, especially in the propagation stage of the polymer generation, the idea was to identify mostly affected residues in order to rationally engineer an appropriate mutant that could show a higher reaction rate.

Since CALB has a narrow and funnel shaped active site, creating more space at the entrance should allow to increase the turnover number. Visual inspection of the enzyme's structure can be useful to identify the residues suitable for mutation. Nevertheless, a more quantitative approach was preferred.

In order to evaluate the position of the hot spots for the generation of the desired mutant, the tetrahedral intermediate of the substrates were built inside the active site and the systems were subjected to molecular dynamics simulations. This approach leads to the identification of the residues that are mostly under the effect of the bulky substrates and are therefore the best candidates for the rational redesign of the enzyme. The mentioned effect was quantified by a novel class of descriptors, specifically developed during this work.

Summarizing, the scope of the work was to understand the behaviour of the glycolide derived substrate as compared to the other two, predict the mutations that could lead to a better lipase for
polylactide synthesis and finally the conception of molecular descriptors for rational redesign of natural enzymes in case of bulky substrates.

2.3. Results and discussion

2.3.1. In silico simulation of enzyme-substrate complexes

Three different monomers were selected for the present investigation due to the advantages offered by the biocatalytic production of the respective polymers. The structures are shown in Figure 2.6.

![Monomer structures](image)

**Figure 2.6.** Monomer structures that were used in the present work. A) Glycolide, B) DD-lactide and C) LL-lactide.

Enzymatic ring opening polymerization consists of three different stages, as already discussed in the introduction, namely activation, initiation and propagation. During the activation and the initiation step a nucleophile (water or alcohol) is needed for the opening of a lactide unit. In the propagation step, the hydroxy group of the open lactide acts as nucleophile for the ring opening of a new lactide unit.

It has been previously shown by Joan et al. that *Candida antarctica* lipase B has a good activity for the ring opening step of both DD-lactide and LL-lactide. ⁴¹ No experimental evidence has been published about glycolide. This can indicate that the initiation step proceeds without problems
and that the low activity of CALB towards the ROP of lactides is due to the propagation step. Therefore, in order to understand the phenomenon at a molecular level, the tetrahedral intermediate (TI) of the propagation step for the three substrates has been built. As an example, in Figure 2.7 the TI correspondent to DD-lactide is shown. In this case, a DD-lactide unit is the acyl donor, whereas a benzyl DD-dilactate is the acyl acceptor. The benzyl unit comes from 1-phenylethanol that is intended to be the initiator in the reaction.

![Diagram of tetrahedral intermediate](image)

**Figure 2.7.** Tetrahedral intermediate with benzyl DD-dilactate after its nucleophilic attack on the acylated enzyme. Essential hydrogen bonds for efficient catalysis are shown in green dotted lines.

Similarly to the DD-lactide, the other two transition states of the reactions of interest were simulated by modelling the tetrahedral intermediates formed in the nucleophilic attack of the free hydroxyl group of the open monomers (glycolide and LL-lactide) on the acyl enzyme.

### 2.3.2. Glycolide, DD-lactide and LL-lactide effects on the enzyme's structure

Once the models were prepared, the correct protonation of ionizable residues was calculated and an energy minimization was performed to release excessive energy strengths present in the structures. One nanosecond long molecular dynamics simulations of the three systems, as well as...
of the empty enzyme were then performed. Since the intention was to reproduce the experimental conditions that were known to yield the best polymerization results reported in the literature, especially the solvent contribution and the temperature, toluene and 60°C were the chosen conditions. The solvent contribution was simulated by implementing an implicit toluene solvation in all the stages of the study, whereas the temperature was set to 333 K corresponding to 60°C.

It is important to say that the tetrahedral intermediates of the substrates have been built inside the active site based on the theoretical knowledge about the physical chemical properties of the active site. In other words, the acyl donors have been placed in the acylic subsite while the acyl acceptors occupy the alcoholic subsite of the enzyme's active site. This implicates that some of the built conformations might be forced in the simulated position which wouldn't be naturally achieved by the substrates in the reaction mixture. This is in particular the case for LL-lactide derived TI since LL-lactide is known to have difficulties in the propagation stage of the polymerization catalysed by CALB. Nevertheless, performing MD simulations and comparing the results with the other two substrates can give important hints about the residues that are mostly disturbed and could possibly be mutated.

*C. antarctica* lipase B is an enzyme made of 317 amino acid residues. In order to evaluate and understand the enzyme's behavior during the molecular dynamics, the RMSD (root mean square deviation) of every amino acid during the different simulations was calculated. Histogram-like plots correlating the amino acid number (X axis) and its specific RMSD value (Y axis) were chosen to represent the results (Figure 2.8).
Figure 2.8. Root mean square deviations of CALB's amino acids under the influence of the three substrates. Values are reported on the Y axis and expressed in Ångströms. Values under 1Å have been omitted for clarity.

It is evident from the graphs shown in Figure 2.8 that the MD simulation of LL-lactide derived substrate is the one that generates most conformational changes in the enzyme's structure. Several residues show a deviation higher than 2.5Å whereas in the glycolide and DD-lactide derived substrates it is only few amino acids that move as much during the simulation. It is also interesting to see that motions of higher entities always include the movement of several residues together, corresponding to zones (helices, loops) of the enzyme. This translates into the fact that non-natural bulky substrates such as the ones considered in this study can trigger changes of the secondary or even the tertiary structure of the enzyme. This makes the engineering of a suitable mutant more complex and not as immediate as in cases of smaller substrates.

Residue movements below 1Å can be considered as natural, internal vibrational motions of the protein at room temperature \(^{42}\) and were in fact omitted from the plots. Nevertheless, the plots are also characterized by quite high noise, and all three of them show several residues having RMSD values higher than 1.5Å. One reason for that, besides the direct substrate effect, can also be the rather high temperature (\(60^\circ\text{C}\)) used for the simulations, which obviously increases the kinetic motions of the simulated systems.

The analogous simulation and RMSD plot (Figure 2.9) of the CALB without any substrate in its active site can elucidate this uncertainty. Since this MD was performed in the exact same
conditions as the previous three, the differences observed in the plotted RMSD values will be due only to the substrate effect.

**Figure 2.9.** RMSD profile of the empty enzyme. The highest peak corresponds to residues forming the enzyme's lid.

During the MD simulation of the empty enzyme, the energetic equilibrium was reached within the first 10 ps and remained stable. From the conformational point of view, there are two evident modifications (clusters with high values in the plot) and correspond to the enzyme's lid\textsuperscript{43} (residues 140-147) and a small loop at the surface of the protein (residues 89-94). In the present simulation, many amino acids show RMSD values higher than 1.5Å as well, indicating the existence of the postulated temperature effect.

Lipases are enzymes that show a phenomenon called interfacial activation which is responsible for the correct functioning of this enzyme family. The enzyme's lid is usually a big domain that shows high flexibility and controls the access of the substrates into the active site.\textsuperscript{44} When a lipase encounters its hydrophobic substrates, the lid opens and the active site becomes accessible. In *C. antarctica* lipase B, this domain is much smaller than in other lipases and cannot fully cover the active site. Nevertheless, in the performed simulation it shows a considerable structural deviation which is most likely due to the anhydrous solvent used for the simulation. The apolar nature of the solvent resembles the natural lipid substrates of the lipase which causes the opening of the lid and the full exposure of the active site to the environment.
A more thorough inspection of the simulations trajectories has also been done. Several similarities have been found between the glycolide and DD-lactide derived substrates. Both simulations reached the energetic equilibrium rather fast in the simulation indicating that there is a good recognition of the substrate and no persistent steric clashes are present. By the end of the simulations, both substrates adopt a similar position inside the catalytic site. The acyl acceptor part of the substrates shows a movement inward resulting in a better fit to the alcohol sub site.

Furthermore, in both MD simulations there is a progressive separation of two ribbon turns (aa 186-189 and aa 275-279) at the borders of the alcoholic part of the active site due to the movements of the acyl acceptor parts of both substrates. Such separation is not evident in the empty enzyme nor under the effect of the LL-lactide derived substrate. At the left side of Figure 2.9 the red arrow indicates the green line representing the measured separation as the distance between the alpha carbon of Glu188 and Lys278. The DD-lactide derived substrate is visible in the active site. The right hand side of the figure shows three graphs in which this distance over the simulation time is plotted. The upper most graph is measured for the enzyme with the LL-lactide derived substrate, the middle one corresponds to the enzyme with the glycolide derived substrate, whereas the lowest graph shows the plotted distance for the DD-lactide derived substrate.
Figure 2.9. The separation of two turns delimiting the lower border of the active site (green line) and the correlation plots of this distance over time. LL-lactide derived substrate (top), glycolide derived substrate (middle), DD-lactide derived substrate (bottom).

From the graphs in Figure 2.9, it is evident that the two above mentioned loops are progressively moving away from each other under the effect of the glicolide and DD-lactide derived substrate. This, along with the previously mentioned similarities observed in the simulations regarding these two substrates indicate that glycolide shows a behavior comparable to the DD-lactide and it can therefore be expected that CALB has similar substrate specificity for the two substrates. This translates to the fact that ring opening polymerization reactions starting from glycolide should be possible using *C. antarctica* lipase B as far as enzyme-substrate recognition is concerned.

On the other hand, the simulation of the enzyme in complex with the LL-lactide derived tetrahedral intermediate shows a different behavior. The system internal energy is rather high and unstable through most part of the simulation which is in correlation with the RMSD plot of Figure 2.8. It is normal that residues belonging to the protein surface show higher deviations. The same is true for residues forming secondary structures such as turns as loops that are not
characterized by specific constraints such as hydrogen bond interactions. On the other hand, pronounced deviations of residues forming highly structured enzymatic motives such as helices or ribbons generally indicate the unfolding of a protein. From the analysis of the calculated trajectory of the third analysed enzyme-substrate complex it was evident that the enzyme underwent a progressive loss of secondary structures during the advancement of the dynamics, which progressively unfolded into turns (Figure 2.10). These conformational changes caused the loss of energetically favorable hydrogen bonding patterns and therefore led to a general destabilization of the enzyme-substrate complex.

This energy destabilization indicates bad substrate recognition by the enzyme due to the S configuration of the secondary alcohol obtained after ring opening of the LL-lactide. The substrate in this configuration doesn't fit to the geometrical requirements of the active site which causes the observed effect. The inability of CALB to catalyse the polymerization starting from the LL-lactide can be at least partially attributed to this phenomenon.

![Figure 2.10. CALB's structure at its initial conformation (left) and after 1 ns of MD simulation with the LL-lactyllactic acid in its tetrahedral intermediate (right; substrate not shown).](image)
2.3.3. Differential RMSD calculation

As already discussed, the goal of the present study was to understand the molecular basis of substrate recognition, similarities and differences between the effect of the three substrates on the enzyme structure and identification of possible hot spots for \textit{in vitro} mutagenesis. To observe the pure substrate effect, a comparison with the empty enzyme has to be done. The presence of the substrate in the active site generates different trajectories as compared to the empty enzyme, mostly due to steric and electrostatic interactions between the bulky substrate and the residues of the active site. Therefore an adequate approach to quantify these differences is to calculate a differential RMSD plot by subtracting the amino acidic deviations of the empty enzyme from those of the enzyme simulated with the covalently bound tetrahedral intermediates.

Three different profiles were obtained corresponding to the subtraction of the RMSD plots of the enzyme-substrate complexes and the RMSD plot of the empty enzyme (see Figures 2.11, 2.12, 2.13). The obtained differential RMSD (dRMSD) are expressed in their absolute values.

The interesting feature of the so obtained differential plots is the evident similarity between the empty-glycolide and the empty-DD-lactide plots. This fact indicates that the effects caused by the presence of these two substrates in the active site are comparable and that in respect to the empty enzyme roughly the same zones of the enzyme are affected. In both plots there is a predominant peak which corresponds to the lid and derives from the high RMSD values the components of this domain show in the simulation of the empty enzyme. In the case of the enzyme-substrate complexes, the lid does not move as much, indicating that the enzyme has lower conformational freedom when bound to the substrate. Therefore for all three plots, the central peak (corresponding to the lid) is not indicating a large movement under the effect of the substrates but rather derives from the subtraction of RMSD values.

In both DD-lactide and glycolide cases, there are two other narrow and high peaks formed by residues 91-97 and 193-200 and a third, larger peak with lower dRMSD values (266-291). This overall resemblance of the two plots is an additional confirmation that these two substrates show similar behaviour during the performed simulations and corroborate the conclusion that glycolide is a good candidate for the ROP reaction catalysed by CALB.
Figure 2.11. Differential RMSD plot; the substrate is derived from glycolide. On the X axis the residue's number, on the Y the deviation expressed in Ångström.

Figure 2.12. Differential RMSD plot; The substrate is derived from DD-lactide. On the X axis the residue's number, on the Y the deviation expressed in Ångström.
On the other hand, the differential plot obtained by the subtraction of the RMSD plot of the empty enzyme and the RMSD of the enzyme bound to the tetrahedral intermediate formed with LL-lactyllactic acid has substantially different appearance (Figure 2.13). The most evident peak is formed by amino acids 243-253 forming a superficial loop at the back side of the active site. There are also other relevant modifications mainly regarding single amino acids or groups of 3-4 residues. It is interesting to notice that the plot does not show any evidence of the loss of secondary structures observed in the MD. This is probably due to the fact that the simulated unfolding is still in its initial state and the globular shape of the protein is still conserved by the end of the simulation. A longer simulation would be necessary to translate the unfolding event into RMSD.

![Figure 2.13](image)

**Figure 2.13.** Differential RMSD plot obtained by subtracting the RMSD of the residues of the empty enzyme to the RMSD of the residues of the enzyme-substrate complex. The substrate is derived from LL-lactide.

A closer look has been given to the residues forming the active site of the enzyme. Hardly any of the amino acids directly involved in the substrate recognition shows significant dRMSD values in the plots. Most of the values are lower than 0.5Å. Nevertheless, relevant conformational changes of single amino acids have been observed at the entrance of the active site in the glycolide and DD-lactide analysis. The fact that this movement is not visible from the dRMSD plots is certainly
due to a lower extent of the movement as compared to highly flexible zones well described by the plots. Moreover, only the side chain or part of the side chain changes position by the end of the simulation. Several of these apparently unimportant changes occur during the simulation and their concerted motion results in the partial occlusion of the entrance of the acylic part of the active site. This hinders the prolongation of the polymer as the opened monomers cannot access the active site and be added to the growing chain.

The only strategy to obviate this problem is to enlarge the space of the binding pocket by reducing the volume of the occluding residues. Therefore, the amino acids that have to be mutated to avoid this phenomenon are those situated at the entrance of the acylic subsite, namely Leu140, Ile189 and Ile285 (fig. 2.14) into smaller amino acids (e.g. Ala or Gly). In the case of the proposed mutants, there would be more space for the bulky substrates allowing the enzyme to have a greater turnover number (k\text{cat} value) while improving both the reaction rate and the degree of polymerization.

**Figure 2.14.** The position of the glycolide-derived substrate (atom type color code) inside CALB’s active site by the end of the molecular dynamics simulation. The residues occluding the acylic part of the active site are colored in blue (Leu 140, Ile 189, Ile 285).
2.3.4. Caprolactone as a model substrate

Since it is well known that CALB is able to catalyse ring opening polymerization reactions starting from caprolactone\(^1\), the aim of this part of the work was to investigate what type of interactions the enzyme establishes with a substrate that doesn't show lactide characteristic problems. The focus was once again on the enzyme-substrate interactions. In the present case, the simulated substrate consists of three units of caprolactone which makes it less bulky as compared to the previous three substrates. Therefore less steric hindrance is expected. The substrate is also achiral and problems related to chirality in lactides will not be present.

The starting point of the performed simulation was the model of the tetrahedral intermediate with caprolactone acting both as acyl donor and acceptor as shown in Figure 2.15. In order to obtain data comparable with those previously generated using lactides, the molecular dynamics simulation was performed in the same conditions and implementing the exact same parameters as mentioned above. The RMSD values calculated for each amino acid during the simulation were used to quantify and represent the effects of the substrates.

**Figure 2.15.** Covalently bound tetrahedral intermediate inside CALB’s active site. The substrate is formed of three caprolactone units.
During the simulation, it was noticed that the entire system is energetically stable and no major perturbations were evident. The equilibrium was reached within the first few ps of the simulation and remained constant during the whole simulation time. As expected, the system was stable and the calculated RMSD plot of the enzyme's residues is clearly different from the previously calculated ones. The RMSD are almost all lower than 1 Å for any amino acid as shown in the plots of Figure 2.16.

**Figure 2.16.** Two representations of the same RMSD plot showing the deviation of each amino acid during the MD simulation of CALB and caprolactone derived substrate. In the plot at the right, values under 1 Å have been omitted for comparison with previous graphs.

The results of this simulation indicate that the enzyme in complex with a good substrate behaves differently from all previous cases. The substrate does not cause any distortion of the enzyme structure. Moreover, the favorable interactions that are formed between the enzyme's active site and the substrate stabilize the system that shows only minor fluctuations of the amino acids, making it even more rigid than the empty enzyme. The temperature effect was not observed in this simulation. The obtained results indicate a good correlation with the catalytic efficiency of the enzyme, thus suggesting the validity of the dRMSD approach for pointing out differences in enzyme specificity.
The MD performed with the caprolactone was utilized for a different purpose as well. The calculated RMSD values were taken as a fingerprint of an energetically stable system and as a reference to estimate the different effect of the three lactides on the CALB structure. Therefore, the idea was to find the effect of lactide substrates on the enzyme residues not compared to the empty enzyme but compared to a polymerization system that actually works. The differential RMSD plots were calculated as a difference in the RMSD of the amino acids in the lactides simulations and the same amino acids in the caprolactone simulation (see Figures 2.17, 2.18, 2.19).

Once again, there are similarities between the DD-lactide and glycolide. The corresponding plots have a similar appearance and the clusters of highly flexible amino acids are comparable and differ only in their absolute values, indicating that DD-lactide and glycolide have a similar effect on the enzyme.

Interestingly, the previously identified residues which movement was seen to occlude the entrance of the active site in the new plots show significant dRMSD values (Leu140, Val154, Ile285). The only exception is Ile189 that doesn't appear in the plots. Another active site component shows a significant dRMSD value in the case of DD-lactyllactic acid, namely Gln157 (Figure 2.18). Being situated in the acylc part of the active site, mutating this residue into a smaller one should also improve polymerization results.

**Figure 2.17.** Differential RMSD calculated in a subtraction procedure of the residue's RMSD values under the caprolactone effect subtracted to the RMSD values under the glycolide effect.
These results indicate that a comparison with a well-established, functioning system is better than the comparison with the empty enzyme and should be used whenever data are available. This is especially true when the residues of the active site are concerned. The simulations indicated that in free, empty enzymes these amino acids are highly flexible and subjected to greater fluctuations than the inner, core residues. The active site of an enzyme is in direct contact with the environment and resents the effect of its surrounding, even in cases when the solvent effect is implicitly simulated. On the other hand, when a substrate establishes favorable interactions with the active site, the residues are more rigid due to the generation of weak forces such as electrostatic and hydrophobic interactions and hydrogen bonds.

**Figure 2.18.** Differential RMSD calculated in a subtraction procedure of the RMSD values under the caprolactone effect subtracted to the values of the RMSD under the DD-lactide effect.

In Figure 2.19 the differential RMSD values of caprolactone and LL-lactide systems are shown. This plot shows more perturbations in respect to the other two. Some of the interesting residues show visible dRMSD values (Leu140 and Ile285) but still lower than 1 Å. Nevertheless, the severe deformations shown in Figure 2.10 probably wouldn’t be obviated by mutating some amino acids of the active site. Perhaps a redesign of enantioselectivity would be necessary as well.
2.3.5. *In vitro* production of mutants and experimental validation

The production of mutants was handled by the group of *Hult et al*. Five different mutants were produced *in vitro* by site directed mutagenesis and screened towards the hydrolysis of tributyrin, transacylation of ethyl octanoate and ROP of DD-lactide. The proposed mutagenesis hot spots were Gln157, Ile189 and Leu278 based on the group’s own modelling study. The produced mutants were a combination of single, double and triple mutants. All the substitutions were done changing the parent amino acid to alanine.

Alanine in positions 157 and 189 caused a decrease in the hydrolysis activity towards tributyrin as compared to the wild type-enzyme (WT). The L278A mutant showed a high increase in the activity towards tributyrin. The double mutant I189A, L278A and the triple mutant Q157A, I189A, L278A confirmed the positive effect of the alanine in position 278 and the negative effect of having alanine in positions 157 and 189, by showing intermediate values that were similar to those of the WT (Table 2.1). The activity of the mutants for the transacylation of ethyl octanoate

**Figure 2.19.** Differential RMSD calculated in a subtraction procedure of the RMSD values under the caprolactone effect subtracted to the values of the RMSD under the LL-lactide effect.
with 1-hexanol, were all lower as compared to the WT (Table 2.1).

**Table 2.1.** Rates of the hydrolysis of tributyrin and transacylation of ethyl octanoate with 1-hexanol catalysed by CALB wild type and mutants.

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Tributyrin&lt;sup&gt;b&lt;/sup&gt; Rate (s&lt;sup&gt;-1&lt;/sup&gt;)</th>
<th>Ethyl octanoate&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>330</td>
<td>200</td>
</tr>
<tr>
<td>Q157A</td>
<td>10</td>
<td>92</td>
</tr>
<tr>
<td>I189A</td>
<td>26</td>
<td>15</td>
</tr>
<tr>
<td>L278A</td>
<td>1400</td>
<td>16</td>
</tr>
<tr>
<td>I189A, L278A</td>
<td>180</td>
<td>5.5</td>
</tr>
<tr>
<td>Q157A, I189A, L278A</td>
<td>230</td>
<td>28</td>
</tr>
</tbody>
</table>

<sup>a</sup> (mole product). (mole catalyst)<sup>-1</sup>. s<sup>-1</sup>.  
<sup>b</sup> Hydrolysis of tributyrin determined by pH-stat.  
<sup>c</sup> Transacylation of ethyl octanoate with 1-hexanol determined by GC.

Initial screening of the mutants and WT towards the ring opening polymerization of DD-lactides revealed that mutants Q157A and Q157A, I189A, L278 had the highest improvements as compared to WT. These were chosen for a detailed kinetic study. Mutant Q157A showed 4-fold improvement in the initiation step and 90-fold in the propagation step as compared to the WT. The space created in the acyl donor pocket by replacing glutamine with alanine made a large impact on the propagation step. This was also confirmed by the mutant (Q157A, I189A, L278A) as it showed a 83-fold rate increase as compared to the WT.

Preparative ring opening polymerization reactions were run using the WT and mutants (Q157A) and (Q157A, I189A, L278A). The reactions were allowed to run for 48 hours in D8-toluene at 60°C using 1-phenylethanol as initiator. The synthesised polymers were characterized by NMR, MALDI-TOF and SEC. As shown in Table 2.2, the highest conversion (89%) was achieved with mutant Q157A, I189A, L278A and polymers with molecular weight (Mn) of 780 Da were produced corresponding to a degree of polymerization (DP) of 4.5 lactide unites (9 lactic acid unites) and a polydispersity index (PDI) of 1.6. Mutant Q157A gave a conversion of 70% and polymers with Mn of 680 Da, DP of 4 and PDI of 1.6 (Table 2.2). The WT showed 11% conversion and polymers with Mn of 280 Da, DP of 1 and PDI of 1.1 (Table 2.2) which indicate
that the WT has catalysed only the initiation step. The DP of the mutants was lower than expected probably due to water initiation. MALDI-TOF analysis of the polymers produced by the two mutants (Q157A and Q157A, I189A, L278A) showed a main group of peaks. The mass difference between the peaks was 72 Da, which corresponds to one lactic acid unit. No peaks corresponding to polylactide were detected by the MALDI-TOF spectrum of the WT reaction.

**Table 2.2.** Synthesis of poly(DD-lactide) by wild type CALB and its mutants.

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>$M_n$ (Da)</th>
<th>$PDI^a$</th>
<th>$DP^b$</th>
<th>Conv. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>220</td>
<td>280</td>
<td>1.1</td>
<td>1</td>
</tr>
<tr>
<td>Q157A</td>
<td>560</td>
<td>680</td>
<td>1.6</td>
<td>4</td>
</tr>
<tr>
<td>Q157A, I189A, L278A</td>
<td>740</td>
<td>780</td>
<td>1.6</td>
<td>4.5</td>
</tr>
</tbody>
</table>

* Determined by SEC. ** Determined by $^1$H NMR

The experimental results published by Hult *et al.* demonstrate the validity of the developed approach and the deriving predictions of the hot spots to obtain more active mutants. Two out of three selected hot spots were identified by the work described in this chapter. The produced mutants showed a higher rate in the ROP reactions and longer polymers were obtained, as expected.
2.4. Experimental section

The protein structure used for this study was retrieved from the Protein Data Bank (Id:1TCA). This initial structure was pretreated in MOE\textsuperscript{46} by removing the crystallographic water molecules and two molecules of NAG (N-acetyl glucosamine) present in the pdb file. Hydrogen atoms were added and their position was optimized with an energy minimization procedure using the protonate 3D module of the software.

All tetrahedral intermediates were sketched bonding the hydroxyl oxygen of the catalytic serine (Ser105) and the reactive carbonyl carbon of the substrates. This carbon atom changes to an sp\textsuperscript{3}-hybridized configuration. Once the tetrahedral intermediates were built, the protonation procedure was done once again to ensure the correct partial charges of the substrates.

The molecular dynamics simulations were performed using the DYNAMICS module of MOE. All the dynamics were performed in a NPT environment simulating the temperature of approximately 333K. The dielectric constant was set to 2.38, according to the tabulated value for toluene. An integration time of 2 fs was used and a frame of the trajectory was saved every 10 fs.

The root mean square deviations of every amino acid were calculated by a specifically developed script. Once the values were obtained for all the simulated systems, open office spread sheet and its functions were used to calculate the differential RMSD values. First the corresponding values were subtracted, and then the absolute value was used to generate the plots.
2.5. Conclusions

The increasing concern of our society has driven scientists to seek for materials prepared from renewable sources as alternatives for synthetic petrochemical-based materials. The combination of well suited physical properties and biodegradability makes polylactic acid a promising substitute for synthetic petrochemical based polymers in a wide range of applications, e.g. single use packaging materials and commodities.

In this chapter, the development and application of a novel class of molecular descriptors have been described. Intended to point out plausible modifications in *Candida antarctica* lipase B for ROP reactions, differential RMSD descriptors represent an effective tool for studying enzyme-substrate complexes. In cases when hindrance of substrate can affect severely the enzyme conformation, these descriptors can be used to monitor and quantify these effects. The method is relatively low demanding from the computational point of view and the descriptor itself is simple to understand and interpret.

Two different strategies have been developed and their combined use gives comprehensive information on the quality of substrate acceptance by the enzyme:

1- subtracting RMSD values from the corresponding values calculated for the empty enzyme
2- subtracting RMSD values from the corresponding values calculated for an energetically stable system (caprolactone in the case of polylactate synthesis)

The combination of the two approaches allows differentiating between normal enzyme flexibility as a reaction to substrate binding and actual structural distortions caused by unfavorable fitting. Visualization of RMSD plots allow the facile identification of the conformationally altered zones and can be used for the rational identification of hot-spots for mutagenesis. The quality of the concept was assessed by the experimental results of enzyme engineering done by the group of Hult *et al.*
3. J. Berzelius, Rapport annel de l’Institut geologique de Hongrie. **1847**
4. M. M. Berthelot, C. R., **1853**, 37, 398 – 405
43. M. Skjot, L. De Maria, R. Chatterjee, A. Svendsen, S. A. Patkar, P. R. Ostergaard, J. Brask, *Chembiochem*, **2009**, 10, 520-527
46. MOE v.2008.10, Chemcomp, Montreal, Canada.
3. Converting a hyd oxynitril lyase into an esterase; a computational study

3.1. Summary

Hyd oxynitrile lyases (HNL) and esterases are an excellent example of divergent evolution since some representatives of these classes of enzymes derive from a common ancestor. It is the case of enzymes having the same, α/β-hydrolase fold but catalysing very different types of reactions. The complicated and still not well understood relationships between sequence, structure and function indicate that a protein function is not easily predictable and therefore difficult to engineer. Nevertheless, there have been examples of engineering reaction promiscuity, achieved with the aim of molecular modelling and simulations and a subsequent rational redesign of the enzyme’s active site.

In the present chapter the HNLs belonging to the alpha/beta hydrolase fold have been compared to esterases of the same superfamily. By means of structural superposition and a physico-chemical description of the active sites applying the GRID-PCA technique, similarities and differences between enzymes belonging to the two classes have been described. Four different enzymes were selected for the study, namely HNL from Hevea brasiliensis, HNL from Manihot esculenta, arylesterase from Pseudomonas fluorescens, and carboxylesterase from Alicyclobacillus acidocaldarius. The selection of these enzymes was based on concepts of sequence alignments, structural similarities and evolutionary relationships.

The results indicate that the structural similarities between different enzymes can be in correlation with their evolutionary history; the phylogenetic tree obtained by sequence alignment and the structure based phylogenetic tree of the four enzymes have very similar appearances. The application of the GRID-PCA highlighted the differences of the enzyme's active sites in terms of shape and physico-chemical properties. Several concepts were developed which were inevitably correlated to the type of reactions catalysed by the two classes as well as to the characteristics of their natural substrates.

Following the obtained results, several residues of the Manihot esculenta HNL's active site were selected for mutation in order to reproduce the features encountered in the carboxylesterase from...
A. acidocaldarius. The aims were i) Redesign of substrate specificity of the HNL and ii) The conversion of general acid/base catalysis to a covalent one. The resulting four-fold mutant generated in silico shows all the necessary requirements for ester hydrolysis.
3.2. Introduction

3.2.1. Alpha/Beta-Hydrolase fold enzymes

The α/β-hydrolase fold family of enzymes is rapidly becoming one of the largest groups of structurally related enzymes with diverse catalytic functions. Members in this family include acetylcholinesterase, dienelactone hydrolase, lipase, thioesterase, serine carboxypeptidase, proline iminopeptidase, proline oligopeptidase, haloalkane dehalogenase, haloperoxidase, epoxide hydrolase, hydroxynitrile lyase and others. The enzymes all have a Nucleophile-His-Acid catalytic triad evolved to efficiently operate on substrates with different chemical composition or physico-chemical properties and in various biological contexts.

The three-dimensional fold of a polypeptide backbone provides the architectural framework on which the enzymes function and specificity strictly rely. There are two models for the evolution of enzyme structure and function. In convergent evolution, protein structures or folds evolve independently to perform the similar function. An illustrative example is the trypsin$^2$ and subtilisin$^3$ families of serine proteases having two completely different folds, both of which are distinct from the α/β-hydrolase fold, but sharing the active site catalytic machinery designed for efficient peptide hydrolysis. The more common, divergent evolution, involves the development of proteins with different function from a common ancestor protein. It is well known that a large number of different protein sequences can conform to a similar protein fold. It appears that nature has preferentially opted for stable protein cores, or ancestor folds, which have served as templates in an evolutionary process of molecular adaptation and fine tuning of enzyme function. Given a particular fold, this process involving amino acid substitutions, loop insertions or deletions, may in fact lead to enzymes with diverse catalytic functions and with apparently only a few sequential features in common. The α/β-hydrolase fold family of enzymes represent such an example$^4$. 
Figure 3.1. Canonical fold of the α/β-hydrolase. a: Topology diagram, with the strands indicated by red arrows and the helices by cyan cylinders. b–d: Three examples of esterase structures

In the beginning of the 1990’s the elucidation of the 3D-structures of five apparently unrelated hydrolytic enzymes revealed what was called the α/β-hydrolase fold. The canonical, α/β-hydrolase fold is an eight-stranded mostly parallel α/β structure (Figure 3.1). The spatial positions of the catalytic triad atoms involved in catalysis are remarkably well conserved in all enzymes in the family. An important part of the active site of many α/β-hydrolases is the spatial arrangements of hydrogen bond donors, the oxyanion hole, that stabilize the oxyanion transition state of the catalysed reaction.

This protein architecture is among the most promiscuous protein folds described. It tolerates large insertions into the single-domain protein. The resulting enzymes range from 25-65 kDa in molecular mass, operate on substrates with completely different chemical or physicochemical properties, and act in diverse biological contexts. The α/β-hydrolases do not require any co-factor for their function. As mentioned, all the enzymes have a Nucleophile-His-Acid catalytic triad with its residues positioned in loops that are the best conserved of the fold. Only the histidine in the triad is invariant in the enzymes. The nucleophile can be a Ser, Cys or Asp residue while the
acid loops accommodate an Asp or Glu amino acid residue. This contrast the belief of an invariant Ser-His-Asp catalytic triad motif first identified four decades ago in the trypsin and subtilisin protease families.

3.2.2. Reaction promiscuity and evolutionary implications

Enzymes are traditionally referred to as remarkably specific catalysts. Yet the notion that many enzymes are capable of catalyzing other reactions and/or substrates, in addition to the ones for which they are physiologically specialized, or evolved, is not new. Early examples of enzyme promiscuity include pyruvate decarboxylase,\(^5\) carbonic anhydrase,\(^6\) pepsin,\(^7\) chymotrypsin\(^8\) and L-asparaginase.\(^9\) Nonetheless, the notion of “one enzyme - one substrate - one reaction” dominated, and still dominates the textbooks, and until recently, the wider implications of enzyme promiscuity were largely ignored.

The idea of nature as an opportunistic modifier of preexisting suboptimal functions is also relatively old and has been formulated by Jacob in his classical note “Evolution and Tinkering”\(^{10}\). The first direct connection between promiscuity and protein evolution was made in 1974 by Jensen.\(^{11}\) Jensen boldly forwarded the hypothesis that, unlike modern enzymes that tend to specialize in one substrate and reaction, the primordial, ancient enzymes possessed very broad specificities. Thus, relatively few rudimentary enzymes acted on multiple substrates to afford a wider range of metabolic capabilities. Divergence of specialized enzymes, via duplication, mutation, and selection, led to the current diversity of enzymes and to increased metabolic efficiency.\(^{12}\)

New enzymes have constantly emerged throughout the natural history of this planet. Enzymes that degrade synthetic chemicals introduced to the biosystem during the last decades,\(^{13}\) enzymes associated with drug resistance,\(^{14}\) and enzymes in plant secondary metabolism\(^{15}\) provide vivid examples of how fast and efficient the evolution of new enzymatic functions can be. Indeed, extensive research since Jensen’s article provided ample evidence for the notion that promiscuity is a key factor in the evolution of new protein functions.

An enzyme superfamily combines dozens to thousands of enzymes that, although distant in sequence and catalyzing different chemical transformations of many different substrates, share the same fold and a common catalytic strategy (e.g., abstraction of a proton from a position alpha
to a carboxylate, and stabilization of the resulting enolate intermediate, in the enolase superfamily). Understanding the role of promiscuity in nature can help to improve the rational design of novel catalytic functionalities as well as to predict the effect of given mutations.

Fortunately, the analysis of enzyme families and superfamilies provides ample evidence for the role of promiscuity in the evolution of new functions. Specifically, the identification of promiscuous activities between different members of the same enzyme family or superfamily and the ability to evolve these promiscuous activities in the laboratory provide important hints regarding evolutionary, structural, and mechanistic relationships within enzyme superfamilies.

Conclusions supported by these data are summarized below:

1. The primary, or native, function of one family member is often identified as a promiscuous activity in other family members. This overlap may reflect the common catalytic strategy that underlines these families and superfamilies as well as a common evolutionary origin.\(^\text{16}\)

2. The same promiscuous activity is often shared by more than one family member

3. The magnitude of promiscuous functions varies significantly between family members.\(^\text{17}\)

Given these notions, it is evident that the starting point for introducing reaction promiscuity into a given enzyme should be a component of the same superfamily. An extra advantage would be starting from an enzyme that already shows a weak activity for the reaction of interest, in other words improving what is already there. This characteristic is called evolvability, \(i.e.\) the capability of biological systems, whether they are organisms, cells or proteins, to evolve. It comprises two elements: plasticity and robustness. Plasticity is the introduction of novel traits by a relatively small number of mutations and has been demonstrated by many laboratory evolution experiments. However, plasticity is in conflict with the fact that most mutations are deleterious giving organisms (and proteins) the necessity to endure mutations (robustness) and resistance to the effects of mutations.

Ultimately, the acquisition of a proficient new activity must come at the expense of the old one. Yet, the relative rate by which a new function is gained, and the old one is lost can be different (Figure 3.2). In those cases, where the negative trade-off is initially weak (convex route), the divergence of new function proceeds via a generalist intermediate exhibiting broad specificity. This route suggests that, under selection for increasingly higher proficiency, specialists might evolve spontaneously (\(i.e.\) without an explicit selection against the original function) because at a certain point increases in the new function will be accompanied by large losses in the original
one. At present, however, laboratory evolution experiments (directed evolution) demonstrate that generalist intermediates re-specialize primarily upon dual selection for an increase in the newly evolving activity and a decrease in the original activity.\textsuperscript{18,19}

\textbf{Figure 3.2.} Possible routes to the divergence of a new function. The introduction of a new activity into a protein with a given primary function (blue dot) yields a new protein with a function that replaced the original one (green dot). The dynamics of this divergence may vary. The gain-loss of the activities may trade-off linearly (dotted line) or follow either the convex or the concave route.

Mutation incorporated through rational and semi-rational design show larger trade-offs than mutations obtained from random repertories. This difference can be largely ascribed to the location of the function altering mutations. Evolutionary processes (both \textit{in vivo} and \textit{in vitro}) usually involve mutations in the active site periphery (second or third shell mutations) with more subtle effects. Rational design aims at the replacement of key active site residues (first shell) and such changes yield more drastic changes of specificity.\textsuperscript{20} An important facet of the trade-off that is not reflected in the kinetic parameters is the effect of mutations on stability. Most mutations are destabilizing, and mutations that affect function often exhibit even higher destabilizing effects.\textsuperscript{21} Destabilization usually results in reduced cellular enzyme levels, owing to misfolding and aggregation, proteolytic digestion or clearance.
directed evolution experiments many mutations appear without any obvious role in the new function, but rather exert compensatory stabilizing effects. Rational design, on the other hand, has no way of predicting such mutations. Nevertheless there are several examples in which catalytic promiscuity has been achieved by just one to few mutations, indicating that in such cases the robustness of the parental enzyme was strong enough to fight the possible destabilizing effects.

The summarizing points regarding the mechanistic and evolutionary perspective of enzyme promiscuity are the following:

1. Promiscuity regards activities, substrates, and/or reactions that an enzyme performs, although it never evolved to do so (as opposed to its original, native activity).
2. Promiscuous activities are not rare exceptions but are rather widely spread, inherent features of enzymes, and proteins in general.
3. Specificity and promiscuity can reside within the same active site. Promiscuous enzymatic functions may utilize different active-site conformers, and their mechanisms can overlap, partly overlap, or differ altogether from the mechanism by which an enzyme performs its native function.
4. Promiscuous enzyme functions provide immediately accessible starting points for the evolution of new functions via a gradual mutational path that eventually converts a weak, promiscuous function into the primary, native function.
5. A promiscuous function of an enzyme can be a vestige of the function of its ancestor. Promiscuous activities shared by members of the same enzyme family and/or superfamily correlate with their divergence from a common ancestor.
6. Mutations that increase a promiscuous activity and have little effect on the primary, native function (weak trade-off) underlie the divergence of a new enzymatic function via a generalist intermediate.
3.2.3. Hydroxynitrile lyases and esterases

Hydroxynitrile Lyases (HNL’s, also known as Oxynitrilases) are enzymes that catalyse the release of hydrogen cyanide (HCN) from natural cyanohydrins in several higher plants. This reaction is a defense mechanism of these plants against several herbivores, commonly known as cyanogenesis. Since the discovery of these versatile enzymes, much effort has been put into research on their application as catalysts for the reverse reaction, the enantioselective synthesis of cyanohydrins.

The HNL’s from *Prunus amygdalus* (PaHNL), *Sorghum bicolor* (SbHNL), *Manihot esculenta* (MeHNL), *Linum usitatissimum* (LuHNL) and *Hevea brasiliensis* (HbHNL) are currently the five major HNL’s that are being used in asymmetric synthesis of enantiopure cyanohydrins derived from several aldehydes and some ketones.

PaHNL can be found in almonds, HbHNL comes from the rubber tree, SbHNL from cane like grass, MeHNL is found in cassava and LuHNL in flax. Out of these five HNL’s, PaHNL and LuHNL are (R)-selective. The three other HNL’s are all (S)-selective. The accessible amount of the respective HNL’s depends on the natural sources they are isolated from and is often a bottleneck in HNL usage. Some sources have been shown to contain sufficient amounts of HNL’s, while others only contain a small amount of the wanted enzyme. Lately, the production of some HNL’s, especially MeHNL and HbHNL, has been optimized by cloning and over-expressing them in different bacterial and yeast strains, such as *Escherichia coli* and *Pichia pastoris*. Nevertheless, the relatively high demand for these enzymes due to their ability to synthesise C-C bonds yielding enantiopure products makes the research in this field appealing. In particular, to achieve the possibility of discovering promiscuous HNL activity in other enzyme classes.

As mentioned, hydroxynitrile lyases play an important role in the synthesis of chiral cyanohydrins which are valuable intermediates in the follow-up synthesis of pharmaceuticals, agrochemicals and other chemicals. Possible applications of cyanohydrins is summarized in Figure 3.2.
Figure 3.2. HNLs condense aldehydes or ketones with a cyanide source and synthesise $\alpha$-hydroxynitriles, versatile chiral intermediates for ranges of chemicals and pharmaceuticals.

The functional groups of cyanohydrins provide a wide variety of functionalized units for further transformations in follow-up chemical or chemo-enzymatic reactions for synthesis of valuable intermediates or pharmaceuticals and agrochemicals (Figure 3.3).

Figure 3.3. General structure of a cyanohydrin molecule.
**Nitrile Group.** This group in cyanohydrins can undergo hydrolysis, solvolysis, and reduction; therefore, ranges of products are synthesized by this property.

**Hydroxyl Group.** To suppress or avoid instability, degradation and racemization, O-protection can be used for the synthesised cyanohydrins. It can be also used for conformation inversion.

**Reactions at the carbon center.** Inversion of the carbon center also provides a range of opportunities for follow-up reaction for the cyanohydrins.

Hydroxynitrile lyases originated from various ancestors among the protein classes. They have different primary structures, and all catalyse the same reaction; therefore, they have been considered as an example of the convergent evolution.\(^{25,26}\)

Amino acid sequence alignment of the known HNLs classifies them into four superfamilies:

- Esterase-lipase superfamily including HNLs from *Manihot esculenta* (S-MeHNL), *Hevea brasiliensis* (S-HbHNL), *Baliospermum montanum* (S-BmHNL), and *Arabidopsis thaliana* (R-AtHNL);
- Peptitades-S10 superfamily: serine carboxypeptidase, including HNL from *S. bicolor* (S-SbHNL)
- Medium-chain reductase/dehydrogenase (MDR) superfamily, including HNL from *L. usitatissimum* (R- LuHNL)
- Glucose-methanol-choline (GMC) oxidoreductases N and C superfamily including HNLs from *P. dulcis* or *P. amygdalus* (R-PaHNL), *Prunus serotina* (R-PsHNL), and *Prunus mume* (R-PmHNL)

Despite more than 3000 species of known cyanogenic plants and hundreds of cyanogenic compounds,\(^ {27}\) only 40 different hydroxynitrile lyases have been identified. On the other hand, the number of HNL species which crystal structure has been solved is only five and is shown in Figure 3.4. This fact restricts the possibility of studying these enzymes by molecular modelling techniques to only a fraction of existing HNLs.
Esterases, on the other hand, are enzymes in the α/β-hydrolase fold family that catalyse the hydrolysis of esters via the addition of water. The α/β-hydrolase fold is a superfamily of enzymes that includes different types of esterases, namely carboxylesterases, cholinesterases, epoxide hydrolases, aryl esterases and phosphotriesterases. They are ubiquitous enzymes with important physiological and biotechnological roles in the synthesis or hydrolysis of ester-containing compounds. However, esterase expression and activity are tissue- and organism dependent, with levels and activities varying widely. In contrast to lipases, which display maximal activity towards water insoluble long chain triglycerides, esterases hydrolyse at least partly soluble fatty acid esters with acyl chain lengths of less than 10 C atoms. Till now, esterases have been widely used in industry to catalyse the stereospecific hydrolysis, transesterification and conversion of a variety of amines as well as primary and secondary alcohols. Thus, newly found esterases are
potentially useful for industrial processes.

Many esterases contain the pentapeptide motif GXSXG, however not all. The GDSL motif is seen in one subfamily of hydrolytic/lipolytic enzymes\textsuperscript{32} that exhibit multifunctional properties and broad substrate specificity due to the flexibility of the active site that undergoes an induced fit style conformational change upon substrate binding. The GDSL superfamily is further classified based on the strict conservation of the catalytic residues S, G, N and H (Ser, Gly, Asn and His) into SGNH hydrolases. The catalytic triad of this family is shaped by a Ser in the GDSL motif and the Asp and the His in conserved blocks.\textsuperscript{33} While SGNH hydrolases are well represented in the eukaryotic organisms, the isolation and characterization of SGNH hydrolases from bacteria remains limited.

The publication of crystal structures of mammalian carboxylesterases in 2003 (Figure 3.5) has greatly contributed to the understanding of the enzyme catalytic mechanism. Carboxylesterases cleave carboxylic esters via a two-step process that involves the formation and degradation of an acyl-enzyme intermediate. The catalytic or nucleophilic serine is first activated to generate the oxygen nucleophile, which then attacks the carbonyl carbon of the ester substrate, leading to formation of the acyl-enzyme intermediate. The alcohol hydrolysis product is then released to undergo nucleophilic attack by water, leading to the release of the carboxylic acid and return of the catalytic amino acids to their original state.
Figure 3.5. Crystal structure of human liver carboxylesterase 1 (hCE1) complexed with homatropine. The α/β-hydrolase fold is clearly visible.

3.2.4. GRID-PCA and its application in rational enzyme engineering

Sequence alignment, and phylogenetic analysis are important tools for the comprehension of evolutionary paths within enzyme superfamilies. Computational analysis based on bioinformatics methods offer a full platform of mathematical instruments to extract information from protein sequences, but understanding the intimate of enzyme promiscuity and developing strategies for enzyme engineering need the structural chemical information to come into play. The comparative structural analysis of enzyme superfamilies is a much more complex task than simple sequence analysis. The substantial increase of the dimensionality of the problem, with the introduction of the Cartesian space and the chemical information, requires higher level of sophistication in the
analytical approaches.

The comparison of two or more three dimensional structures of enzymes presents several hurdles:

- the structures, often with a limited sequence identity despite a common fold pattern, must be efficiently superimposed;
- the relevant chemical differences must be identified among a big number of small, negligible differences. Visual inspection of such a complex system cannot be efficient, even with state-of-the art visualization software;
- the differences must be evaluated, possibly using quantitative parameters and must be interpreted in the context of the biological/chemical problem of interest.

GRID-PCA is a computational method based on the combination of a force field-based calculation of Molecular Interaction Fields (MIFs) and a statistical analysis tool (Principal Component Analysis, PCA) for extracting relevant information from the data. It has been specifically developed to identify differences among correlated proteins, as a tool to support the design of selective inhibitors in drug discovery, but it has demonstrated to be efficient also in the comparison of enzyme active sites, to outline the molecular basis of substrate specificity differences.

The method is a two-steps procedure:

1. Calculation of MIFs by GRID computational procedure
2. PCA of the generated MIFs

GRID, developed by Goodford in 1985, detects energetically favorable binding sites on molecules of known three-dimensional structure. It calculates the interaction between a small chemical probe, mimicking a diverse set of chemical entities (such as the common functionalities of organic compounds) and a target molecule, an enzyme, or its active site in this case. The calculation is performed establishing a three-dimensional array of grid points throughout and around the target and simulating the interaction between the probe(s) and the target in each of these Cartesian spots. The final result of the procedure is a three-dimensional matrix of interaction energy values, often several thousands, each of them belonging to a determined position in the Cartesian space. These matrices are called Molecular Interaction Fields (MIFs).
The MIFs are valuable descriptors of the system, because they are quantitative, reflect the three-dimensional structure of the target and intrinsically carry chemical information.

Similar proteins will have similar MIFs and a simple arithmetical operation can point out quantitatively their differences. However, MIFs are made of thousands of variables (and their three-dimensional information), thus requiring a method for the simplification and rationalization of the information contained in them and in any mathematical operation made on them.

This is the role of the Principal Component Analysis, the second step of the GRID-PCA.

PCA is a variable reduction procedure. It is useful when systems made of a large number of variables contain some redundancy in those variables. Redundancy means that some of the variables are correlated with one another, possibly because they are measuring the same construct. An example could be a group of GRID points generated by a H-donor probe, grouped in a small, well-defined Cartesian space. They all account for a H-acceptor counterpart in the protein in that specific zone, and although they have slightly different numerical values, they are actually carrying the same information. Because of this redundancy, it is possible to reduce the observed variables (the group of H-donor GRID points) into a smaller number of principal components (artificial variables) that will account for most of the variance in the observed variables.

In other words, PCA is a method for reducing the dimensionality of the problem, making easier its interpretation. It is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences.

The use of different probes in the GRID-PCA analysis, for the calculation of different types of enzyme-substrate interactions (hydrophobic, electrostatic, H-bonding), gives a very detailed three-dimensional quantitative picture of the catalytic differences of a number of correlated enzymes and the ease of information interpretation is due to the PCA statistical engine.

This information can be exploited to understand phylogenetic correlations among enzyme superfamilies and to design strategies to engineer existing scaffolds.
3.2.5. Aim of the work

Enzymes in their simplicity, in respect to organisms or species, can be a great topic to study evolution related mechanisms. The effects that the selective pressure engages in enzymes, in terms of insertions, deletions or mutations, translate into exquisite catalytic efficiency and selectivity. This fact indicates the necessity of looking at the enzyme's evolutionary path when searching for the interconversion of catalytic activities. On the other hand, it also makes the engineering of enzyme properties a delicate task and the perception of enzymes far from the canonical “just another catalyst” point of view.

The superfamily of α/β-hydrolases is a good example of divergent evolution with its components belonging to many distinct classes, catalyzing different chemistries. Nevertheless, promiscuous activity is to be expected in ancient, primordial components of the superfamily. The hydroxynitrile lyase and the esterase activity were the point of interest of this chapter, together with the study of the relationships between amino acid sequence, folding, catalytic machinery and activity.

The clarification of structural and biochemical features of HNLs and esterases was the main scope of the work described in this chapter. Developing the focal concepts responsible for the appearance of distant catalytic mechanisms in enzymes sharing the fold and the active triad can be of great utility to understand the nature's course in diverging a common ancestor and applying the same concepts to generate artificially evolved enzymes with a rational design.

This scope was put into practice through the development of a rationally engineered mutant of *M. esculenta* HNL which possesses all the features identified as crucial for esterase activity. This fact indicates that the correct bioinformatic, modelling and statistical tools were used for the selection of the right candidates, together with the feasibility of the approach.
3.3. Results and discussion

As mentioned in the introduction, several representatives of both families (HNLs and esterases) share a common tertiary structure, belonging to the α/β-hydrolase superfamily, and a very similar catalytic machinery (Ser-His-Asp triad) yet catalyzing two very distinct chemical reactions through different reaction mechanisms. Structural, as well as physico-chemical differences of the active sites are therefore inevitable.

Given the common origin of structurally related HNLs and esterases, the comparison of their active sites and a careful analysis of their mechanisms of action should allow to convert the two functionalities. Through the introduction of the proper characteristics (i.e. through mutations) into an appropriate scaffold we can artificially simulate the nature's divergent course in developing new, specialized enzymes. The concepts obtained from such a study can be later exploited to design new or more efficient enzymes with properties different from their natural ones that could serve not only for the introduction of reaction promiscuity, but also substrate promiscuity or enhanced enantioselectivity.

Figure 3.6. Proposed mechanism of *H. brasiliensis* hydroxynitrile lyase.
From the reaction mechanism point of view, HNLs with α/β-hydrolase fold have initially been thought to act through the accepted hydrolase mechanism, in the form of a covalent intermediate of the substrate (aldehyde or ketone) with the side chain of the active serine. However, subsequent structural studies on the complexes of the *H. brasiliensis* enzyme with substrates and inhibitors together with modelling studies shed doubt on this hypothesis. In the revised mechanism (fig. 3.6), residues of the catalytic triad act as a general acid/base with the central histidine residue as the active species. Both experimental and modelling studies ascribed a pivotal role to an active site lysine residue, whose positive charge stabilizes the complementary negative charge that is built up on the cyano group. This electrostatic stabilization facilitates cyanohydrin cleavage and, in the reverse reaction, the deprotonation of the hydrocyanic acid necessary to increase its nucleophilicity. Ultimately, a side chain of a conserved threonine residue has been shown to play an important role in the process of catalysis. In respect to covalent catalysis encountered in many other α/β-hydrolases (e.g. esterases), the HNL active site lacks the presence of a real oxyanion hole consisted of two/three hydrogen bond donors. The abovementioned threonine residue has been proposed to perform a similar role based on the superposition of an HNL and an acetylcholinesterase, but it is unlikely to suffice for the stabilization of an oxyanion in case the HNL would be screened for esterase activity. The lysine residue present in the HNL active site is also a hurdle for the occurrence of esterase activity due to its rather bulky and charged side chain. Anyway, a detailed comparative analysis of the active sites belonging to the two classes of enzymes is indispensable for making such considerations.

### 3.3.1 Data set selection, sequence and structural similarity

Given their limited number (Figure 3.4), all available crystal structures of hydroxynitrile lyases belonging to the α/β-hydrolases were downloaded from the protein data bank (PDB) together with their sequences (*H. brasiliensis* HNL, *M. esculenta* HNL and *A. thaliana* HNL). FAD-dependent enzymes (*Sorghum bicolor* HNL) and metalloenzymes (*Linum usitatissimum* HNL) were not taken into account due to their more complex catalytic mechanisms. In order to identify the esterases that could be used for the comparison of the active sites of the
two classes, a BlastP search in the EBI SRS server\textsuperscript{41} has been done. Since the structural similarity was the key factor for the comparison of the enzymes, bioinformatic tools had limited effectiveness in the study. Nevertheless, a sequence similarity search followed by sequence alignment had to be performed to identify candidate enzymes possessing esterase activity. The databases were searched for esterases homologous to HNLs of known crystal structure despite the risk of encountering enzymes with high sequence similarities, but different folding and therefore little structural similarities. Eventually, this was not the case.

The search yielded dozens of possibly related esterases based on the sequence similarities to the HNLs. The results of the search were then further refined, selecting esterases of available crystal structure, possessing an $\alpha/\beta$-hydrolase folding. Finally, six esterases were selected for further analysis based on their structures and catalytic mechanisms, shown in Table 3.1. In this case as well, cofactor dependent enzymes or metalloenzymes were avoided. Interestingly, most of the esterases resulting from the search were enzymes belonging to bacteria, ancient organisms in the evolutionary time scale.

Table 3.1. Details of selected esterases after the BlastP search, folding and mechanism verification.

<table>
<thead>
<tr>
<th>PDB entry</th>
<th>Resolution (Å)</th>
<th>Chemistry</th>
<th>UniProtKB</th>
<th>Organism</th>
</tr>
</thead>
<tbody>
<tr>
<td>1esc</td>
<td>2.10</td>
<td>Esterase</td>
<td>P22266</td>
<td>Streptomyces scabies</td>
</tr>
<tr>
<td>1ci9</td>
<td>1.80</td>
<td>Carboxylesterase</td>
<td>Q9KX40</td>
<td>Burkholderia gladioli</td>
</tr>
<tr>
<td>1evq</td>
<td>2.60</td>
<td>Carboxylesterase</td>
<td>Q7SIG1</td>
<td>Alicyclobacillus acidocaldarius</td>
</tr>
<tr>
<td>1va4</td>
<td>1.80</td>
<td>Arylesterase</td>
<td>P22862</td>
<td>Pseudomonas fluorescens</td>
</tr>
<tr>
<td>3fcx</td>
<td>1.50</td>
<td>Esterase</td>
<td>P10768</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>1r1d</td>
<td>2.00</td>
<td>Carboxylesterase</td>
<td>Q06174</td>
<td>Bacillus stearothermophilus</td>
</tr>
</tbody>
</table>

Structures of all esterases were retrieved from the Protein Data Bank and a comparative study with the hydroxynitrile lyases was done by means of structural alignment. The superposition itself was performed on the basis of the conserved catalytic triad as well as by superposing the $\alpha$-carbons of the two (or more) candidates, thus estimating both the comparability of the active sites and the overall similarity. Finally, the overall structural identity was evaluated by calculating the RMSD (alpha carbons) of the superposed couple as well as by visual inspection. After a careful analysis of every possible superposition, four enzymes were selected for the study of active
site/mechanistic similarities between HNLs and esterases, namely HNL from *Hevea brasiliensis* (HbHNL), HNL from *Manihot esculenta* (MeHNL), arylesterase from *Pseudomonas fluorescens* (PFE), and carboxylesterase from *Alicyclobacillus acidocaldarius* (EST2), Figure 3.7.

**Figure 3.7.** The three-dimensional structures of single domains of the four selected enzymes. A neat superposition of the core structure (mostly beta sheets) is evident; the surrounding helices show more structural variability.

The multiple sequence alignment, shown in Figure 3.8, calculated an extremely low sequence identity (1.613%) between the four enzymes and only 48 similar positions. This fact indicates that the enzymes are very distantly related which is also evident from their taxonomies and the phylogenetic tree shown below in Figure 3.9. The two HNLs are not very distant on the evolutionary time scale (blue branches) since they differ only in genre and species, whereas the bacterial esterases are both distant from each other (host organisms belonging to different *Phyla*) as well as from the HNLs (different Kingdoms).
Figure 3.8. Multiple sequence alignment of the selected proteins. The first two chains belong to the HNLs, the second two to the esterases. The fifth row indicates the degree of similarity: "*" for identical, ":" for conserved substitutions, "." for semi-conserved substitutions.

Figure 3.9. Phylogenetic tree of the selected enzymes. The order is the same as in Figure 3.7. Green lines represent bacterial proteins, blue lines represent eukaryotic ones.
It is well known that both protein and RNA structure is more evolutionarily conserved than sequence,\(^4^2\) thus structural alignments can be more reliable for discovering homologous traits between sequences that are very distantly related and that have diverged so extensively that sequence comparison cannot reliably detect their similarity. As a matter of fact, the selected enzymes for the present work, show an extremely well conserved core folding (Figure 3.7) and an overall structural similarity value of 5.385 Å calculated as the RMSD of the enzymes backbones (Figure 3.10).

The structure based phylogenetic tree shown in the Figure 3.10, has been built according to the calculated RMSD values after the superposition. Interestingly, the relationships of the structural similarities between the enzymes respect the evolutionary relationships and sequence similarities shown in Figure 3.9. In other words, the sequence-based phylogenetic tree and the structure-based phylogenetic tree are very similar, despite the structural analysis indicated a much higher degree of similarity than the sequence alignment. The HNLs are structurally closely related, while the esterases show more distinctions. PFE is however more similar to the HNLs (RMSD of 4.76 and 4.32 Å) in respect to EST2 that shows higher RMSD values after superposition (6.36 Å and 6.89 Å).

**Figure 3.10.** The left part of the figure shows the RMSD (Å) values between the backbones of the superposed enzymes with color-coded legend on side. On the right, the structure based phylogenetic tree of the four enzymes. The numbers on the Y axis are expressed in Ångström.
3.3.2. GRID-PCA analysis and differences identification

After the structural superposition during which the common catalytic triads were constrained to share the same position in the Cartesian space, the GRID maps of the active sites were calculated. Four different probes were selected, namely water (OH2), hydrophobic probe (DRY), H-bond acceptor, carbonyl C=O (O::) and H-bond donor, amide N-H: (N1). Thus, the MIFs contained the information about the four active sites' physico-chemical properties, accounting for hydrophobic, electrostatic and H-bonding interactions.

Statistical treatment of the outcome data and principal component analysis have been used to point out the most relevant structural differences emerging from the calculated molecular interaction fields, following the GRID-PCA scheme. The study filtered out the hundreds of small, not significant differences that are obviously correlated to such a diverse ensemble of low sequence identity enzymes and allowed to identify a few structural determinants of esterase-HNL differentiation.

The GRID-PCA proceeds independently for each MIF, then the outcome differences for each MIF are combined together with an appropriate weight function to compensate the intrinsically different absolute values of each interaction type (i.e. enthalpy of van der Waals interactions is 4-5 times lower than H-bonds). Subsequently the statistical analysis points out the most relevant of the identified differences, which can be three-dimensionally visualized together with the protein structures.

For clarity the discussion is divided in hydrophobic and hydrophilic interactions.

Hydrophobic interactions

The esterases' active sites appear to be significantly more hydrophobic than the HNLs ones, which are on the contrary distinctly hydrophilic. In particular the entrance of esterases' binding pockets show some remarkable hydrophobic zones, not present in the HNLs which are determined by a number of phenylalanine residues and other hydrophobic residues.

Figure 3.11 represents these differences as they appear after the GRID-PCA analysis. GRID-PCA identifies the MIFs' grid nodes where the differences are statistically significant and visualize them in their original three-dimensional coordinates. In the figure, they appear as small red cubes.
The MIF for the corresponding probe is visualized as isopotential surface, light blue in the figure. It can be noted that the hydrophobic MIF covers completely the red cubes in the Esterase active site (left in Figure 3.11), while there is no trace of hydrophobicity in correspondence of the red cubes in HNL active site (right in Figure 3.11). The visual analysis of the area pointed out by GRID-PCA can easily identify the aminoacids correlated to the differences, information that can be used to engineer the active site properties.

In particular, in the esterase by *A. acidocaldarius* (EST2), that shows the highest degree of hydrophobicity, the amino acids mostly responsible for the differentiation from the HNLs are Phe30, Leu36, Phe37, Phe214, Phe287. Figure 3.12 reports a different graphical representation of the GRID-PCA result for the hydrophobic interactions.

**Figure 3.11.** The same perspective of the EST2 (left) and the MeHNL (right) active sites for comparison. The red cubes are the grid nodes of the hydrophobic MIF pointed out by GRID-PCA as statistically significant in the structural difference between the enzymes. EST2 is evidently hydrophobic, while MeHNL is more hydrophilic. The zone responsible for the difference is mostly at the entrance of the active site pocket.
Figure 3.12. EST2 enzyme's structure. The residues identified by GRID-PCA as responsible for the hydrophobicity of the active site entrance are emphasized in balls and sticks rendering and molecular surfaces colored according to lipophilicity (magenta for hydrophilic, green for lipophilic). The arrow indicates the entrance to the active site.

**Hydrophilic interactions**

The same logic used for the analysis of the differences in terms of active site hydrophobicity was applied for the hydrophilic interactions, based on the MIFs generated by the H-bond donor, H-bond acceptor, and Water probes.

The analysis pointed out that the inner parts of the active sites of either esterases and HNLs are distinctly hydrophilic without notable differences in terms of geometrical organization of the interactions pattern in the binding pockets.

Molecular interaction fields generated by hydrophilic probes set at a low value of interaction energy are a good representation of the active site shape. This is how the differences among
esterases and HNLs appear clear. As it can be seen in Figure 3.13, HNLs show a globular active site, while esterases’ pocket appears to be more funnel-shaped, longer and narrower.

Figure 3.13. On the left the esterase PFE, on the right the HbHNL for comparison. The HNL shows a globular, almost spherical hydrophilic interaction space within the active site, while the same molecular interaction fields in the esterase’s active site shows a significantly different shape. This is the most important structural difference pointed out by principal component analysis.

The corresponding relevant residues in the esterase from *P. fluorescens* are Trp28, Phe125, Phe158, Phe162, Phe198, Ile224 (see Figure 3.14). Some of them correspond to the residues determining the hydrophobicity of the active site entrance, identified above in Figure 3.11 for esterase from *A. acidocaldarius* (same residues, but different numbering).
Figure 3.14. The structure of the PFE enzyme with the residues responsible for the difference in the shape of the active site, in the comparison with HNLs. Clearly visible the funnel-shaped hydrophilic region of the pocket determined by the surrounding hydrophobic residues highlighted in balls and sticks and corresponding molecular surfaces colored by lipophilicity.

Summarizing the GRID-PCA results, it is possible to conclude that:

- Esterase’s active site is significantly more hydrophobic at the entrance, which is also wider
- The inner parts of either esterases and HNLs are both hydrophilic
- The shape of the active sites of the two classes is significantly different. HNLs show a globular active site while the esterases’ pocket is funnel-shaped, longer and narrower.

These results are in accordance with the type of substrates accepted by the two classes of enzymes. As a matter of fact, both HNLs hydrolyse acetone cyanohydrin as natural substrate which is a rather small, globular and polar molecule. On the other hand, the natural substrates of
selected esterases are carboxylic or aryl esters which are hydrophobic, quite big and flexible molecules.

3.3.3. Rational identification of hot-spot for conversion of HNL into Esterase and *in silico* mutant generation

The initial idea of converting one of the selected esterases into a hydroxynitrile lyase was aborted after a similar work by Kazlauskas *et al.* was published, therefore the opposite route was preferred.\(^{43}\)

The candidates to be used as templates for such computational enzyme engineering experiment were EST2 and MeHNL, resulted to be the most similar in terms of active site similarity (with the best superposition of the active triad), GRID-PCA and general compatibility with ester substrates (for the HNLs). Moreover, the EST2 enzyme has a shorter evolutionary history that classifies it into the ancient enzymes category. This could lead to a higher probability to encounter a natural reaction promiscuity, as discussed in Section 3.2.2 and, accordingly, mimicking its active site properties should allow to approach an esterase activity with a lower trade-off and by fewer mutations.

The results of the comparative study were used to develop four concepts as a strategy to introduce esterase activity into the *M. esculenta* hydroxynitrile lyase:

1. Increase the size of active site entrance
2. Create a funnel-shaped active site in order to accommodate bulkier substrates
3. Induce the generation of an oxyanion hole that could stabilize the tetrahedral intermediate for an efficient covalent catalysis
4. Mutate the Lys 237 residue into a neutrally charged residue

A close-up of the wild-type MeHNL's active site is shown in the figure below (Figure 3.15). The snapshot was taken from the inside, placing the active site serine in evidence. The narrow entrance channel is visible on the back, whereas the acetone cyanohydrin is present in the middle of the active site. The clearly globular shape of the active site is also evident, as well as its buried position.
Several different mutants were produced following the rationale described above. All the generated mutants were subjected to a 1 ns long molecular dynamics with explicit solvent (water and ions) to relax the structure and verify the effect of the mutations on the backbone of the enzyme.

It was immediately evident that for increasing the active site entrance tryptophane 128 had to be changed into a smaller residue since most of the entrance was occluded by its bulky side chain (Figure 3.15). Another key point was the elimination of the HNL’s catalytically important lysine, but the appropriate mutation was not as straightforward. Different mutations were tried and it was seen that when mutated to a much smaller residue (alanine or glycine), the overall structure of the active site was severely changed. The best solution for this position was a Lys237Leu mutation, since the shape of the introduced leucine was able to replace the steric properties of the mutated lysine (limiting the conformational freedom of the respective active site portion), allowing at the same time to loose the undesirable positive charge.
The introduction of an oxyanion hole was the most difficult task. The GRID-PCA results were not able to discriminate the two enzyme classes on the basis of the presence/absence of the oxyanion hole. Therefore, from this point of view, no information was available on how to proceed. The only possible way was to mimic the oxyanion present in the active site of the esterase. Therefore the structure of every generated and equilibrated mutant was superposed to the esterase’s structure to evaluate the good or bad superposition of the components of the oxyanion hole. Different mutations, as well as deletions, were tried. Apart from the superposition with the esterase, the goodness of the mutants was checked by docking of several commercially available esters, used in esterase activity screening tests (Table 3.2). The aimed result was the successful docking of all the compounds into the mutant's active site.
Eventually, the Thr11Gly mutation gave satisfying results, since the elimination of the threonine side chain allowed the amino group of a neighbor residue (Ile12) to act as one component of the oxyanion hole. The other component was the nitrogen atom of cysteine 81. The quality of this mutant was initially checked by comparing it to the EST2 and by docking. Later, the docked
substrates were covalently linked to the catalytic serine to simulate the reaction transition state through the construction of the tetrahedral intermediate. All the enzyme-substrate complexes were further relaxed in a short MD simulation (100 ps) to check the persistency of the oxyanion stabilization. The resulting appearance of the active site is shown in Figure 3.16.

The best results were obtained for a four-fold mutant of MeHNl, i.e. W128A, L153A, T11G and K237L (Figure 3.16). The mutant has all of the requirements for showing esterase activity.

First of all, the active site entrance was transformed from a blocked substrate channel to a wide-open cleft. The importance of mutating tryptophane 128 has already been shown by a study of Effenberger et al. in 2002. The group generated a single mutant Trp128Ala and studied its crystal structures in the presence and absence of the substrates. The resulting structural modification caused higher accessibility of the active site to bulkier substrates.44

Secondly, the entrance of the active site now contains several hydrophobic amino acids, in accordance with the esterase’s features and results obtained from the GRID-PCA. Three of them are highlighted in Figure 3.16 by dark green color, i.e. Leu121, Phe125 and Leu179. The active site shape is successfully altered as well, appearing as a wider, funnel shaped environment.

Finally, the oxyanion hole has been successfully engineered. The mutant shows excellent results with seven out of eight substrates. The S conformer of substrate 7 was however not compatible with the mutant's active site as the docking simulation was unable to calculate any productive conformation of the substrate. This might indicate an R-specificity of the mutant, but further verifications of this fact should be done.
Figure 3.16. Active site of MeHNL mutant W128A, L153A, T11G, K237L. The substrate (S1 in Table 3.2) is covalently linked to Ser80 and highlighted in balls and sticks mode. The mutated residues are shown in dark red color, whereas residues responsible for the active site hydrophobicity are shown in dark green. Hydrogen bonds are depicted as black dotted lines.
3.4. Experimental section

The crystal structures of all used enzymes in different stages of the work were retrieved from the Protein Data Bank (PDB). The amino acid sequences of HNLs were also downloaded from the PDB in fasta format. The sequences were then used to perform a BlastP search in the EBI SRS server.

The results were filtered on the basis of the folding and reaction mechanisms, structures and sequences of interesting esterases were downloaded from the PDB. The multiple chain alignment as well as structural superposition were both performed using Molecular Operating Environment (MOE) version 2009.10 version, with the align and superpose algorithms using the Blosum 30 amino acid substitution matrix with a tree-based method.

Once the four candidate enzymes were selected, the UNIPROTKB Align tool was used to perform an alignment of their sequences. The sequence identity percentage and the phylogenetic tree were calculated as well using the tools available at the UNIPROT web page.

After this compatibility verification, the proteins were further refined by removing water molecules, ions and inhibitors from their structures. In case of dimers or tetramers, only single domains were used for the backbone superposition and the superposition of the active sites. These operations were all performed using MOE software.

GRID calculations were performed on the proteins using a 4560Å³ box completely surrounding the superimposed active sites, with a grid spacing of 0.5Å. OH2, DRY, O:: and N1 probes were used for calculating the molecular interaction fields. For the PCA, a data pretreatment was performed by selecting a cutoff to positive variables, by setting to zero all variables with an absolute value lower than 0.01 and having a standard deviation below 0.03.

The mutants were generated using the mutate tool of MOE. The energy minimization and the molecular dynamics simulations were performed using the GROMACS 4 software with the GROMOS-96 53a5 force field. The mutant structures were solvated with explicit SPC water model in a cubical box of 10 nm side. All the MD simulations were performed in an NPT environment, with a simulated temperature of 300K. The pressure was kept constant using the Berendsen thermostat. For the minimization, the PME (particle mesh ewald) algorithm was used for the calculation of the electrostatic interactions.

The output structures of the dynamics simulations were used as the template for performing the
molecular docking. Every mutant was protonated according to a neutral pH value with the Protonate3D module of MOE since the structures presented only polar hydrogens. All the substrates were manually sketched in MOE and used as substrates for the docking which was performed using the AlphaPMI placement algorithm and the Affinity dG algorithm for the scoring. After placement, the poses were further refined with a forcefield minimization (MMFF94x forcefield).

For the best mutant, the best scored placement of every substrate was used to generate the tetrahedral intermediates by generating a covalent bond between the Oγ of Ser 80 and the esteric carbon atom with the builder module of MOE. The charges of the tetrahedral intermediates were adjusted and an initial minimization of the catalytic serine and the TI was done. A second minimization of the whole system was performed using the AMBER99 force field and a gradient of 0.05. The short molecular dynamics (300 ps) of the enzyme-substrate complexes were performed in MOE in an NPT environment and an implicit treatment of the solvent by implementing a dielectric constant of 80 for water. The temperature was set to 300K and the used forcefield was AMBER99.
3.5. Conclusions

It is certainly true that α/β-hydrolase fold enzymes accept a variety of substrates differing in structure, chemical composition and phisico-chemical properties. The fact that this superfamily of enzymes shares a common folding but has completely different catalytic properties indicates its divergent evolution. From the reaction point of view, however, many enzymes of this superfamily show a promiscuous behaviour and this fact can be exploited to artificially introduce reaction promiscuity by rational enzyme engineering.

In this work, a hydroxynitrile lyase has been modified by in silico rational engineering with the aim of transforming it into an esterase. Four specific mutations of the active site yielded a mutant enzyme showing all the prerequisites for ester hydrolysis. The spatial and phisico-chemical properties of the wild-type active site have been changed and the oxyanion hole has been introduced. This way, the type of reaction mechanism was modified as well, switching from an acid/base catalysis to a covalent one.
Chapter 3

13. L. P. Wackett, *Curr. Opin. Microbiol.*, 2009, 12, 244–251
41. www.uniprot.org
4. Development of an automatic tool for *in silico* enzyme engineering

4.1. Summary

The multitude of available simulation techniques that are currently being used in the field of computational biocatalysis are very different in terms of philosophy, level of theory, computational demand and the way input and output files are handled, formatted and produced. These facts significantly limit the integration of different approaches and therefore the exploitation of the potential residing in merging them, which would translate into much faster development cycles of new, improved biocatalysts. The tool developed in this part of the work allows to handle inputs coming from different software and can be seen as a hybrid environment. *modeFRONTIER* software has been used to develop different automated workflows that combine different software to generate, simulate and score series of enzyme mutants. The introduction of amidase activity into the *C. antarctica* lipase B scaffold has been taken as the reference study. Two different strategies for scoring the quality of the mutants have been developed and implemented. One of them is based on the 3D-QSAR approach, the second is a stabilization of the transition state for amidase reaction through H bond pattern analysis. Consequently, two separate methods for designing the generation of the mutants were conceived and tested. One is particularly useful for reducing the dimensionality of the problem and can be regarded as a virtual evolution; the approach is based on genetic algorithms. The second method was a factorial approach suitable to screen problems limited to few dozens of possible designs. Although some difficulties were encountered in the development of the 3D-QSAR scoring, both final workflows were successfully designed. The workflow based on the H bonding type of scoring allowed to project a triple CALB mutant bearing the necessary characteristics for the introduction of the amidase activity.
4.2. Introduction and aim of the work

The present chapter describes the development of a new, innovative tool for *in silico* protein engineering and screening. It is well known that the development cycles of new and improved enzymes are too long, and can be as long as 15 years.\(^1\) This problem can be partly ascribed to an incomplete knowledge of biotechnological techniques necessary for the large scale production of the enzymes but substantial efforts, in terms of amount of resources and time, are also spent in the initial phase of the development cycle, *i.e.* for the engineering of the desired properties. Therefore, in order to shorten those extremely long time lines, fast design strategies for production and screening of mutants are necessary.

In this respect, computational approaches have a huge advantage over the experimental strategies for generation and screening of mutants (described in section 1.2. of the introduction). A rational treatment of the problem together with the capability to study the engineering at atomic detail, allow the development of concepts and strategies that are beyond the reach of the experimentalists:

- rational design of efficient biocatalysts to be produced through engineering
- fast and efficient *in silico* screening of available enzymes/mutants to fully exploit catalytic potential of existing biocatalysts
- providing quantitative parameters for describing enzyme’s efficiency
- fast substrate-screening and rational substrate engineering
- understanding molecular basis of biocatalyst’ action and properties

The idea behind the work described in this chapter was the design of a framework able to exploit the integration of several different modelling approaches; molecular mechanics (MM) methods, such as docking and molecular dynamics, QM-QM/MM methods, such as *ab initio* and semiempirical calculations, and optimization strategies based on 3D-QSAR concepts. The overall strategy can therefore go beyond modern state of the art computational techniques offering a hybrid environment based on concepts coming from different levels of theory, both directly and indirectly.
Moreover, the developed computational infrastructure contains all the different steps for biocatalyst (more specifically mutants) design, modelling and scoring. It is extremely flexible and can be easily tuned based on the specific case study. Thus, several problematics can be implemented, ranging from substrate and reaction engineering, to enhanced enantioselectivity or increased thermostability.

Another fundamental advantage of the developed method is the ability to choose the appropriate way to simulate the generation and scoring of mutations, that can be either focused on the automation of the procedure or on the virtual evolution of the system. The former is particularly useful when the mutations that want to be performed and tested are known (deriving from rational conceptualization or previous knowledge) and are not many in number. In such cases, the tool can significantly reduce the implicit time costs of the simulations, due to its automatized nature and the limited requirements of intervention by the operator.

The virtual evolution of the system can, on the other hand, be very useful in cases when a high number of possible hotspots are of interest and a saturation mutagenesis-like procedure wants to be adopted, *i.e.* every possible mutation could be of interest. In such cases the problem can have an extremely high dimensionality since only 5 hotspots translate into $5^{20}$ possible mutants. It is obvious that not all the possibilities can be tested since it would require huge efforts (computational power and time costs). Fortunately, with the use of appropriate optimization algorithms (e.g. genetic algorithms) the dimensionality of the problem can be significantly reduced. The advantage of relying on an optimization algorithm is to generate an environment able to learn, generation after generation, the correlation between mutations and their efficiency. By the end of the work, the parental enzyme is endowed with a specific property of interest through the simulation of a virtual evolution of the system.
4.2.1. The case study

The selected case study for the generation of this automatic tool was the introduction of amidase activity into *Candida antarctica* lipase B. The topic of reaction promiscuity and related problematics have been extensively described in chapter 3 and are not going to be discussed again here. However, in respect to the engineering done in chapter 3, the present case is much more challenging.

In fact, the differences in catalytic properties of lipases and serine proteases are not as yet clear since both classes share not only the fold (α/β-hydrolases), but also the catalytic machinery as well as the reaction mechanism. However, proteases are able to catalyse both amide and ester hydrolysis while lipases generally do not cleave amides.

The catalytic machinery of lipases consists of a serine, histidine and an aspartate/glutamate together with an oxyanion hole generally formed by backbone amide protons. This spatial arrangement is topologically very similar to those found in serine proteases which have a mirror image arrangement of the catalytic triad compared to lipases,\(^2\) as visible in Figure 4.1.

![Figure 4.1. Catalytic triad of a lipase (on the left) and a serine protease (on the right)](Image)
Given the high similarity of the overall spatial arrangement of catalytically important residues, it is not surprising that the two classes of enzymes show an almost identical catalytic mechanism. In fact, lipases catalyze the hydrolysis of esters via the same double displacement mechanism through an acyl-enzyme intermediate as observed within the serine protease family. Figure 4.2 shows the reaction mechanism of *C. antarctica* lipase B.

![Reaction mechanism of C. antarctica lipase B.](image)

**Figure 4.2.** Reaction mechanism of *C. antarctica* lipase B.

The catalytic triad of CALB consists of Asp187, His224, and Ser105 while the oxyanion hole is formed by the backbone amide protons of Thr40 and Gln106 and the side chain of Thr40. First, a substrate reversibly complexes to the free enzyme (Figure 4.2, top left) thereby forming a Michaelis-Menten complex. After correct positioning of the substrate, a nucleophilic attack of Ser105 onto the substrate carbonyl group occurs and a first tetrahedral intermediate is formed (Figure 4.2, top right). In this tetrahedral intermediate, the negative charge on the former substrate carbonyl oxygen is stabilized by three-fold hydrogen bonding interaction with the oxyanion hole, whereas the positive charge on His224 is stabilized by the interaction with Asp187.
Subsequently, proton transfer from His224 to the substrate alkyl oxygen happens and the alcohol part of the residue is liberated from the enzyme. As a result, a covalently bound acyl enzyme intermediate is formed at the end of the acylation step (Figure 4.2, bottom right). Subsequently, the acyl enzyme intermediate is deacylated by an incoming nucleophile (R³OH), which is generally water, an alcohol or an amine. A second tetrahedral intermediate is formed by the attack of the nucleophile onto the acyl enzyme carbonyl group (Figure 4.2, bottom left). In this process, the proton is transferred from the nucleophile to the His224 residue and the positive and negative charges are effectively stabilized by Asp187 and the oxyanion hole, respectively. Then, the proton is transferred from the His224 residue to the Ser105 alkyl oxygen while restoring the carbonyl bond of the bound substrate. As a result, a weakly bound enzyme-product complex is formed and the free enzyme is regenerated after release of the reaction product.³

Serine proteases’ mechanism of action has been extensively studied and is analogous to the one just described for CALB. Nevertheless, there are some additional facts in the case of amide hydrolysis that has to be considered. All proteases must overcome three obstacles to hydrolyze a peptide bond: i) amide bonds are very stable due to electron donation from the amide nitrogen to the carbonyl. Proteases usually activate an amide bond via the interaction of the carbonyl oxygen with a general acid, and may also distort the peptide bond to disrupt resonance stabilization; ii) water is a poor nucleophile; proteases always activate water, usually via a general base; and iii) amines are poor leaving groups; proteases protonate the amine prior to expulsion. Serine proteases perform these tasks very efficiently: the rates of peptide hydrolysis by serine proteases are ~1010-fold greater than the uncatalyzed reactions. Obviously, these mechanisms of catalysis are not confined to peptide hydrolysis; serine proteases also readily hydrolyze other acyl compounds, including amides, anilides, esters, and thioesters.

Figure 4.3 displays the generally accepted mechanism for chymotrypsin-like serine proteases.⁴ The same described double displacement mechanism is evident.
Despite these indisputable similarities, esterases and amidases show clear differences towards the hydrolysis of amides. A better knowledge about the optimal environment for amide (or ester) hydrolysis is therefore important for the explanation of such behaviour. In other words, the physico-chemical properties of the active sites cannot be neglected.

Following these considerations, a parallel study was performed by our group (unpublished results) on an elevated number of serine hydrolases; lipases (10 different enzymes), esterases (11 different enzymes), proteases (9 different enzymes) and general amidases (11 different enzymes).

The work was concentrated on the description of the differences encountered between the four classes, in terms of active site properties, with the main focus on the hydrophobic/hydrophilic balance of the active sites.

The results showed specific characteristics related to the catalysed reaction and a clear distinction between lipases/esterases and proteases/amidases active sites in following terms.

1. Lipases active sites are characterized by:
   - Large and continuous hydrophobic areas
   - Low hydrophilic/lipophilic ratio and few, small hydrophilic regions
   - Distribution of the hydrophilic and lipophilic regions compatible with the amphiphilic nature of their natural substrates
2. Proteases/amidases active sites are characterized by:

- High hydrophilic/lipophilic ratios
- Several, well distributed strong interactions with water, indicating that the active sites are more hydrated as compared to lipases
- Much more extended Hydrogen bonding capabilities (compared to lipases)

Due to these intrinsic properties lipases are generally more stable in harsh operative conditions such as higher temperatures and anhydrous organic solvents which make them more suited for industrial applications. Amidases generally require higher amounts of water in the reaction media with a minimum water activity coefficient of around 0.7 and tolerate lower temperatures. For this reasons, endowing a stable lipase scaffold with amidase activity would be of great interest for creating a completely novel enzyme family with features much more suited for an industrial use. Moreover, the engineered enzyme(s) would probably have different substrate specificities in respect to naturally occurring amidases since the engineered lipase's active site should accept and clave more hydophobic amides.
4.3. Results and discussion

The automation/evolution tool was developed using a preexisting software offering a multi-objective optimization and design environment written to couple computer aided engineering tools. The software is called modeFRONTIER and has originally been developed to cope with mechanical engineering research problems by solving numerical equations. The optimal solution is found through modifying the values of the input variables, to find one or several optimal solution for the desired output. Only recently the software started to be used in the field of life sciences. It has never been applied for an enzyme engineering purpose which implicated the necessity to conceive a general strategy for *in silico* protein engineering. The initial work was based on the design of a single strategy for automation and integration of different software required to perform a computational enzyme engineering task, to obtain a ready to use method for high-throughput screening of mutants. Different strategies were finally developed and tested, all sharing the same conceptual workflow, Figure 4.4.

Figure 4.4. A general scheme of the concepts of the optimization cycle. Actions are grouped into three categories; those performed by modeFRONTIER (purple), those performed by all the integrated modeling software (yellow) and design specific parameters, determined by the user (red).
The general computational strategy (Figure 4.4.) for designing a new enzyme mutant is based on the following steps:

- Identification of hotspots
- Choice of amino acid change for each hotspot
- Protein modeling: creation of the mutant model
  - Mutation and placement of side chains
  - Calculation of protonation states
  - Calculation of the effect of mutation(s) on backbone conformation through MD, molecular docking, etc.
- Scoring of the enzyme mutant. Some scoring examples are:
  - Calculation of the interaction with substrates measuring the BFE (binding free energy)
  - Simulation of transition states and calculation of correlated properties (kinetic parameters, enantioselectivity, etc.)
  - Superposition to crystallographic structures measuring the RMSD value, etc.

An overview of all the different steps will be made, describing the advantages offered by the developed strategy together with possible drawbacks. The steps comprise the variables selection, the generation of the objects (mutants), the integrated molecular modelling simulations and the scorings of the mutants together with the objectives of the mutant design. In the scoring section, the case specific scorings will be described in detail. The focus will be both on the rationale behind the selected scorings and on the description of the development procedures. Finally, a detailed description of two different workflows will follow.
4.3.1. The variables

In a protein engineering study, the variables are obviously the amino acidic residues composing the protein. The variable identity is the position of the residue selected for the mutation (e.g. position 24), whereas the variable associated value is the specific mutation decided for the respective position (e.g. alanine). The starting point of the simulation is an enzyme structure that wants to be engineered. Usually, the above mentioned steps are made manually, by means of many different software tools. The manual intervention inevitably limits the number of models to be generated, thus limiting dramatically the exploration of the experimental space, which is huge also in case of a relatively small number of selected hotspots. It has to be noted that in the case of 4 hotspots, the number of possible mutants is 1.6x10^5, for 8 hotspots it raises up to 2.56x10^10. Moreover, the manual generation of a big number of models is operatively tedious and time consuming. The developed tool is a framework that can be considered as a container where software and methods can converge and operations are automatized. This fact significantly reduces the time costs. The dimensionality of the problem is a major constrain for the “brute force” approach as well, where all the possibilities are automatically generated by the computer and then screened for scoring their properties in silico. For instance, even if it was possible to make the whole cycle of mutation-modelling-scoring in one second, it would still be necessary to invest almost 1200 days for screening 10^10 mutants, corresponding to 8 hotspots only. The best possible solution to the dimensionality problem is looking for an optimization algorithm, able to reduce drastically the number of mutants to be generated, while still being able to hit the design objective.

4.3.2. The generation of the objects

This step is peculiar for the selected case study. The developed tool offers the possibility to select the appropriate way of generating the mutants by the implicit modeFRONTIER modules. The DOE module (Design Of Experiments) gives the possibility to select the logic by which the initial mutations will be picked; it can be a completely random approach (no preference for the
introduced mutation) or it can follow an ordinate schedule of wanted mutations. Moreover, even
in case of the random generation, the operator is free to select whether to consider all 20 possible
mutations or just a smaller group. In cases when only few (2-4) hotspots and few mutations are of
interest, a full factorial design can also be considered.

Figure 4.5. A workflow in its appearance inside modeFRONTIER. The logic flux (operations)
unfolds horizontally (from left to right), the data flux vertically.

Figure 4.5 exemplifies the appearance of a workflow inside the modeFRONTIER software. On
the horizontal dimension the logic flux is represented, while the vertical dimension represents the
data flux. The logic flux is the ensemble of all the operations that are made for each generation
cycle and obviously the logic flux relies on the input data and generates the output data, which
together constitute the data flux.

In cases when performing an *in silico* evolution of mutants is of interest, the software has several
optimization algorithms that can perform the evolution based on the output variables (scoring).
For protein engineering tasks, genetic algorithms are the best suited optimization procedure. On
the other hand, when the tool wants to be used only for its automation advantages, the
optimization part can be skipped and only mutants specified in the design of experiment will be generated and scored.

4.3.3. Simulations

Simulations comprise all the steps that are implemented for the mutation, modelling and scoring of the mutant in each generation (optimization) cycle and the integration of the different software required to perform the different steps. This part is also extremely flexible as it can be tuned in many ways, depending on the type of information available and the desired method to simulate/measure the effect of the mutation on the mutant structure and on its activity. Obviously, more detailed and complex are the simulations involved, longer will be the time required for each cycle. This issue is however not of secondary importance since the number of mutants to be generated and scored is potentially very high. For example, if a docking step is involved, the size of the grid (Figure 4.6) for substrate placement has to be selected carefully.

**Figure 4.6.** Three different grid sizes used for the same docking procedure. The length of the grid side is 10Å in the left image, 12Å in the central image and 15Å in the right image.

The placement of the substrate will be more accurate if a bigger area of the active site is considered for docking. Nevertheless, a difference of only 2Å in grid side can translate into days of extra calculations when hundreds of mutants are screened. Therefore, a correct balance between accuracy and time costs has to be found, which requires experience in the field.
One aspect that cannot be neglected or underestimated is modelling the conformation of side chains in proteins. Knowledge of protein structure is of prime importance in the effort to understand their function. In fact, the investigation of the correct structural relationships among the elements of a protein is of high importance to assign them a precise role which is one of the critical steps in the interpretation of the effect of amino acid substitutions.

The simulation of side chain conformations is critical as the introduction of a mutation can have important impacts on its structure and function (depending on the physico-chemical properties of the residue and its position). While the protein backbone is formed by a standardized and iterative sequence of a small number of atom types covalently bound to each other, side chains are constituted by different atoms and their conformation is modulated through non-covalent interactions. These interactions are dependent not only upon the sequence, but especially upon the spatial proximity of other residues. Furthermore, interactions with other elements such as the solvent or ligand may affect residue conformation. Finally, side chains generally have higher conformational freedom than the backbone.

During time, several approaches have been developed to predict side chain conformations. Usually, to solve the side chain packing problem, the arrangement of side chains conformations that minimizes the value of particular force field energy is sought. To do this, a good software for side chain modelling should perform the mutagenesis on the following criteria:

- a rotamer library
- an algorithm to perform the conformational search
- an energy function to find the best conformation

The introduction of rotamer libraries is one of the most widely adopted strategies to reduce the combinatorial problem of placing side chains on a backbone structure. A rotamer, short for “rotational isomer”, is a single side chain conformation represented as a set of values, one for each dihedral angle degree of freedom. Because bond angles and bond lengths in proteins have rather small variances, they are usually not included in the definition of a rotamer.

A rotamer library is, on the other hand, a collection of rotamers for each residue type. Rotamer libraries usually contain information about both the conformation and the frequency of a certain
conformation. Often, libraries also contain information concerning the variance about dihedral angle means or modes, which can be used in sampling.

Rotamer libraries can be backbone-independent, secondary structure-dependent or backbone-dependent. The distinctions are made according to whether the dihedral angles of the rotamer and/or their frequencies depend on the local backbone conformation or not. Backbone-independent rotamer libraries make no reference to backbone conformation and are calculated from all available side chains of a certain type. Secondary structure-dependent libraries present different dihedral angles and/or rotamer frequencies for \( \alpha \) helix, \( \beta \) sheet or coil secondary structures. Backbone-dependent rotamer libraries present conformations and/or frequencies that are dependent on the local backbone conformation, as defined by the backbone dihedral angles.\(^7\)

All three types of rotamer libraries are popular and widely used. In his review, Dunbrack suggests that backbone-independent libraries should be used for applications such as X-ray and NMR structure refinement, the development of entropy scales and for representing properties of unfolded states of proteins. Instead, backbone-dependent libraries are more suitable for predictions of side chain conformations in homology modelling or in protein design since the explicit dependence of dihedral angles from backbone conformation can avoid extensive sampling of the conformational space resulting in less expensive calculations.\(^7\)

It has been shown that only for two residues (Asp and Asn) there is a strong backbone-dependency of their conformation due to their strong and specific interaction with the local environment.\(^8\) Another study showed that the performance of the backbone-dependent library was significantly better to predict side chain conformation of Asp, Asn, Ser, Val, Thr and Pro, whereas the advantages were less significant for other residues.\(^9\)

Following all these consideration, the developed tool for automatic mutant generation and screening relies strictly on backbone-dependent rotamer libraries.
4.3.4. Scoring and objectives

The final step of *in silico* mutant generation/optimization is the measurement of their quality in respect to the sought properties, *i.e.* the scoring. This step requires perhaps the most difficult choice since the rates of biocatalysed reactions cannot be directly measured by computational techniques.\textsuperscript{10} Alternative scorings has to be found and they have to be appropriate for the measurement of the desired mutagenesis effect. Moreover, the implicit requirements of modeFRONTIER to work on numbers do not allow the scoring to be a qualitative information, thus restricting the possible choices.

For the selected case study, the modification of a lipase into an amidase, three different scorings were initially identified and tested. First of all, the docking of amide substrates into the mutant's active site was considered. Docking is a suitable scoring tool because a general fit between the newly generated active site and a substrate of interest can be tested. Furthermore, the output is nicely represented as the free energy of binding (BFE), it can be easily parsed by modeFRONTIER and incorporates the information about the quality of the mutant; it is not just a yes or no answer.

On the other hand, considering solely docking results for scoring the mutants introduces a high possibility of measuring the wrong effect. In fact, molecular docking calculations account only for the binding of the substrate and are not able to detect the catalytic efficiency. On one hand this means potentially active mutant could be missed by the calculated BFE value, on the other hand, an extremely good BFE result could mean that inhibition might occur. Either way, this scoring method introduces a risky bias to the mutants screening and brings along the risk of identifying both false negatives and false positive hits.

The second scoring was an in house developed three dimensional quantitative structure activity relationship (3D-QSAR) model. A good 3D-QSAR model can be valuable for understanding the physico-chemical properties of enzymes, and a guide for understanding which changes due to mutations are correlated with the increase of CALB amidase activity and which are not.

3D-QSAR has proven its validity under many different aspects, predicting biocatalytically important properties. Braiuca *et al.* reported studies on substrate selectivity,\textsuperscript{11} enantioselectivity,\textsuperscript{12} and thermostability\textsuperscript{13} that could be quantitatively predicted by QSAR models based on training
sets with a small number of enzymes and their corresponding experimentally measured properties.

The fundamental concept of 3D-QSAR is the application of multivariate statistics to find the correlation between structural features of a set of molecular models and their functional properties. The structural features must be described quantitatively by appropriate molecular descriptors, that can be based on the GRID approach (see sections 3.2.4), as in the works of Braiuca et al.

GRID-based 3D-QSAR models are suitable as scoring tools in a modeFRONTIER generation/evolution workflow. Their advantages as scoring tools are fivefold: they give numerical predictions; account for virtually any kind of interaction within the active site; do not calculate direct interaction energy between specific substrates and the enzyme active site; can be applied to any enzyme mutant structural model; are computationally inexpensive.

For the present study an equivalent, GRID based, approach was used to calculate a 3D-QSAR model starting from a preexisting set of CALB mutants and their associated amidase activities. The data set was based on in vitro produced mutants and their experimentally measured amidase activity, coming from a collaboration with Novozymes A/S. The results are summarized in Table 4.1 (unpublished data). These data could be used to understand the structural characteristics of the active site correlated to the amidase activity, to generate a predictive model suited for the scoring of in silico produced mutants, inside a modeFRONTIER workflow.
Table 4.1. Data set used for the generation of the 3D-QSAR model. The amidase activity is normalized against the CALB wild-type.

<table>
<thead>
<tr>
<th>Mutant</th>
<th>Amidase activity ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>A225E</td>
<td>0.32</td>
</tr>
<tr>
<td>A225F</td>
<td>0.75</td>
</tr>
<tr>
<td>A225I</td>
<td>1.55</td>
</tr>
<tr>
<td>A225K</td>
<td>1.99</td>
</tr>
<tr>
<td>A225L</td>
<td>1.19</td>
</tr>
<tr>
<td>A225M</td>
<td>1.11</td>
</tr>
<tr>
<td>A225V</td>
<td>1.22</td>
</tr>
<tr>
<td>A281S</td>
<td>0.45</td>
</tr>
<tr>
<td>A283K</td>
<td>0.69</td>
</tr>
<tr>
<td>D145S</td>
<td>0.69</td>
</tr>
<tr>
<td>D223G</td>
<td>1.26</td>
</tr>
<tr>
<td>D223K</td>
<td>0.86</td>
</tr>
<tr>
<td>D223N</td>
<td>1.15</td>
</tr>
<tr>
<td>D223R</td>
<td>0.79</td>
</tr>
<tr>
<td>D265P</td>
<td>0.57</td>
</tr>
<tr>
<td>F71R</td>
<td>0.78</td>
</tr>
<tr>
<td>G39A</td>
<td>2.76</td>
</tr>
<tr>
<td>G39A L278A A281G</td>
<td>1.67</td>
</tr>
<tr>
<td>G39A Q191R</td>
<td>1.49</td>
</tr>
<tr>
<td>G39A W104F</td>
<td>4.15</td>
</tr>
<tr>
<td>G39S</td>
<td>0.55</td>
</tr>
<tr>
<td>I189D</td>
<td>0.09</td>
</tr>
<tr>
<td>I189E</td>
<td>0.1</td>
</tr>
<tr>
<td>I189H</td>
<td>1.11</td>
</tr>
<tr>
<td>I189N</td>
<td>0.95</td>
</tr>
<tr>
<td>I189Q</td>
<td>0.41</td>
</tr>
<tr>
<td>I189S</td>
<td>0.08</td>
</tr>
<tr>
<td>I189T</td>
<td>0.08</td>
</tr>
<tr>
<td>I189Y</td>
<td>1.14</td>
</tr>
<tr>
<td>L140E</td>
<td>0.55</td>
</tr>
<tr>
<td>L278A</td>
<td>2.5</td>
</tr>
<tr>
<td>Q191R</td>
<td>1.28</td>
</tr>
<tr>
<td>R249Q</td>
<td>0.74</td>
</tr>
<tr>
<td>T103A</td>
<td>1.1</td>
</tr>
<tr>
<td>T103G</td>
<td>3</td>
</tr>
<tr>
<td>T256K</td>
<td>0.75</td>
</tr>
<tr>
<td>T40A</td>
<td>0.12</td>
</tr>
<tr>
<td>T40G</td>
<td>0.13</td>
</tr>
<tr>
<td>T40S</td>
<td>0.5</td>
</tr>
<tr>
<td>T42G</td>
<td>1.87</td>
</tr>
<tr>
<td>T42S</td>
<td>1.21</td>
</tr>
<tr>
<td>W104F</td>
<td>2</td>
</tr>
<tr>
<td>Wild-type CALB</td>
<td>1</td>
</tr>
</tbody>
</table>
The mutant structures were produced in silico starting from a CALB wild-type structure which was preliminarily equilibrated and simulated by molecular dynamics for 6 ns. Each mutant structure was generated by a mutagenesis software relying on backbone-dependent rotamer libraries (see previous section) and subjected to 300 ps of molecular dynamics. The simulations were all performed in explicit water (Figure 4.7), to relax the mutant structure of any possible strain given by the introduction of new residues. Moreover, the scope of the introduced MD steps was to evaluate the effect that a given mutation has on the overall conformation of the enzyme, backbone included. As a matter of fact, even when the placement of side chains is based on a rotamer library, it was seen that a short molecular dynamics simulation performed after the mutagenesis is useful for a correct conformational sampling, as the mutation usually causes small deviations of the backbone structure.

Figure 4.7. The explicit treatment of solvent. The figure shows an enzyme in water, arranged around the protein in a cubic box.

After the completion of the simulations a clustering analysis of the generated trajectories was done. The meaning of such approach is to evaluate the overall flexibility of the protein (given by the number of clusters present) and to obtain the conformation that the enzyme adopts during
most of the simulation time (the energetic minimum).

All the generated mutants showed only few conformational clusters present during the simulation and the most populated cluster had a much higher number of members in respect to other clusters. This fact was a clear indication that all the systems have reached an energetic equilibrium. Moreover, working with a mutant structure belonging to the mostly populated cluster ensures that the mutation effects are simulated correctly and the extracted conformation is actually a good representation of the protein's energetic minimum. For performing a 3D-QSAR, this is an issue of prime importance.

The energetic minimum conformations for all mutants and the wild-type CALB were collected, superposed and used as input structures for the 3D-QSAR model. The superposition was performed on the backbone $\alpha$-carbons. The first step of the QSAR was the calculation of the GRID-based molecular interaction fields (MIFs) between the mutants and 3 different probes; the DRY probe to map the hydrophobic regions, the N1 probe to map the hydrogen bond acceptors and the O probe to map the hydrogen bond donors. Once the MIFs have been calculated, these data were used as input for the statistical analysis.

The variables were pretreated by means of a variable grouping algorithm called SRD (smart region definition). This algorithm performs the grouping of variables that represent the same chemical and statistical information and are neighboring in the 3D space. An initial number of seeds is selected (number of groups to extract) and only variables having the highest correlation with the activity are maintained. Then the variables around the seed are maintained based on the cutoff distance selected. In this procedure all the variables that are far away from any of the seeds are discarded and therefore only important variables are kept. Following this grouping procedure every operation on the variables acts on the generated groups and not on single variables.

The grouping of variables is important in case of structure based QSAR analysis, especially when the entire protein structure is considered, as in this case. The grouping allows to treat areas of the protein instead of single variables. This allows to focus the analysis on the areas where the effective differences are located, eliminating scattered, isolated variables spread around the protein that might result from a classical variable pretreatment procedure. In fact, keeping the treatment of the variables as single points would make the model very difficult to interpret since most of the identified variables would result as a defective superposition of the objects.
After variable pretreatment, used to remove the groups of variables that were non informative or showed a low correlation with the experimental activity, the final predictive model contained roughly 10000 variables and showed a correlation coefficient ($r^2$) of 0.89 and a predictive coefficient ($q^2$) of 0.71 on the third principal component.

**Figure 4.8.** Correlation plots (experimental activity vs. predicted activity) of three calculated models. A: Full data set, B: Full data set, logarithmic activity range, C: Data set with 4 removed outliers, logarithmic activity range.

The predicted versus the experimental activity plot (Figure 4.8, A) shows a bad correlation of the data, even if the predictive coefficient is high. As a matter of fact, the activities of the bad mutants are overestimated whereas the activities of the good mutants are underestimated.

To overcome this problem, the activity values have been expressed as the logarithmic values of the experimentally measured data, but the observed trend in the correlation plots was still present (Figure 4.8, B). In respect to the first model, it appears clear that the objects having the lowest activity values are outliers of the model (the four right-most objects of the B figure) and since they are also the mutants which experimental values contain the highest experimental error, they were removed and another model was calculated. The removal of the outliers however did not solve the problem of the prediction (Figure 4.8, C).

The reason for the bad performances of the model can be found in the experimental data. The training set is characterized by an amidase activity ratio, normalized against the wild-type
activity. Although this is formally correct, the problem resides in the low absolute values of amidase activity for all the mutants of the data set. Wild-type CALB is able to reach 0.12\% of conversion after 18h and the whole data set has a range of activity ratio between 0.1 and 4. This means that the worst mutant hydrolyses 0.01\% of the reference amide in 18h and the best mutant of the series reaches less than 0.5\%. With such a small level of hydrolysis the precision of experimental measurements is critical. The experimental error in this system is significantly higher than usual because of the low reactants' concentration variation and it is big enough to cover the actual variability of the data set.

What the 3D-QSAR is actually trying to correlate is consistently faded into experimental error, therefore the predictivity cannot be satisfactory and there is nothing that can be done computationally to get real improvement. A further refinement of the variables would bring to a much higher predictive coefficient and apparently allow to explain and predict the measured amidase activity; it would however be only due to overfitting.

A much bigger range of amidase activity would be necessary to solve the existing problem and make the 3D-QSAR scoring method feasible. A similar experimental set of data is unfortunately not available, thus impairing the application of this strategy.

Despite the bad performances of the model, it has been decided to use it as a scoring tool. The purpose was to test the methodological feasibility of the approach and the performances of modeFRONTIER's genetic algorithms.

The third developed scoring tool was based on a specific concept deriving from literature data and from the work performed in collaboration with other research groups, i.e. the existence of a hydrogen bond in the transition state of protease/amidase catalysed hydrolysis of peptides and amides. The H bond is donated by the scissile HN group of the substrate whereas the hydrogen bond acceptor can be found either in the enzyme (enzyme assisted) or in the substrate (substrate assisted).\textsuperscript{14}

The rational design of a mutant able to stabilize the transition state of an amide hydrolysis reaction could therefore be the introduction of an H bond acceptor in the CALB scaffold that is not present in the wild-type of the enzyme. The presence of such an additional H bond would certainly lower the activation energy of the reaction, and the hypothesis is that the extra H bond would allow the mutant to cleave the amide bond.

An energy profile diagram for the reaction of amide hydrolysis has been calculated by an \textit{ab}
*initio* geometry optimization of a minimal enzyme model (see section 1.3) and is displayed in Figure 4.9. The calculations were done at the Royal Institute of Technology in Stockholm.

![Energy profile diagram for the acylation step in amide/ester hydrolysis](image)

**Figure 4.9.** Reaction energy profile for an amide hydrolysis. The relative energies of the different steps are indicated on the Y axis.

As it can be seen from the figure, the transition state for the nitrogen inversion (TS\textsubscript{inv}) is the highest energy point along the reaction coordinate. Transition states are not stable chemical species and therefore cannot be simulated by molecular mechanics methods used in the automatic generation of the mutants. For this reason, the corresponding tetrahedral intermediates (TI\textsubscript{1} before the nitrogen inversion and TI\textsubscript{2} with the inverted nitrogen) should be used as models of the transition state. In particular, stabilizing TI\textsubscript{2} by a hydrogen bond with the amide proton would mean lowering the activation energy after the nitrogen inversion when the free electron pair is pointing towards His\textsubscript{224} and the proton transfer can occur.

The size of the system that can be treated by *ab initio* molecular orbital approaches is very limited, although very accurate. Calculations on active site models can, on the other hand,
examine interactions between groups at the active site, and can provide useful models of transition states and intermediates. These model have to be treated by semiempirical or density-functional theory methods, as discussed in section 1.3. For the present case study, the amide hydrolysis reaction with a semiempirical treatment has been calculated at the University of Copenhagen. The study yielded, among other important information, the transition states geometry both before and after the nitrogen inversion (TI1 and TI2 in figure 4.8). The results are presented in Figure 4.10.

![Figure 4.10. Optimized geometries of the wild-type CALB active site and tetrahedral intermediates before (left) and after (right) the nitrogen inversion. The NH group of the substrate is circled in red.](image)

These calculations allowed to identify the position of the active site where an enzyme assisted hydrogen bonding might occur after appropriate mutations. Glycine 39 was individuated as the best candidate for a mutagenesis experiment, as visible in the right-hand picture of Figure 4.9. Following the above mentioned rationale, the generated lipase mutant able to cleave amide bonds has to be able to stabilize the transition state for nitrogen inversion by forming a hydrogen bond with the substrate amidic NH group. These considerations were used as a scoring strategy. The scoring was based on the verification of this phenomenon by counting the number of H bonds present in the mutant's active site.
4.3.5. Workflow description: automation and evolution

In this section the developed workflows (cycles) inside modeFRONTIER will be described. Integrated software, simulations, time expenses, results and possible drawbacks will be discussed. The distinction between an automation based approach and a virtual evolution will be described through real examples of two different cycles that use different modes of mutants generation. Finally, the integration of the two scoring methods discussed in the previous section will be described, as the concrete correlation with the selected case study of endowing a lipase with an amidase activity.

4.3.5.1 3D-QSAR scoring

The first developed workflow was based on the following rationale:

- Concept: reproducing the physico-chemical properties of an amidase
- Scoring: 3D-QSAR model
- Objective: maximize the amidase activity predicted by the model

As mentioned in the scorings section, the 3D-QSAR model did not show the desired performances. Nevertheless, QSAR models can be of relevant importance for enzyme engineering once a reliable data set exist because of the statistical treatment of the structural descriptors. This would allow to highlight hidden, coexisting effects of multiple mutations and structural features positively correlated with the activity that are otherwise impossible to individuate and predict. For these reasons an evolution cycle based on 3D-QSAR would be a powerful tool for enzyme engineering. Therefore, despite the poor predictivity of the specific lipase/amidase model, a workflow integrating the calculated 3D-QSAR model has been developed. The scope was testing the feasibility of the approach as well as developing a ready-to-use tool once a better QSAR exists.

The statistical model serves as a predicting tool; every generated mutant is treated as a new object of the model and an external prediction of its amidase activity is calculated. Generally, in order to
obtain a reliable prediction, the new object has to be simulated the same way as the other components of the existing model. This meant that the same simulation steps used for the QSAR generation had to be incorporated in the automatic procedure. Therefore the integrated simulation steps and software were the following:

1. Generation of the mutant by SCAP software
2. Energy minimisation by GROMACS software
3. 300 ps long MD simulation by GROMACS software
4. Clustering and calculation of the energy minimum mutant structure by GCLUSTER module of GROMACS
5. Superposition to the objects of the QSAR data set by PYMOL software
6. Calculation of GRID descriptors by means of UNIX shell scripts based on GRID software
7. External prediction of the mutants amidase activity by in house developed UNIX shell scripts based on GOLPE software (and the existing 3D-QSAR model)

The initial structure of the enzyme was the CALB wild-type simulated by molecular dynamics for 6 ns. The mutants were produced using the same *in silico* mutagenesis software (namely SCAP) used for the generation of the QSAR model (see previous section), followed by 300 ps of molecular dynamics in explicit water and the clustering procedure. The energy minimum structure for each mutant was obtained and the GRID molecular descriptors were calculated using the same probes as before (DRY, N1, O). The variable pretreatment was also the same as before. After the completion of all these steps, the new mutant was projected on the calculated model and the resulting predicted activity was recorded and used as the output variable of the workflow. The objective of modeFRONTIER was to maximize this precise value.

All the above described steps were integrated into the workflow and performed automatically by modeFRONTIER; no human intervention was required once the generation-simulation-scoring of the mutants have started.

Five different input variables were selected, *i.e.* position 39, 104, 42, 103 and 278. Some of them were hotspots showing beneficial effect upon mutation in the existing experimental sets of CALB mutants; the others were identified by visual inspection of the enzyme-substrate complex.

The objective of the cycle was a virtual evolution, so the idea was starting from a random situation and let the system to evolve spontaneously using a kind of evolution algorithm. No
limitation on the possible mutations was imposed, so the number of possible solutions for the problem was $5^{20}$. The DOE consisted of 20 randomly selected mutants chosen by the software modeFRONTIER, whereas a genetic algorithm (50 generations) was used to perform the optimization of the activity. This way each generation consisted of 20 mutants, that were evaluated, selected and mixed following a genetic evolution scheme.

![Diagram of genetic algorithms](image)

**Figure 4.11.** A general scheme of genetic algorithms

The virtual evolution performed by genetic algorithms usually starts from a population of randomly generated individuals and happens in generations. In each generation, the fitness of every individual in the population is evaluated, multiple individuals are stochastically selected from the current population (based on their fitness), and modified (recombined and possibly randomly mutated) to form a new population. The new population is then used in the next iteration of the algorithm. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population (Figure 4.11).

The calculations performed in the workflow required around 35 minutes for each cycle. Most of the calculation time is spent on the conformational analysis (energy minimization and molecular
dynamics). To keep the time expenses as low as possible, the conformational analysis could be launched on external clusters or workstations which would significantly cut the simulation times. However, the preliminarily runs of the automatic workflow confirmed the feasibility of the approach as well as the huge flexibility and utility of modeFRONTIER. Some designs of mutants failed, mostly due to the high internal energy generated by the introduction of specific mutations which were not sufficiently minimized using the parameters implemented for the minimization procedure. This fact caused the MD simulation to crash. The modeFRONTIER cycle is designed not to be aborted in case of simulation failure and this specific case is simply not taken into account. Changing the minimization parameters would surely allow to overcome this inconvenience.

The discussion of the results has no scientific relevance, as the 3D-QSAR based scoring was not efficient enough to let the system to evolve correctly. Nevertheless, the evolution cycle could be completed since it represents a robust framework for similar case studies, once a predictive QSAR model can be obtained.

4.3.5.2 Hydrogen bond scoring

The second workflow was based on the following rationale:

- Concept: introduction of a H bond acceptor
- Scoring: number of H bonds in the active site
- Objective: maximize the number of H bonds

The screened network of hydrogen bonds consisted of the catalytic triad, the oxyanion hole and the amide proton. This way, the scoring can take into account the overall quality of the mutant in terms of catalytically indispensable interactions plus the occurrence of the enzyme assisted hydrogen bonding for the stabilization of the transition state.

The design of the workflow has evolved through 5 different versions.
The first version of the workflow was based on genetic algorithm and the integration of only 2 different softwares:

1. SCAP for generating the mutant structures
2. CHIMERA for the detection of the H bonding pattern

Two different input variables (hotspots) were selected by visual inspection of the optimized geometry of the tetrahedral intermediate inside CALB's active site, with the aim of introducing the H bond of interest. The selected positions were 39 and 104. Every possible mutation has been considered, basically to test the software's performances, as well as the efficiency of the genetic algorithm (MOGA-II).

Once the choice of the mutation has been done by the DOE (initially) or MOGA-II (in optimization cycles), the input files for SCAP are generated and the mutant is produced. After that, Chimera handles the scoring by counting the number of existing H bonds and generates an output file with the number of these bonds. The objective of the project was to maximize this precise number. Results are sent to the optimization algorithm which consequently decides which is the next mutant to be produced and tested. This type of implementation takes around 30 seconds for a complete cycle.

Different runs of the cycle have been carried out for testing the software settings. Initially some SCAP parameters had to be tuned in order for the mutant to be produced i.e. the type of rotamer library had to be changed. The medium size library (initially tried) was too small in some cases and the mutant could not be produced. Switching to the large library solved the problem.

Nevertheless, the results were unsatisfactory. Since the starting system is a covalent enzyme-substrate complex that is screened for the introduction of a single H bond, the effect of all mutations produced an identical result in terms of the H bonds that were formed. More specifically, the placement of the newly introduced residues adapts to the conformation of the existing tetrahedral intermediate without modification of the enzyme backbone, and without interacting with the catalytically essential H bond network. The number of detected H bonds was therefore always the same. This inconvenience reflects back to the genetic algorithm which is unable to learn the effect of specific mutations, repeating unsuccessful designs several times.
This problem was obviated by introducing some molecular dynamics steps prior to the count of the number of H bonds. The second version of the workflow therefore needed to integrate 3 different softwares:

1. SCAP for generating the mutant structures
2. MOE for protonation and MD
3. CHIMERA for the detection of the H bonding pattern

MOE software was used for the MD calculation. Specific scripts were written to perform the protonation of the mutant and the subsequent MD. Molecular dynamics simulations relax the conformation of the generated mutants and take into account the effect the mutations have on the enzyme backbone and on the catalytically essential H bond network. This way the scoring of the mutant is no longer based on the generation of a single additional H bond, but is rather the number of the existing H bonds that can go from zero to six.

Due to the molecular dynamics simulations expense in terms of computational power and time, these calculations had to be done on an external workstation and not on the notebook initially used for the automatic generation and screening of mutants. The software used, modeFRONTIER, has the possibility to handle such need, as specifically developed nodes are available for the user within the software.

The performed MDs were 300 ps long for each generated mutant. It was assured that the simulation time is long enough to relax the conformation of the generated mutant through the stabilization of the potential energy of the system. As previously described, some newly introduced residues in the mutant can have significant effects on the backbone and the overall structure of the protein which are visible only if the generated strains are relaxed and the conformational freedom is explored through MD. Given that the analyzed mutants differ for a small number of mutations (two or three) from the wild-type and that the initial wild-type structure is extensively relaxed, the simulation length of 300 ps is appropriate. Moreover the rotamer library based algorithm for generating the mutants takes into account the backbone conformation producing an optimally energy minimised structure.
This second version cycle was based on the same hotspots Glycine 39 and Tryptophan 104 and genetic algorithm MOGA-II. All the amino acid substitutions were allowed.

The analysis of the results indicated that in some cases a suitable H bond acceptor in position 39 interacts with Threonine 42 through electrostatic interactions as shown in Figure 4.12. When this phenomenon occurs, the residue in position 39 is no longer in the correct position to stabilize the transition state of the amide. Therefore this undesirable interaction had to be obviated.

**Figure 4.12.** Tetrahedral intermediate of the amide substrate in G39S mutant’s active site. The Serine in position 39 (black carbon atoms) is forming H bonds with Threonine 42 and is no longer pointing towards the amide proton.

For this reason, Threonine 42 had to be mutated into a residue with aliphatic sidechain that would no longer be able to establish the interactions shown in figure 4.12 with the residue in position 39. In *C. antarctica* lipase B, Threonine 42 is placed in a highly structured, inner zone of the enzyme's structure; therefore only small amino acids could be considered for mutation, such as alanine, glycine and valine.

The third round of mutant generation and screening was then based on the hotspot 39 and 42, limiting the possible substitutions for each position.
The following residues were chosen:

- Position 39: Asp, His, Ser, Glu, Gln, Thr, Asn, Tyr
- Position 42: Ala, Gly, Val

The substitutions for position 39 were rationally selected among those amino acids that could stabilize the transition state of an amide hydrolysis through H bonding (polar and charged residues).

Due to the low number of variables and the overall number of mutants to be screened in this case, the genetic algorithm was not used in this case. The simulations were based on a user defined design of experiments which in this case was a full factorial design. It led to the generation and screening of 24 mutants. The focus of the use of modeFRONTIER in the third version of the workflow switched from the virtual evolution to the automation. In other words, the limited number of possible solutions did not require the reduction of the number of experiments to run, but the automation of the simulations allowed substantial time saving.

As in the previous version of the workflow, the scoring of the mutants was based on the number of H bonds present in the active site after the molecular dynamics simulation. The results of the screened mutants are shown in Figure 4.13.
Figure 4.13. The results of the scoring of generated mutants. On x axis the residues in the position 42 and in position 39; on y axis the number of H bonds present in the network.

The results clearly indicate that none of the mutants yielded a good scoring. The highest number of H bonds detected is 4, whereas the complete network of an optimal mutant would require the presence of 6 H bonds. This indication, along with some visual inspection of the MD trajectories showed that the Tryptophan 104 in the wild-type enzyme causes steric clashes with the residues in position 39. For this reason, some of the mutants presented steric hindrance in their structures that caused the molecular dynamics to crash (most mutants with 0 H bonds in figure 4.13). It was deduced that position 104 is a necessary hotspot. The new mutant needed to have more space in the area of Tryptophan 104, but the new residue had to stabilize the orientation of the H bond acceptor in position 39 towards the substrate's proton. 4 amino acids in position 104 appeared to be suitable for this purpose, Ser, Glu, Thr and Asn.

Another important information that emerged from the screening results was the information about the suitability of the 8 candidate mutations for position 39. In other words, the results were used to discard substitutions that were deleterious for catalysis. At this respect, only Asp, Ser, Thr and Asn seemed to be valuable substitutions for Glycine in position 39 and were therefore kept in the next round of screening for amidase activity.
The fourth version of the workflow was then based on 3 input variables corresponding to position 39, 42 and 104. Different mutations were chosen for these hotspots, and particularly the interplay between the mutations was of interest. The following mutations were selected:

- Position 39: Glu, Ser, Asn, Thr
- Position 42: Ala, Gly, Val
- Position 104: Ser, Glu, Thr, Asn

Once again, a full factorial design of experiments was utilized. The appearance of the final version of the workflow is shown in Figure 4.14.

**Figure 4.14.** Scheme of the work flow in its final version. Nodes for handling the MD simulation on an external workstation are visible in the middle of the scheme.

The overall results (shown in Figure 4.15) of this last round of mutant generation and screening were better than the previous for two reasons:

- The choice of reducing steric hindrance in position 104 proved to be correct because it allowed to solve the issue of the faulty MD simulations experienced in the previous
version of the workflow. As a matter of fact none of the performed simulations crashed.

- A higher number of mutants were satisfactory in terms of detected number of H bonds and in particular one of the 45 generated mutants showed the desired 6 H bonds.

![Bar chart showing the number of H bonds for different residues at position 39.](image)

**Figure 4.15.** Results in terms of number of H bonds present in the active site of the screened mutant. On X axis the number of the screened mutant and the corresponding residue in position 39.

From Figure 4.15 it is clear that different residues in position 39 affect differently the geometry of the active site of the mutant, e.g. Threonine in position 39 yielded poor results in terms of H bonds. Following the same principle, it is also quite evident that Aspartate in position 39 is not a good choice. On the other hand, Serine and Asparagine in that position gave some interesting results.

The workflow generated three very interesting mutants, namely those that showed 5 or 6 H bonds in the network. These were further analysed by a 1ns MD. The purpose of this longer simulation was to observe the evolution of the system and check whether the H bond network would be stable during the whole simulation time.
The triple mutant A39S T42V W104E, as well as the A39N T42G W104S maintained the 5 hydrogen bonds in the active site during the whole trajectory, meaning the catalytic properties of the mutant should be intact as compared to the wild-type enzyme. Unfortunately, the stabilization of the amide proton did not occur at all.

The most interesting result of the screening was obtained with the triple mutant G39N T42G W104T, which showed 6 H bonds in the network. Even in this case all the hydrogen bonds were present until the end of the simulation with a regular time of appearance.

A close-up of the active site with the network of H bonds is shown in Figure 4.16. It is possible to notice the H bond stabilization of the amide.

![Figure 4.16. Active site of the best scored mutant. The amide substrate in its tetrahedral intermediate is visible and it is stabilized by a hydrogen bond network (dotted lines). In particular, the amide proton's position is stabilized by a H bond with Asn39.](image)
There are still some aspects that could be improved in the present version of the approach. For example, the output structure of the MD simulation is taken from the last snapshot of the trajectory which can lead to some unwanted biases. It would be better to perform some kind of clustering and select the most represented structure or calculate a mean structure for the simulation.

Nevertheless the approach in its current implementation permitted to predict and generate a good mutant which has all the necessary requirements to act as a good amidase through enzyme assisted hydrogen bonding.
4.4. Experimental section

3D-QSAR models for correlating mutant structure and amidase activity

The preliminary model was generated from a data set of 42 CALB mutants and the wild-type enzyme. The analysis of the structural differences was performed by means of GRID molecular descriptors.

The strategy necessary for this approach consists in 3 main steps:

- Construction of structural models of mutants (mutation, protonation and 300 ps of MD)
- GRID analysis of each mutant and extraction of molecular descriptors
- Statistical analysis

MD conformational analysis of the mutants and wild-type template

The mutants models were constructed starting from CALB wild type structure PDB id: 1TCA.

The structure was treated by the following procedure:

- hydrogen atoms were added
- the structure was minimised
- 6ns MD was run as a conformational search
- the energy minimum structure taken from the trajectory is used as a template for mutations

The software GROMACS 4 was used for all the above mentioned simulation steps. The energy minimization and the molecular dynamics simulations were performed using the GROMOS-96
53a5 force field. The structure was solvated with explicit SPC water model in a cubical box of 8 nm side.

For the minimization, the PME (particle mesh ewald) algorithm was used for the calculation of the electrostatic interactions.

The MD simulation was performed in an NPT environment, with a simulated temperature of 300K. The pressure was kept constant using the Berendsen thermostat.

The same procedure was used for all the mutants generated during the construction of the 3D-QSAR model and during the modeFRONTIER workflow. The only difference is the length of the MD run, 300ps.

**Superposition of enzyme mutants**

PYMOL software was used to superpose each newly generated mutant structure to the wild type template. A specific PYTHON script was developed to automatically launch the superpose PYMOL function without using the software graphical user interface.

**Mutants generation**

The mutants were generated by software SCAP. It generates side chain predictions based on the result of three different components of the search algorithm:

- the rotamer library
- the algorithm to perform the conformational search
- the energy function to find the best conformation

SCAP allows setting different parameters such as the force field to be used, the number of initial conformations to be tested, and the type of rotamer library.
The mutagenesis performed by the software is carried out using the following parameters:

1. Amber force field
2. One initial conformation having the lowest interaction energy with the backbone
3. Minimization of the side chain conformation
4. Large rotamer library

This means that there is an iterative sampling of all rotamers in the library until convergence. The final lowest-energy conformation is then minimized by refining the side chain conformation with 2 degrees rotation on each bond to search for the lower energy conformations around the rotamer.

*Clustering of the MD trajectories snapshots*

The clustering was performed by GROMACS 4 module G_cluster. The following parameters were set:

- GROMOS method
- cutoff value of 0.1nm
- minimum number of structure per cluster: 5.

The rest of parameters were left at their default values.

The output structure was the mean structure of the most populated cluster.

*GRID calculations*

The calculations were performed by GRID version 21, using the following parameters:

- grid dimension: automatic
grid spacing: 1 PLA

Probes: DRY, N1, O

list: -2

**QSAR model generation**

The 3D-QSAR model was generated by GOLPE version 4.5. The variables were pre-treated by the following procedure:

- BUW scaling: default parameters
- max cutoff: 0.0 kcal/mol
- zeroing negative below 0.1kcal/mol absolute value
- SD cutoff: 0.15
- N-level variable selection

Following the generation of PLS with 5 PCs, an SRD variable selection was applied. The criteria for this step were:

- 3rd latent variable
- critical distance: 2Å
- collapse distance: 10Å

Finally a full factorial variable selection procedure was applied.

With the remaining X variables a 3 PCs PLS was generated and validated by LOO procedure.
The use of 3D-QSAR as a scoring tool for modeFRONTIER workflow

GOLPE is not a scriptable software and its output cannot be redirected to a text file that can be parsed by modeFRONTIER for the workflow automation. Moreover GOLPE is not an open source type of software, meaning its code cannot be accessed and modified. To make the generated QSAR a suitable scoring tool for the modeFRONTIER based workflow the PLS equation was extracted from the GOLPE output files and a custom FORTRAN script was developed to calculate the predictions for new object (mutant) and redirect the output to a modeFRONTIER readable text file.

H bond scoring

A batch version of the Chimera software was used to evaluate the quality of the generated mutant by estimating the generation of a specific H bond assumed to be essential for the hydrolysis of a given amide. The software gets the instructions to search, count and record the number of catalytically important H bonds through a specifically developed script in the python programming language to launch the program and redirect the output towards a modeFRONTIER readable text file.

The following software parameters were used to relax the H bonding geometric criteria:

- tolerance: 0.4Å + 20°

MOE based MD

MOE software was used for the MD calculations of the tetrahedral intermediate species. Specific scripts were written to perform the protonation of the mutant and the subsequent MD in UNIX batch mode.

The following parameters were used:

- protonate 3D
Chapter 4

- MD length: 300ps
- AMBER99 force field
- NPT environment
- Implicit water treatment – keeping crystallographic water molecules in the structure to preserve precision in H bonding calculations

The other parameters were set at their default values.

An SVL (Scientific Vector Language) script was used to check the correct charge of the tetrahedral intermediate oxyanion and of the amide nitrogen and to launch the MD in batch mode.

**modeFRONTIER evolution workflow**

Input variables: Hotspots.

The variables were defined as “catalogue variables” with 20 discrete possible values (the 20 natural amino acids). The introduction of a specific modeFRONTIER calculator node for the codification of the AA string values into numerical values usable by evolution algorithms was necessary.

The genetic algorithm used for the evolution workflow was MOGA-II. The DOE for the initial cycle and for all the subsequent generations was made of 20 objects. The initial DOE was randomly generated.

The automation workflow was based on a FULL FACTORIAL design DOE without any evolution algorithm.

The software launched during each cycle is discussed in the results and discussion section and summarized in Figures 4.4, 4.5, and 4.14. The parameters used are the same discussed in the previous paragraphs of the Experimental Section.
The simulations were launched on several different workstations, depending on the installed software and CPUs availability. Different simulation times have been recorded on the different machines directly correlated to their computational power and free resources available at the time of the calculations.

The criteria and objectives for the scoring, evolution and termination of the workflow were different in the different cases and are discussed in the results and discussion section. They were managed by the objective node(s) of modeFRONTIER.
4.5. Conclusions

It has been shown how the automatic approach for mutants generation and screening can be a valuable tool for enzyme engineering. Not only it saves a lot of time and allows to work in an organized, precise way, but it also offers an environment in which all the produced results are collected and can be compared and analyzed. The different stages of the development of the workflow based on the H bonds as a scoring parameter led to the generation of a triple mutant in position 39, 42 and 104 that could not be designed by any other strategy. The approach demonstrated to be useful for the identification of the hotspots and for the choices of the mutations performed.

The 3D-QSAR based workflow could not be successfully applied to the design of new mutants because of the previously described limitations of the specific model. Nevertheless, a fully functional workflow was designed and tested and it can be used when future predictive QSARs will be available.

In summary, the modeFRONTIER workflow approach demonstrated to be a fast and accurate automated design strategy for production and screening of enzyme mutants.
5 ESTECO S.r.l., www.esteco.com
Conclusions

5. Conclusions and future perspectives

The topic of this PhD thesis was the development of computational methods for rational screening and engineering of enzyme properties. It has been shown that different research topics require the application of different computational approaches. The appropriate simulation method has to be found on the basis of available knowledge about the problem as well as on the type of information one is searching for, which is often not trivial even for an experienced scientist. Moreover, in some cases, the appropriate simulation method simply does not exist and has to be conceived upon the comparison of a particular requirement. In this respect, the work performed during this PhD program was based on current, hot research topics in the field of industrial biocatalysis that were treated by state of the art computational techniques.

In Chapter 2 the biocatalysed synthesis of polyesters was treated. The scientific relevance of the topic can be seen from the ongoing research and the number of publications per year by many different research groups all over the world. During this project, a rational redesign of C. antarctica lipase B for the synthesis of polylactides has been treated. The wild-type lipase was seen to poses low catalytic activity for the reaction of interest and was therefore supposed that an engineering of its active site could improve the reaction rate. Two out of three substrates of interest were tested experimentally and the results were unsatisfactory. It was however impossible to demonstrate the reason for the occurrence of this phenomenon, which is out of reach for the experimental approaches. Molecular modelling simulations of the enzyme-substrates complexes allowed to visualize the involved interactions which were quantified by the development of a novel class of molecular descriptors, namely the differential RMDS. It was demonstrated that the descriptors can be used to monitor and quantify steric hindrance due to the intrinsic properties of the studied substrates. The difference between using the empty enzyme and an energetically stable enzyme-substrate system as the reference also emerged. In the case when the RMDS values for the empty enzyme were used as template, the method was not able to highlight the structural distortion caused by the bulky substrate due to its implicit flexibility. When the RMSD values of the enzyme complexed with a well-accepted substrate were used, the effect of the substrates on the active site residues
Conclusions

was easily monitored. This fact together with the visual inspection of the molecular dynamics trajectories allowed to identify three different hotspots for mutagenesis experiments. The validity of the approach was proved indirectly by the work performed at the Stockholm's Royal Institute of Technology. The experimentally produced and tested mutants of CALB were very similar to the ones proposed by the study described in Chapter 2. Two out of three hotspots were the same as well as the proposed mutations. The best of all the generated mutants showed a 83-fold increase of the reaction rate.

In Chapter 3 the issue of reaction promiscuity in α/β-hydrolase fold enzymes was treated. The family is formed by a high number of catalytically distinguished enzymes that show different degrees of reaction promiscuity, demonstrating quite clearly the divergent nature of the family's evolution. The research was based on the introduction of esterase activity into a wild-type hydroxynitrile lyase. The selection of the four enzymes (2 esterases and 2 hydroxynitrile lyases) used for the comparison of their active site properties was based on their structural and sequence similarities together with the consideration of their evolutionary pathways. A GRID-PCA analysis of the active sites was performed and used to point out the structural features mainly responsible for the physico-chemical differences among the two classes. Following the obtained results, the Manihot esculenta HNL was engineered to mimic the structural and biochemical properties found in the active site of a carboxylesterase from Alicyclobacillus acidocaldarius. All the necessary features for the inversion of the catalytic mechanism were introduced and tested by docking procedures and short molecular dynamics of the enzyme-substrate complexes.

Chapter 4 extensively describes the step by step development of a novel tool for enzyme engineering and screening of mutants. The work developed in this part of the project is predominantly methodological. Several computational and theoretical advancements are laid out in the chapter. Innovative scoring tool have been developed and their utility was demonstrated by the ability to integrate them into an automatic workflow for mutant generation and screening. The hydrogen bonding theory of amidase catalytic mechanism has been de facto used for the transformation of Candida antarctica lipase B into an amide cleaving enzyme. The activity of the predicted mutant has not however yet been tested experimentally, nevertheless, the structural model of the mutant has been shown to possess all the postulated requirements for amidase activity.
The strategy has a general validity and can be easily tuned and adapted to different enzyme engineering purposes. The flexibility of the method allows the integration of a virtually infinite number of different softwares, this way the enzyme design concepts as well as simulations can relate to different levels of theory. Certain prerequisites of the platform used for the automation procedure cannot however be bypassed, e.g. any integrated software has to be able to be launched in UNIX batch mode, at least so far.

Currently the experimental generation of several mutants designed and discussed in this thesis is ongoing in collaborations with different research groups in European public and private institutions. The outcome of their characterization (unless unexpected problems related to the production and purification of the mutants) will be an important milestone for the validation and a finer tuning of the computational approaches and theoretical knowledge this thesis was based on.