COMBINING IN SILICO DOCKING AND MOLECULAR DYNAMICS SIMULATIONS TO PREDICT THE IMPACT OF MUTATIONS ON THE SUBSTRATE SPECIFICITY OF BTL2 LIPASE

By

Onur Yükselen

Submitted to the Graduate School of Engineering and Natural Sciences

in partial fulfillment of

the requirements for the degree of

Master of Science

Sabancı University

July 2012

COMBINING IN SILICO DOCKING AND MOLECULAR DYNAMICS
SIMULATIONS TO PREDICT THE IMPACT OF MUTATIONS ON THE
SUBSTRATE SPECIFICITY OF BTL2 LIPASE

APPROVED BY:

Assoc. Prof. Uğur Sezerman   ....

(Thesis Supervisor)

Prof. Dr. Canan Atılgan   ....

Assoc. Prof. Batu Erman   ....

Assoc. Prof. Devrim Gözüaçık   ....

Assoc. Prof. Levent Öztürk   ....

DATE OF APPROVAL: ....

COMBINING IN SILICO DOCKING AND MOLECULAR DYNAMICS
SIMULATIONS TO PREDICT THE IMPACT OF MUTATIONS ON THE
SUBSTRATE SPECIFICITY OF BTL2 LIPASE

Onur Yükselen

Biological Sciences & Bioengineering, Master Thesis, 2012

Thesis Supervisor: Assoc. Prof. Dr. Osman Uğur Sezerman

Keywords: Molecular dynamics, docking, scoring function, BTL2, Bacillus
thermocatenulatus, lipase, specific activity, mutation, specificity, triglycerides

Abstract

Lipases are enzymes that hydrolyze the ester bond between acyl groups and glycerol in triacylglycerides which gives the products of glycerol and fatty acids. *Bacillus thermocatenulatus* lipase (BTL2) has shown highest activity toward tributyrin (C4) as substrate. While broad selectivity on the chain length of the fatty acids has a key role in waste water treatment, and laundry formulations; short chain length specificity can be used in the food and cosmetic industry. In order to predict its chain length substrate specificity (tributyrin (C4)/tricaprylin (C8)) upon mutation, we developed a scoring function which combines *in silico* docking and molecular dynamics tools. After calibration on experimentally validated mutants, our scoring function is able to discriminate substrates specificities and predict the impact of a mutation (whether it enhances or reduces) in a rapid and accurate manner (overall correlation r=0.7930, p=0.0007). Also ranking of substrate specificities within the mutants were 100% correct. This method can be powerfully adapted to other protein families to predict the effect of a mutation for the one specific substrate or multiple substrates.

BİLGİSAYAR ÜZERİNDE DOK VE MOLEKÜLER DİNAMİK SİMÜLASYONLARI BİRLEŞTİREREK MUTASYONLARIN BTL2 LİPAZ SUBSTRAT SPESİFİSİTESİ ÜZERİNDEKİ ETKİSİNİN TAHMİN EDİLMESİ

Onur Yükselen

Biyolojik Bilimler ve Biyomühendislik, Yüksek Lisans Tezi, 2012

Tez Danışmanı: Doç. Dr. Osman Uğur Sezerman

Anahtar Kelimeler: Moleküler dinamik, dok, skor fonksiyonu, BTL2, Bacillus thermocatenulatus, lipaz, spesifik aktivite, mutasyon, spesifisite, trigliserit

Özet

Lipazlar trigliseritlerdeki açil grup ile gliserol arasındaki ester bağını keserek gliserol ve yağ asidi oluşturmaktadır. Bacillus thermocatenulatus lipaz (BTL2) en yüksek aktivitesini tributyrin (C4) substratı üzerinde göstermektedir. Zincir uzunluklarına göre geniş substrat selektivitesi göstermesi atık su arıtımında ve deterjan formüllerinde kullanılırken, kısa zincirlere spesifik aktivite göstermesi gıda ve kozmetik sanayilerinde önemli rol oynamaktadır. Mutasyonların yağ asidi spesifitesine (tributyrin (C4)/trikaprin (C8)) etkisini tahmin edebilmek için dok ve moleküler dinamik araçları birleştiren bir skor fonksiyonu geliştirdik. Skor fonksiyonu, mutant enzimlerin deneysel aktivitesine göre kalibre edildikten sonra, substrat spesifisitelerini ayırt edebilmekte ve mutasyonun aktivite üzerindeki azaltıcı ya da arttırıcı etkisini hızlı ve doğru bir biçimde tahmin edebilmektedir (tüm datanın korelasyonu r=0.7930, p=0.0007). Bununla beraber mutantların kendi içindeki spesifisite sıralaması %100 doğru sonuç vermiştir. Bu metot diğer protein ailelerine uyarlanabilir olmakla beraber, mutasyonun etkisini bir substrat için ya da birden fazla substratı karşılaştırarak tahmin edebilmektedir.

TABLE OF CONTENTS

List of Figures

List of Tables

List of Abbreviations

BTL2: Bacillus thermocatenulatus lipase

E_{Elec}: Electrostatic Energy

E_{vdW}: Van der Waals Energy

MD: Molecular Dynamics

QM: Quantum Mechanics

PBSA: Poisson–Boltzmann Surface Area

GBSA: Generalized-Born Surface Area

FEP: Free Energy Perturbation

TI: Thermodynamic Integration

PDB: Protein Data Bank

Å: Angstrom

RMSD: Root Mean Square Distance

W.t.: Wildtype

CHAPTER I

1. INTRODUCTION

1.1 Motivation

This century is the century of green technology. One of the most challenging issues for green technology is the development of alternative solutions to chemicals that are used in industrial processes. Enzymes are the key organic molecules that can replace the role of chemicals in industry. Since enzymes are naturally found molecules they work efficiently at certain temperature, pH range and are specific to certain types of substrates. It is crucial to tailor enzymes that can solve the needs of industrial processes using protein engineering methods. One of the applications of enzyme engineering is to modify the substrate specificity of native enzymes.

Lipases are an important group of enzymes for biotechnology, as they accept surprisingly wide range of substrates and perform different reactions in various temperature, pH and solvents. In aqueous conditions, lipases hydrolyze the ester bond in triacylglycerides, whereas under micro-aqueous conditions, lipases can do the reverse reaction as esterification, alcoholysis and acidolysis [2,9]. This huge potential provides a variety of biotechnological solutions for the food, dairy, detergent and pharmaceutical industries and make the lipases second largest group of industrial biocatalysts, after proteases [23,24]. One of the key aims of protein engineering for the improved industrial applications is enhancing the specificity of a stable and active enzyme for a particular substrate.

In the interest of predicting valuable mutations, numerous computational procedures have been proposed to predict the enzyme specificity for particular substrate which can accelerate the returns of mutagenesis experiments. Free energy methods such as thermodynamic integration (TI), free energy perturbation (FEP), potential of mean force calculations (e.g. Umbrella Sampling), and steered molecular dynamics (SMD) can

reach relatively accurate free energy values for substrate binding under different conditions, but their computationally expensive nature cannot allow them to be used as common practice in screening of large mutation and ligand libraries [53]. Therefore, computationally inexpensive strategies have been introduced by docking algorithms that provide slightly inaccurate, but fast and simple solutions. Alternatively, the combination of the molecular dynamics and docking methods would give an alternative and effective solution to this problem [53].

We propose an easy-to-implement computational procedure that can predict the impact of mutations to the enzyme specificity for a substrate, where the accuracy of molecular dynamics simulations and high speed of docking algorithms are combined. In order to reduce computational cost of quantum mechanical (QM) calculation for every enzyme-substrate complex, we present an alternative method for predicting substrate specificity. Considering the mean interaction energy, desolvation energy of the enzyme substrate complex and the conformational entropy of the ligand, we were able to predict the impact of mutations to the substrate specificity of BTL lipase.

One goal of the project is the prediction of the specific activity upon a mutation that should be taken to account with complex protein reorganization. This conformational rearrangement of our enzyme examined with molecular dynamics simulations. Besides, initial binding pose of the ligand is another key component of the binding energy, this component requires searching the vast conformational space, and is achieved by using docking algorithm. Overall, our scoring function involves electrostatic energy, Van der Waals energy, desolvation energy and conformational entropy.

Since our application procedure provides a fast and accurate binding energy calculation; it can be used for rational design of enzymes, drug design, or other biomolecule designing purposes.

Bacillus thermocatenulatus lipase (BTL2) has shown the highest activity toward tributyrin (C4) as a substrate. In order to predict its substrate specificity upon triacylglycerides, our computational procedure was used. Our scoring function was calibrated by the specific activity of the wild type and two experimentally verified mutants and finally our scoring function was able to predict the specificity of other three mutants.

2

1.2 Outline

The organization of the thesis as follows: Chapter 2 presents a brief biological background and an overview of the related works. In Chapter 3, we explain our approach in detail. Chapter 4 discusses the experiments and the results. Lastly in Chapter 5, the conclusions and the future works are given.

CHAPTER 2

2. BACKGROUND AND RELATED WORK

2.1. Lipases

2.1.1. Lipase structure and function

Lipases (E.C. 3.1.1.3) are enzymes and in aqueous conditions, hydrolyze the ester bond between acyl groups and glycerol in triacylglycerides which gives the products of glycerol and fatty acids. However, under micro-aqueous conditions, lipases change its reaction to reverse as esterification, alcoholysis and acidolysis.

All lipases are members of the α/ß hydrolase fold family which involves eight hydrophobic ß-sheets at the center and covered with amphipathic α-helices [1,3,4,5]. α/ß hydrolases have conserved catalytic machinery with the consistent geometry that composed of serine, histidine, and aspartic (or glutamic) acid residues on top of the ß-sheet. Histidine acts as a general acid/base catalyst for the nucleophilic reactions involving serine, and aspartic (or glutamic) acid stabilizes the doubly protonated histidine that is formed during the reaction. Catalytic activity is performed in five subsequent steps [6]. Binding of a substrate ester initiates the formation of a first tetrahedral intermediate by the attack of serine on the sp2 carbon atom of the substrate ester. Generated oxyanion is held by hydrogen bonds at the oxyanion hole (Fig. 2.1, transition state). The ester bond is cleaved and the alcohol moiety leaves the enzyme. The final step is the hydrolysis of the acyl enzyme with the aid of water [2,6].

Most of the lipases have a lid structure which is closed in aqueous medium. Because triglycerides are not soluble in water, catalytic activity is performed in water–lipid interface which leads to non-classical enzyme kinetics [20]. The inner surface of

the lid is generally comprised of non-polar residues which covers and protects hydrophobic active site cleft from aqueous solvent [21]. However, the presence of lipid/water interfaces or organic media lead to conformational changes such as displacement of the lid region [7]. Generally, mobile lid region is formed by one or two α-helices whose motions are controlled with flexible structural elements [22]. Lipase activated with interfacial activation as lid of lipase adsorbs at the lipidic interface, and binding site becomes accessible to the substrate [8]. Whereas all lipases accept short (C4) to long chain (C16) fatty acids, only particular lipases able to hydrolyze longer chain fatty acids such as C22 triglyceride [6].
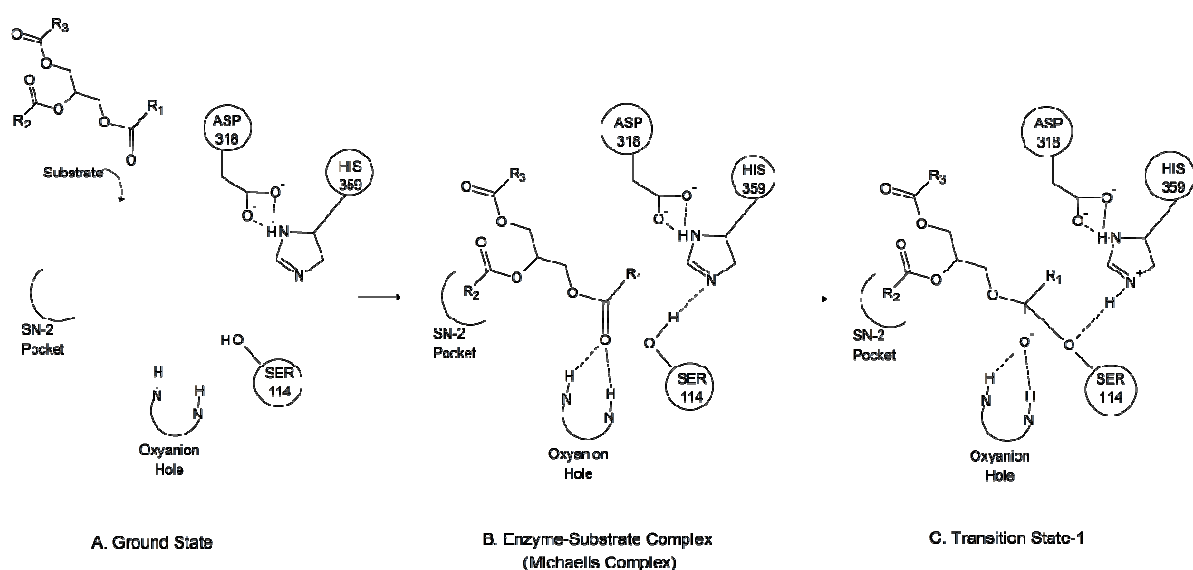
Figure 2.1: First three steps of the triglyceride hydrolysis on BTL2 lipase [6]

2.1.2 Triglyceride selectivity of lipases

Substrate specificity of lipases could be classified into three categories as nonspecific, regiospecific and fatty acid-specific. Nonspecific lipases target any of the ester bonds which ends with complete hydrolysis of triacylglycerides molecule [9]. However, regiospecific lipases target the hydrolysis of 1,3-ester bonds which would give 1,2(2,3)-diacylglyceride and 2-monoacylglyceride. The final group of lipases specifically hydrolyze triacylglycerides of particular length. For instance, Bacillus sp. [10], P. Alcaligenes EF2 [11] and P. Alkaligenes 24 [12] show specificity for long chain length triacylglycerides, whereas other lipases show preference for small and medium chain length triacylglycerides such as lipase from B. Subtilis 168 [13], Bacillus sp. THL027

[14], P. Aeruginosa 10145 [15], P. Fluorescens [16], Pseudomonas sp. ATCC 21808 [17], C. Viscosum [18] and Aeromonas hydrophila [19].

2.1.3 Lipase applications

Lipases are a very important group of enzymes for biotechnological applications, because of their surprising capacity in accepting wide range of substrates and performing different reactions in various solvents. In aqueous conditions, lipases hydrolyze the ester bond in triacylglycerides, whereas in organic solvents, lipases can do the esterification, alcoholysis and acidolysis [2].

Besides their wide range of selectivity (nonspecific, regiospecific and fatty acid-specific) towards particular substrates, their tolerance for the broad range of environmental conditions such as temperature, pH and solvents, and efficient immobilization unlike other enzymes, are the key properties for various types of applications [23]. Therefore, in the food [24], dairy, detergent (in combination with proteases), pharmaceutical industries (fine organic synthesis, racemic mixtures), paper pulp processing, and leather industry, lipases serve various biotechnological solutions [2,9]. The central aim of lipase engineering for the improved industrial applications is the enhanced stability and high specificity for a particular substrate as well as high turnover rate.

For instance, the specificity for the short chain triglycerides can be efficiently used in the production of flavours in cosmetics and food industry. Particularly in food industry, fats and oils are modified in order to achieve higher nutritional value, and improved texture/physical properties, and also bread and cheese are enhanced for better flavor and texture [2]. Broad selectivity for the triglycerides provides advantage for the application areas in waste water treatment and in laundry formulations. On the other hand, in pharmaceutical industry, high enantioselectivity is become crucial factor medical practice. For example, BTL2 shows excellent enantioselectivity (E>100) while performing the hydrolysis of 1-phenylethyl acetate and the acylation of 1-phenylethanol and 1-phenylpropanol with vinyl acetate [2].

2.1.4. BTL2 Lipase

2.1.4.1 BTL2 Lipase structure and function

Bacillus thermocatenulatus lipase (BTL2) consists of 389 residues and has an unusual (α/ß) hydrolase fold which involves seven ß-sheet at the center and encircled with α-helices (Figure 2.3). The crystal structure has two different ions, zinc and calcium. Four residues coordinate each ion and these residues are His82, His88, Asp65, Asp239 for Zn^{2+} ion and Glu361, Asp366, Gly287, Pro367 for Ca^{2+} ion [25].
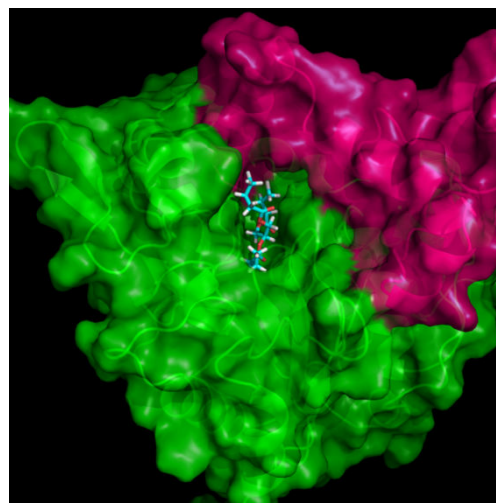
Figure 2.2: Surface of the open BTL2 structure with tributyrin. Lid region (169–239) is colored in purple and the rest of the protein in green. Tributyrin is shown with stick representation.

2.1.4.1.1 Catalytic mechanism

The characteristic property of thermoalkalophilic lipases is the catalytic triad and the oxyanion hole in order to perform their catalytic activity. General consensus for the catalytic triad is the Ser-His-Asp and, it is Ser114, His359, Asp318 for the BTL2 and the catalytic machinery is completed [25] with oxyanion hole which is formed by backbone nitrogen atoms of Phe-17 and Gln-115 (Figure 2.1) (Table 2.1).
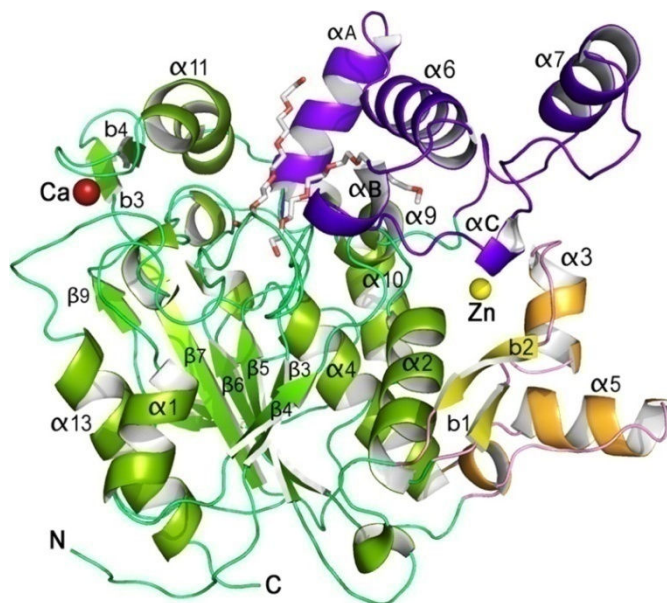
Figure 2.3: Labelled cartoon structure of BTL2. (α/ß) hydrolase core is shown in green, lid domain in purple, the zinc ion domain in yellow, and calcium ion in red. Two Triton X-100 molecules are in stick representation at the center (Figure is taken from Carrasco-López C et al. "Activation of bacterial thermoalkalophilic lipases is spurred by dramatic structural rearrangements" [25])

Catalytic serine is the key residue for the substrate binding. Despite most of the lipases having Gly-X-Ser-X-Gly motif, thermoalkalophilic lipases and BTL2 share Ala-X-Ser-X-Gly motif around their catalytic serine [26].

The crystal structure of L1 lipase, which is a close homolog of BTL2, was determined in the closed state, and shows a firm residue packing (His113, Phe17, Ile320, Thr270, and Met326) around the catalytic serine which leads to stabilization of the serine loop [27]. At this state, the catalytic serine is packed and not exposed to solvent.

However, the crystal structure of BTL2 was determined in the opened state; again catalytic serine has shown tight packing but excluding the Phe17. O_Υ group of Phe17 is rotated (100° torsion angle for x1 side chain) and opens area to catalytic serine for substrate binding (Figure 2.4). Catalytic serine becomes exposed to the solvent and available for substrate binding, once the lipase lid is opened by the interfacial activation [25].

Structural Role	Key residues
Catalytic triad	Ser114, His359, Asp318
Oxyanion hole	Phe17, Gln115
Zn^{2+} coordination	His82, His88, Asp65, Asp239
Ca^{2+} coordination	Glu361, Asp366, Gly287, Pro367
Stabilization of the serine loop and contributing to the lipase thermostability	His113, Phe17, Ile320, Thr270, Met326
Stabilizing role in the oxyanion-binding pocket and lid opening	Arg63
The hyperexposure of aromatic side chains upon activation	Phe28, Phe177, Phe181, Phe182, Tyr200, Tyr205, Phe207, Phe222, Phe226, Phe299
Conserved motif Gly- Phe/Leu/Ile -X-Gly	16Gly 19Gly
Conserved motif around Ser114 Ala-X-Ser-X-Gly	Ala112- X - Ser114- Gly115
Lipid interface interaction	Phe177, Phe181, Phe182
SN-1 Pocket	Ile320, Val321, Leu171, Val175, Leu184, Met174, Phe291, Val295
SN-2 Pocket	Ile363, Trp20, Phe28, Met25, Leu360, Val365
SN-3 Pocket	Phe17, Leu184, Val188, Leu189, Leu57, Leu209,Leu214, Trp212
Lid region	Thr169 to Asp239

Table 2.1: Structural Role of Important Residues on BTL2 Lipase [25]
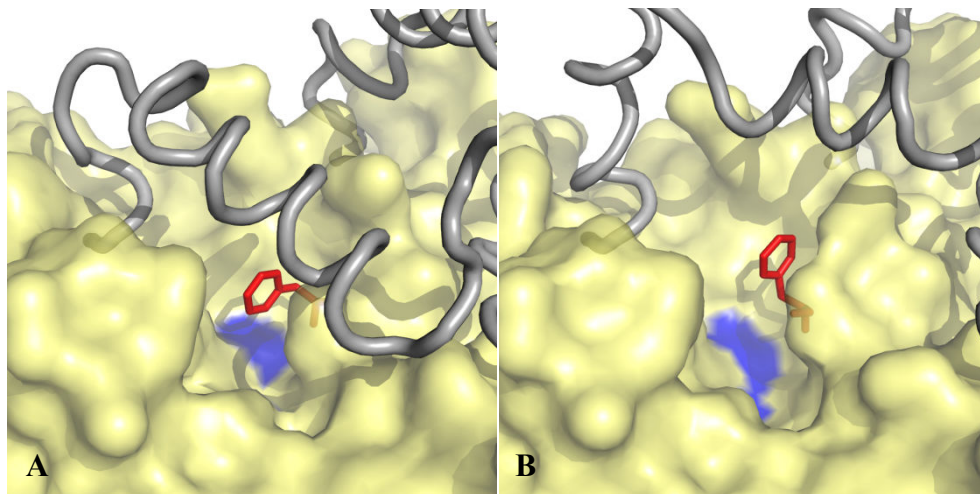
Figure 2.4: (A) Closed and (B) open lid structure of BTL2. 100 degree rotation of Oγ group of Phe17 showed in red stick and catalytic serine showed in blue surface.

2.1.4.1.2. Oxyanion

Lipid cleavage reaction is maintained with the formation of a tetrahedral intermediate which is stabilized by the oxyanion hole. In bacterial lipases (in families I.1 and I.2) and BTL2, intermediate oxyanion is stabilized with main chain amide groups which have identical positions in structural alignment [28]. In BTL2, backbone nitrogen atoms of Phe17 and Gln115 form the oxyanion hole. Phe17 is located in a conserved motif in I.5 family bacterial lipases as the second residue in Gly-Phe/Leu/Ile-X-Gly motif. Moreover, the conserved and buried Arg63 residue connects the oxyanion hole to the loop between strand ß3 and helix α2 which enhances the stability of the oxyanion hole and support the stability of the lid while opening [28].

2.1.4.1.3. The Active Site Cleft and Substrate Binding

In the crystal structure of BTL2, at the active site, two Triton detergent molecules were present whose positions illustrate the binding conformation of the substrate. Similarly, the crystal structure of *Pseudomonas aeruginosa* lipase (PAL) with triglyceride-like inhibitor [28] shows that two chains (sn-1 and sn-3) of the triacylglycerols substrate are well-fitted in superposition with two Triton X-100 detergent molecules of BTL2 (Figure 2.5). Triton X-100 detergent mimics the actual substrate and opens the lipase lid. It has an inhibitory effect at concentrations higher than 1mM in which a competitive inhibitory effect could stop the substrate catalysis [25]. Carrasco-López et al. couldn't crystallize BTL2 with actual substrate; instead they used a detergent that can mimic binding [25].

Three binding pockets for the three branches of the actual substrate have been defined and are shown in Figure 2.5. Branches are mostly surrounded with hydrophobic and aliphatic residues. First, sn-1 branch (HB binding pocket) is lined by Ile320, Val321, Leu171, Val-175, Leu184,Met174, Phe291, Val295; second sn-2 branch (HH binding pocket) is lined by Ile363, Trp20, Phe28, Met25, Leu360, Val365; and final sn-3 branch (HA binding pocket) is lined by Phe17, Leu184, Val188, Leu189, Leu57, Leu209, Leu214, Trp212 [25].
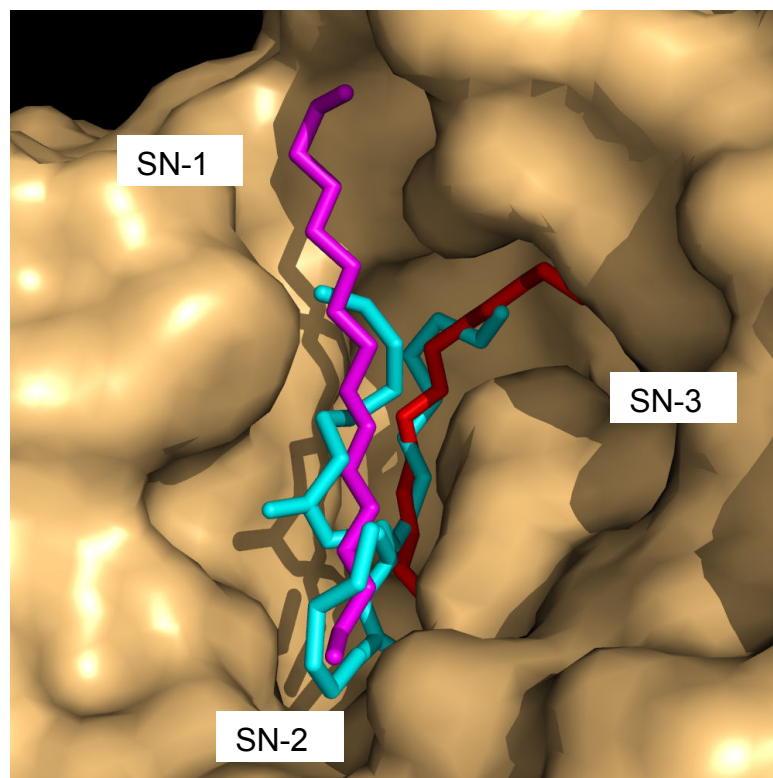
Figure 2.5: Surface of sn-1, sn-2 and sn-3 binding pockets, tricaprylin (docked structure) and two Triton X-100 molecules (X-ray structure). Tricaprylin is shown in cyan stick, and the two triton molecules are in red and magenta.

2.1.4.1.4. Activation Mechanism

Lid opening involves two helices (α6 and α7 helix) and a flexible loop at the end of the α6-helix (Figure 2.3). The lid region covers a total of 71 residues from Thr169 to Asp239. While, the α7-helix moves around the hinge (residues 211 to 239) with classical hinge motion, α6-helix opens itself with both lateral displacement and partial unfolding; the core of the BTL2 stays intact upon opening. Seven N-terminal residues of the α6-helix are integrated into the αA-helix and the loop is reshaped between two helices. Eventually the opening of the lid creates a large hydrophobic cavity within three pockets [25].

Activation causes N-terminal residue of the α6-helix, Phe177, to rotate out towards the solvent upon lipid interaction. Similarly, Phe181, and Phe182 are exposed to the solvent, additionally they stop αA-helix reconstruction with their loop formation (Figure

2.3). Also Asp-179 forms a salt bridge with Arg-242 which is located at the (α/ß) hydrolase core and anchors the lid to the core [25].

The open structure of BTL2 shows that Zn^{2+} binding domain is highly conserved in I.5 family of lipases and is crucial for stabilization of this large lid movement in lipases [25]. As α7-helix moves around the hinge of a region 211 to 239, starting residue Gln211 makes a hydrogen bond with Asp-62 which coordinates Zn^{2+} cation and finally residue Asp239 has a direct role in Zn^{2+} coordination [25].

For the case of α6-helix lid, residues from 208 to 211 act as a hinge; and Lys208 and Asp210 connect the lid regions to the core of BTL2 (Figure 2.3). All of these interactions are conserved on both open and closed configurations of BTL2 and this conservation may let us infer that any mutation on these critical sites would probably affect stability, thermostability and activity negatively [25].

2.1.4.2 Triglyceride specificity of BTL2 lipase

In 2003, Quyen et al. [29] published the relative activity of BTL2 towards various triglycerides. The pH-stat assays were performed at 65°C and pH 7.5 or pH 8.5 and BTL2 lipase shows the highest activity towards tributyrin (C4). As shown in Figure 2.6, at pH 7.5 and 8.5, substrate specificity profiles show very little difference [29]. In Figure 2.6, tributyrin activity at pH 8.5 is fixed as 100% activity for comparison with other substrates. Except for the substrates with chain length C2, C6 and C14, other substrates show slightly (1-3%) higher activity at pH 8.5 than at pH 7.5. The relative activities for acyl groups C8, C6, C10, C16, C2 are 40%, ~20%, ~20%, ~20%, and ~4%, respectively [29].
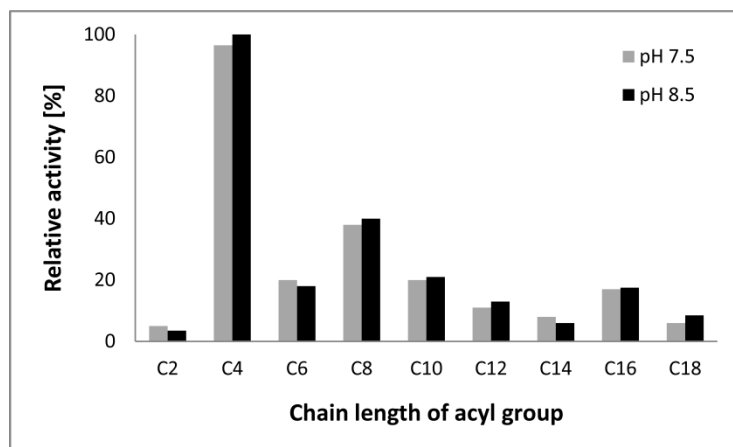
Figure 2.6: Substrate specificity of BTL2. The pH-stat assays were performed at 65°C and pH 7.5 or pH 8.5 [29].

2.1.4.3. Applications of BTL2 lipase

Thermophilic lipases are used in various industries, such as detergent additives, biodiesel production, waste water treatment, removal of oils and fats from fabrics and stereospecific synthesis of compounds for pharmaceuticals, cosmetics and perfumery, with respect to their extreme stability at high temperatures and in organic solvents. Therefore, they have become the focus of many protein engineering structural studies. Thermophilic BTL2 is specific for short length triglycerides and show low activity with medium and longer substrates. Broad selectivity on the chain length of the fatty acids has a key role in waste water treatment, and laundry formulations [2]. Also, long chain length specificity of lipase would give advantage for fat liquefaction. In contrast, short chain length specificity towards esters can be used in the food (cheese, bread, etc.) and cosmetic industry, specifically for the production of particular flavours [30]. These applications show the importance of the mechanisms that alters the chain length specificity which would be used in various particular applications. For instance, various lipases were tested in the hydrolysis of natural oils and show very promising results. This indicates the potential usage of BTL2 lipases in the food industry as enhancing the texture and nutritional value of natural oils. Another example is the potential usage in the laundry formulations, as BTL2 lipase is highly stable at elevated temperatures and in alkaline mediums [2]. Also, positional specificity is another attractive property for BTL2 which is sn-1,3 specific for triglycerides and can be used in the production of structured triglycerides that are used in clinical nutrition [29].

13

2.2 Computational Methods

2.2.1 Scoring Functions

The main aim of the scoring functions is achievement of rapid and accurate predictions with respect to experimental results. However, most of the time there is a trade-off between speed and accuracy.

The usage of scoring functions can be categorized into three major applications. First, determination of binding conformation between protein and ligand provide structural analysis of the complexes. Docking methods search different reasonable conformations according to their searching strategy and these conformations ranked by scoring functions. In order to select the experimentally determined orientation of the ligand and receptor, scoring functions are optimized according to the experimental binding mode. Accurate prediction of the binding configuration would explain the structural mechanism of binding and will lead to design of new drugs or modify the binding site for enhancement of binding to a particular substrate [31].

The second application is the prediction of absolute binding affinity with respect to the experimental results. This goal most efficiently is achieved by *ab initio* quantum methodologies which are computationally expensive. Alternatively, simplified scoring functions use several components of binding interaction energy, which give an approximate binding score. This prediction is mostly used in lead optimization or mutation selection problems that require accurate score predictions between sets of ligands or receptors. Reliable detection of a particular lead or mutation can decrease the potentially high cost of experimental procedures and the synthesis of new ligands.

The third application is the commonly used technique which is called virtual screening that involves detection of potential drug molecules in large sets of ligand databases. In this application, both binding affinity and the conformation prediction have important role for ranking of potential drugs. Therefore ligands that are experimentally known to show high affinity for the given receptor should give high scores at the initial screening.

All these applications require a reliable docking methodology and scoring function. There are three basic categories of scoring functions which are force field-based, empirical, and knowledge-based [31].

2.2.1.1. Force field scoring function

There are three mostly used force fields, CHARMM, AMBER, and GROMACS which are basically derived from experimental data and *ab initio* quantum mechanical calculations. The main idea is to express the potential energy of a system of particles with the parameters of mathematical functions that are based on the physical atomic interactions [32]. These interactions are essentially classified as non-bonded (van der Waals (vdW), and electrostatic interactions) and bonded (bond, angle, and torsion) interactions. Based on these interaction types, force field scoring functions are developed. Most popular example is the scoring functions of Autodock 4. Autodock 4 developed by Morris [70] and uses a genetic algorithm to search the poses of the ligand. It utilizes the Lamarckian version of genetic algorithm where the variations in conformations are used to generate new offspring poses after optimization [43]. All poses evaluated for vdW, hydrogen bonding, electrostatics, and desolvation terms as shown at Equation 2.1. In addition, based on the ligand, a conformation entropy term is added to Autodock 4 scoring function, which is explained in section 2.2.1.1.3.

$$V = W_{vdw} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + W_{\text{hbound}} \sum_{i,j} E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right)$$
$$+ W_{\text{elec}} \sum_{i,j} \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} + W_{\text{sol}} \sum_{i,j} \left(S_i V_j + S_j V_i \right) e^{(-r_{ij}^2 / 2\sigma^2)}$$

$$(2.1)$$

In equation 2.1, W is the weight constants for each term which was calibrated with experimental binding affinity data. The first term is a 6/12 potential for vdW interactions; the second term is a hydrogen bonding term based on a 10/12 potential. Third is the Coulomb potential for electrostatic interactions and the final term is a desolvation potential which is explained in section 2.2.1.1.2 [43].

The key challenge is the calculation of the solvation effect in the force field scoring functions [31]. Instead of using explicit water molecules which requires high computational cost [31,33,34], a number of methods provide rapid and reasonably

15

accurate solutions for the solvation energy problem for specific circumstances, such as Poisson–Boltzmann surface area (PBSA) model [35], the generalized-Born surface area (GBSA) model [36] and Autodock desolvation term [43].

2.2.1.1.1. Poisson–Boltzmann and the generalized-Born surface area model

One practical assumption is the treatment of water molecules implicitly as a continuum dielectric medium. In this way, computational cost is reduced and results are used for relative comparison and virtual screening studies [40]. Two common examples for the implicit solvent method are the Poisson–Boltzmann surface area (PBSA) model [35] and the generalized-Born surface area (GBSA) model [36] the latter is faster and basically an approximation of the PBSA method. In general, implicit solvent methods have boundaries and their capacity is limited by only non-specific interactions. Therefore, comprehensive interactions between solvent and solute, such as strong solvent–solute interactions [37], and strong solvent effects by ions [38,39], are not efficiently calculated by these methods.

Besides, one comparison study showed that solvation energy results of generalized-Born (GB) is not necessarily proper for binding calculations which should be calibrated with empirical parameters to get very accurate Born radii. This limitation restrains the speed of virtual screening and turn GBSA into an impractical solution [41].

Apart from the speed and accuracy of solvation effect, combining each energy term is also another challenging issue. Usually, weighting coefficients have to be used because each energy component is calculated from unrelated methods. For instance, electrostatic part of the solvation comes from Coulombic, PB or GB and hydrophobic part is approximated by the change of solvent-accessible surface area. Therefore, they cannot be simply summed up; instead, individual weighting coefficients should be calibrated for the selected protein or the protein family. This concept indicates the challenge of finding a universal set of parameters for the most of the protein complexes [31].

2.2.1.1.2. Autodock Desolvation Term

The desolvation term is developed based on the Wesson and Eisenberg method [42]. The main assumption of this method is that the change in the surface area accessible to solvent is proportional to the desolvation energy. In addition, each atom type should contribute differently, according to their polarity and hydrophobicity [43].

In order to calculate desolvation score, two parameters are used. First, according to the atom type, solvation parameters are assigned, and second is the amount of desolvation according to ligand. In detail, atom based solvation parameter describes the energy for transferring an atom from solution to buried state. In addition, the amount of desolvation is calculated with the percentage of the volume around an atom which is not occupied. This volume-summing method (similar to Stouten et al. method [44]) effectively represents the atomic degree of exposure which is linearly correlated with solvation energy.

In addition, solvation parameters for each atom are represented by Equation 2.2 [43].

$$S_i = (\text{ASP}_i + \text{QASP} \times |q_i|)$$

(2.2)

In this approach S is the solvation parameter for given atom whereas ASP and QASP are the calibrated parameters. In detail, ASP changes according to six atom types such as aliphatic carbons (C), aromatic carbons, (A), nitrogen, oxygen, sulfur, and hydrogen; however, QASP is constant for each atom and is calibrated according to each sets of atomic charges (Table 2.2). In this way, atomic charge is incorporated into solvation parameter which decreases the number of atom types since it prevents using separate oxygen and nitrogen types according to their charge. If an electron is delocalized within the molecule such as in carboxylate, the charge is located on the most accessible atom. Therefore a new map of interaction potentials is prevented for each new atom type in Autodock [43].

	ASP (std error)
C	-0.00143 (0.00019)
A	-0.00052 (0.00012)
N	-0.00162 (0.00182)
O	-0.00251 (0.00189)
H	0.00051 (0.00052)
S	-0.00214 (0.00118)
QASP = 0.01097 (0.00263)	

Table 2.2: Calibration of the ASP and QASP parameters for desolvation model for aliphatic carbons (C), aromatic carbons (A), nitrogen (N), oxygen (O), hydrogen (H) and sulphur (S).

Since the amount of shielding upon binding correlate with the desolvation energy, solvent accessible surface area difference is calculated between the bound and separate states. Atomic volumes are defined as a sphere with radius equal to the contact radius of each atom (C/A, 2.00; N, 1.75; O, 1.60; S, 2.00) and maximum ΔV is calibrated for each amino acid type over 188 proteins (from the Ligand–Protein Database [63]) according to,

$$\Delta V_i = \sum_{k \neq i} V_k \times e^{(-r_{ik}^2 / 2\sigma^2)}$$

(2.3)

In Equation 2.3, k is the number of atoms in the protein and, i is the atoms of the selected amino acid residue for ΔV_i calculation, r_{ik} is the distance between the centers of atoms i and k; and the distance weighting factor, σ is set to 3.5 Å, based on the original paper. Finally, a least-squares fit method is used for calibrating atomic ASP and QASP parameters and final values are shown at Table 2.2 [43].

2.2.1.1.3. Autodock Conformational Entropy Term

The entropic component of a binding energy is particularly difficult to calculate for docking purposes. Autodock 4 scoring algorithm uses the sum of the torsional degrees of freedom of the ligand for predicting conformational entropy. This approximation is

based on the idea that loss of torsional entropy upon binding is proportional to the number of rotatable bonds in the ligand. The term ΔS_{conf} represent the conformational entropy, N_{tors} is the number of rotatable bonds and W_{conf} is the calibrated weight parameter for entropy term [43],

$$\Delta S_{conf} = W_{conf} N_{tors}$$

(2.4)

2.2.1.2 Empirical scoring function

Empirical scoring functions use sets of weighted energy terms such as vdW, electrostatics, hydrogen bond, desolvation, entropy and hydrophobicity, in order to achieve the best correlation with the experimental data [31]. In some cases, other features can be taken into account such as the number and geometry of hydrogen bonds, the size of the contact surface, the electrostatic potential, the size of the binding cavities, and the flexibility of the ligand [45]. In contrast to the complexity of the force field scoring functions, the empirical ones provide simplicity for each energy term which reduces the computational cost. Eventually, sets of protein–ligand binding affinity data are used to calibrate the weight of the energy terms.

Addition of several terms to accurately represent the binding affinity brings with it multiple-counting problem. Sometimes one term is included in another term in a different way which may lead to miscalculation and double counting of the energy. Also, the capacity of the empirical scoring functions is limited with the size and the nature of the training of data set [31].

The new version of Autodock 4, Autodock Vina [68], reached an about two orders of magnitude speed-up compared to Autodock 4, while improving the accuracy of the binding mode predictions. Autodock Vina combines advantages of knowledge-based potentials and empirical scoring functions since it used empirical data from both the conformational binding modes of the receptor-ligand complexes and the experimental affinity measurements.

2.2.1.3 Knowledge-based scoring function

Knowledge-based scoring functions use the experimental structural data to generate statistical energy potentials. In this method, the occurrence frequency of atom pairs turn into atomic potentials with the conversion of Boltzmann relation [46]. Because of the simple Boltzmann relation, knowledge-based scoring functions provide as fast calculations as empirical scoring functions. Nevertheless, the use of structural information from training dataset is different; it gives an ability to use large and diverse training datasets without fitting parameters to specific type of dataset. Therefore, calculations are slightly independent from their training data set which gives an ability to perform on different subjects with similar accuracy [47]. However, assigning a reference state for the occurrence frequency calculation is a challenging task for this method [48]. Two alternative solutions have been introduced as choosing a randomized state or a physical approximation, but neither one improved the accuracy of the method [48]. Another well-known problem is the limited capability in discriminating the wrong binding modes [48]. As the predicted potential is derived from the pairwise atomic potentials of the ideally-bound structures, a little difference from the ideal pose can drop the overall accuracy of the knowledge-based scoring function [48].

2.2.2 Molecular Dynamics

Biological macromolecules have been recently investigated by one of the most handy and widely applied computational techniques called molecular dynamic (MD) simulations. At various timescales, protein can be simulated for fast internal motions to slow conformational changes. Molecular dynamics allows studying the solvent effect explicitly, and calculating other parameters such as the stability, density, dipole moment, entropy, enthalpy, interaction energy, potential energy and kinetic energy [49]. Therefore experimental calculations may be validated with molecular dynamics and several studies have shown the good correlation between molecular dynamics and experimental results [65,66,67].

Molecular dynamics uses simplified energy terms in order to simulate real molecular motions. These energy terms are parameterized to fit experimental data and QM calculations and all these parameters are called a force field. Several force fields are

20

commonly used for biomolecular simulations are Amber, NAMD, CHARMM, and GROMOS [53]. Although their parameterization methods are different, retrieved results are generally similar. All parameters in these force fields can be divided into two categories as bonded and non-bonded energy terms. The bonded terms include bond, angular and dihedral bond potentials and non-bonded terms involves van der Waals and electrostatic interactions.

For instance, NAMD [50] uses bond potential term that describes a spring between a pair of bonded atoms which is defined as $U_{\text{bond}} = k(r_{ij} - r_0)^2$, where $r_{ij} = \|\vec{r}_j - \vec{r}_i\|$ describes the distance between the pair of atoms, r_0 is the given equilibrium distance, and k is the spring constant.

Angular bond potential defined in triple covalently bonded atoms as $U_{\text{angle}} = k_\theta(\theta - \theta_0)^2 + k_{\text{ub}}(r_{ik} - r_{\text{ub}})^2$ where θ is the angle between vectors, θ_0 is the equilibrium angle, and k_θ is the angle constant. Second term defines non-covalent spring between outer atoms which is similar to bond potential with k_{ub} as the spring constant and r_{ub} as the equilibrium distance.

Torsion (dihedral) angle potential defined in the sequentially bonded four atoms as

$$U_{\text{tors}} = \begin{cases} k(1 + \cos(n\psi + \phi)) & \text{if } n > 0 \\ k(\psi - \phi)^2 & \text{if } n = 0 \end{cases}, \tag{2.5}$$

where n indicates the periodicity, ψ is the angle between the two defined planes, φ is the phase shift angle and k is the multiplicative constant. Given two equations are used to define the torsion term, which gives ability to express complex angular variation, which is truncated from Fourier series.

Van der Waals interactions are defined with the Lennard-Jones 6-12 potential as

$$U_{\text{LJ}} = (-E_{\text{min}}) \left[\left(\frac{R_{\text{min}}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{\text{min}}}{r_{ij}} \right)^6 \right], \tag{2.6}$$

where r_{ij} is the distance between two atoms within the cut-off distance, $E_{\text{min}} = U_{\text{LJ}}(R_{\text{min}})$ is the minimum of the energy term, at the distance R_{min} (Figure

2.7). As r increases, potential energy function reaches zero quickly, and cut-off term is needed to truncate these long distance (frequently above 12Å) interactions.
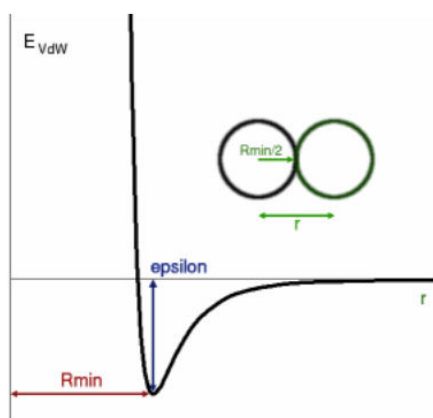
Figure 2.7: The van der Waals energy function [50]

Electrostatic (charged) interactions modelled with Coulomb's law as repulsive or attractive force according to atomic charges. Electrostatic potential defined with $U_{\text{elec}} = \epsilon_{14} \dfrac{C q_i q_j}{\epsilon_0 r_{ij}}$, where r_{ij} is the distance between two atoms within the cut-off distance, q_i and q_j are the atomic charges, C is the Coulomb's constant, ε_0 is the dielectric constant, ε_{14} is the scaling factor [50].

Molecular dynamic studies allow simulating an explicit solvent, ions, and even complex membrane structure with more convenient force fields, improved algorithms to control periodic boundary conditions, temperature and pressure. However, despite all these improvements, non-standard molecules like covalently bound ligands require time-consuming Quantum Mechanical (QM) calculations, in order to generate new force field parameters [51,52]. Also the reaction intermediates are high energy molecules and cannot be precisely represented by the MD force fields. This drawback makes MD calculations impractical for every different enzyme/substrate intermediate and become practically impossible to be used in the selectivity calculations for vast numbers of enzyme/substrate pairs [51,53].

2.2.2.1. Combining Docking With Molecular Dynamics

The computational cost and the accuracy of predictions are the most important factors that reveal the effectiveness and the applicability of computational methods. Docking programs generally serve simple solutions to binding energy problem with fast, slightly inaccurate and inexpensive algorithms. Therefore, they can search the huge conformational space of ligands quickly. The main disadvantage comes from the lack, or limited flexibility of the protein, especially upon binding. Whereas, molecular dynamic simulations are computationally costly, they expand the conformational space and permit flexible movements/rotations for both the protein and the ligand. This strategy supplies induced fit conformations around the binding site which increases the accuracy of binding energy calculations. In addition to conformational variety, solvation effect comes from the explicit water molecules which enhance the accuracy of trajectories and calculated binding energies [53].

Consequently, the combination of the two methods gives an advantage where docking is used for rapid screening of vast conformational space and molecular dynamic simulations optimize the complex structure and increase the calculated accuracy of binding energy [53].

In numerous studies, MD and docking have been used for examining the dynamic properties of the binding process. In the study of human cytochrome P450 2A6 [55], the mutation effect have been investigated on the enzymatic pathway, in another study, inhibitor of tyrosine kinase EphB4 screened by high-throughput docking and continued with 45 ns MD simulation for detailed binding mode exploration [56]. Also combination of MD with covalent docking gives structural information of enzyme-substrate complex as in the study of glycoside hydrolase and monosaccharides [57].

2.2.3 Free Energy Methods

Free energy methods are computationally expensive and usually take a long computational time to present results. Frequently used examples are thermodynamic integration, potential of mean force calculations and steered molecular dynamics.

2.2.3.1. Thermodynamic integration

Thermodynamic integration (TI) is a method used to compute the free energy difference between two states, usually the initial and the final state. In order to calculate the free energy difference, thermodynamic parameters are slowly changed between the states and at each stage system should be in equilibrium. Typically molecular dynamics or Metropolis Monte Carlo simulations are used for sampling each stage. Finally, along the defined reversible path, the integration is performed over thermodynamic parameters such as the energy, temperature, and the specific heat [58]. While TI is an accurate and flexible approach, it has some limitations on the conformational changes of the whole protein which may prevent the convergence over long simulation times. Also mutagenesis studies may become impractical since controlling the size and shape of the active site for each state is difficult and the structural variance of two microstates is too large for feasible integration [59].

2.2.3.2 Potential of Mean Force calculations

Potential of mean force (PMF) is simply determining free energy between two states of reaction using the Boltzmann-weighted average over all degrees of freedom [60]. One of the popular sampling techniques, Umbrella sampling, provides effective calculation of PMF from MD trajectories [61]. This technique adds the biasing potential to the Hamiltonian in order to lead the simulations toward a particular target, to cause a particular conformational transition. Therefore, large energy barriers are dealt with more efficiently. Finally, the data from all simulations are combined to get accurate and unbiased free energy prediction. The disadvantage of the Umbrella sampling is in deciding a conformational target of the system in order to receive a successful umbrella potential [61].

2.2.3.3 Steered Molecular Dynamics

Steered molecular dynamics (SMD) is an MD simulation method to mimic the idea of atomic force microscopy (AFM). In standard SMD, a force is applied to the ligand with spring in the chosen direction, and the overall work done by the ligand over the

trajectory is calculated according to Jarzynski's equality [72] to retrieve the absolute free energy of binding [62]. SMD has already been used to examine the pathway of the ligand along binding which also explains the induced fit changes of the protein. However, randomly or guessed pulling direction of the spring may change the pathway and affect efficiency. Therefore, the calculated energy may belong to an unfavourable pathway which leads to erroneous result [62].

2.2.4. Scoring strategy based on the factors that affect triglyceride specificity

In general, the prediction of binding affinity is a challenging task because it is not only the result of collective non-covalent interactions as performed by most docking algorithms. The main reason for failure is the inability of the scoring functions to discriminate between native and non-native substrate conformations. In particular, docking algorithms tend to bend substrates excessively in order to increase their scores [31].

Docking algorithms mostly assign a common set of weights to the individual energy terms that contribute to the overall energy score; on the other hand, these weights should be protein family dependent [31]. In addition, they wrongly assume that individual interactions that contribute toward the total binding affinity in an additive linear manner. In nature, non-covalent interactions often contribute in a nonlinear manner [31,40].

Prediction of binding affinity in a catalytic manner is dependent on several factors including the ability of the ligand to access the binding site, the desolvation free energy of the ligand and the binding site. Also entropy and enthalpy changes in the ligand, protein, and solvent, transition-state stabilization, steric complementarity of the enzyme to both substrate and intermediate state are critical factors for binding energy and catalytic activity [31,40].

Lipases perform their activity at a water-substrate interface which may lead to complex explanation for the chain length dependency [6]. Simply decreasing the shape of the substrate binding site would lead to steric blockage for the substrate, while increasing its size would give additional space leading to alternative binding of substrates and therefore decreasing k_{cat} [63]. In addition to the substrate binding site, other structural

features such as structure and hydrophobicity of the lid should be taken into account for more accurate prediction [6].

CHAPTER III

3. METHODOLOGY

3.1. Computational Method

3.1.1. System setup

X-ray structure of BTL2 was obtained from the Protein Data Bank (PDB entry 2W22) and used in MD simulations and docking runs. Water and ligand molecules in the crystal structure were removed. All hydrogen atoms were added using the AUTOPSF plugin of VMD. Especially, active site and oxyanion hole protonations were checked for appropriate catalytic state coordination (Figure 2.1B). Structure of tributyrin and tricaprylin (Figure 3.1) substrates were generated with the CHARMM27 topology and parameter files.
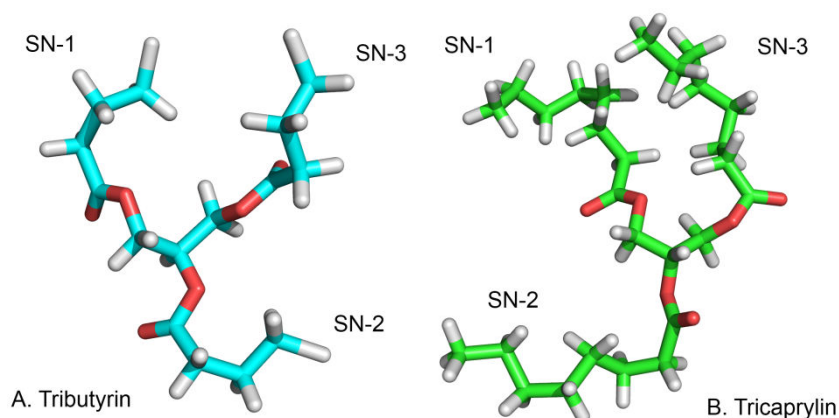
Figure 3.1: Structures of A. Tributyrin (C4) and B. tricaprylin (C8) and their sn-1, sn-2, sn-3 chains

3.1.2. Docking setup

Protonated structure of BTL2 is used for Autodock Vina [68] simulation. Both the ligand and the receptor molecules were converted in pdbqt format (Autodock4 format) in which the Gasteiger charges were assigned and non-polar hydrogens were merged by AutoDockTools (ADT). Size of the grid box is selected so as to contain the active site and binding pockets and also provide enough space for the ligand translation and rotation. In docking runs, the exhaustiveness parameter defines the time spent on the search and a higher value decreases the probability of not finding the global minimum [68]. In our runs, exhaustiveness parameter was selected as 100 (the default is 8) and the number of binding modes (generated output poses) was selected as 20 (the default is 9) for detailed exploration of the ligand conformational and orientational space as in the work of Azoia et al. [71].

3.1.3. Docking pose selection

For the first equilibration run, only tributyrin (C4) substrate was docked by Autodock Vina [68]. Poses were selected based on three criteria; i) ligand chains should be placed on correct clefts (see Figure 3.2 for tributyrin and Figure 2.5 for tricaprylin), ii) the distance between attacking serine oxygen atom and sp2 carbon atom of the substrate ester should be 3.2 Å maximum, iii) double bonded oxygen of substrate should point out towards oxyanion hole. If these three criteria hold for multiple poses, the pose that has the smallest distance for the second criteria is selected for subsequent analysis.
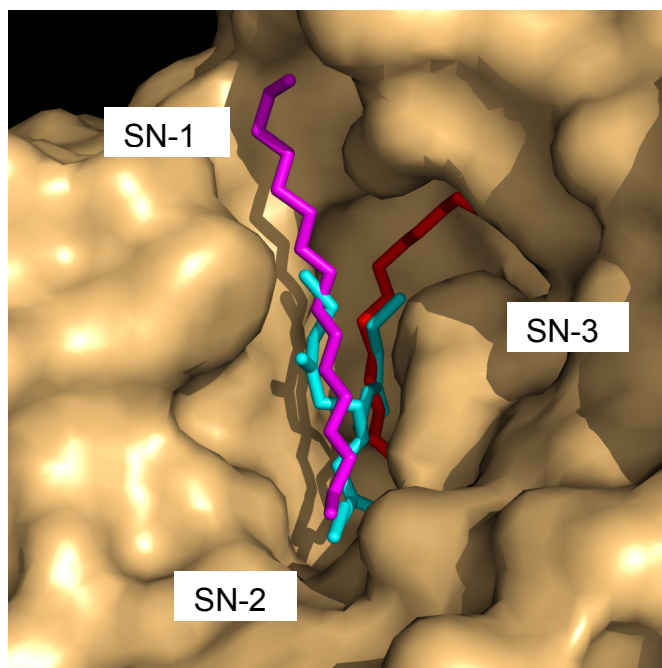
Figure 3.2: Surface of sn-1, sn-2 and sn-3 binding pockets, tributyrin (C4) (docked structure) and two triton molecules (X-ray structure). Tributyrin shown in cyan stick, and two triton molecules are in dark red and magenta.

3.1.4. Equilibration

Selected docking pose is used for assigning ligand coordinates. All MD simulations were performed using NAMD 2.7 with periodic boundary conditions, particle-mesh Ewald (PME) method and a 12 Å nonbonded cutoff. Protein and tributyrin were embedded in TIP3P water box with the "solvate plugin" in VMD, in a way to leave a 1 nm space around the solute. After that Na^+ and Cl^- ions were added into the water box to neutralize the system with final 0.15 M NaCl concentration. The system was then subjected to energy minimization (combination of conjugate gradient and line search algorithm) with harmonically constraints on protein and substrate atoms. Along minimization procedure, the force constants were gradually decreased every 100 steps from 10, 9, 8, 7, 6, 5, 4, 3, 2 to 1 kcal/mol, then every 200 steps from 1, 0.8, 0.5, 0.3 to 0.1 kcal/mol and followed by a 500 steps of non-restrained simulation. The energy-minimized structure was heated gradually from 0 to 298 K in 0.01 ns and then equilibrated for 1 ns at 1 atm with Langevin piston pressure method and Langevin

dynamics temperature control. All MD runs was performed with 2 fs time-step with the SHAKE algorithm.

3.1.5. Mutants and Re-docking

The equilibrated structure was used for subsequent mutant generation. All experimentally examined mutants (F17A, V175A, V175F, D176F, T178V, F181A, F182A, I320F, L360F, Figure 3.3) were subjected into "mutate residue" algorithm from the visualization package VMD. Tributyrin (C4) and tricaprylin (C8) were docked on the generated mutated proteins while mutated side chains were allowed to be flexible. Docking parameters and pose selection criteria were the same as in the previous docking run.
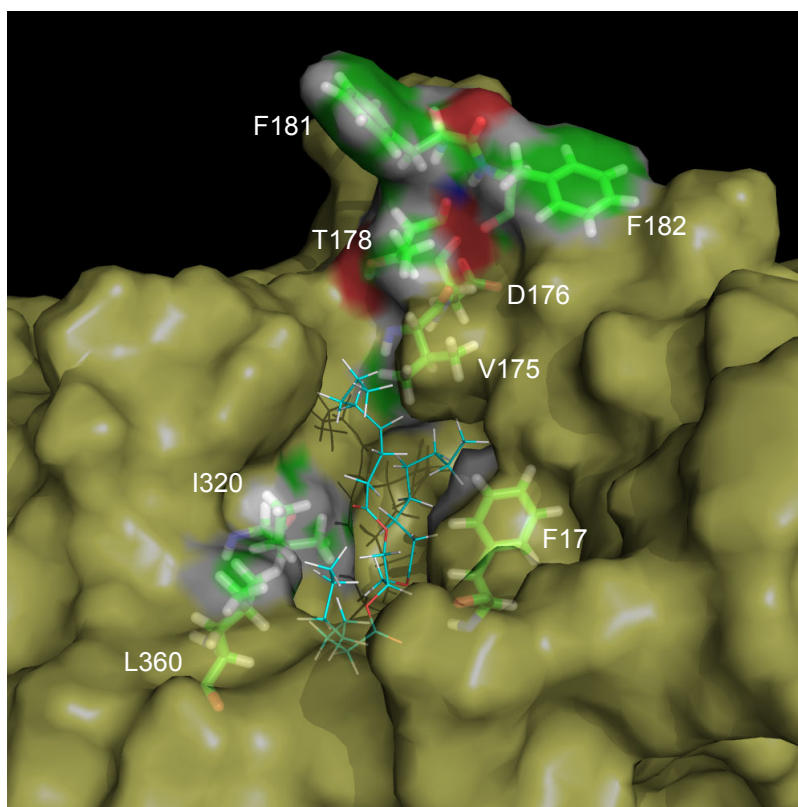
Figure 3.3: Surface of the wildtype BTL2 with mutated residues and tributyrin. Mutated residues F17, V175, D176, T178, F181, F182, I320, and L360 are labelled and shown in green stick. Tricaprylin (C8) is shown in blue lines at the center.

3.1.6. Production Phase

For consistency both the mutated and the wild-type structures were solvated in a water box which leave a 1 nm space around the solute and systems were neutralized with 0.15 mol/L NaCl ions. The system was again subjected to 2500 step energy minimizations with harmonic constraints same as the previous MD run. The energy-minimized structures were again heated gradually and then subjected to 4 ns MD run. First 2 ns part of the run was used for RMSD calculation and the final 2 ns part used for scoring calculations. All simulations were performed at 1 atm with Langevin piston pressure method and 298 K with Langevin dynamics temperature control.

3.1.7. Scoring Methodology

The final 2 ns part of the 4 ns production trajectories were used for scoring calculation. The electrostatic (E_{Elec}) and van der Waals (E_{vdW}) interaction energies were calculated using NAMDEnergy in NAMD2.7 at 2 ps intervals and using a cutoff distance of 12 Å for energy calculations. Desolvation energy (E_{desolv}) and loss of torsional entropy upon binding (ΔS_{conf}) for each snapshot of the MD trajectory were calculated with "compute_AutoDock41_score" python script. Weights of the four parameters in Equation 3.1 were calibrated with linear regression (by using MATLAB) in order to maximize the Pearson and Spearman correlations between experimental activity and computational scores. Pearson coefficient, r defines the strength of linear dependence, whereas Spearman describes the ranking power. In this way, BTL2 dependent scoring scheme is optimized.

$$\text{Score} = a(E_{Elec}) + b(E_{vdW}) + c(E_{desolv}) + d(\Delta S_{conf}) \qquad (3.1)$$

Equation 3.1: Scoring scheme. Electrostatic energy (E_{Elec}), van der Waals energy (E_{vdW}), desolvation energy (E_{desolv}) and conformational entropy (ΔS_{conf}) terms with their weights.

3.1.8. RMSD analysis

To evaluate the stability, RMSD (root mean-square deviation) of the protein backbone was calculated with RMSD Visualizer tool in VMD 1.9.1. Prior to the analysis, backbone atoms (N, C, O, CA) of protein were aligned, to the initial structure. Afterwards, RMSD of the protein was calculated based on backbone atoms (N, C, O, CA) to evaluate the stability of the protein during the simulation. Also, RMSD of the substrates were calculated based on overall atoms to track the structural rearrangements of the substrates during the simulation.

3.2. Experimental Method

3.2.1. Site-directed Mutagenesis and Expression

All experiments were performed by Sezerman's group. BTL2 gene (1,167-bp) is cloned into pMCSG-7 vector using the ligation independent cloning sites which are shown with bold at below on the primer sequences.

Wild_Forward: 5'-**tacttccaatccaatgaa**gcggcatccccacgcg-3'
Wild_Reverse: 5'-**ttatccacttccaatgaa**aggccgcaaactcgccaa-3'

In the experimental stage, seven point mutations (F17A, V175A, V175F, D176F, T178V, I320F, L360F) which were introduced by Overlap Extension PCR. Primers that were used for introducing the mutation are listed at below.

T178V_Forward: 5'-gtcaatatggtcgatttcgtggatcgcttctttgacctg-3'
T178V_Reverse: 5'-caggtcaaagaagcgatccacgaaatcgaccatattgac-3'
D176F_Forward: 5'-gacgacacttgtcaatatggtcttttttcactgatcgcttctttgac-3'
D176F_Reverse: 5'-gtcaaagaagcgatcagtgaaaaagaccatattgacaagtgtcgtc-3'

For each mutant, two PCR reactions were performed. At first, wild_forward primer and mutant_reverse were used, in the second PCR, mutant_forward and wild_reverse were used. These mutant fragments were cloned into pMCSG-7 using ligation independent approach. The positive clones were chosen after sequence verification.

In the cloning stage, Escherichia coli (E. coli) XL1-Blue was used with E. coli SHuffle® Express host. E. coli cells were cultivated at 37°C with suitable antibiotics. 1 mM isopropyl-β-d-thiogalactopyranosideat was used for induction.

3.2.2. Purification and Enzyme Assays

Polyhistidine tags were used to purify the protein on to nickel coated beads (GenScript). For purifications 20 mM Potassium Phosphate buffer, and for elution 500 mM Imidazole was used. 10 kDa Filters (Millipore) were used in buffer-exchange method which uses 0.1 M Tris pH 7.3 versus 500 mM Imidazole.

Direct titration of fatty acids was used for measuring lipase activity and one unit is described as the amount of enzyme to release one μmole fatty acid in one minute. Specific activities of lipases towards tributyrin (C4) and tricaprylin (C8) was measured and replicated two times at 55°C, and pH 8.0 in terms of units/mg.

CHAPTER IV

4. RESULTS

4.1. First Docking

Before the first equilibration simulation (1 ns), tributyrin (C4) and tricaprylin (C8) were docked; all of the obtained poses are shown at Figure 4.1. As expected, valid conformations of both substrates were obtained and our pose selection criteria were satisfied (Figure 4.2). In table 4.1, typical output of Autodock Vina is shown with our criteria checklist (for details see section 3.1.3 under methods).
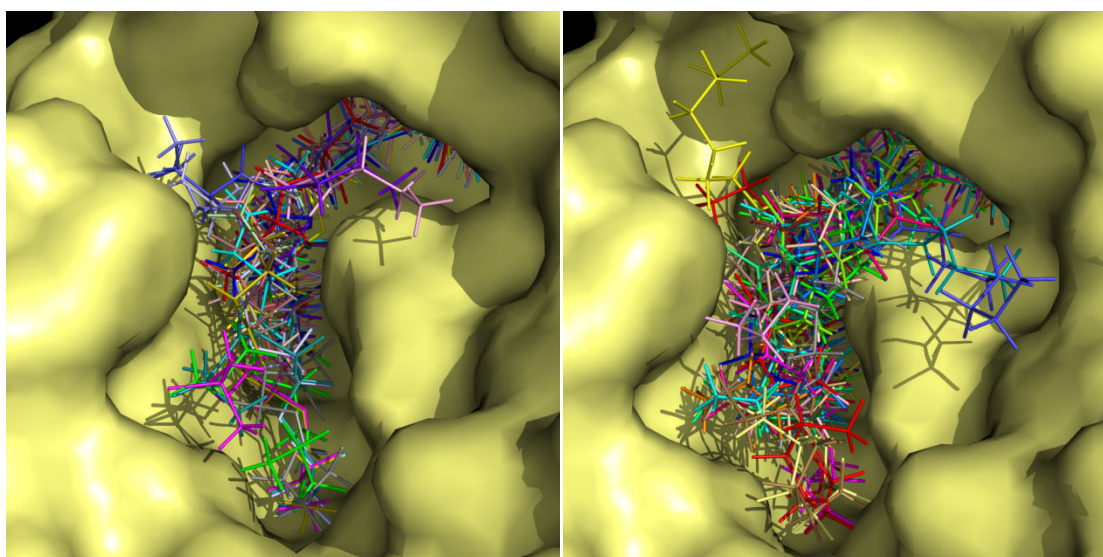

Figure 4.1: All docked poses of tributyrin (C4) and tricaprylin (C8) on the X-ray structure of BTL2
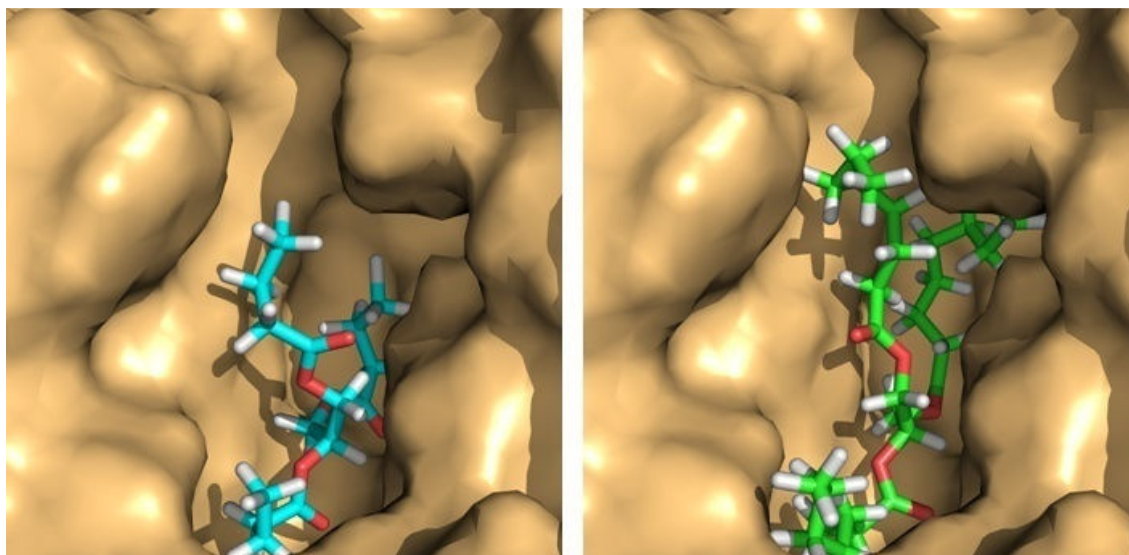
Figure 4.2: Docked poses of tributyrin (C4) and tricaprylin (C8) at the active site of the
BTL2 before the first 1 ns equilibration

All docked poses of tributyrin (C4)						All docked poses of tricaprylin (C8)					
Pose	affinity (kcal /mol)	RMSD lower bound	RMSD upper bound	Criterion 1	Criterion 2	Criterion 3	Pose	affinity (kcal /mol)	RMSD lower bound	RMSD upper bound	Criterion 1	Criterion 2	Criterion 3
1	-5.8	0.000	0.000	x	x	x	1	-6.1	0.000	0.000	√	√	√
2	-5.7	1.471	7.033	x	x	x	2	-6.0	1.297	4.869	x	x	x
3	-5.7	1.211	1.929	x	x	x	3	-6.0	1.968	8.121	x	x	x
4	-5.6	1.840	4.573	x	x	x	4	-6.0	2.027	5.814	x	x	x
5	-5.6	2.661	7.997	x	x	x	5	-6.0	1.887	2.811	√	x	x
6	-5.6	1.635	7.042	x	x	x	6	-6.0	1.814	8.253	x	x	x
7	-5.6	1.785	4.309	x	x	x	7	-6.0	2.002	4.899	x	x	x
8	-5.6	1.734	2.843	x	x	x	8	-6.0	1.877	8.033	x	x	x
9	-5.6	1.613	2.228	√	√	x	9	-5.9	1.857	8.221	x	x	x
10	-5.6	1.806	7.136	x	x	x	10	-5.9	2.179	4.789	x	x	x
11	-5.5	1.804	5.606	x	x	x	11	-5.9	1.828	6.891	x	x	x
12	-5.5	0.911	6.788	x	x	x	12	-5.9	1.528	7.774	x	x	x
13	-5.5	1.331	6.437	√	√	√	13	-5.9	1.070	7.378	x	x	x
14	-5.5	1.425	1.760	x	x	x	14	-5.9	1.337	7.108	x	x	x
15	-5.5	1.121	5.283	x	x	x	15	-5.9	2.070	8.695	x	x	x
16	-5.5	1.441	5.804	x	x	x	16	-5.9	2.023	7.512	x	x	x
17	-5.4	1.334	6.681	x	x	x	17	-5.9	1.836	6.664	x	x	x
18	-5.4	2.686	5.295	√	√	x	18	-5.8	1.916	8.842	x	x	x
19	-5.4	1.962	5.173	x	x	x	19	-5.8	1.824	6.249	x	x	x
20	-5.3	1.373	3.129	x	x	x	20	-5.8	0.999	4.981	x	x	x

Table 4.1: Autodock Vina output [affinity (in kcal/mol), RMSD lower bound, RMSD
upper bound)] and our criteria checklist (criterion 1, 2 and 3). RMSD values are calculated
relative to the best predicted pose and based on the heavy atoms. Two alternative RMSD values are given
differing in how the atoms are matched in the RMSD calculation. Criterion 1: substrate branches placed
on correct clefts, criterion 2: the distance between attacking serine oxygen and sp2 carbon atom of the
substrate is maximum 3.2 Å, criterion 3: double bonded oxygen of substrate point out to oxyanion hole.

4.2. Equilibration Simulation

Before introducing mutations, a 1 ns equilibration simulation was performed with tributyrin (C4). To evaluate the stability of the protein structure during the simulations, a root mean-square deviation (RMSD) was calculated on the backbone atoms after aligning the trajectory on the backbone atoms. During the simulation, the RSMD reaches a plateau around 1 Å after 300ps for protein (Figure 4.3). Similarly RMSD of the ligand shows a stabilized pattern between 1.5-2.5 Å after 300 ps (Figure 4.3).
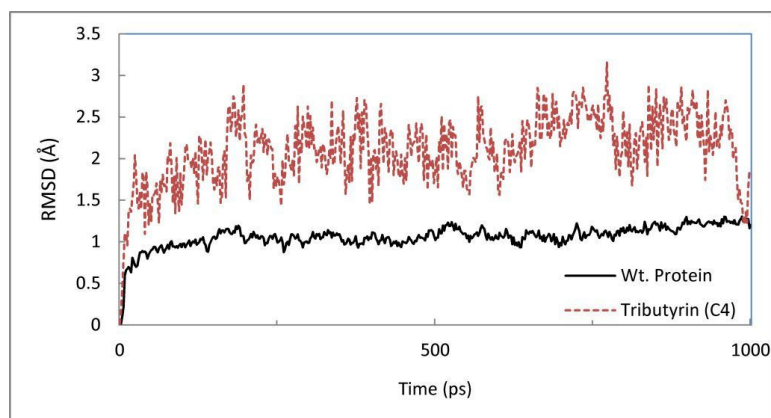
Figure 4.3: RMSD of protein backbone and tributyrin (C4) from first equilibration MD

4.3. Production Phase

During the 4 ns production simulation, the mutant protein and substrate stability were checked with RMSD (Figure 4.4). As shown in Figure 4.5, the distance between attacking serine oxygen atom and sp2 carbon atom of the substrate (d1); between oxyanion hole and double bonded oxygen of substrate (d5, d6); and stabilized catalytic triad interactions between serine, histidine and aspartic acid (d2, d3, d4) that should be conserved during the simulation are measured and their average values over the second 2 ns part of the production MD simulation are shown on Table 4.2. Only one MD simulation (C8-Leu360Phe) was not generated as the substrate could not be stabilized during the first 1 ns of the simulation despite 10 repeated simulations.
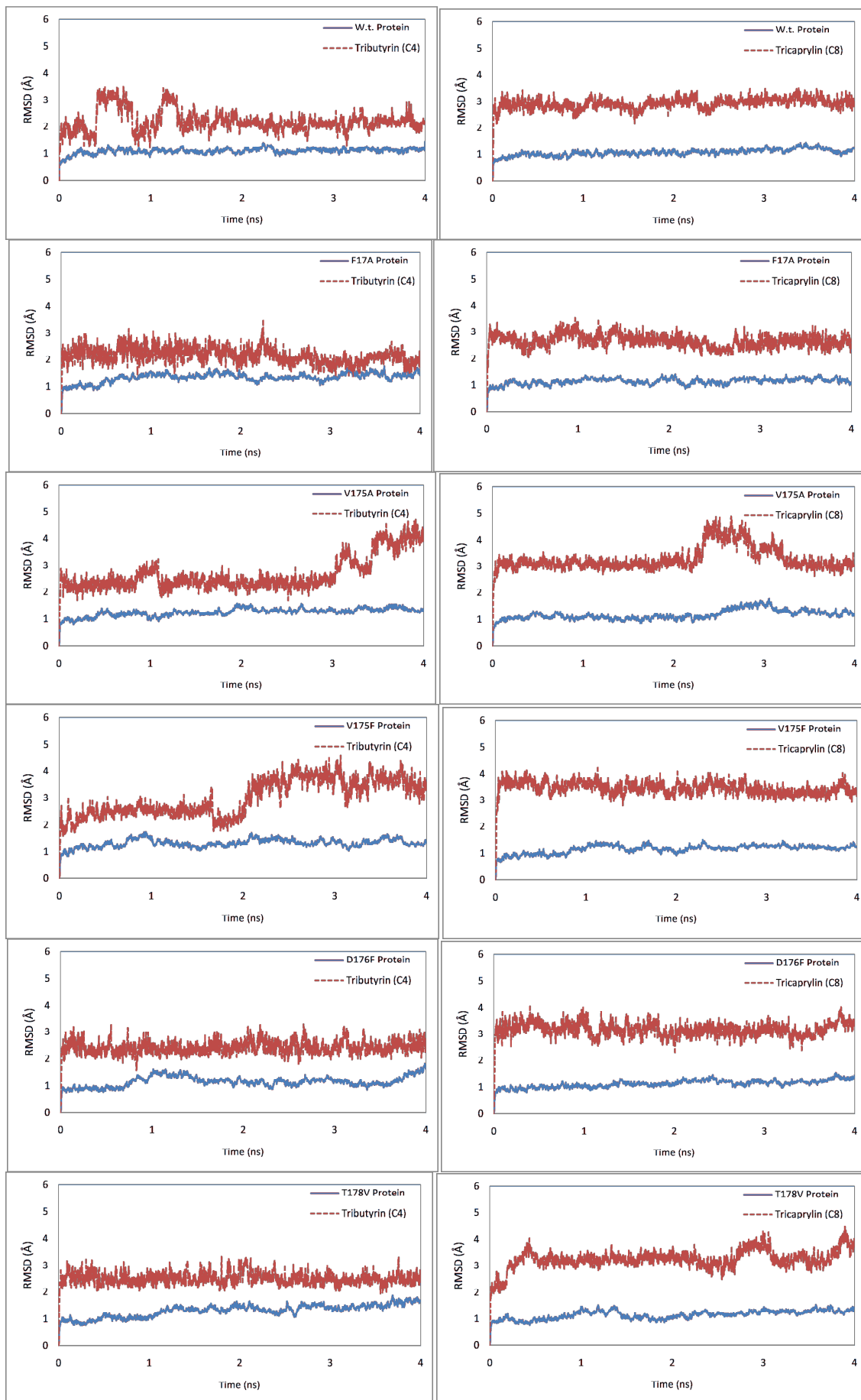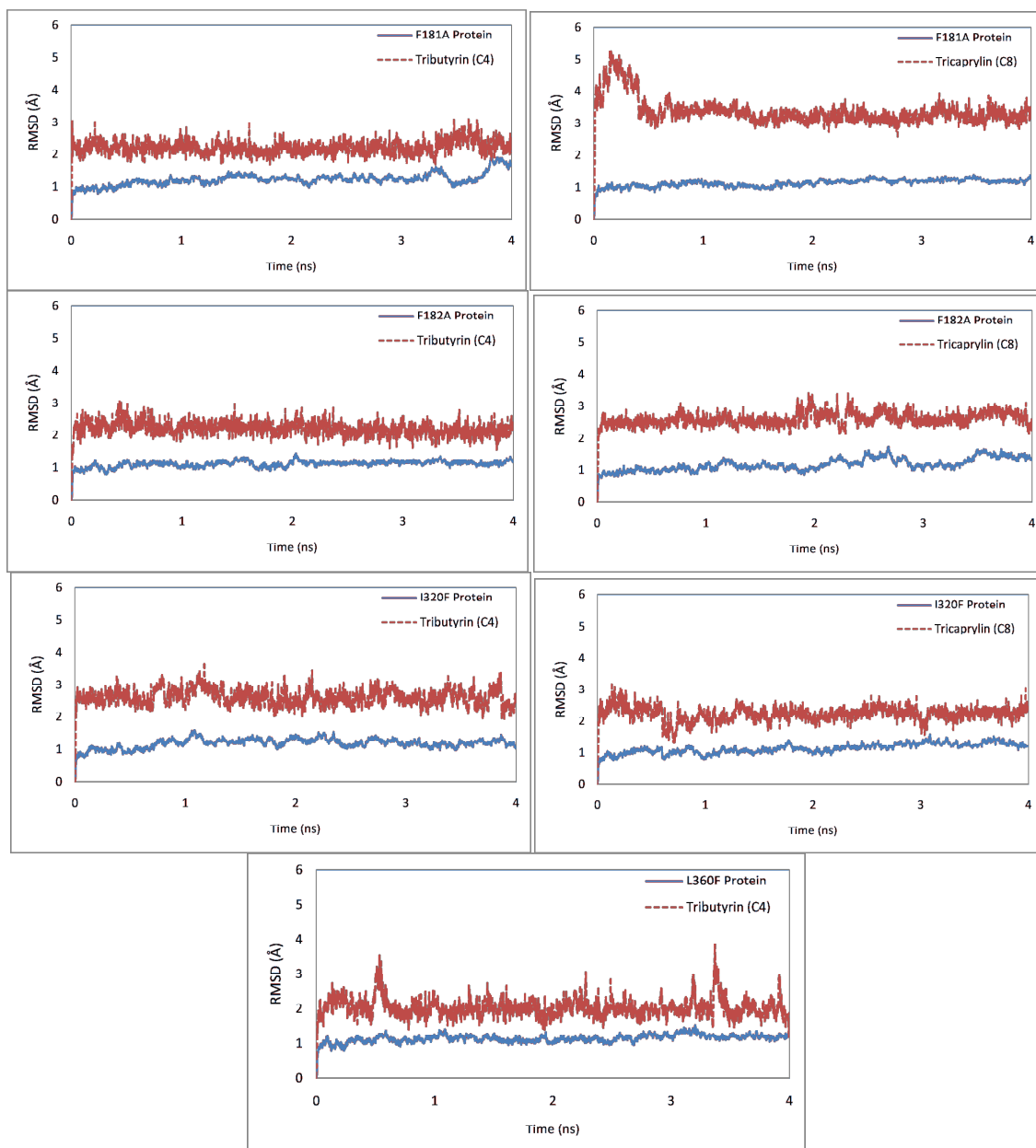
Figure 4.4: RMSD of protein (wildtype, F17A, V175A, V175F, D176F, T178V, I320F and L360F) in blue lines and substrate (tributyrin or tricaprylin) in red lines during 4 ns MD simulation.
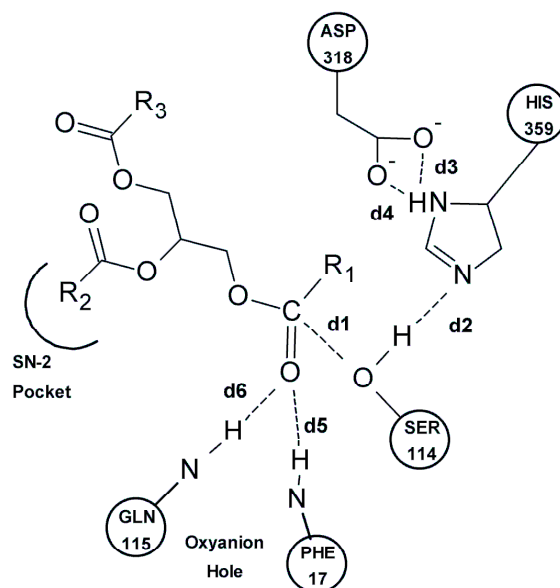
Figure 4.5: Distances in the binding pocket as d1, d2, d3, d4, d5 and d6.

	d1	d2	d3	d4	d5	d6
4C wildtype	3.160	2.058	2.816	1.784	2.366	2.876
4C F17A	3.176	2.419	2.770	1.798	1.970	2.302
4C V175A	3.074	2.192	2.805	1.783	2.007	2.490
4C V175F	3.117	1.968	2.941	1.777	2.027	2.614
4C D176F	**3.059**	2.122	2.759	1.782	**1.990**	**2.488**
4C T178V	3.101	2.106	2.767	1.795	2.037	2.491
4C F181A	**3.050**	2.281	2.762	1.768	**1.957**	**2.412**
4C F182A	**3.066**	2.222	2.826	1.789	**1.986**	**2.385**
4C I320F	3.098	2.103	2.829	1.775	2.148	2.764
4C L360F	3.086	2.258	2.754	1.766	2.021	2.476
8C wildtype	3.084	2.111	2.819	1.750	2.023	2.658
8C F17A	3.131	2.252	2.734	1.805	1.933	2.451
8C V175A	**3.062**	2.094	2.776	1.783	2.014	2.630
8C V175F	3.078	2.050	2.836	1.793	1.991	2.561
8C D176F	3.100	2.209	2.792	1.789	2.037	2.601
8C T178V	3.079	2.154	2.915	1.811	2.056	2.622
8C F181A	**3.059**	2.172	2.860	1.788	2.021	2.580
8C F182A	**3.063**	2.246	2.812	1.774	**2.009**	**2.389**
8C I320F	3.085	2.268	2.761	1.769	1.989	2.414

Table 4.2: Average distances during the second 2 ns part of the production MD simulation between substrate and protein atoms which are defined in Figure 4.5 as d1, d2, d3, d4, d5 and d6.

The final 2 ns part of the 4 ns production simulations were used for scoring calculation. Average values of the 2 ns simulation were calculated based on the information given under the methods section and results shown in Table 4.3 and 4.4.

Substrate -Protein	Electrostatics	vdW	Desolvation	Conf. Entropy
C4-Wildtype	-23.3996	-32.7005	34.19415	17
C4-Phe17Ala	-21.6883	-32.1534	36.164	17
C4-Val175Ala	-25.8687	-32.8823	36.95152	17
C4-Val175Phe	-27.438	-32.8815	37.70971	17
C4-Asp176Phe	-26.27	-32.123	35.08078	17
C4-Thr178Val	-26.452	-31.551	34.44336	17
C4-Phe181Ala	-23.4172	-34.6931	36.69598	17
C4-Phe182Ala	-22.5677	-33.5853	36.13598	17
C4-Ile320Phe	-25.8829	-33.8285	38.45747	17
C4-Leu360Phe	-24.553	-36.2894	39.87395	17

Table 4.3: Average values of electrostatics, vdW, desolvation energy from 2 ns production phase simulation with tributyrin and conformational entropy according to substrate.

Substrate -Protein	Electrostatics	vdW	Desolvation	Conf. Entropy
C8-Wildtype	-23.6097	-48.6684	45.80038	29
C8-Phe17Ala	-27.2476	-52.4624	50.24974	29
C8-Val175Ala	-25.9222	-49.246	46.79106	29
C8-Val175Phe	-27.9326	-51.0876	45.35592	29
C8-Asp176Phe	-26.947	-50.5465	46.50245	29
C8-Thr178Val	-26.4045	-48.9606	43.91499	29
C8-Phe181Ala	-27.9076	-49.9145	45.185	29
C8-Phe182Ala	-25.4707	-51.17	47.04334	29
C8-Ile320Phe	-26.1359	-52.7581	47.75737	29
C8-Leu360Phe	N.A.*	N.A.	N.A.	N.A.

Table 4.4: Average values of electrostatics, vdW, desolvation energy from 2 ns production phase simulation with tricaprylin and conformational entropy according to substrate. *N.A.: Not available

Substrate-protein simulation of C8-Leu360Phe was not generated as its substrate is not stabilized during the first 1 ns of the simulation, and its averaged values are defined as "not available" in Table 4.4. This substrate-mutant pair is directly classified as "non-binding" and removed from the calibration data.

4.4. Purified Lipases and Experimental Activity Assays

For the purified two mutant lipases (T178V and D176F), SDS gel electrophoresis was performed and related bands were shown at Figure 4.6 with the prestained plus molecular weight marker (Fermentas).
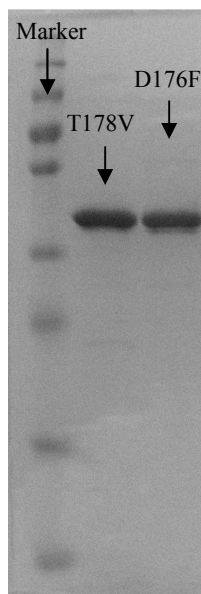
Figure 4.6: Result of SDS gel analysis. Prestained plus molecular weight marker (Fermentas), and T178V and D176F mutant lipases were highlighted.

Specific activities of all lipases (wildtype, F17A, V175A, V175F, D176F, T178V, I320F and L360F), towards tributyrin (C4) and tricaprylin (C8) are measured and replicated two times to evaluate consistency. Measured specific activities (U/mg) and their standard deviation are summarized in Figure 4.8A.

4.5. Coefficient Calibration for Both Tributyrin (4C) and Tricaprylin (8C)

Coefficients of scoring function were calibrated with linear regression for every possible training set that has at least three different mutation data points. Both tributyrin and tricaprylin data were used for every selected mutation and wild type data should exist in all training sets.

By using calibrated coefficients, scores were calculated for each case and correlation of training set, test-set and overall data were calculated. Several important calibrated coefficients with Pearson and Spearman correlation coefficients are shown at Table 4.5

Training set	Test Set	a	b	c	d	Training set Corr.*	Test Set Corr.*	Overall Corr.*	Accuracy of Predicted Mutation Impact
W.t. F17A V175A	V175F D176F T178V I320F	-101.92	-113.14	-100.56	-102.34	0.836 (0.0380) 0.771	0.713 (0.0470) 0.762	0.793 (0.0007) 0.895	9/12
W.t. F17A T178V	V175F V175A D176F I320F	-78.74	-71.69	-56.89	-80.93	0.850 (0.0322) 0.886	0.861 (0.0060) 0.881	0.812 (0.0004) 0.881	9/12
W.t. F17A I320F	V175F V175A D176F T178V	-132.46	-95.61	-112.80	-75.28	0.924 (0.0085) 0.771	0.797 (0.0179) 0.833	0.791 (0.0008) 0.851	8/12
W.t. V175F I320F	F17A V175A D176F T178V	-114.97	-71.77	-87.40	-57.60	0.973 (0.0011) 0.829	0.778 (0.0231) 0.929	0.788 (0.0008) 0.851	8/12
W.t. D176F I320F	F17A V175F V175A T178V	-123.91	-37.59	-54.25	-61.98	0.921 (0.0092) 0.943	0.751 (0.0317) 0.810	0.762 (0.0015) 0.785	8/12
W.t. F17A V175A I320F	V175F D176F T178V	-102.31	-94.88	-81.35	-103.15	0.847 (0.0080) 0.810	0.706 (0.1173) 0.543	0.816 (0.0004) 0.868	9/12
W.t. V175A D176F I320F	F17A V175F T178V	-143.26	-119.28	-120.37	-116.92	0.838 (0.0094) 0.833	0.874 (0.0230) 0.771	0.813 (0.0004) 0.846	8/12
W.t. V175F D176F I320F	F17A V175A T178V	-134.02	-79.95	-96.05	-76.50	0.880 (0.0040) 0.929	0.958 (0.0026) 1.000	0.799 (0.0006) 0.855	9/12
W.t. F17A V175F I320F	V175A D176F T178V	-126.86	-95.45	-110.18	-74.43	0.951 (0.0003) 0.857	0.769 (0.0742) 0.886	0.789 (0.0008) 0.824	8/12

Table 4.5: Training set, test set, calibrated coefficients (a, b, c, d) (Equation 3.1) of both substrates, correlation coefficients for each dataset and accuracy of directional prediction (enhanced or reduced) of the mutation effect according wildtype. *Values are

given for Pearson correlation coefficient (Pearson p-value) and Spearman correlation coefficient.
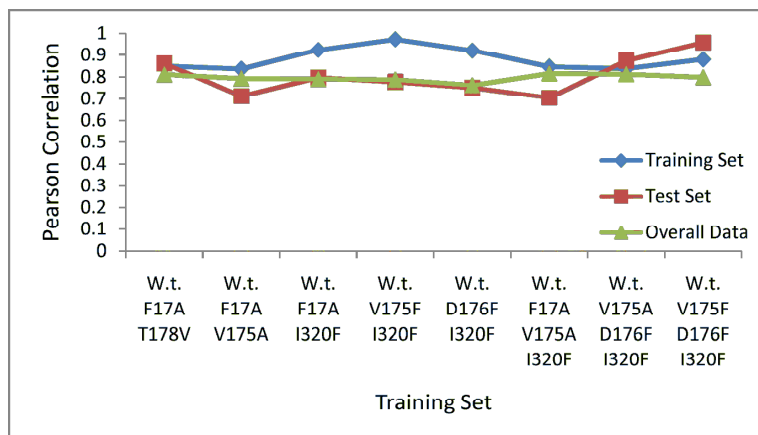
Figure 4.7: Pearson correlation for training, test and overall datasets with respect to training set.

In order to predict the impact of mutation on specificity, one of the best correlated coefficients are selected from "wildtype-F17A-V175A" training set and coefficients a, b, c, d are determined as -101.92, -113.14, -100.56, -102.34, respectively. Predicted scores versus experimental specific activity are shown in Figure 4.8.
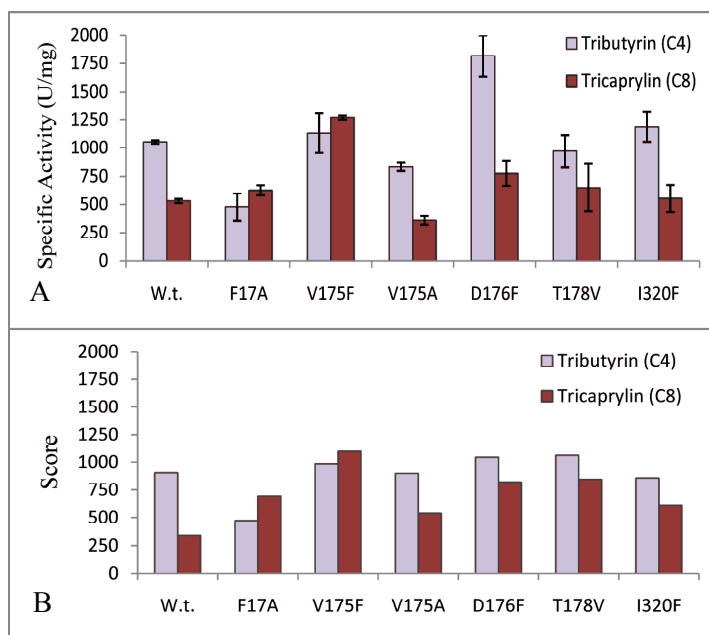
Figure 4.8: (A) Specific activities of BTL2 mutants towards tributyrin (C4) and tricaprylin (C8) in terms of units/mg. (Standard deviation shown with error bars)

(Unpublished results from Sezerman's Lab.) (B) The impact of mutations on the computational score of tributyrin (4C) and tricaprylin (8C). Wildtype, F17A and V175A selected as training set for calibration.

4.5.1. Specificity Prediction of Previously Published Mutations

In the previously published article [64], pH-stat ($55°C$ in pH 8.5) assay were used for specific activity calculations. One specific activity unit was described as the amount of enzyme that released 1.0μmol of fatty acid per minute. Experimental specific activities are shown in Figure 4.9A and predicted scores by our scoring function are shown in Figure 4.9B.
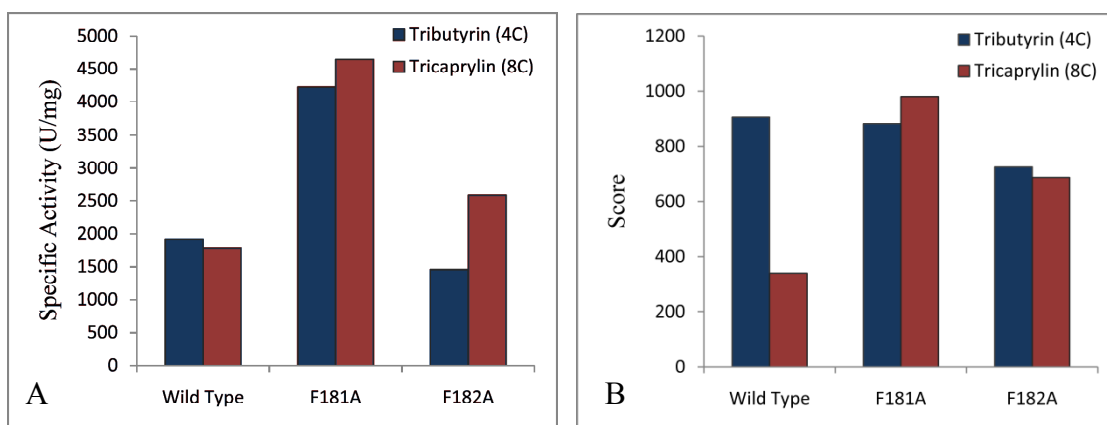
Figure 4.9: The effect of 181Ala and 182Ala on the (A) specific activity and (B) computational score. Previously calibrated coefficients (-101.92, -113.14, -100.56, -102.34) were used in the scoring function.

4.6. Coefficient Calibration for Tributyrin (4C) Only

Coefficients of tributyrin (4C) specific scoring function were calibrated with only the tributyrin data. For the possible training sets that has at least three different mutations and wildtype data points, scoring function coefficients and correlation of training set, test set and overall data were calculated. Several important calibrated coefficients with Pearson and Spearman correlation coefficients are shown in Table 4.6.

Training set	Test Set	a	b	c	d	Training set Corr.*	Test Set Corr.*	Overall Corr.*	Accuracy of Predicted Mutation Impact
W.t. F17A V175F V175A L360F	D176F T178V I320F	-113.87	-190.62	-136.15	-189.04	0.953 (0.0122) 0.900	0.278 (0.8204) 0.500	0.710 (0.0486) 0.476	5/7
W.t. F17A V175F I320F L360F	V175A D176F T178V	-133.40	-158.10	-113.99	-197.27	0.965 (0.0080) 0.700	0.578 (0.6080) 0.500	0.709 (0.0489) 0.405	5/7
W.t. F17A V175F V175A I320F L360F	V175F D176F T178V	-113.46	-176.13	-112.54	-209.61	0.885 (0.0457) 0.500	0.581 (0.6052) 0.500	0.699 (0.0535) 0.548	5/7
W.t. F17A V175F T178V L360F	V175A D176F I320F	-104.05	-173.14	-106.68	-204.16	0.959 (0.0101) 0.500	0.947 (0.2076) 1	0.691 (0.0578) 0.571	5/7
W.t. V175F V175A T178V L360F	W.t. F17A D176F I320F	-104.91	-193.53	-122.10	-213.53	0.546 (0.3408) 0.500	0.957 (0.1882) 1	0.689 (0.0587) 0.571	5/7
W.t. F17A V175F L360F	V175A D176F T178V I320F	-127.21	-175.38	-132.69	-183.61	1 (0) 1	0.443 (0.5573) 0	0.716 (0.0460) 0.381	5/7

Table 4.6: Tributyrin training set, test set, calibrated coefficients (a, b, c, d), correlation coefficients for each dataset and accuracy of directional prediction (enhanced or reduced) of the mutation effect according wildtype. *Values are given for Pearson correlation coefficient (Pearson p-value) and Spearman correlation coefficient.

One of the top correlated coefficients is selected from "wildtype-F17A-V175F- L360F" training set and coefficients a, b, c, d is assigned as -127.21, -175.38, -132.69, and 183.61 respectively. Predicted scores and experimental specific activity values are shown in Figure 4.10.
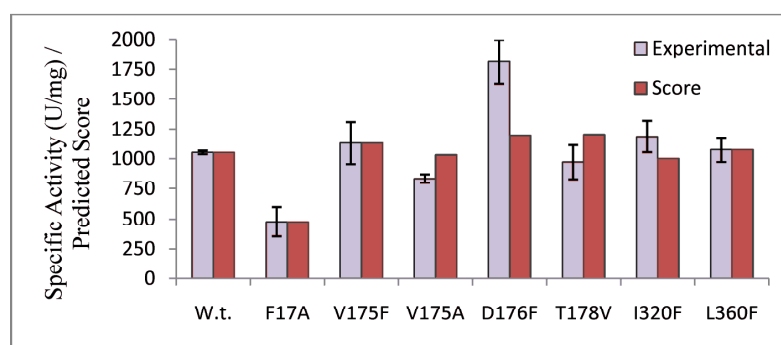

Figure 4.10: Specific activity of tributyrin (4C) and computational score. Wildtype, F17A, V175F and L360F selected as training set for calibration.

4.7. Coefficient Calibration for Tricaprylin (C8) Only

Coefficients of tricaprylin (8C) specific scoring function were calibrated with only the tricaprylin data. Training sets were generated with criteria that have at least three different mutations and wildtype data points. Then coefficients were calibrated and correlation of the training set, test set and overall data were calculated. A number of calibrated coefficients with significant Pearson and Spearman correlation coefficients are shown in Table 4.7.

Training set	Test Set	a	b	c	d	Training set Corr.*	Test Set Corr.*	Overall Corr.*	Accuracy of Predicted Mutation Impact
W.t. F17A D176F I320F	V175F V175A T178V	-54.67	-61.45	-75.30	-10.37	1 (0) 1	0.917 (0.2618) 1.0	0.814 (0.0258) 0.857	5/6
W.t. F17A V175F I320F	V175F D176F T178V	-85.94	-123.27	-153.58	-15.95	1 (0) 1	0.798 (0.4123) 0.5	0.808 (0.0279) 0.750	5/6
W.t. F17A T178V I320F	V175F V175A D176F	-30.73	-14.12	-15.37	-6.10	1 (0) 1	0.999 (0.0267) 1.0	0.776 (0.0402) 0.857	5/6
W.t. F17A V175F T178V I320F	V175A D176F	-87.96	-101.60	-102.09	-66.59	0.852 (0.0669) 1	1 (N.A.) 1	0.848 (0.0158) 0.929	5/6

Table 4.7: Tricaprylin training set, test set, calibrated coefficients (a, b, c, d), correlation coefficients for each dataset and accuracy of directional prediction (enhanced or reduced) of the mutation effect according wildtype. *Values are given for Pearson correlation coefficient (Pearson p-value) and Spearman correlation coefficient.

One of the best correlated coefficients is selected from "wildtype-F17A-D176F-I320F" training set and coefficients a, b, c, d is defined as -54.67, -61.45, -75.30, and 10.37, respectively. Predicted scores versus experimental specific activity are shown in Figure 4.11.
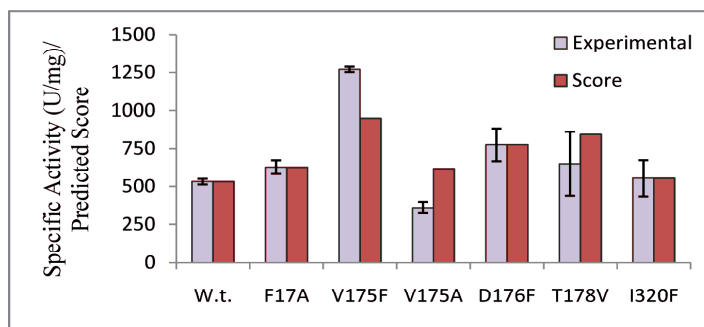
Figure 4.11: Specific activity of tricaprylin (8C) and computational score. Wildtype, F17A, D176F and I320F selected as training set for calibration.

CHAPTER V

5. DISCUSSION

We developed a scoring function to predict the impact of a mutation on substrate specificity. Calibrated coefficients of electrostatics, vdW, desolvation and conformational entropy were able to distinguish the reducing and enhancing mutations in a robust and rapid way. When coefficients were calibrated for both substrates, correlation coefficient of training set was 0.836, test-set was 0.713 and overall data was 0.793. In this training set, wildtype and two mutants (F17A and V175A) were able to generate powerful coefficients that can predict the impact of other four mutations (V175F, D176F, T178V, I320F) with correlation r=0.713 (p=0.047). Apart from Pearson correlation, Spearman correlation coefficient for the test set was r=0.762 (p=0.0368) and for the overall data was r=0.895 (p=0) which indicate the robustness of our method in ranking. After calibration, scoring method correctly predicted the enhancement of tributyrin-V175F, tributyrin-D176F, tricaprylin-F17A, tricaprylin-V175F, tricaprylin-D176F, tricaprylin-T178V, tricaprylin-I320F and the reduction of tributyrin-F17A. As shown in Figure 4.8, ranking of substrate specificities within the mutant were 100% correctly predicted by our scoring function. Therefore our method would be used for discriminating substrates specificities and deciding on the impact of a mutation (whether it enhances or reduces) in a rapid and accurate manner (overall correlation r=0.793, p=0.0007). In addition, regardless of the change in training set, the overall correlation stayed secure between 0.76 and 0.81 as shown in Figure 4.7. This fact reveals the data-independent nature of our scoring function which allows adapting it into other enzyme-substrate problems.

In 2009 [64], pH-stat assay was performed at 55°C in pH 8.5 which is somewhat different from our experimental procedure as we performed assays at 55°C, and pH 8.0. Because of the different assay conditions (pH or temperature) or variation in

emulsification of triglycerides, specific activity of tributyrin (4C) and tricaprylin (8C) shows variation from the 2009 article; therein 4C/8C specific activity ratio is close to one which is dissimilar from our measurements. Nevertheless, our scoring function correctly predicts the tricaprylin (8C) specificity which is enhanced by F181A and F182A mutations and especially higher for F181A against F182A (Figure 4.9). However, predictions for specificity of tributyrin (4C) may be negatively affected by the variation in experimental values and lead to wrong results (Figure 4.9).

When coefficients were calibrated for tributyrin (4C), the correlation coefficient of the training set was 1, test-set was 0.443 and overall data was 0.716. It shows a lower correlation when compared with the correlation values that come from both substrates. This impact originated from the low amounts of data points which negatively affected both the training and the test set correlations. In order to generate more accurate scoring function coefficients, sampling data should be enlarged. However, when coefficients were calibrated for tricaprylin (8C), correlation coefficient of the training set was 1, test-set was 0.917 and overall data was 0.814. This high correlation continued in the ranking correlations as the overall Spearman correlation was 0.857. These results show the high accuracy of our method even with a low amount of training and test samples. In this way our method can be easily adapted into only single substrate problems with high accuracy. Because using the same substrate in all simulations would decrease the variation in binding pattern that comes from the intrinsic factors of the specific substrates, we suggest this method would be used more efficiently to predict the impact of a mutation for the one specific substrate.

Since our scoring function is not an absolute free energy method, it requires calibration of each scoring parameter for the selected protein or protein family in order to get more accurate results. Other factors that can affect the specific activity are the internal energy of ligand, size/shape of the contact surface, hydrophobicity of binding pocket, and volume of binding cavity. These parameters can be adapted into our scoring function to improve the overall correlation with experimental data.

Four nanosecond MD simulations give critical information about substrate stabilization, complementarity and binding affinity. Favoured substrates were stabilized near the catalytic triad. On the other hand, unstable tricaprylin (8C) − 360F mutant simulation

gives insights about the negative effect of mutation into the binding modes, as substrate specificity was also shown to be decreased experimentally (Figure 4.8A).

According to the RMSD analysis, except for the simulations of tributyrin-V175A and tricaprylin-V175A, other simulations showed stabilization pattern for both protein and ligand after 2 ns (Figure 4.4). The alternative binding modes of V175A mutant may decrease the specific activity for both substrates. Our scoring function didn't explain the decrease of specific activity and gave better score. Nevertheless for the both cases, calculated desolvation energy increased by 1-2 kcal/mol which gives a signal for unfavourable binding modes.

The average distances of important interactions (Table 4.2) have provided significant information about the specific activity of the substrate. In particular, enhanced specific activities were seen with the decreased distances of d1, d5 and d6 when compared to the wildtype. Largest decrease in d1, d5 and d6 distances were seen in D176F, F181A and F182A mutants for tributyrin (C4) and 175A, F181A and F182A mutants for tricaprylin (8C). Since these entire six mutants experimentally showed the biggest improvements in specific activity, the distance between attacking serine oxygen atom and sp2 carbon atom of the substrate (d1) and distance between oxyanion hole and substrate oxygen (d5, d6) are good candidates for predicting the impact of a mutation.

Substrate specificity is associated with the shape of the binding site. As decreasing its size would cause steric blockage with the substrate, while increasing its size would create additional space which may lead to alternative binding modes and thus decrease substrate specificity [6]. This idea is supported in the mutants of D176F and I320F in which the volume of the binding pocket is decreased with bulky phenylalanine mutation and their specificity is enhanced towards tributyrin (4C). In these mutants, short-length substrate (4C) can be more efficiently stabilized and alternative binding modes in the large cavity can be avoided.

It has been proposed that tight packing of the active site with hydrophobic residues would contribute to thermoactivity and stability of the enzyme at high temperatures [69]. For the case of F17A mutation, bulky hydrophobic residue (F17) around the catalytic serine mutated to small residue alanine would negatively affect thermoactivity and stability of the enzyme at elevated temperatures [69]. Our results showed that specific activity of tributyrin (4C) is decreased by half and tricaprylin (8C) is increased

by 97 U/mg. The decrease of specific activity for tributyrin (4C) may arise from the destabilization of the active site which leads to the reduction of interaction energy in MD runs (Table 4.3).

Substrate binding is mainly directed by the head part of the substrate which forms crucial interactions for catalysis. Any mutation to catalytic residues (Ser-His-Asp) or their nearby residues would probably decrease the catalytic activity and make major impact on substrate selectivity. However, our aim was to change chain length specificity of the substrates by mutating the residues that are located at the binding pockets, and interact with substrate branches (F17, D176, V175, T178, I320, and L360). We proposed that these sites are directly related to chain length specificity of the triglycerides and we did not expect to damage the mechanism of the catalysis. As expected, most of the mutations did not negatively affect the catalytic activity, but instead changed its specificity towards particular substrate [F175F and F17A for tricaprylin (8C) and D176F for tributyrin (4C)].

CHAPTER VI

6. CONCLUSIONS AND FUTURE WORK

We developed a scoring function that can be used for discriminating substrates specificities and deciding on the impact of a mutation (whether it enhances or reduces activity) in a rapid and accurate manner (overall correlation r=0.793, p=0.0007), before doing time-consuming and laborious experimental assays. Our predictions are experimentally verified for BTL2 lipase. After calibration of scoring parameters, our method correctly predicted the enhancement of tributyrin-V175F, tributyrin-D176F, tricaprylin-F17A, tricaprylin-V175F, tricaprylin-D176F, tricaprylin-T178V, tricaprylin-I320F and the reduction of tributyrin-F17A substrate-mutant pairs. Also, ranking of substrate specificities within the mutants were 100% correct. This method would be efficiently adapted to other protein families to predict the impact of a mutation for the one specific substrate or multiple substrates.

Currently, our algorithm uses vdW energy, electrostatic energy, desolvation energy and conformational entropy terms, however, with small improvement prediction accuracy can be improved by adding other features, such as internal energy of ligand, size/shape of the contact surface, hydrophobicity of the binding pocket, and volume of binding cavity.

Theoretically, this methodology is expected to give good results within the same protein family. In order to verify this assumption, different lipase-substrate pairs should be analyzed in future work. Also these studies would investigate the nature of our scoring function for its universality or dependence to protein/substrate. In addition, the experimental data should be enlarged with other probable mutants to increase the confidence of calibration. In this manner, the overall correlation would be explored and confidence of our scoring function would be tested.

7. REFERENCES

[1] Schrag JD, Li Y, Cygler M, Lang D, Burgdorf T, Hecht HJ, et al. The open conformation of a Pseudomonas lipase. Structure 1997;5:187-202.

[2] Guncheva, M. and D. Zhiryakova (2011). "Catalytic properties and potential applications of Bacillus lipases." Journal of Molecular Catalysis B: Enzymatic 68(1): 1-21.

[3] Brzozowski, A. M., U. Derewenda, et al. (1991). "A model for interfacial activation in lipases from the structure of a fungal lipase-inhibitor complex." Nature 351(6326): 491-494.

[4] Derewenda U, Brzozowski AM, Lawson DM, Derewenda ZS. Catalysis at the interface: the anatomy of a conformational change in a triglyceride lipase. Biochemistry 1992;31:1532-41.

[5] Uppenberg J, Ohrner N, Norin M, Hult K, Kleywegt GJ, Patkar S, et al. Crystallographic and molecular-modeling studies of lipase B from Candida antarctica reveal a stereospecificity pocket for secondary alcohols. Biochemistry 1995;34:16838-51.

[6] Pleiss J, Fischer M, Schmid RD. Anatomy of lipase binding sites: the scissile fatty acid binding site. Chemistry and physics of lipids 1998;93:67-80.

[7] Berger M, Schneider MP. Lipases in organic solvents: The fatty acid chain length profile. Biotechnology Letters 1991;13:641-5.

[8] Verger R. 'Interfacial activation' of lipases: facts and artifacts. Trends in Biotechnology 1997;15:32-8.

[9] Gupta R, Gupta N, Rathi P. Bacterial lipases: an overview of production, purification and biochemical properties. Applied microbiology and biotechnology 2004;64:763-81.

[10] Lanser AC, Manthey LK, Hou CT. Regioselectivity of new bacterial lipases determined by hydrolysis of triolein. Current microbiology 2002;44:336-40.

[11] Gilbert EJ, Cornish A, Jones CW. Purification and properties of extracellular lipase from Pseudomonas aeruginosa EF2. Journal of general microbiology 1991;137:2223-9.

[12] Misset O, Gerritse G, Jaeger KE, Winkler U, Colson C, Schanck K, et al. The structure-function relationship of the lipases from Pseudomonas aeruginosa and Bacillus subtilis. Protein engineering 1994;7:523-9.

[13] Lesuisse E, Schanck K, Colson C. Purification and preliminary characterization of the extracellular lipase of Bacillus subtilis 168, an extremely basic pH-tolerant enzyme. European journal of biochemistry / FEBS 1993;216:155-60.

[14] Dharmsthiti S, Luchai S. Production, purification and characterization of thermophilic lipase from Bacillus sp. THL027. FEMS microbiology letters 1999;179:241-6.

[15] Finkelstein AE, Strawich ES, Sonnino S. Characterization and partial purification of a lipase from Pseudomonas aeruginosa. Biochimica et biophysica acta 1970;206:380-91.

[16] Sugiura M, Oikawa T, Hirano K, Inukai T. Purification, crystallization and properties of triacylglycerol lipase from Pseudomonas fluorescens. Biochimica et biophysica acta 1977;488:353-8.

[17] Kordel M, Hofmann B, Schomburg D, Schmid RD. Extracellular lipase of Pseudomonas sp. strain ATCC 21808: purification, characterization, crystallization, and preliminary X-ray diffraction data. Journal of bacteriology 1991;173:4836-41.

[18] Horiuti Y, Imamura S. Purification of lipase from Chromobacterium viscosum by chromatography on palmitoyl cellulose. Journal of biochemistry 1977;81:1639-49.

[19] Anguita J, Rodriguez Aparicio LB, Naharro G. Purification, gene cloning, amino acid sequence analysis, and expression of an extracellular lipase from an Aeromonas hydrophila human isolate. Applied and environmental microbiology 1993;59:2411-7.

[20] Arpigny JL, Jaeger KE. Bacterial lipolytic enzymes: classification and properties. The Biochemical journal 1999;343 Pt 1:177-83.

[21] Secundo F, Carrea G, Tarabiono C, Gatti-Lafranconi P, Brocca S, Lotti M, et al. The lid is a structural and functional determinant of lipase activity and selectivity. Journal of Molecular Catalysis B: Enzymatic 2006;39:166-70.

[22] Palomo JM, Fernandez-Lorente G, Ortiz C, Segura RL, Mateo C, Fuentes M, et al. Conformational Engineering of Lipases via Directed Immobilisation: Improving the Resolution of Chiral Drugs. Medicinal Chemistry Reviews - Online 2005;2:369-78.

[23] Sharma R, Chisti Y, Banerjee UC. Production, purification, characterization, and applications of lipases. Biotechnology Advances 2001;19:627-62.

[24] Aravindan R, Anbumathi P, Viruthagiri T. Lipase applications in food industry. Indian Journal of Biotechnolgy. 2007;6:141–158

[25] Carrasco-Lopez C, Godoy C, de Las Rivas B, Fernandez-Lorente G, Palomo JM, Guisan JM, et al. Activation of bacterial thermoalkalophilic lipases is spurred by dramatic structural rearrangements. The Journal of biological chemistry 2009;284:4365-72.

[26] Kim HK, Park SY, Lee JK, Oh TK. Gene cloning and characterization of thermostable lipase from Bacillus stearothermophilus L1. Bioscience, biotechnology, and biochemistry 1998;62:66-71.

[27] Jeong ST, Kim HK, Kim SJ, Chi SW, Pan JG, Oh TK, et al. Novel zinc-binding center and a temperature switch in the Bacillus stearothermophilus L1 lipase. The Journal of biological chemistry 2002;277:17041-7.

[28] Nardini M, Lang DA, Liebeton K, Jaeger KE, Dijkstra BW. Crystal structure of pseudomonas aeruginosa lipase in the open conformation. The prototype for family I.1 of bacterial lipases. The Journal of biological chemistry 2000;275:31219-25.

[29] Quyen DT, Schmidt-Dannert C, Schmid RD. High-level expression of a lipase from Bacillus thermocatenulatus BTL2 in Pichia pastoris and some properties of the recombinant lipase. Protein expression and purification 2003;28:102-10.

[30] Bradoo S, Saxena RK, Gupta R. Two acidothermotolerant lipases from new variants of Bacillus spp. World Journal of Microbiology and Biotechnology 1999;15:87-91.

[31] Huang SY, Grinter SZ, Zou X. Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. Physical chemistry chemical physics : PCCP 2010;12:12899-908.

[32] Huang N, Kalyanaraman C, Irwin JJ, Jacobson MP. Physics-Based Scoring of Protein-Ligand Complexes:? Enrichment of Known Inhibitors in Large-Scale Virtual Screening. Journal of Chemical Information and Modeling 2005;46:243-53.

[33] Wang W, Donini O, Reyes CM, Kollman PA. Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. Annual review of biophysics and biomolecular structure 2001;30:211-43.

[34] Adcock SA, McCammon JA. Molecular dynamics: survey of methods for simulating the activity of proteins. Chemical reviews 2006;106:1589-615.

[35] Rocchia W, Sridharan S, Nicholls A, Alexov E, Chiabrera A, Honig B. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. Journal of computational chemistry 2002;23:128-37.

[36] Still WC, Tempczyk A, Hawley RC, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. Journal of the American Chemical Society 1990;112:6127-9.

[37] Bhattacharyya SM, Wang Z-G, Zewail AH. Dynamics of Water near a Protein Surface. The Journal of Physical Chemistry B 2003;107:13218-28.

[38] Figueirido F, Del Buono GS, Levy RM. Molecular mechanics and electrostatic effects. Biophysical chemistry 1994;51:235-41.

[39] Yu Z, Jacobson MP, Josovitz J, Rapp CS, Friesner RA. First-Shell Solvation of Ion Pairs:? Correction of Systematic Errors in Implicit Solvent Models†. The Journal of Physical Chemistry B 2004;108:6643-54.

[40] Dong F, Olsen B, Baker NA. Computational methods for biomolecular electrostatics. Methods in cell biology 2008;84:843-70.

[41] Liu HY, Grinter SZ, Zou X. Multiscale generalized born modeling of ligand binding energies for virtual database screening. The journal of physical chemistry B 2009;113:11793-9.

[42] Wesson L, Eisenberg D. Atomic solvation parameters applied to molecular dynamics of proteins in solution. Protein science : a publication of the Protein Society 1992;1:227-35.

[43] Huey R, Morris GM, Olson AJ, Goodsell DS. A semiempirical free energy force field with charge-based desolvation. Journal of computational chemistry 2007;28:1145-52.

[44] Stouten PFW, Frömmel C, Nakamura H, Sander C. An effective solvation term based an a atomic occupancies for use in protein simulations. Molecular Simulation 1993;10:97–120

[45] Bohm HJ. Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. Journal of computer-aided molecular design 1998;12:309-23.

[46] Koppensteiner WA, Sippl MJ. Knowledge-based potentials--back to the roots. Biochemistry Biokhimiia 1998;63:247-52.

[47] Huang SY, Zou X. An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function. Journal of computational chemistry 2006;27:1876-82.

[48] Thomas PD, Dill KA. Statistical potentials extracted from protein structures: how accurate are they? Journal of molecular biology 1996;257:457-69.

[49] Durrant J, McCammon JA. Molecular dynamics simulations and drug discovery. BMC Biology 2011;9:71.

[50] Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, et al. Scalable molecular dynamics with NAMD. Journal of computational chemistry 2005;26:1781-802.

[51] Ji L, Xiaoling T, Hongwei Y. Prediction of the enantioselectivity of lipases and esterases by molecular docking method with modified force field parameters. Biotechnology and bioengineering 2010;105:687-96.

[52] Otte N, Bocola M, Thiel W. Force-field parameters for the simulation of tetrahedral intermediates of serine hydrolases. Journal of computational chemistry 2009;30:154-62.

[53] Alonso H, Bliznyuk AA, Gready JE. Combining docking and molecular dynamic simulations in drug design. Medicinal research reviews 2006;26:531-68.

[54] Yuriev E, Agostino M, Ramsland PA. Challenges and advances in computational docking: 2009 in review. Journal of molecular recognition : JMR 2011;24:149-64.

[55] Li W, Ode H, Hoshino T, Liu H, Tang Y, Jiang H. Reduced Catalytic Activity of P450 2A6 Mutants with Coumarin: A Computational Investigation. Journal of Chemical Theory and Computation 2009;5:1411-20.

[56] Lafleur K, Huang D, Zhou T, Caflisch A, Nevado C. Structure-Based Optimization of Potent and Selective Inhibitors of the Tyrosine Kinase Erythropoietin Producing Human Hepatocellular Carcinoma Receptor B4 (EphB4). Journal of Medicinal Chemistry 2009;52:6433-46.

[57] Moura-Tamames SA, Ramos MJ, Fernandes PA. Modelling ß-1,3-exoglucanase–saccharide interactions: Structure of the enzyme–substrate complex and enzyme binding to the cell wall. Journal of Molecular Graphics and Modelling 2009;27:908-20.

[58] H. B. Callen, Thermodynamics and An Introduction to Thermostatistics. John Wiley & Sons, New York, 1985.

[59] Meirovitch H. Recent developments in methodologies for calculating the entropy and free energy of biological systems by computer simulation. Current Opinion in Structural Biology 2007;17:181-6.

[60] Roux B. The calculation of the potential of mean force using computer simulations. Computer Physics Communications 1995;91:275-82.

[61] Mills M, Andricioaei I. An experimentally guided umbrella sampling protocol for biomolecules. The Journal of chemical physics 2008;129:114101.

[62] Liu X, Wang X, Jiang H. A steered molecular dynamics method with direction optimization and its applications on ligand molecule dissociation. Journal of Biochemical and Biophysical Methods 2008;70:857-64.

[63] Protein-ligand database. http://lpdb.chem.lsa.umich.edu/ (accessed July 5, 2010).

[64] Karkhane AA, Yakhchali B, Jazii FR, Bambai B. The effect of substitution of Phe181 and Phe182 with Ala on activity, substrate specificity and stabilization of

substrate at the active site of Bacillus thermocatenulatus lipase. Journal of Molecular Catalysis B: Enzymatic 2009;61:162-7.

[65] LaConte LE, Voelz V, Nelson W, Enz M, Thomas DD. Molecular dynamics simulation of site-directed spin labeling: experimental validation in muscle fibers. Biophysical journal 2002;83:1854-66.

[66] Showalter SA, Brüschweiler R. Validation of Molecular Dynamics Simulations of Biomolecules Using NMR Spin Relaxation as Benchmarks:? Application to the AMBER99SB Force Field. Journal of Chemical Theory and Computation 2007;3:961-75.

[67] Markwick PRL, Cervantes CF, Abel BL, Komives EA, Blackledge M, McCammon JA. Enhanced Conformational Space Sampling Improves the Prediction of Chemical Shifts in Proteins. Journal of the American Chemical Society 2010;132:1220-1.

[68] Trott O, Olson AJ. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of computational chemistry 2010;31:455-61.

[69] Jeong ST, Kim HK, Kim SJ, Chi SW, Pan JG, Oh TK, et al. Novel zinc-binding center and a temperature switch in the Bacillus stearothermophilus L1 lipase. The Journal of biological chemistry 2002;277:17041-7.

[70] Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, et al. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. Journal of computational chemistry 1998;19:1639-62.

[71] Azoia NG, Fernandes MM, Micaêlo NM, Soares CM, Cavaco-Paulo A. Molecular modeling of hair keratin/peptide complex: Using MM-PBSA calculations to describe experimental binding results. Proteins: Structure, Function, and Bioinformatics 2012;80:1409-17.

[72] Jarzynski C. Nonequilibrium Equality for Free Energy Differences. Physical Review Letters 1997;78:2690-3.