Modeling and Engineering Proteins Thermostability

THÈSE N^{O} 6637 (2015)

PRÉSENTÉE LE 16 JUIN 2015
À LA FACULTÉ DES SCIENCES DE LA VIE
UNITÉ DU PROF. DAL PERARO
PROGRAMME DOCTORAL EN BIOTECHNOLOGIE ET GÉNIE BIOLOGIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Hasan PEZESHGI MODARRES

acceptée sur proposition du jury:

Prof. G. Fantner, président du jury
Prof. M. Dal Peraro, directeur de thèse
Prof. M. Cascella, rapporteur
Prof. P. De Los Rios, rapporteur
Dr M. De Vivo, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2015

Acknowledgements

I would like to sincerely thank Professor Matteo Dal Peraro, my PhD advisor, for his guidance, enthusiastic encouragement and constructive critiques throughout my Thesis research.

I am also grateful to all of my lab mates for creating a friendly and stimulating scientific environment in the Laboratory for Biomolecular Modeling.

It was a great pleasure to collaborate and learn from Professor Melanie Blokesch and Dr. Nicolai Ravin.

Last but not least, I thank my family and friends for their constant support and encouragement.

Abstract

Enzymes have evolved during millions of years to become efficient catalysts for special biochemical reactions within a specific range of working conditions in the cellular environment. The activity of an enzyme is directly related to its folded structure and even slight changes in its 3D conformation may cause irreversible, negative effects on its activity. The structure of enzymes are very sensitive to the environmental conditions and changes from their optimal conditions, like higher temperature, salt concentration, and pH, might result in denaturation and subsequently inactivation. In the past decades, natural enzymes extracted from different organisms have found a wide range of applications in the industrial and biotechnological setting. For the majority of the applications their activity at high temperatures is more favorable, however enzymes have evolved, in most of the cases, to optimally work in limited range of temperature of their native cellular environment. Therefore, enhancing enzyme thermostability will not only increase their application range, but could also shed light into new aspects of their evolution and chemical activity. The physical chemical principles underlying enzymatic thermostability are keys in fact to understand the way evolution has shaped proteins to adapt to a broad range of temperatures. Understanding the molecular determinants at the basis of protein thermostability, using both *in silico* methods and *in vitro* experiments, is also an important way for engineering more thermo-resistant enzymes to be used in the industrial setting, as for instance DNA ligases, which are important for DNA replication and repair and have been long used in molecular biology and biotechnology.

In this thesis I used *in silico* techniques, like molecular modeling and simulation coupled with bioinformatics analyses, to assess existing methods and predict potential thermo-stabilizing mutations for target proteins. First, I studied a thermophilic protein and after exploring the origins of its thermostability I proposed mutations to further increase its thermostability. Then, I took advantage

of what learned from this study to explore further thermostability engineering methods in order to develop faster, accurate, and easy-to-use methods that can be generally used for a broad array of proteins.

1. Understanding and engineering thermostability in the DNA ligase from *Thermococcus sp. 1519* (*LigTh1519*). In this thesis, I first addressed the origin of thermostability in the thermophilic DNA ligase from archaeon *Thermococcus sp.* 1519, and identified thermo-sensitive regions using molecular modeling and simulations. In addition, I predicted mutations that can enhance thermostability of the enzyme through bioinformatics analyses. I showed that thermo-sensitive regions of this enzyme are stabilized at higher temperatures by optimization of charged groups on the surface, and predicted that thermostability can be further increased by further optimization of the network among these charged groups. Engineering this DNA ligase by introducing selected mutations (i.e., A287K, G304D, S364I and A387K) produced eventually a significant and additive increase in the half-life time of the enzyme when compared to the wild-type.

Then, based on what I learned from thermostability analyses and improvement of *LigTh1519*, my aim was to design a general-purpose protein thermostability engineering protocol that can enable thermostability engineering on different protein families. Therefore, I examined different, commonly used sequence- and structure-based methods in the protein thermostability engineering field.

2. Structure-based protein thermostability engineering. I introduced and assessed two effective structure-based methods: (i) the first one, which wants to be an alternative technique to costly molecular dynamics simulations, is using elastic network models to find thermo-sensitive regions on a protein structure. Comparing results of this method with available experimental data indicated that detected regions are potential positions which can be targeted with thermo-stabilizing mutations; (ii) the second strategy relies on the definition of simple rules able to optimize hydrophobic interactions at the protein core for inducing higher thermostability. These rules are shown to be robust in finding thermo-stabilizing and destabilizing mutations.

3. Sequence-based protein thermostability engineering. Despite the clear advantages of structure-based methods, a 3D structure of the target protein is not always available. Thus, to take advantage of the available large-scale data for protein sequences, I examined the most commonly used sequence-based protein thermostability engineering method, namely *Consensus Concept*, and assessed its capability for the first time against a large dataset of published reports. I showed that this method is not very robust and accurate in finding thermo-stabilizing mutations and proposed some simple, but effective, strategies to enhance its predictive power for protein thermostability engineering.

4. A general integrative protocol for protein thermostability engineering. Altogether, my analysis supported the idea that for efficient protein thermostability engineering a combination of sequence- and structure-based methods are needed to design a rational, integrative thermostability engineering protocol. I therefore proposed and evaluated a protocol that takes advantage from the integration of all available data concerning the thermostability properties of a target protein and is able to suggest a limited set of mutations, which, based on my benchmarks, are effective in enhancing thermostability.

In conclusion, during this doctoral work I developed a fast and accurate framework, which combines sequence, structural, and thermostability data with an array of molecular modeling and bioinformatics methods to design more thermostable proteins. I also showed that using different data sources the robustness of the engineering protocol can be significantly enhanced, as for the case of the *LigTh1519* enzyme. The protocol developed in this thesis is the foundation for more rational protein thermostability engineering, which can allow a routinely production of enzymes effectively working at higher temperatures for applications in science and technology.

Keywords: proteins, enzymes, mutation, thermostability, protein engineering, molecular dynamics simulations, elastic network models, modeling, bioinformatics, *Thermococcus sp. 1519*, DNA ligase.

Résumé

Les enzymes ont evolué pendant des millions d’années pour devenir d’efficaces catalyseurs pour des reactions biochimiques spécifiques à divers environnements cellulaires. L’activité d’une enzyme est directement liée à sa structure si bien que même un léger changement dans sa conformation tridimensionnelle peut causer d’irreversibles changements dans son activité. La structure des enzymes est aussi très sensible aux conditions environementales et à tout changement de leur condition optimale, tel que des temperatures ou concentrations salines élevées, changement de pH, qui peuvent provoquer des dénaturations et ainsi des inactivations.

Auparavant, les enzymes naturelles extraites d’organismes differents ont été utilisées de plusieurs façons dans le domaine académique aussi bien que dans celui industriel. En general, ces applications requièrent des températures élevées. Cependant les enzymes ont evolué pour être efficace seulement dans une zone restrainte de température qui correspond à leur environement naturels. C’est pour cela qu’ameliorer leur thermostabilité peut non seulement augmenter leur number d’application mais aussi élucider des nouveaux aspects de leur évolution et activité catalytique.

Les principes physico-chimiques qui gouvernent la thermostabilité des enzymes doivent être pris en compte pour comprendre comment l’evolution a adapté la forme des proteines enzymatiques aux différentes temperatures environementales.

Comprendre les éléments moléculaires qui contribuent à la thermostabilité, en usant des méthodes aussi bien *in silico* qu’*in vitro*, est un des moyens les plus efficaces pour concevoir des enzymes plus résistantes thermiquement qui pourront être ensuite usée pour l’industrie, comme la ligase de l’ADN, enzyme qui est importante pour la réplication et la réparation de l’ADN et qui a été utilisée dans la biologie moléculaire et la biotechnologie.

Dans cette thèse, J'utiliserais donc des techniques in silico, comme la modélisation moléculaire et simulation avec des analyses bioinformatiques, pour évaluer les méthodes existantes et prédire des mutations thermo-stabilisantes potentielles pour des protéines clefs. En premier temps, J'ai étudié une protein thermophile et après avoir exploré les origines de sa thermostabilité, j'ai proposé des mutations pour augmenter sa thermostabilité. Ainsi, j'ai utiliser les connaissances de cetter analyse pour investiguer de façon plus profonde les methodes pour l'ingénierie de thermostabilité dans le but de developer des méthodes plus rapides, faciles et précises qui peuvent être utiliser pour une plus grande variete de protéines enzymatiques.

Contents

Acknowledgements .... **v**

Abstract .... **vi**

Résumé .... **ix**

Contents .... **xi**

Chapter 1 Introduction **1**

- 1.1 Computational strategies for protein thermostability engineering 5
- 1.2 Objectives of this thesis 7

Chapter 2 Computational Methods **11**

- 2.1 Molecular Modeling for Structure-Based Protein Engineering 11
  - 2.1.1 Potential Energy Functions for Modeling Biomolecular Interactions 11
  - 2.1.2 FoldX 14
  - 2.1.3 Molecular Dynamics Simulation 15
  - 2.1.4 Elastic Network Models 23
- 2.2 Sequence-Based Bioinformatics Methods 25
  - 2.2.1 Protein Sequence Alignment 25
  - 2.2.2 Databases and Servers 26
  - 2.2.3 Statistical Analysis of Predictions 28

Chapter 3 Understanding and Engineering Thermostability in the DNA Ligase from *Thermococcus sp. 1519* **31**

- 3.1 Introduction 31
- 3.2 Materials and Methods 34
  - 3.2.1 Computational Procedure 34
  - 3.2.2 Plasmids, Bacterial Strains and Culture Conditions 35
  - 3.2.3 Expression and Purification of Ligth1519 and Its Mutants 36
  - 3.2.4 DNA Ligation Assay 36
- 3.3 Results 38
  - 3.3.1 Origin of LigTh1519 Thermostability 38

3.3.2 Thermostability Engineering in LigTh1519 42
3.3.3 Experimental Validation and Characterization of LigTh1519 Mutants 45
3.4 Discussion 47
3.5 Conclusions 51
Chapter 4 Structure-Based Analyses for Protein Thermostability Engineering 53
4.1 Introduction 54
4.1.1 Finding Thermo-Sensitive Regions 54
4.1.2 Improving Hydrophobic Interactions 56
4.2 Materials and Methods 58
4.2.1 Finding hot spots using a Gaussian Network Model (GNM) 58
4.2.2 Characterizing HB Interactions 62
4.3 Results and Discussion 63
4.3.1 Detection of Thermo-Sensitive Regions 63
4.3.2 Improving HB Interactions 71
4.4 Conclusions 76
Chapter 5 Sequence-Based Protein Thermostability Engineering 79
5.1 Introduction 79
5.2 Materials and Methods 80
5.2.1 Consensus Concept 81
5.2.2 Analysis of the Thermophilic Sequences 83
5.3 Results and Discussion 83
5.3.1 Consensus Concept 83
5.3.2 Analysis of Thermophilic Sequences 88
5.4 Conclusions 90
Chapter 6 A General Framework for Protein Thermostability Engineering 91
6.1 Introduction 91
6.2 Materials and Methods 91
6.2.1 Details of the Protocol 91
6.2.2 Dataset and Protein Sequences 95
6.3 Results and Discussion 97

6.4 Conclusions 102

Chapter 7 Conclusions and Perspectives 103

References 107

Appendix A1: Supplementary Information for Chapter 3 119

Appendix A2: Supplementary Information for Chapter 4 123

Appendix A3: Supplementary Information for Chapter 5 135

Appendix A4: Supplementary Information for Chapter 6 147

Appendix A5 151

Chapter 1 Introduction

Cells need to survive a vast variety of proteins functioning in a highly orchestrated manner. The function of proteins, specifically enzymes, is optimized through evolution to match their growing environmental condition, like temperature and pH [1-4]. While the majority of organisms live in environments where the temperature excursion is fairly limited, mainly at temperatures close to ambient conditions, there is a vast number of microorganisms that have adapted to survive in extreme temperature conditions. Based on their native living temperatures, microorganisms can be categorized into four classes: *psychrophilics*, living at temperatures between 5 °C and 20 °C; *mesophilics*, living at temperatures between 20 °C and 40 °C; *thermophilics*, living at temperatures as high as 40 °C to 80 °C; and *hyperthermophilics* that prefer to live at temperatures higher than 80 °C [5-7]. During the past two decades, remarkable experimental analyses have addressed the characteristics of thermophilic or hyperthermophilic microorganisms. For the first time, in 1967, the existence of hyperthermophilic microorganisms (*Thermus aquaticus*), growing in the Yellowstone National Park hot springs, was reported by Thomas D. Brock [5] and then more microorganisms were discovered living in such harsh conditions [5]. In fact, hyper/thermophilic microorganisms are represented only by Archaea and Bacteria, which have been isolated from different environments like hot springs and geothermally heated oil-containing stratifications [5]. Enzymatic systems of such microorganisms have been interesting for their unique features, especially for their high sustainability at elevated temperatures; while they maintain the functionality at high temperatures, they preserve their folding. Such ability has attracted great attention from scientists to find out the key modulating mechanisms of stability of enzymes [5, 7]. However, the difficulties in large-scale culturing of those microorganisms postponed the purification of the first hyperthermophilic enzymes to the late 1980s [5, 7]. Eventually people figured out that thermostable enzymes can be expressed using mesophilic hosts while keeping their activity and thermal stability preserved, opening new ways to study key features of proteins thermostability [5, 8]

Nowadays, industrial applications for enzymes, like amylases for food processing, proteases for washing powder formulations, cellulases and xylanases for pulp and paper processing are very well established [9] and other applications involving chemical synthesis [10], bioremediation [11], and biosensors [12] are continuously expanding the application of natural enzymes in the industrial setting. But, for an enzyme to be optimally suitable for the industrial setting a number of features are essential to meet including high activity and stability against high temperatures, organic solvents, and high concentration of substrates or products [8, 13]. Specifically, working at high temperatures is favorable for industrial applications because of increased substrate solubility, minimized microbial contamination risk, and increased reaction rates with longer catalyst half-life [5, 7, 14]. Therefore, thermophilic enzymes, which are also resistant to other harsh conditions like chemical denaturant compounds, were the first candidates for this purpose [5, 7]. For example, in the starch processing industries, starch needs to be gelatinized at high temperature to make the polysaccharide accessible for the enzymes [7, 15]. For this purpose, starch processing enzymes like amylases, pullulanases, and glucoamylases have been extracted from thermophilic microorganisms like *Pyrococcus woesei, Thermotoga maritima,* and *Clostridium thermosaccharolyticum* and were used in starch processing industries. Some of thermostable enzymes and their applications are summarized in Table 1-1.

However, the majority of characterized enzymes with potential applications get denatured only at a few degrees above their normal active temperature [16]. In addition, in most cases, the naturally available enzymes or even commercially available ones are not optimal for the usual conditions encountered in industrial and biotechnological chemical processes [8].

Therefore, to take the most out of an enzyme, as a precious treasure given to us by nature, different protein engineering schemes have been developed and applied to modify their properties in a controlled manner. Protein engineering aims to optimize a specific feature of an already available protein with a known function. The features of interest for protein engineering are usually stability and function [17]. Because of different advantages of enzyme activity at elevated temperatures, as discussed above, protein thermostability engineering (PTE) has been the subject of a number of researches to prepare enzymes with lower thermostability for high-temperature applications by introducing a minimum number of mutations [18-26].

Table 1-1. Some thermophilic enzymes with industrial applications. This table is adapted from reference [7]

Enzyme	**Origin**	**Application**
Taq polymerase	*T. aquaticus*	PCR technologies
Deep Vent DNA polymerase	*P. furiosus*	PCR technologies
Tcs DNA ligase	*Thermus scodoductus*	Ligase chain reaction
Carboxypeptidase	*S. solfataricus*	C-terminal sequencing
α -Amylase	*Pyrococcus woesei*	Starch processing
Pullulanase	*Thermotoga maritima*	Starch processing
Glucoamylase	*Clostridium thermosaccharolyticum*	Starch processing
Endo-1,4-b-glucanase	*T. maritima*	Cellulose degradation
Endoxylanase	*Thermotoga* sp. strain FjSS3-B.1	Paper pulp bleaching
β-Mannanase	*Rhodothermus marinus*	Softwood pulp bleaching; coffee bean treatment and coffee extraction
b-Xylosidase	*Thermoanaerobacterium saccharolyticum*	Carbohydrate synthesis; xylose Production
Esterase	*P. furiosus*	Transesterification and ester synthesis
Secondary alcohol dehydrogenase	*Thermoanaerobacter ethanolicus*	Chemical synthesis
Pectate lyase	*Thermoanaerobacter italicus*	Fruit juice clarification
b-Galactosidase	*T. maritima*	Production of lactose-free dietary milk products

Protein engineering, besides all valuable applications, still faces several limitations. The main obstacle is the extremely vast sequence space that is needed to be sampled and understood in order to optimize a desired feature of a protein [17]. For example, for a protein of 300 residues, the sequence space has 20^{300} possibilities, assuming that we only use the 20 standard amino acids. Therefore, to decrease the enormous size of this mutations library, efficient strategies are needed to make the sequence space as small as possible. Aiming at thermostability engineering of proteins, different *in silico* techniques have been developed and used as guidelines for mutation library design and mutagenesis experiments. The *in silico* PTE approaches are mainly based on the analysis of the protein sequence and structure [8, 17]. These techniques use the key principles/factors of thermo-stabilizing proteins and propose mutations that can make a protein stronger against high temperatures by satisfying the given principles [18-22]. To find protein thermo-stabilization key factors, two sources of data are usually used: available protein sequences/structures and mutagenesis experiments that target thermostability of proteins by single or multiple mutations. Using these sources one can decode the strategies that nature has used to enhance thermostability of proteins by taking advantage of rapidly growing number of protein sequences and structures. Comparing proteins belonging to species with different preferred living temperatures can reveal indeed the principles of protein thermostability [8].

Comparison of sequence/structures of the isolated proteins from thermophilic and hyperthermophilic species with those of mesophilic ones has provided valuable insights about how thermostability is developed and encoded in the protein sequence [8, 27-31]. Hydrophobic core packing improvement, increasing of hydrogen bonding network, surface charge–charge interactions optimization, increasing salt-bridges and disulfide bonds, more favorable helical dipolar interactions, improving secondary structure propensity, increasing rigidity, and entropic stabilization are some of the factors that mainly provide larger thermostabilization to proteins and have thus been used in a number of rational PTE studies [5, 6, 18-22, 27-30, 32-34]. It is noteworthy that some issues, like the correlation of protein rigidity and the thermostability, have been controversial [35-40].

Although these are general rules found by exploring datasets of proteins, they should be considered carefully for PTE [41-43]. For example, although reports indicate the substitution of glycine with proline is one of the most thermo-stabilizing strategy [43, 44], not all glycine to proline mutations result in thermo-stabilization of proteins and other factors like thermal fluctuation of residues should be taken into account as well [43]. Furthermore, reports show that these strategies may vary significantly among protein families [45]. In other words, for each individual protein one should look for specific working strategies. However, reports also indicate that the mentioned general strategies are still valid guidelines for protein thermostability engineering [46]. Altogether, although a large number of thermo-stabilizing features have been reported, it is now clear that there are no universal rules that univocally modulate the thermostability of enzymes [5].

1.1 Computational strategies for protein thermostability engineering

Computational PTE approaches have shown promising results in several studies [47, 48]. From a computational point of view, there are two main strategies to attack in a rational way the PTE problem: on one hand one can use uniquely information originating from the protein sequence, on the other, if available, the 3D structure of the target protein and related homologous proteins can greatly help the PTE process [47, 48].

Structure-based engineering techniques mainly rely on the native target structure to find flexible/weak regions by B-factor analysis or molecular dynamics (MD) simulations and stabilize them by suggesting appropriate mutations [18-26]. To stabilize the thermo-weak regions on a protein structure, different strategies have been utilized including increasing hydrogen bonding, salt-bridges, and disulfide bonds and optimization of hydrophobic interactions [22-25].

However, these approaches have always to face the issue of the limited number of available protein structures [47, 48]. In addition, a tradeoff between thermostability and activity, by disruption of catalytic or other functional residues upon thermo-stabilizing mutations, has been reported [49]. On the other hand, sequence-based engineering takes advantage of the huge, and still growing, amount of protein sequence data. This approach to PTE, which is also known as data driven engineering, mainly focuses on statistical analysis of the amino acids distribution at each point position on the target sequence over a multiple sequence alignment of homologous sequences [48, 50]. Its main advantage, compared to structure-based engineering, is that rich number of information can be retrieved from the available sequence ensemble. In addition, using this approach all the analysis is focused on a specific protein family with the result to minimize the risk to perturb the main activity of the target protein preserving functional residues within the protein family. However, sequence based techniques usually suffer from the limited ability to improve thermostability compared to when structure-based methods are used [48, 50]. It has been shown that using sequence-based methods usually multiple mutations are needed to get the same thermostability enhancement that could be achieved by only a single mutation predicted using structure-based engineering [48, 50]. Therefore, it appears as an obvious conclusion that combining these two complementary approaches to PTE would be the optimal solution to find appropriate sites and their corresponding mutations for PTE.

To develop any computational PTE method or protocol, a final assessment is needed to evaluate the accuracy of the method and compare it with other methods. Fortunately, there is a growing number of mutagenesis experiments that evaluate different thermodynamic properties, like changes in the unfolding midpoint temperature (ΔT_m), by single or multiple mutations. The accumulated experimental studies in freely available databases, like ProTherm [51], are valuable resources for assessment of newly developed computational methods, or training models to predict changes in protein stability upon mutations. Based on such experimental data, it is important to consider that the point mutation effects on protein thermostability can be well approximated as independent and additive [52]. Such an approximation is ideal for protein thermostability engineering because design of different effective thermo-stabilizing mutations can result in a significant thermo-stabilizing effect [52]. This observation puts more emphasis on the importance of taking advantage of different protein engineering strategies to design mutations with the most probable increase in thermo-stabilization [16, 52].

1.2 Objectives of this thesis

Understanding the origin of thermostability of thermophilic proteins is a valuable key to conduct rational PTE. Although different methods and tools have been developed to predict the effect of mutations on stability of proteins, systematic investigation of the robustness of each method with comprehensive exploration of their strengths and weaknesses to develop high level rational PTE platform is still weakly addressed. In addition, the tradeoff between thermostability and activity produced by mutations aiming at enhancing thermostability is not clarified yet [49]. Then, more investigations are needed to find appropriate combinations of different tools and methods to increase the chance of thermostability enhancement and simultaneously decreasing the risk of compromising native activity. In the ideal situation one would tend toward the simultaneous increment of both thermostability and catalytic activity of enzymes [53]. To address these problems, in this thesis I investigated thermostability affecting factors for a specific protein target and, starting from this knowledge, I developed a general protocol for PTE.

In particular, following these premises and after giving a general introduction in Chapter 2 about the molecular modeling and bioinformatics methods used throughout this work, the thesis is organized as follow:

Chapter 3. Understanding and engineering thermostability in the DNA ligase from *Thermococcus sp. 1519* (*LigTh1519*). In this chapter I focused on the in-depth in silico analysis of thermostabilization properties of DNA ligases, which are in charge of repairing DNA strands by joining the adjacent 3’-hydroxide and 5’-phosphate groups in a nicked single strand DNA by formation of a phosphodiester bond [54]. Engineering DNA ligases to gain resistance against higher temperature can result in valuable applications like optimization of PCR protocols [55]. Here, I first studied the origin of thermostability of a thermophilic DNA ligase, namely *LigTh1519,* from *Thermococcus sp. 1519.* Then, by analysis of its sequence and structure, I rationally engineered it to enhance its thermostability by proposing four distinct mutations. Experimental assessments demonstrated a significant thermo-stabilization effect of the mutations by increasing the half-inactivation time of the enzyme at 94°C from 8 to 41 minutes without a significant reduction of specific catalytic activity.

Following the knowledge gained from this study on the *LigTh1519* enzyme, I continued in the following chapters to study, develop, and assess simple and fast methods in order to combine them in a general PTE protocol, which would be applicable to different protein families.

Chapter 4. Structure-based protein thermostability engineering. In this chapter I mainly focused on two effective structure-based engineering approaches: (i) finding thermo-sensitive regions on the structure and (ii) the optimization of hydrophobic (HB) interactions. In the study on *LigTh1519* (Chapter 3) I used MD simulations to find thermo-sensitive regions. But, MD simulations are usually costly. Therefore, faster tools are needed to find thermo-sensitive regions. I thus used elastic network models as a simple and fast strategy to find thermo-sensitive regions on protein structures. In this chapter, I also addressed the importance of engineering HB interactions mainly stabilizing the core of the target protein. For this purpose, I developed and tested a simple and fast protocol that could find and optimize HB interactions accurately. Combining these two approaches together one can cover the most important challenges and limitations in structure-based protein engineering and is in the position to suggest several promising mutation candidates.

Chapter 5. Sequence-based protein thermostability engineering. Although there is a huge amount of sequence data and several studies have used this in PTE, a systematic analysis of the methods used to assess their robustness and increasing their performance is still missing. I here focused on the most popular method, called *consensus concept* method, and assessed its performance. Besides the general limitation of sequence-based methods, which usually can increase T_m only by less than 3°C, the suggested mutations by the consensus concept method are usually a mixture of stabilizing, destabilizing, and neutral effects [48, 50]. Therefore, I suggested straightforward and simple way to cover this limitation by using complementary data available for the target sequences in order to select only the stabilizing mutations for PTE.

Chapter 6. A general protocol for protein thermostability engineering. My observations from engineering the *LigTh1519* enzyme and analyses on structure- and sequence-based methods strongly supported the idea that one needs to take advantage of all available inputs to achieve actual thermostabilization for a protein target. In this chapter, I used rational combinations of sequence and structure-based PTE methods to develop a framework able to study and predict the most effective thermostability enhancing factors for a protein target. This framework, assessed against published experimental data, is able to effectively predict thermo-stabilizing mutations for PTE, filtering out at the same time thermo-destabilizing mutations for different protein families.

Finally, I also report in the appendix the results of work done on a subject that differs from the main objective of my thesis. In this work, I conducted protein modeling and simulation for better understanding the biological function of the competence protein ComEA from *Vibrio cholera*. My contribution, within the collaboration led Prof. M. Blokesch at Laboratory of Molecular Microbiology – EPFL, was key for the finding of the DNA binding mode of ComEA and for proposing a possible DNA uptake mechanism [56].

Chapter 2 Computational Methods

This chapter describes the theoretical background of the techniques used in this thesis to address the problem of protein thermostability engineering. First, I focus on techniques that are relying on the knowledge of the molecular structure of the protein target, such as molecular dynamics (MD) simulation and its analysis based on classical principles of molecular mechanics. In the second part I discuss bioinformatics techniques, which are instead useful to analyze the broad body of protein sequence data available in current databases.

2.1 Molecular Modeling for Structure-Based Protein Engineering

2.1.1 Potential Energy Functions for Modeling Biomolecular Interactions

Potential energy functions describe the molecular interactions within the system of interest. In this thesis, I always used an all-atom (AA) representation of biomolecular systems, in which atoms are described as spheres linked by springs and interactions are represented by specific potential energy functions. This set of functions along with the associated set of parameters is termed *force field*. In this thesis I mainly used the AMBER force field [57] that has been shown over the past years to be one of the best for simulation of proteins and nucleic acids. The AMBER force field is parameterized using high level quantum mechanical calculations and the strategy of parameterization is described in ref. [58]. The functional form of the AMBER force field is composed of bonded and non-bonded terms. Following the assumption that these contributions are additive and transferable one can write the functional form as:

$$\begin{aligned} U_{MM} &= U_{bonded} + U_{non-bonded} \\ &= U_{bond} + U_{angle} + U_{dihedral} + U_{improper} + U_{vdW} + U_{Coulomb} \end{aligned} \qquad \text{Eq. 2.1}$$

where the first four terms describe the bonded potential energy, and the last two terms the non-bonded potential, namely van der Waals (vdW) and Coulomb interactions.

In particular, the bond energy, U_{bond}, represents the energy associated with covalent bonds between atoms and is described as:

$$U_{bond} = \sum_{bonds} \frac{k_b}{2}(r - r_0)^2 \qquad \text{Eq. 2.2}$$

where k_b is the force constant and r_0 is the equilibrium distance of the bond. The angle potential term, U_{angle}, models the angles between three consecutive atoms as:

$$U_{angle} = \sum_{angless} \frac{k_\alpha}{2}(\alpha - \alpha_0)^2 \qquad \text{Eq. 2.3}$$

where k_a is the force constant and α_0 is the equilibrium angle.

The dihedral energy term, $U_{dihedral}$, describes the torsional angles as:

$$U_{dihedral} = \sum_{dihedral} k_\phi[1 + \cos(\mathrm{n}\,\emptyset - \emptyset_0)] \qquad \text{Eq. 2.4}$$

where k_ϕ is the force constant and ϕ_0 is the equilibrium torsion and *n* is the multiplicity.

The improper potential term, $U_{improper}$, represents improper torsions of four out of plane atoms and is represented as:

$$U_{improper} = \sum_{improper} k_{\psi}(\psi - \psi_0)^2$$ Eq. 2.5

where k_{ψ} is the force constant and ψ_0 is the equilibrium value. This potential term is commonly used to force the chirality or the planarity of molecules such as the aromatic ring in phenylalanine amino acid.

The vdW non-bonded potential term, U_{vdW}, can be approximated by Lennard-Jones potentials between two non-bonded atoms at the distance *r* as:

$$U_{vdW} = \sum_{pairs} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]$$ Eq. 2.6

where $\varepsilon_{\mathrm{ij}}$ is the potential well depth and σ_{ij} is the distance where the Lennard-Jones potential is equal to zero, and r_{ij} is the distance between the two atoms *i* and *j*.

The vdW potentials are representing short-ranged interactions, and for this reason it is possible to introduce a cutoff along with a switching distance to decrease the computational costs of calculating energies and forces from these pairwise potentials. Meaning that, atoms interacting at shorter distances than the cutoff distance are considered in vdW calculations and a switching function is used on the U_{vdW} for the interactions that are occurring at distances longer than the cutoff distance to guarantee the continuity of the potential. A switching function is a polynomial function of the distance by which the potential energy function is multiplied. It has value of 1 at *r=0* and a value of 0 at $r=r_c$, the cutoff distance.

The Coulomb energy term, $U_{coulomb}$, models the electrostatic interactions between two atoms *i* and *j* with partial charges of q_i and q_j and distance r_{ij} as:

$$U_{coloumb} = \sum_{pairs} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \qquad \text{Eq. 2.7}$$

where ε_0 is the electrical permittivity of vacuum. Since these interactions are long-ranged, the electrostatic calculation is computationally the most costly step in MD simulations. To reduce its computational cost, two methods have been utilized. In one method, a cutoff distance is applied during the simulations like what was described for short-range interactions. But, entry or exit of charges in the excluded area in polar systems can cause significant errors. Then, to solve this problem the cutoff should be elongated enough that consequently will increase the simulation cost again with a complexity of $O(N)^2$, where N is the number of particles in the system. As an alternative, Particle Mesh Ewald (PME) method is introduced that uses Ewald summation [59]. In PME the calculation is divided into a short-range and a long-range terms, reducing the complexity to $O(NlogN)$. The short-range term the computations are performed in real space, while for the long-range term the calculation is performed in the reciprocal space using Fast Fourier Transforms. Nevertheless, to use PME the periodic system should be neutralized using counter ions.

2.1.2 FoldX

In Chapter 3, we use a simplified force field scheme to select candidate mutations among a list of possible mutations to increase thermostability. The empirical FoldX force field was used for this task [60]. FoldX is a fast and quantitative algorithm that uses AA description of protein structures to estimate important interactions for the stability of proteins. It uses the following energy function to calculate unfolding free energy of the protein of interest [60]:

$$\Delta G = W_{vdw} \times \Delta G_{vdw} + W_{solvH} \times \Delta G_{solvH} + W_{solvP} \times \Delta G_{solvP} + \Delta G_{wb} + \Delta G_{hbond} + \Delta G_{el} + W_{mc} \times T \times \Delta S_{mc} + W_{sc} \times T \times \Delta S_{sc} \qquad \text{Eq. 2.8}$$

where ΔG_{vdw} is the sum of the van der Waals contributions, ΔG_{solvH} and ΔG_{solvP} are solvation free energy differences for apolar and polar groups respectively, ΔG_{hbond} is difference in the free energy between the formation of an intra-molecular hydrogen bond, ΔG_{wb} is the water bridges' extra stabilizing free energy, ΔG_{el} is the charged groups' electrostatic contribution, ΔS_{mc} is the entropy cost needed for fixing the backbone, ΔS_{sc} is the entropic cost needed for fixing a side chain, and W_{vdW}, W_{solvH}, W_{solvP}, W_{mc}, and W_{sc} correspond to the weighting factors applied to the raw energy terms [60]. Using this strategy, FoldX can predict changes in protein stability for single and multiple mutations quickly. However, it should be considered that its accuracy is still limited [61].

2.1.3 Molecular Dynamics Simulation

For a finite number of particles, the aim of MD simulation is to reproduce the temporal evolution of the system, so called trajectory, by solving the Newton's equation of motion within a classical mechanics representation of the potential energy associated with the molecular interactions of the system. Using an integrator to solve the Newton's equation of motion, the evolution of the coordinates and velocities of all particles in the system can be calculated in a time interval of δt. In other words, coordinates and velocities for each particle in the system gets updated from time t to $t+\delta t$ [62].

The classical equation of motion or Newton's second law is at the basis of MD simulation. Newton's equation of motion states that force ***F*** applied to a group of atoms at time *t* and position ***X***(t) is equal to the negative gradient of the potential *U* experienced by the group of atoms:

$$\boldsymbol{F}\big(\boldsymbol{X}(t)\big) = -\nabla U(\boldsymbol{X}(t)) \qquad \text{Eq. 2.9}$$

and whenever atoms change their position their experienced force will change. To solve this set of equations an iterative procedure is used. In MD simulation the integration is broken down into multiple steps separated by time interval of δt. At each time step the forces on each atom is calculated based on its interactions with other atoms in the system. The new position and velocity of atoms at time $t + \delta t$ is calculated from the acceleration, position and velocity of each atom at time *t* and the forces are assumed to be constant during time steps. The Verlet integration algorithm is commonly used in MD simulation, which uses the position ***X*** of atoms at times t and $t - \delta t$ and acceleration ***A*** at time *t* to calculate the position at time $t + \delta t$. This algorithm does not directly deal with velocities but the velocities can be computed by derivative of the calculated positions.

$$\boldsymbol{X}(t + \delta t) = 2\boldsymbol{X}(t) - \boldsymbol{X}(t - \delta t) + \delta t^2 A(t) \qquad \text{Eq. 2.10}$$

$$\boldsymbol{V}(t) = \frac{[\boldsymbol{X}(t-\delta t) - \boldsymbol{X}(t+\delta t)]}{2\delta t} \qquad \text{Eq. 2.11}$$

Defining the value of the timestep δt is an important aspect in MD simulations. Based on the Nyquist-Shannon sampling theorem, the frequency of sampling should be at least two times faster than the fastest vibration frequency event of the system to guarantee the stability of the system. In normal biological systems that are usually studied by MD simulations, the C-H covalent bond has the highest vibration frequency with period of approximately 10 fs. Therefore, in general a $\delta t = 1$ fs is used in MD simulations. However, it is always convenient to increase the timestep as much as possible to increase the sampling accessible via MD simulations by constraining the bond length with the fastest vibrations, algorithms like SHAKE [62] make possible to increase the δt to up to 2 fs.

The Shake Algorithm

Increasing the δt is practically obtained by eliminating the high frequency vibrations (usually for covalently bond H atoms to heavy atoms) using the SHAKE algorithm [62] to constraint the distance of the relative chemical bond. The constraint distance is defined as follows:

$$\sigma_a(r) = r_{ij}^2 - d_{ij}^2 = 0 \qquad \text{Eq. 2.12}$$

where r_{ij}^2 and d_{ij}^2 are instantaneous vector and reference distance between two atoms i and j. Using the SHAKE algorithm one can increase the timestep up to 2 fs for MD simulation of biological systems.

Molecular Dynamics Simulations at Constant Temperature

To explore the physicochemical properties of system at finite temperature, which is the common condition experienced by biological systems, MD simulations are run in the NVT ensemble using extended Hamiltonian system methods to modify Newton's equation of motion by the addition of certain additional variables. One common methodology used to control temperature T makes use of the following formulation, called Nose´-Hoover equation [63]:

$$H_{N-H} = \sum_i \frac{\boldsymbol{p}_i^2}{2m_i} + U(\boldsymbol{r}) + \frac{p_\gamma^2}{2Q} + Lk_B T\gamma \qquad \text{Eq. 2.13}$$

From which one can derive the following equation of motions:

$$\dot{\boldsymbol{r}}_i = \frac{\boldsymbol{p}_i}{m_i} \qquad \text{Eq. 2.14}$$

$$\dot{\boldsymbol{p}}_i = \boldsymbol{F}_i - \frac{p_\gamma}{Q}\boldsymbol{p}_i \qquad \text{Eq. 2.15}$$

$$\dot{\gamma} = \frac{p_\gamma}{Q} \qquad \text{Eq. 2.16}$$

$$\dot{p_\gamma} = \sum_i \frac{\boldsymbol{p}_i^2}{m_i} - Lk_B T \qquad \text{Eq. 2.17}$$

where $\{\boldsymbol{r}_i\}, \{\boldsymbol{p}_i\}$ are the coordinates and momenta of the N particles in the system, with masses m_i and L is a parameter to be determined. The two non-physical variables γ and

p_γ regulate the fluctuations in the total kinetic energy of the physical variables. The parameter Q controls the strength of the coupling to the thermostat.

Molecular Dynamics Simulations at Constant Temperature and Pressure

Biological systems are however usually exposed to constant temperature as well as constant pressure conditions, which are represented in statistical mechanics by the so-called NPT ensemble [64]. Therefore, in addition to applying constraints to keep the temperature constant, it is convenient to apply constraints that keep the pressure constant during the MD simulations. In the framework of the extended Hamiltonian approach, a MD simulation at NPT can be defined by the following extended Hamiltonian:

$$H_{NPT} = \sum_i \frac{\boldsymbol{p}_i^2}{2m_i} + U(\boldsymbol{r}) + \frac{p_\gamma^2}{2Q} + Lk_BT\gamma + \frac{p_\epsilon^2}{W} + P_{ext}V \qquad \text{Eq. 2.18}$$

The equations of motion corresponding to this Hamiltonian are:

$$\dot{\boldsymbol{r}}_i = \frac{\boldsymbol{p}_i}{m_i} + \frac{p_\epsilon}{W}\boldsymbol{r}_i \qquad \text{Eq. 2.19}$$

$$\dot{\boldsymbol{p}}_i = \boldsymbol{F}_i - \frac{p_\gamma}{Q}\boldsymbol{p}_i - \frac{p_\epsilon}{W}\boldsymbol{p}_i \qquad \text{Eq. 2.20}$$

$$\dot{V} = \frac{dVp_\epsilon}{W} \qquad \text{Eq. 2.21}$$

$$\dot{p_\epsilon} = dV(P_{int} - P_{ext}) - \frac{p_\gamma}{Q}p_\epsilon \qquad \text{Eq. 2.22}$$

$$\dot{\gamma} = \frac{p_\gamma}{Q} \qquad \text{Eq. 2.23}$$

$$\dot{p_\gamma} = \sum_i \frac{\boldsymbol{p}_i^2}{m_i} + \frac{p_\epsilon^2}{W} - Lk_BT \qquad \text{Eq. 2.24}$$

where V is the volume of the system. The momentum p_ϵ is correlated to the $\frac{d\epsilon}{dt}$ that $\epsilon = \frac{1}{d}\ln\left(\frac{V}{V_0}\right)$ and V_0 is the reference volume. The p_ϵ variable acts as a barostat driven by the fluctuations of the internal pressure P_{int} around to the external pressure applied in an isotropic manner on the walls of the simulation box. The internal pressure is therefore given by:

$$P_{int} = \frac{1}{dV}\left[\sum_i \frac{\boldsymbol{p}_i^2}{m_i} + \sum_i \boldsymbol{r}_i \, . \, \boldsymbol{F}_i - (dV)\frac{\partial U}{\partial V}\right] \qquad \text{Eq. 2.25}$$

Where Q and W control the strength of the coupling to the thermostat and barostat, respectively.

Periodic Boundary Conditions

Periodic boundary conditions (PBC) are applied in MD simulations to decrease size effects [65]. In simulations with PBC the atoms of the system are enclosed by a box with specific dimensions and shape, which is then replicated in all directions. The computation of forces for one particle in the central box is performed with atoms localized in the same box or with atoms located in the others boxes with a distance smaller than an assigned cutoff, R_c. When an atom moves out of the central box, it enters at the same time from the opposite side conserving the same velocity. To compute non-bonded interactions, the minimum image convention is applied, that states that the cutoff should not be larger than the half of the shortest box vector. In addition, by setting the cutoff smaller than the half of the shortest box vector guarantees that no particle can see its own image.

$$R_c \leq min(x,y,z)/2 \qquad \text{Eq. 2.26}$$

Calculating Averages from Molecular Dynamics Simulations

To describe a classical system, we can use a classical Hamiltonian *H* that is a function of coordinates **r** and momenta **p**. If the potential energy function is velocity-independent (U_{MM}=U(**r**)) then *H* is equal to the total energy:

$$H = H(\boldsymbol{r},\boldsymbol{p}) \equiv K(\boldsymbol{p}) + U(\boldsymbol{r}) = \sum_i \frac{\boldsymbol{p}_i}{2m_i} + U(\boldsymbol{r}) \qquad \text{Eq. 2.27}$$

where $K(\boldsymbol{p})$ is the kinetic energy, $U(\boldsymbol{r})$ is the potential energy, $\boldsymbol{p}_i$ is the momentum of paricle *i,* and m_i is the mass of particle *i*. Therefore, by a set of values $\{\boldsymbol{r},\boldsymbol{p}\}$ a microscopic state of the system can be characterized corresponding to a point in phase space defined by both a set of coordinates $\boldsymbol{r}$ and momenta $\boldsymbol{p}$.

Having a distribution function $\rho(\boldsymbol{r},\boldsymbol{p})$ it is possible to calculate phase space averages for any interesting dynamic variable $A(\boldsymbol{r},\boldsymbol{p})$**;** these averages are called thermodynamic or ensemble averages and can be expressed as:

$$\langle A(\boldsymbol{r},\boldsymbol{p})\rangle \;=\; \int_V d\boldsymbol{r} \int_{-\infty}^{\infty} d\boldsymbol{p}\rho(\boldsymbol{r},\boldsymbol{p})A(\boldsymbol{r},\boldsymbol{p}) \qquad \text{Eq. 2.28}$$

where angle brackets <> indicate an ensemble average of the property A. Therefore, to calculate ensemble averages we need to know the distribution probability for every state $\{\boldsymbol{r},\boldsymbol{p}\}$. Generally, calculation of these integrals is very difficult because for a system all possible states should be accounted for. As an alternative, MD simulations can be used to calculate the same thermodynamic observables, using instead time averages over the generated trajectory. Following the motion of a single point in the phase space during time evolution in MD trajectories, we can take temporal average over the points that were visited during the trajectory as:

$$\langle A(\boldsymbol{r},\boldsymbol{p})\rangle_{\boldsymbol{\tau}} = \frac{1}{\tau}\int_0^\tau A(\boldsymbol{r}(t),\boldsymbol{p}(t))dt \qquad \text{Eq. 2.29}$$

where τ is the duration of the MD simulation. These temporal averages are strictly equivalent to ensemble averages only in the case the ergodic hypothesis, one of the fundamental axioms of statistical mechanics, is satisfied**.** The ergodic hypothesis states that over a long (at least longer with respect to the timescales characteristic for the observables under investigation) period of time, the temporal averages of a measured quantity and the relative ensemble averages are equal:

$$\langle A(\boldsymbol{r},\boldsymbol{p})\rangle_{\boldsymbol{\tau}} = \langle A(\boldsymbol{r},\boldsymbol{p})\rangle \qquad \text{Eq. 2.30}$$

This hypothesis practically enables the calculation of important thermodynamic quantities using trajectories provided by MD simulations.

Analysis Tools for Molecular Dynamics Simulations

The Root Mean Square Deviation (RMSD) is used to quantify the structural deviation of the structure during the MD simulation from a reference structure. The reference structure is generally the starting point of the simulation. The RMSD is calculated by:

$$RMSD(t_k,t_0) = \sqrt{\frac{1}{M}\sum_{i=1}^{N} m_i\left(r_i(t_k) - r_k(t_0)\right)^2} \qquad \text{Eq. 2.31}$$

where *M* is the total sum of the masses of the atoms:

$$M = \sum_{i=1}^{N} m_i \qquad \text{Eq. 2.32}$$

$r_i(t_k)$ represents the position of the atom i at time t_k, and $r_k(t_0)$ represents the positions of the atom at reference structure at time t_0. This analysis is usually used to determine if the system has reached the equilibrium during a MD simulation. At equilibrium the RMSD of the molecular system usually reaches a plateau value. This analysis was used in Chapter 3.

In addition to RMSD, Root Mean Square Fluctuations (RMSF) is commonly used for MD simulation trajectory analysis. RMSF is used to quantify the atomic fluctuation of atoms (usually the C_α of the residues in proteins). RMSF can be calculated as:

$$RMSF_i = \sqrt{\frac{1}{T}\sum_{t_j=1}^{T}\left(r_i(t_j) - \langle r_i \rangle\right)^2} \qquad \text{Eq. 2.33}$$

where T is the total time of the MD simulation, and $\langle r_i \rangle$ is the average position of C_α carbon i. The RMSF can be directly related to the X-ray crystallography B-factor:

$$B(r_i) = \frac{3}{8}\pi^2 RMSF(r_i)^2 \qquad \text{Eq. 2.34}$$

I used RMSF analyses in Chapter 3 to find flexible residues on the protein from trajectories generated by running MD simulations at different temperatures. This analysis helped to find which residues were more sensitive to temperature.

2.1.4 Elastic Network Models

In Chapter 3 I use MD simulation to find flexible residues on a protein structure by study of their fluctuations using RMSF. However, MD simulations are usually costly and it is beneficial if we can use other simpler and faster models to study dynamic fluctuations of residues. Elastic network models (ENM) are simplified and fast techniques that can help with this task. ENM is a topology-based coarse grained model at single-residue resolution (i.e., only C_α positions are most of the times taken into account), independent on the amino-acid sequence [66]. In such a model, details of residue interactions including salt bridges, hydrogen bonds and water solvation are not taken into account [66-68]. In Chapter 4 I used a variation of ENM called Gaussian network model (GNM) to develop an alternative strategy to MD simulations to find residues located on structurally weak points of a protein structure [67, 68]. GNM has been mainly used to study atomic fluctuations of amino acids in protein structures.

The main assumption behind GNM is that inter-residue distances (fluctuations) around the equilibrium coordinates are following a Gaussian distribution [67-69]. In this model, C_α atoms are used as representative of residues in a protein and their coordinates in PDB files are considered as the equilibrium coordinates [68]. C_α atoms of i^{th} and j^{th} residues are connected by a classical spring with force constant of γ_{ij} if their distance is less than a cut-off radius of r_c.

For a protein, the internal Hamiltonian of the system can therefore be written as [66, 70]:

$$H = \frac{1}{2}\gamma[\Delta R^T(\Gamma \otimes E)\Delta R] \qquad \text{Eq. 2.35}$$

where ΔR represents residue fluctuations represented by the C_α atoms; the ΔR^T is the transpose of ΔR; E is the third-order identity matrix; $\otimes$ is the direct product; and Γ is the Kirchhoff matrix that is defined as follows:

$$\Gamma_{ij} = \begin{cases} -c & if|i-j| = 1 \\ -\gamma & if|i-j| > 1 \; and \; R_{ij} < r_c \\ 0 & if|i-j| > 1 \; and \; R_{ij} > r_c \\ -\sum_{i,j\neq i}\gamma_{ij} & if \; i = j \end{cases} \qquad \text{Eq. 2.36}$$

where R_{ij} is the distance between the i^{th} and j^{th} C_α atoms and r_c is the cutoff distance.

The N×N Kirchhoff matrix (Γ) can describe residue fluctuations and their cross correlations. It is noteworthy that the absolute value of γ does not affect the eigenvectors but uniformly scales eigenvalues [68]. The inverse of the Kirchhoff matrix can be decomposed as:

$$\Gamma^{-1} = U\Lambda^{-1}U^T \qquad \text{Eq. 2.37}$$

where columns of the orthogonal matrix U, u_i ($1 < i \leq N$), are the eigenvectors of Γ, and elements of the diagonal matrix Λ are eigenvalues of Γ, λ_i.

The cross-correlation fluctuations between the i^{th} and j^{th} residues can be calculated by:

$$< \Delta R_i . \Delta R_j > = \frac{3k_BT}{\gamma}[\Gamma^{-1}]_{ij} \qquad \text{Eq. 2.38}$$

where k_B is the Boltzmann constant and T is the absolute temperature. By assuming $i = j$, ($[\Gamma^{-1}]_{ii}$) using the equation 4 one can calculate the mean square fluctuation (MSF) of the i^{th} residue. Then, the Debye-Waller or B-factor of a residue can be calculated by:

$$B_i = \frac{8\pi^2}{3} < \Delta R_i . \Delta R_i > \qquad \text{Eq. 2.39}$$

To calculate the MSF in the distance vector R_{ij} between the residues i and j one can use:

$$\langle (\Delta R_{ij})^2 \rangle = \langle (R_{ij} - R_{ij}^0)^2 \rangle = \langle (\Delta R_i - \Delta R_j)^2 \rangle = \langle \Delta R_i . \Delta R_i \rangle + \langle \Delta R_j . \Delta R_j \rangle - 2\langle \Delta R_i . \Delta R_j \rangle = \frac{3k_BT}{\gamma}([\Gamma^{-1}]_{ii} + [\Gamma^{-1}]_{jj} - 2[\Gamma^{-1}]_{ij}) \qquad \text{Eq. 2.40}$$

where R^0_{ij} and R_{ij} denote the equilibrium and instantaneous separation vectors between residues i and j, respectively.

2.2 Sequence-Based Bioinformatics Methods

2.2.1 Protein Sequence Alignment

Comparing protein sequences is still the simplest, but very powerful, approach to protein analyses [71, 72]. Different schemes of protein sequence alignment have been developed to make this comparison efficient in revealing functional protein properties. Specifically, pairwise sequence similarity analysis (i.e., pairwise alignment) is used to find sequences similar to a target sequence within sequence databases for applications as diverse as function discovery or homology modeling [72]. For pairwise sequence alignment, two methods are usually used, called local and global alignment. Local alignment algorithms identify the local similarities between sequences while global alignment algorithms align the full length of the two sequences [72]. BLAST (Basic Local Alignment Search Tool) is the most commonly used algorithm for pairwise local sequence alignment [71]. BLAST first finds the high similarity sequence pieces in the two given protein sequences and builds the final alignment based on the initial similarity. BLAST is mainly used to find homologous sequences to a query sequence within protein sequence databases, like the NCBI (National Center for Biotechnology Information) database [73].

To compare multiple sequences simultaneously, different multiple sequence alignment algorithms have been introduced. Among them, the most widely used method is called "progressive sequence alignment" [74]. This method starts with the most similar sequences and adds progressively more distant (in the evolutionary sense) sequences to the alignment. Among different implementations of this method, ClustalW [75] is one of the most popular [72]. ClustalW works in tree steps. First, a pair wise alignment is done over all sequence pairs assigning a score for each pair of sequences as distance matrix. Then, this matrix of distances is used to create a dendrogram (i.e., guide phylogenetic tree) among all the sequences. Finally, the dendrogram is used as a basis for constructing the real multiple sequence alignment where the most closely related pairs of sequences are aligned. The quality of the alignment is determined by assigning a positive score to identical aligned residues, a lower or negative score to mismatches using substitution matrix and gap penalty if a gap is introduced [75]. BLOSUM (BLOcks SUbstitution Matrix) matrix is a substitution matrix that can be used to score the sequence alignment. To build it, the BLOCKS database for very conserved regions of protein families (i.e., without any gap in the alignment) was searched and then the relative frequencies of amino acids and their substitution probabilities were counted. Using this approach, a log-odds score for each of the 210 possible substitution pairs of the 20 standard amino acids was calculated [72].

In this thesis, BLAST algorithm has been used to find homologous sequences to a target sequence within NCBI database. CLUSTALW2.0 [75] is also used to perform comparative studies on multiple sequences from the same family.

2.2.2 Databases and Servers

A large amount of protein sequence data (the number of sequence records in GenBank now is more than 181×10^6, see: http://www.ncbi.nlm.nih.gov/genbank/statistics) are available on online databases making possible to extract useful information about specific protein properties solely based on sequences. In addition to these databases, there are several webservers that provide specific online tools for a broad array of analyses. In this thesis I mainly used NCBI [73], Pfam [76], and ProTherm [77] online databases to extract row data needed for my analyses. These databases provide the row data and do not perform the predictions that one may need.

- The *NCBI database* [73] provides sequences and annotated information for different species. At the moment more than 52000 organisms and 44M protein sequences are available in this database I used this resource to find homologuous sequences using the BLAST algorithm as input for our PTE framework (see Chapter 5). All of the BLAST and sequence fetching operations were handled using Biopython modules [78].

- *Pfam* [76] is a database for the identification of protein families. It contains more than 14800 manually curated entries and provides an accurate MSA for each protein family. In chapters 5 and 6, instead of performing MSA for each protein, I first found the corresponding family in the Pfam database for the target sequence and used the provided MSA in the database. This procedure did not only increase the speed of analyses, but also increased the accuracy by providing more accurate MSAs. Pfam is divided into two databases: Pfam-A and Pfam-B. The main difference between them is that Pfam-A entries are of higher quality, being carefully curated manually, while Pfam-B is generated automatically resulting in a general lower quality of the classification of the MSAs. In this thesis I only used the high quality Pfam-A database.

- *ProTherm* (Thermodynamic Database for Proteins and Mutants) database [77] collects thermodynamics data such as Gibbs free energy change, enthalpy change, heat capacity change, transition temperature. The total number of entries in this database corresponds to 25820 mutations for 740 proteins. This database has been the main source of experimental data for developing or training several algorithms that can predict stability changes upon mutations [79, 80]. In chapters 4, 5, and 6 of this thesis, I used this database to feed my benchmarks. Using this database I collected a number of mutations considering their corresponding difference in the midpoint temperature of the thermal unfolding (ΔT_m) compared to the wild type.

As I was interested in study of thermostability changes upon mutations, I often used ΔT_m of mutations and considered increment and decrease as a metric for thermo-stabilizing and destabilizing effects. The value of ΔT_m is usually measured using differential scanning calorimetry technique (DSC), which is a powerful technique to study the thermodynamics of globular protein unfolding. Using DSC one can measures the excess heat capacity of a protein solution as a function of temperature relative to a buffer. Analysis of the heat capacity curves vs. T provides the thermodynamic data. The maximum of the heat capacity curve occurs close to the T_m of the protein. The maximum is precisely at T_m if the ΔC_p of the transition is zero. Figure 2-1 shows a typical unfolding DSC curve.

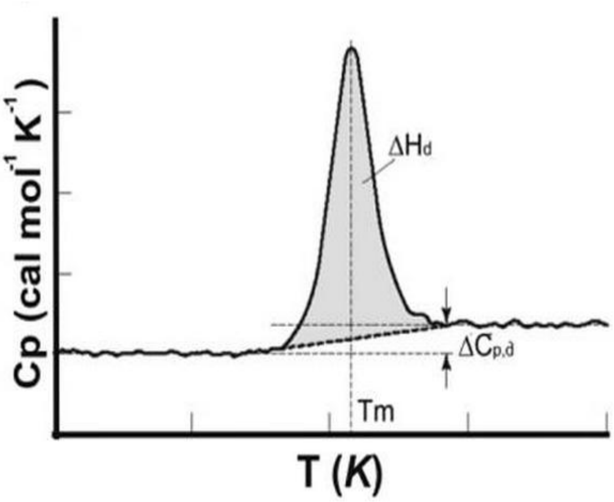


Figure 2-1. A typical DSC curve. Using this plot for wild-type and mutant we can calculate ΔT_m upon mutations. The figure is adopted from reference [81].

Then, I compared results of predictions with experimental data extracted from ProTherm database to assess the methods.

In addition to the databases, I used two online tools, namely I-Mutant2.0 [79] and AUTO-MUTE [80], which predict the effect of mutations on stability of proteins. I compared the protocol developed in Chapter 6 with these currently available algorithms.

I-Mutant2.0 [79] can predict the effect of single mutations on the stability of proteins using Support Vector Machine (SVM) method. This method can predict the stability change starting from the protein structure or from the protein sequence (in this thesis only structural based predictions were used). The accuracy of I-Mutant2.0 is reported to be ~80% for structure-based predictions and 77% for sequence-based predictions.

AUTO-MUTE [80] is another online predictor that was used here. This method uses supervised classification and regression algorithms to predicting changes in stability of a given proteins native structure upon single mutation. Its accuracy is reported to be as high as 80%.

2.2.3 Statistical Analysis of Predictions

The above presented computational prediction methods, which are discussed and used in this thesis especially in Chapters 4 to 6, are assessed by comparison with published experimental data, extracted from ProTherm database [77]. In this context, each prediction could be a *true positive* (TP), a *true negative* (TN), a *false positive* (FP), or a *false negative* (FN) result. In Chapters 4, 5, 6 the aim was to predict the effect of mutations on thermostability of proteins, which could have a positive effect if producing thermo-stabilization or a negative effect by producing thermo-destabilization. Thus, TP, TN, FN, and FP have in this context the following meaning:

TP are thermo-stabilizing mutations that are predicted correctly as thermo-stabilizing.

TN are thermo-destabilizing mutations that are predicted correctly as thermo-destabilizing.

FP are thermo-destabilizing mutations that are predicted incorrectly as thermo-stabilizing.

FN are thermo-stabilizing mutations that are predicted incorrectly as thermo-destabilizing.

Using TP, TN, FN, and FP outputs, the accuracy (ACC), true positive rate (TPR), true negative rate (TNR), the Mattews correlation coefficient (MCC), and the F1 score (F1) were calculated to estimate the robustness of the prediction methods, as well as the framework developed in this thesis [82].

The accuracy is calculated as:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \qquad \text{Eq. 2.41}$$

TPR (or sensitivity) and TNR (or specificity) show how accurately the method could predict positive and negative effects and are simply calculated as:

$$TRP = \frac{TP}{TP + FN} \qquad \text{Eq. 2.42}$$

$$TNP = \frac{TN}{TN + FP}$$ Eq. 2.43

The MCC was introduced to measure the quality of predictions and reports a correlation coefficient between the observed and predicted binary classification. The MCC returns a value from −1 to +1, where MCC = 1 indicates a perfect prediction (i.e., meaning that all prediction results are true and the number of FN and FP predictions are both 0), MCC = 0 indicates that predictions are not better than random, and MCC = −1 indicates a total disagreement between prediction and observation (i.e, meaning that both TP and TN are 0 and all positive effects are predicted to be negative and vice versa). The MCC is defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$ Eq. 2.44

The F1 score is another test for the evaluation of the accuracy of the predictions, Compared with normal accuracy reported in equation 1.42, which considers both TP and TN, F1 gives higher weight to TP results. F1 score can be calculated as:

$$F1 = \frac{2TP}{2TP + FP + FN}$$ Eq. 2.45

In Chapters 4, 5, and 6 all these analyses are used to assess the performance of the predictors.

These analyses usually are used by machine learning based studies where the performance of binary classifiers is investigated. In the analyses, it should be considered that if the goal of predictor is finding more TP results, like in the case of my work, F1 score is more efficient than the traditional definition of accuracy. Because, the traditional accuracy definition takes into account both TP and TN results, while F1 does not consider TN results and instead puts higher weight on TP results.

Chapter 3 Understanding and Engineering Thermostability in the DNA Ligase from *Thermococcus sp. 1519*

This chapter is adapted from the paper published as:

Hassan Pezeshgi Modarres, Boris D. Dorokhov, Vladimir O. Popov, Nikolai V. Ravin, Konstantin G. Skryabin, Matteo Dal Peraro, *Understanding and Engineering Thermostability in the DNA Ligase from Thermococcus sp. 1519*. Biochemistry, DOI: 10.1021/bi501227b.

In this work I conducted all the *in silico* analyses that led to the characterization of the thermostability origin of the enzyme and to the prediction of 4 thermo-stabilizing mutations using molecular simulation and bioinformatics. The experimental part aimed at testing the in silico predictions was instead entirely done by our collaborators at the Russian Academy of Science in Moscow (i.e., Dr. Nicolai Ravin and his colleagues).

3.1 Introduction

Microorganisms are usually divided into four groups based on the temperature conditions they have evolved to live at: psychrophilics, living in cold environments; mesophilics, living in conditions around human body temperature; thermophilics, living at temperatures between 40 °C and 80 °C; and hyperthermophilic that are able to survive at temperatures higher than 80 °C [5, 6]. Since 1967, when Thomas D. Brock published the first report about the discovery of microorganisms with optimal growth temperature (OGT) higher than 75 °C, enzymes extracted from these microorganisms that can withstand high temperatures have been increasingly studied to understand the origin of their improved thermostability, and used thereafter in biotechnological applications [32].

Due to decreased viscosity and increased diffusion coefficient and solubility of substrates at high temperatures enzymes isolated from hyper-thermostable organisms have in fact found extensive applications in various industrial domains, like genetic engineering, food processing (e.g., baking, brewing, dairy, starch hydrolysis, etc.), petroleum bioremediation, chemical processes, and paper and pulp industries [5, 32, 83, 84]. In addition, using enzymes that can survive at elevated temperatures usually reduces the risk of presence of common mesophiles that may contaminate the reaction environment [32, 83, 84]. Thus, it is not surprising that numerous studies aimed at understanding the molecular origin of thermostability of enzymes from mesophiles to hyper-thermophiles have attracted the attention of the scientific community [6, 33, 34, 85-95].

It is now clear that nature does use a vast array of complementary strategies to enhance thermostability, as increasing hydrogen bonding, producing better protein packing, promoting burial of hydrophobic surface area, modifying surface charges, increasing salt-bridges and disulfide bonds, more favorably developing helical dipolar interactions, improving secondary structure propensity, increasing rigidity and entropic stabilization [5, 6, 32-34]. For each enzyme family, decoding the strategies that hyper-thermophiles use to resist at high temperatures based on their sequence and structure can be important to reverse-engineer these principles and in turn enhance thermostability modifying an enzyme with the least number of mutations [41-44, 46].

Computational approaches can help making such procedure faster and more efficient in order to not only produce higher thermostability, but also improve other properties like increased stability in salts solutions [96]. For instance, rational design strategies have been used to capture thermo-sensitive regions, proposing appropriate substitution to enhance thermostability [18-26]. In this context, molecular dynamics (MD) simulation is also a valuable means to identify thermo-sensitive regions in proteins [18, 21, 22, 24, 26], by probing structure at elevated temperatures [18-26]. These insights, along with data extracted by bioinformatics analyses [97], can be then used to engineer protein structures by optimization of unstable residues [18-21, 24, 26].

Following these premises, in this work we studied the origin of thermostability in a thermophilic DNA ligase, which is also attractive for applications in biotechnology and pharmaceutics [54, 98], and based on this analysis, we proposed mutations able to enhance its thermostability. In particular, we focused on the DNA ligase (hereafter called *LigTh1519*) from the thermophilic archaeon *Thermococcus sp.* 1519 isolated from a hydrothermal vent and growing optimally at 85°C [99]. *LigTh1519* has been recently characterized biochemically and solved by X-ray crystallography [99, 100] (Figure 3-1). *LigTh1519*, as most of ATP-dependent DNA ligases, is a multi-domain enzyme composed of three subdomains: the DNA binding domain (DBD), the nucleotide binding domain (NBD), and the OB-fold domain (OBD) (Figure 3-1). DNA ligases catalyze the phosphodiester bond formation between adjacent 3' hydroxide and 5' phosphate groups in a nicked single strand DNA [54, 101-104], which may occur as a result of natural replication, in so-called Okazaki fragments, or by DNA damaging agents [54, 101-104]. Moreover, while *LigTh1519* has an optimum activity above 70 °C, the half-life of this enzyme at 94 °C is only few minutes [99].

In order to understand the molecular origin of these properties we first studied *LigTh1519* using MD simulations at different temperatures starting from its atomistic structure [100]. Results of this initial analysis revealed the most relevant thermo-sensitive regions, which we targeted using a set of rational design strategies in order to predict mutations able to increase *LigTh1519* thermostability. These predictions were eventually produced and tested *in vitro* showing a remarkable and additive increase of the half-life time at 94 °C for the engineered enzyme (namely, up to about 41 min when up 4 mutations were simultaneously engineered), while 50% inactivation of parental *LigTh1519* was observed already upon 8 minutes incubation at this temperature.

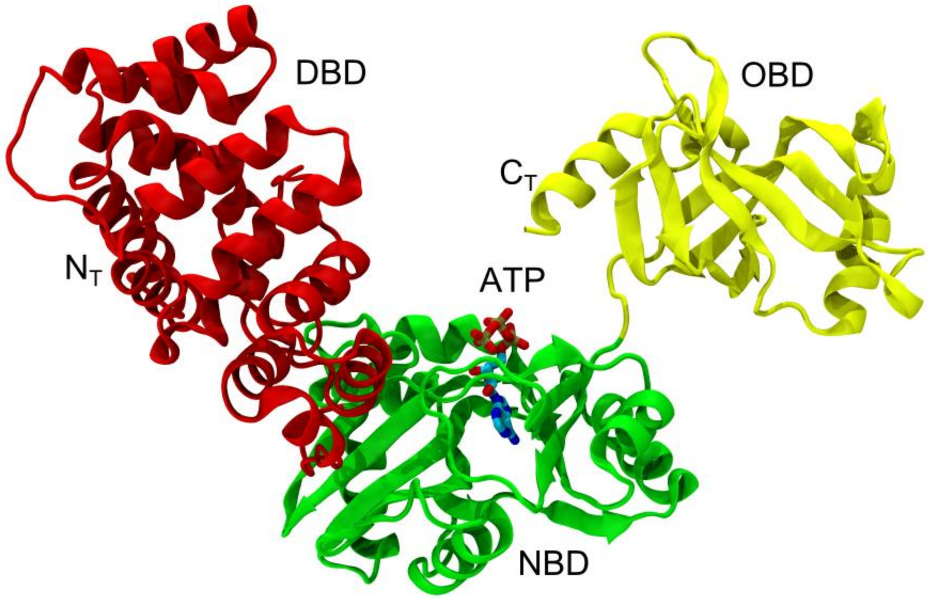


Figure 3-1. LigTh1519 structure. LigTh1519 (PDB ID: 3RR5) [100] is composed of three sub-domains: the DNA Binding Domain (DBD) (in red), the Nucleotide Binding Domain (NBD) (in green), and the OB-fold domain (OBD) (in yellow). The ATP binding site is located on the NBD. ATP is located at the binding site by superimposing the LigTh1519 NBD to that of the 2HIX structure, which shows a co-crystallized ATP molecule.

3.2 Materials and Methods

3.2.1 Computational Procedure

The structure of *LigTh1519* DNA ligase (PDB ID: 3RR5 [100]) at 3 Å resolution was used for all MD simulations and analysis [100]. Selected mutated structures (i.e. A287K, G304D, S364I and A387K) were build based on this X-ray structure using the Modeller package[105]. The X-ray structures were protonated at pH 7 conditions and inserted in a box of water molecules of 11.6×8.6×10.8 nm^3. $MgCl_2$ salt was added to neutralize and adjust the ion concentration to 10 mM. The final system had ~90,000 atoms and ~27,000 water molecules. The systems were first energy minimized with constrained C_α atoms and then without any constraint for 1000 steps. Then, to equilibrate the systems, the temperature was increased gradually up to 300 K in the NVT ensemble and was kept at 300 K for 100 ps using a 1 fs time step. Finally, NPT MD simulations were run at 300 K for 500 ps with a 2 fs time step to complete the equilibration procedure.

The equilibrated structures were thus used as a starting point for production (Figures S3-1 and S3-2).

All MD simulations were performed using NAMD simulation package [106] with Amber99SB force field [107] and the TIP3P water model [108]. Constant temperature was imposed by using Langevin dynamics with a damping coefficient of 5.0 ps. Constant pressure of 1 atm was maintained with Langevin piston dynamics using a 200 fs decay period and 50 fs time constant [109, 110]. All production MD simulations were run at 1 bar with a time step of 2 fs, using the SHAKE algorithm [111] on all bonds, and PME [112] for treating electrostatic interactions. Simulations of the wild-type were run at seven different temperatures, namely 280, 300, 320, 360, 380, 400, and 500 K for more than 10 ns in NPT ensemble.

Salt bridge analysis, visualizations, and rendering of figures were performed using VMD [113]. For root mean square fluctuation (RMSF) analysis ProDy package [114] was used. For bioinformatics analysis BLAST [97] was used to find homolog sequences and ClustalW 2.0 [75] was used to perform multiple sequence alignments. Analysis of multiple sequence alignments were performed using Python [115] and Biopython [78] packages. To evaluate the effect of mutations on the thermodynamic properties of the enzyme we used FoldX, which is a fast and quantitative algorithm that uses an all-atom description of protein structures to estimate important interactions for protein stability [60].

3.2.2 Plasmids, Bacterial Strains and Culture Conditions

For expression of recombinant *LigTh1519* ligase in *Escherichia coli*, the corresponding gene was cloned in pQE30 (Qiagen) yielding expression vector pQE30-*LigTh1519* [99]. This plasmid allowed expression of recombinant enzyme with N-terminal 6-histidine tag. Site-directed mutagenesis of pQE30-*LigTh1519* was performed to introduce four mutations (singularly and altogether) in *LigTh1519*: A287K (GCC→AAA), G304D (GGC→GAT), S364I (AGC→ATT) and A387K (GCG→AAA). Amino acid numbers are given according to the sequence of the native protein (GenBank ACN59570). The corresponding vectors, pQE30-*LigTh1519*mut, pQE30-*LigTh1519-A287K,* pQE30-

LigTh1519-G304D, pQE30-*LigTh1519-S364I,* and pQE30-*LigTh1519-A387K* were used to express mutant enzymes, *LigTh1519*mut, in *E. coli* cells.

Rosetta-gami (DE3) (Novagen) strain was applied for expression of recombinant ligase and its mutant derivatives. The *E. coli* strains were cultured in Luria–Bertani (LB) medium at 37°C with shaking, and ampicillin was added to the medium at final concentration of 100 mg ml^{-1}.

3.2.3 Expression and Purification of Ligth1519 and Its Mutants

The cells were grown at 37°C in 10 ml of LB medium to the mid-exponential growth phase (OD_{600} of 0.5). Isopropyl β-D-1-thiogalactopyranoside was added at final concentration of 1 mM to induce gene expression, followed by an additional 12 h of incubation. The cells were then harvested by centrifugation and resuspended in 1 ml of 50 mM phosphate buffer (pH 8.0) containing 300 mM NaCl and 10 mM imidazole. The cells were disrupted by the treatment by lysosyme (1 mg/ml) for 30 min and sonication, followed by centrifugation (8,000 x g, 30 min, 4°C). The supernatant was then mixed with 200 µl of 50% slurry of Ni-NTA resin (Qiagen) and incubated with gentle agitation for 30 min at 4°C. The mixture was centrifigated for 10 sec at 1,000 x g (4°C) to pellet the resin. It was then washed twice with Wash buffer (50 mM phosphate buffer pH 8.0, 300 mM NaCl, 20mM imidazole). The bound proteins were eluted twice in 200 µl Elution buffer (50 mM phosphate buffer pH 8.0, 300 mM NaCl, 200mM imidazole). The eluates were combined and dialyzed against 50 mM Tris-HCl buffer (pH 7.2) containing 1 mM DTT and 1 mM EDTA at 4°C overnight in Slide-A-Lyser MINI Dialysis Units 3,500 MVCO (Thermo Scientific). The protein concentration was determined by Bradford method using BSA as a standard. Protein purity was then assessed by 10% SDS-PAGE.

3.2.4 DNA Ligation Assay

The substrate used in the ligation assays was a 80 bp DNA duplex, which contained an internal nick. It was produced with the annealing of 35-mer oligonucleotide LA-35bt (5’-CAG AGG ATT GTT GAC CGG CCC GTT TGT CAG CAA CG-3’) and 40-mer oligonucleotide LA-40FAM (5’-CGC ACC GTG ACG CCA

AGC TTG CAT TCC TAC AGG TCG ACT C-3') to a complementary 80-mer oligonucleotide LA-80 (5'-CGT TGC TGA CAA ACG GGC CGG TCA ACA ATC CTC TGG AGT CGA CCT GTA GGA ATG CAA GCT TGG CGT CAC GGT GCG CCA AC-3'). The 35-mer oligonucleotide was labelled with biotin at its 3'-end and phosphorylated at its 5'-end, and the 40-mer oligonucleotide was labelled with fluorescein 6-FAM at its 5'-end. To produce the ligase substrate, the mixture was slowly cooled to room temperature after heating at 95°C for 5 min. The molar ratio of the components in the annealing mixture was 1:1:1.

The assay for the nick-closing activity of DNA ligase was performed as previously described [99, 116] with minor modifications. The 30 pmol of fluorescein/biotin-labelled substrate was added to the assay mixture (final volume 30 µl) containing 40 mM Tris-HCl (pH 7.5), 10 mM $MgCl_2$, 100 µM ATP and DNA ligase. To determine the DNA ligase activity we made a series of dilutions of investigated enzyme samples and used in test assays. DNA ligase activity assay was conducted at 55°C for 1 h. The reactions were stopped by adding EDTA to 50 mM followed by 10 min incubation on ice.

Relative activity of DNA ligase was evaluated by measuring the fraction of duplex substrate in which the single strand break between LA-35bt and LA-40FAM was joined. The product of reaction, 75-bp oligonucleotide, contains both fluorescein and biotin and may be immobilised on streptavidin particles. For each assay sample 0,1 mg of streptavidin magnetic beads (Dynabeads M-270 Streptavidin, Invitrogen) were prepared by washing three times with 1x B&W buffer (5 mM Tris-HCl, pH 7.5, 0,5 mM EDTA, 1M NaCl) and were finally resuspended in 2x B&W buffer. 30 µl of ligase assay sample was mixed with an equal volume of resuspended beads and then incubated at room temperature for 20 min with gentle rotation to bind biotinilated DNA. The beads were separated from the unbound DNA using magnet and washed three times with 1x B&W buffer. Finally immobilised DNA was denatured with 0.1N NaOH, washed and released from the beads by incubation at 65°C for 5 min in 10 mM EDTA with 95% formamide. Fluorescent signal was acquired with TBS-380 Mini-Fluorometer (Turner BioSystems, emission: 515-575 nm, excitation: 465-485 nm) calibrated with 1 pmol of oligonucleotide LA-40FAM.

Thermostability was evaluated by this assay after incubating the enzyme (1 pmol) in 40 mM Tris-HCl buffer (pH 7.5) with 10 mM $MgCl_2$, for various lengths of time at 94°C. The selected temperature of 94°C is the most relevant from practical issues, since the promising area of application of LigTh1519 is ligase chain reaction (LCR). One step of LCR (like PCR) is in fact denaturation of DNA by heating at 94°C. Upon incubation the sample was cooled on ice. The dsDNA template and 100 µM ATP were added and the assay was conducted at 55°C for 1 h followed by stopping the reaction with 50 mM EDTA. Fluorescence signal detected from ligated oligonucleotides provides quantitative evaluation of the activity of ligase adjusted relative to the control samples without DNA ligase. The enzyme half-inactivation time calculation was based on logarithmic plot by determining the time required for 2-fold loss of activity in the linear phase.

3.3 Results

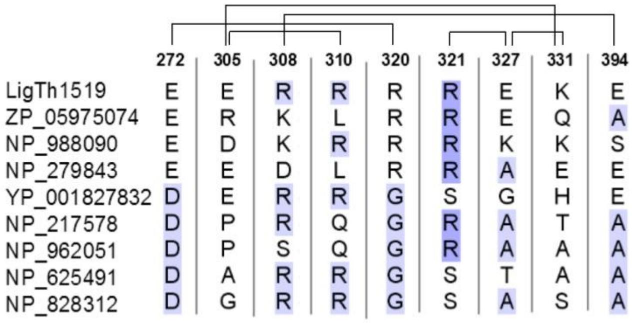
To understand the origin of thermostability of *LigTh1519* MD simulations at different temperatures were first used to identify domains with the highest thermosensitivity. Then, to find how mesophilic sequences have evolved to stabilize thermo-sensitive regions, the *LigTh1519* sequence was compared with those of mesophilic homologues. Next, comparison with hyperthermophilic species was key to propose a set of mutations that could increase thermostability of *LigTh1519,* without decreasing its activity when tested *in vitro*.

3.3.1 Origin of LigTh1519 Thermostability

The NBD contains the catalytic site and it is known to function even without the other two domains [54, 101-104] (i.e., DBD and OBD shown in red and yellow, respectively, in Figure 3-1). Therefore, we mainly focused on this domain to study and engineer *LigTh1519* thermostability. First, to find those regions on the protein that are most sensitive to temperature changes MD simulations at different temperatures were run. These simulations indicated that the NBD is overall fairly stable, showing very low root mean square deviation (RMSD) from the reference X-ray structure (Figure S3-1). RMSF analysis of MD trajectories was used to detect regions more affected by structural variations upon changes in temperature (Figure 3-2). From the RMSF analysis we found that the most thermo-sensitive region is located between residues L300-E330 (Figure 3-3). In addition, residues in the regions A230-Y250 and K380-L400 showed to be thermo-sensitive, even though to a lesser extend (Figures 3-2 and 3-3). Secondary structure analysis of these thermo-sensitive regions indicates that there are a number of residues that are located on turns and are thus naturally flexible. However, residues located on helices (i.e., Q313-F319, E327-I332, and K380-G396) are more relevant and can have an active role in defining the temperature-dependent properties of the enzyme.

With the aim of finding if the optimization of the salt-bridge network is implicated in conferring more elevated thermostability, we next compared salt-bridge forming residue pairs on *LigTh1519* with those of homologous mesophilic DNA ligases, mostly focusing on those pairs located on the detected temperature sensitive regions (Figures 3-2 and 3-3). In order to do that, a multiple sequence alignment (MSA) was performed with 8 mesophilic homologue sequences extracted from the NCBI database [73] (Table S3_1). The salt-bridge analysis and sequence comparison results are summarized in Table 3-1. In addition, Figure 3-3 shows the detected thermo-stabilizing salt bridge network for *LigTh1519* on its structure and their location on the detected thermo-sensitive region. Taken together, we found that an extended network of salt bridges (i.e. E280-R318, E272-R320, E305-K331, E305-R310, E327-R321, E327-K331, and E394-R308,Figure 3-3) is responsible to connect the thermo-sensitive region (residues L300-E330) to the rest of the NBD.

Table 3-1. Salt bridge network in LigTh1519 and substitutions in mesophiles. Extract from a multiple sequence alignments (MSA) of LigTh1519 with mesophilic species (see also Table S3_1) where the salt bridge network connectivity is indicated on the top. The relative recurrent substitutions in the network are reported for selected species.



	272	305	308	310	320	321	327	331	394
LigTh1519	E	E	R	R	R	R	E	K	E
ZP_05975074	E	R	K	L	R	R	E	Q	A
NP_988090	E	D	K	R	R	R	K	K	S
NP_279843	E	E	D	L	R	R	A	E	E
YP_001827832	D	E	R	R	G	S	G	H	E
NP_217578	D	P	R	Q	G	R	A	T	A
NP_962051	D	P	S	Q	G	R	A	A	A
NP_625491	D	A	R	R	G	S	T	A	A
NP_828312	D	G	R	R	G	S	A	S	A

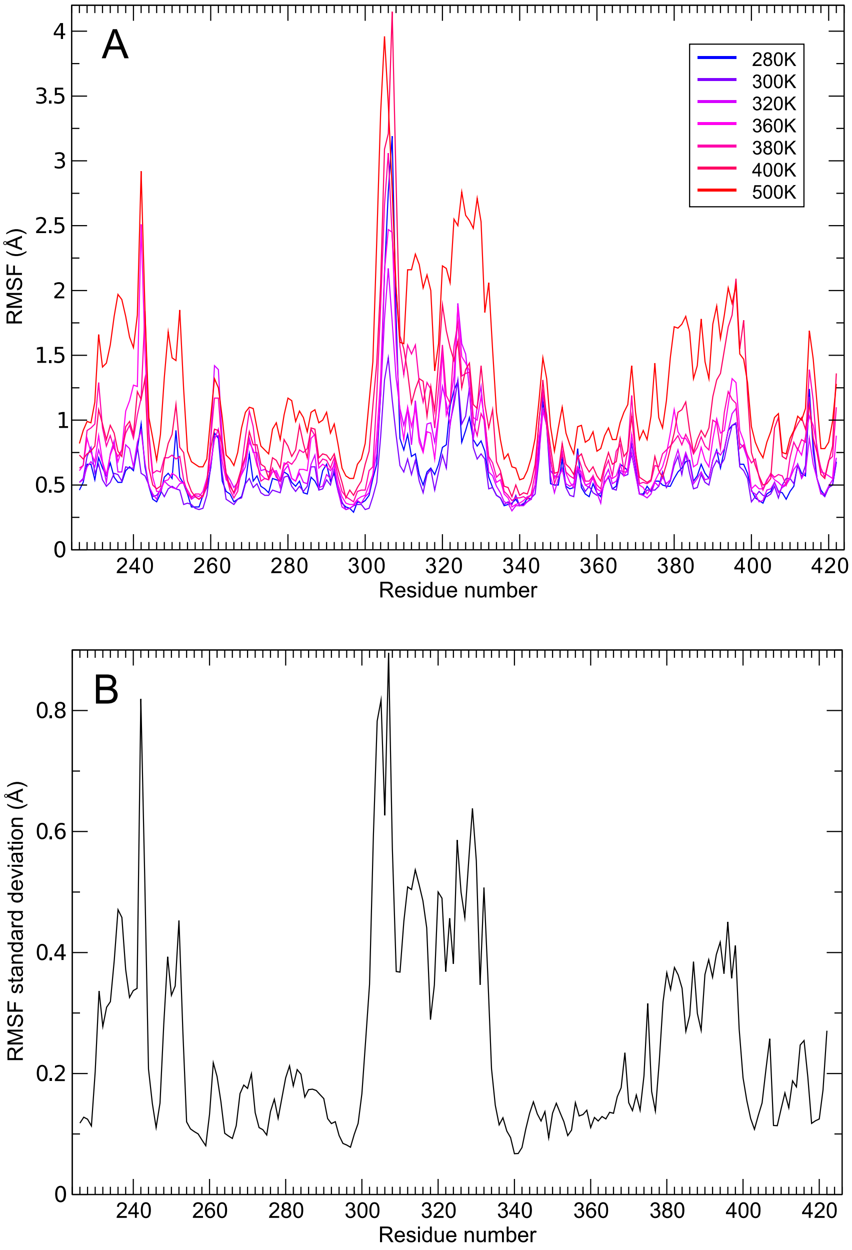


Figure 3-2. Thermosensitive analysis of LigTh1519 based on MD simulations. (A) Root mean square fluctuations (RMSF) of NBD residues at different temperatures simulated by MD. Residues numbers between 300 and 333 in NBD show the highest thermo-sensitivity. (B) Standard deviation of RMSF values at the relative different temperatures.

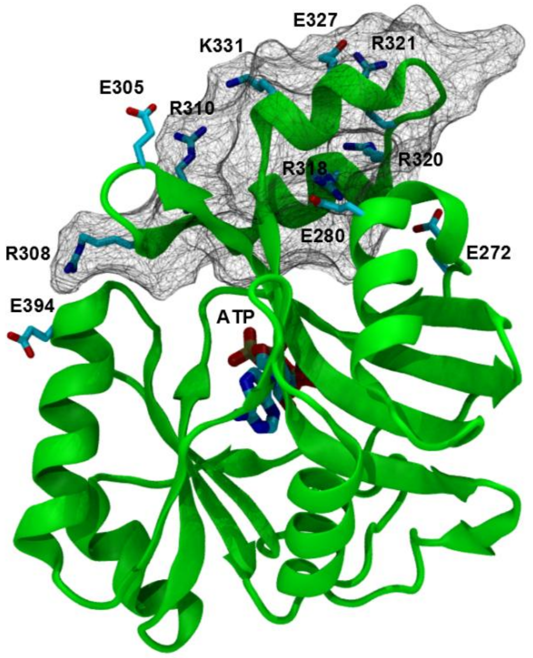


Figure 3-3. LigTh1519 thermo-sensitive regions. Location of the most temperature sensitive region is shown in wire-frame representation. Residues that could preserve integrity of this thermo-sensitive region at high temperatures comparing with mesophiles are shown in licorice representation.

3.3.2 Thermostability Engineering in LigTh1519

Using BLAST we found ~1000 homologous sequences for *LigTh1519* DNA ligase. They were classified as 6 psychrophilic, 883 mesophilic, 56 thermophilic, and 55 hyperthermophilic sequences. As *LigTh1519* DNA ligase is a thermophilic protein, and our goal was to increase thermostability, only hyperthermophilic sequences were chosen for multiple sequence alignment (Table S3_2). The distribution of amino acids at each position on the alignment was used to find key changes functional to enhance thermostability of this family of enzymes. This led us to focus on hydrophobic (or hydrophilic) residues on *LigTh1519* for which their counterparts in hyperthermophiles are hydrophilic (or hydrophobic) with more than 75% probability. Then, to find the most appropriate substitutions leading to enhance thermostability we analyzed the amino acid composition in the alignment at selected positions and finally amino acids with highest frequency on each position were chosen (Table 3-2). Following this screening protocol the best candidate positions for increasing thermostability obtained comparing the target sequence with 55 hyperthermophilic sequences are finally shown in Table 3-2. A287, G304, and A387 were hydrophobic residues for which their counterparts on the alignment were mainly hydrophilic (with probability of hydrophilic amino acids of 89%, 82%, and 75%, respectively). On the other hand, S364 was the only hydrophilic amino acid with probability of hydrophobic counterparts occurrence on the alignment as high as 93% (Table 3-2). In addition, to predict the effect of these mutations on the protein stability in terms of the relative change in free energy, the FoldX force field was used to estimate the ∆∆G contribution for the proposed mutations (Table 3-2).

Table 3-2. Predicted thermostable mutation for LigTh1519. In the first column only hydrophobic (or hydrophilic) residues for which their counterparts in hyperthermophiles are hydrophilic (or hydrophobic) with more than 75% probability are shown. Percentage of hydrophobic or hydrophilic residues is indicated as HB and HL in the second and third column. The forth column reports the distribution of amino acids at the selected site. Selected candidate(s) for mutation and the corresponding calculated free energy of mutation is shown in fifth column.

Residue	HB	HL	Distribution (%)	Mutant candidates	Selected
A287	0.11	0.89	K:35.7; E:28.6; A:10.7; D:8.9; N:7.1; S:3.6; R:3.6; T:1.8	K (ΔG = −0.65 kcal/mol) E (ΔG = −0.03 kcal/mol)	K
G304	0.18	0.82	D:57.1; G:14.3; K:10.7; E: 5.4; N:5.4; H:3.6; I:1.8; V:1.8	D (ΔG = -0.61 kcal/mol)	D
S364	0.93	0.07	I:57.1; V:25.0; L:7.1; A:3.6; S:3.6; T: 3.6T	I (ΔG = -1.06 kcal/mol)	I
A387	0.16	0.75	K:21.4; E:21.4; S:10.7; Q: 8.9; A:7.1; V:7.1; N:5.4; T:3.6	K (ΔG = 0.04 kcal/mol) E (ΔG= -0.06 kcal/mol)	K

Among the proposed position sites, we found that the most probable substitution for G304 is aspartate with 57.1 % (Table 3-2). G304 is located on a turn facing an adjacent β-strand where two arginines (R308 and R310) are located. Moreover, K331 is located on a α-helix next to G304, thus that a G304D appears a convenient choice to create a strong network of salt bridges in that region (Figures 3-3 and 3-4). S364 is another residue that is often mutated in hyperthermophiles DNA ligases, with mutation to an isoleucine being the most probable option (namely, 57.1 %). The molecular reason of this effect has likely to be found in the possibility of I364 to join and further stabilize the hydrophobic core composed by V294, V296, I341, V344, L349, L361, and V365 (Figure 3-3).

Among other possible mutation sites, there are candidate substitutions, which have similar probability to occur, but completely different chemical properties. For example, in the case of A287 two equally good substitutions are possible: lysine (35.7%) and glutamate (28.6%). To select the most appropriate mutations in these cases, the FoldX force field was used to rank the most suitable substitutions accounting for the interactions with neighboring residues. In particular, A287 is located on a helix (Figure 3-4), where it may form a salt bridge with E366. Thus, a lysine substitution is more likely to enhance thermostability, as confirmed by FoldX calculations, which indicate A287K as energetically more favorable than A287E ($\Delta G_{A287K} = -0.65$ kcal/mol *vs.* $\Delta G_{A287E} = -0.03$ kcal/mol, Table 3-2). A387 is another of these cases, with lysine and glutamate as equally possible mutations (both with 21.4%). But, in this case FoldX calculations could not help to discriminate between these two possible substitutions by showing no difference between predicted ΔG_{A387K} and ΔG_{A387E} (Table 3-2). As both methods, MSA analysis and FoldX calculations (Table S3_3), as well as MD analysis of the 2 mutants (see in Discussion), showed equal chances for lysine and glutamate to stabilize the structure, we picked lysine for consequent experimental evaluation. Based on this analysis we therefore predicted a list of point mutations, namely A287K, G304D, S364I and A387K, to be tested experimentally for thermostability enhancement.

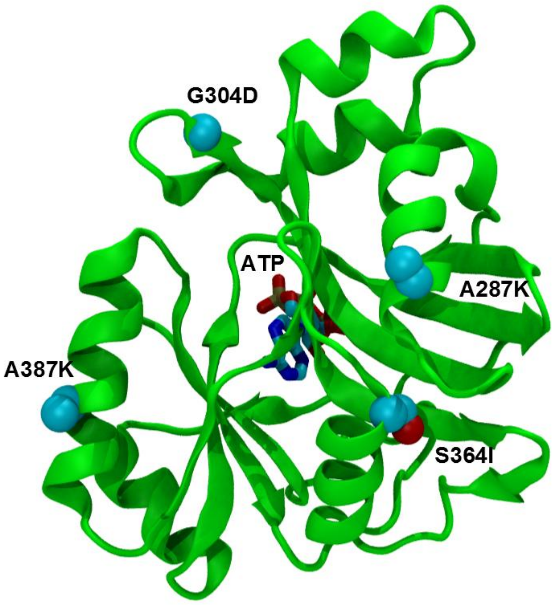


Figure 3-4. Thermostabilizing mutations on the LigTh1519 structure. Residues predicted to increase thermostability are shown in space filling representation. All the proposed mutations are far from ATP binding site and its interacting residues, which are located on the opposite face of the NBD.

3.3.3 Experimental Validation and Characterization of LigTh1519 Mutants

The proposed mutations were introduced in gene encoding *LigTh1519* ligase by site directed mutagenesis. The mutations were introduced either separately or altogether to more precisely dissect their contribution on thermostability. The recombinant proteins, unmodified *LigTh1519,* single mutants *(LigTh1519-A287K, LigTh1519-G304D, LigTh1519-S364I* and *LigTh1519-A387K*) and a mutant with all 4 combined mutations, *LigTh1519*mut (i.e., A287K, G304D, S364I and A387K) were expressed in *E. coli* and purified by metal affinity chromatography on 50% slurry of Ni-NTA resin (Qiagen). The purification of the enzymes was monitored by SDS-PAGE, which revealed a homogeneous 64,8 kDa major protein band, the size expected for the fusion product comprising the

63,38 kDa *LigTh1519* DNA ligase protein, and a 1,4 kDa peptide corresponding to the N-terminal His-tag.

The activity of the *LigTh1519* ligase and all mutants was quantitatively determined by a nick-closing activity assay utilizing a fluorescein/biotin substrate (Figure 3-5A). It was found that the half-inactivation time was 41 min for the enzyme carrying all 4 combined mutations, whereas unmodified enzyme loss half its activity upon 8 min incubation at 94°C. Enzymes with single mutations exhibited moderate, but non-negligible, increase in thermostability in comparison to *LigTh1519mut* with half-inactivation times at 94°C of about 15 min for *LigTh1519-A387K*, 20 minutes for *LigTh1519-A287K* and *LigTh1519-S364I*, and 25 min for *LigTh1519-G304* (Figure 3-5A*)*, hinting to an quasi-additive effect of the individual mutations on thermostability*.* The data on thermostability of *LigTh1519* at 94°C are in agreement with the ones reported in a previous study [99].

It is known that thermostable enzymes often has lower specific activity than their mesophilic counterparts since increased rigidity of the structure may reduce catalytic efficiency. Particularly *LigTh1519* mutant was designed to make more salt bridges stabilizing the enzyme structure. Therefore, we evaluated the relative activities of the *LigTh1519* ligase and the mutant in nick-closing activity assay. Equal amounts of purified enzymes (five different concentrations) were assayed and it was found that *LigTh1519*mut retained about 80% activity (Figure 3-5B). Activities of other individual mutants were in the range of 70-90% from wild type, except for *LigTh1519-G304* exhibiting about 60% of the maximal activity (Figure 3-5B).

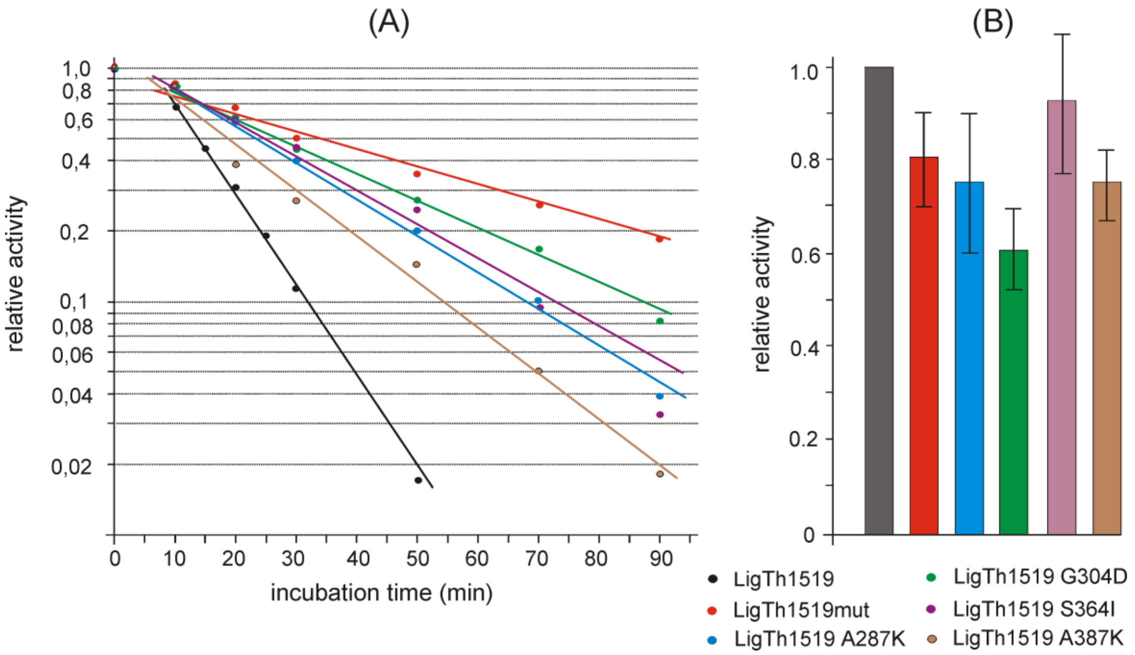


Figure 3-5. Thermostability and activity of unmodified and mutated LigTh1519. (A) Thermostability was evaluated after incubating the enzyme for various durations at 94°C. The residual ligase activity of the treated samples was then determined by the aforementioned nick-closing activity assays. (B) Relative activities of equal amounts of purified enzymes were determined by the nick-closing activity assays. Data represent the means of six experiments (serial dilutions of enzymes) and error bars represent standard deviation.

3.4 Discussion

As high temperature is a favorable condition for many industrial processes, the issue of enzymatic thermostability is a sensitive topic in this domain. Thus, hyperthermophilic enzymes isolated from hyperthermophilic species provide a naturally evolved and efficient tool for a variety of industrial and biotechnological applications. Nonetheless, the ability of computational and experimental techniques to further enhance thermostability would be beneficial to meet wider types of industrial demands even when naturally isolated enzymes are lacking.

Among isolated enzymes DNA ligase has been long used in biology and biotechnology, specifically for PCR and LCR [55, 117]. It was shown, for instance, that amplification of long DNA fragments could be enhanced if a thermostable DNA ligase is used to add a step of ligation in PCR cycles. Based on this evidence a new design was introduced for PCR amplification of circular plasmids using a thermostable DNA ligase [55, 117]. Although there have been some reports on isolation and purification of DNA ligases from hyper/thermophilic species [100, 116] to our knowledge the work presented here is the first report that addresses the thermostability of a thermophilic DNA ligase and aimed at enhancing its thermophilic properties via computational-based protein engineering.

To study and enhance thermal stability of an enzyme a deep knowledge of its unfolding mechanism is needed. Mechanically weak points on a protein structure are the most prone to initiate and lead to unfolding at high temperatures. Thus, the determination and optimization of these weak regions have been the focus of several studies analyzing thermostability of proteins [18-26]. B-factor analysis of crystallographic structures has shown to be useful to identify flexible residues that can be targeted for thermostability engineering [19, 118-123]. The dynamic information included in B-factors may however be affected by artifacts of crystal packing and does not always reflect the native dynamics in solution. Nonetheless, the integration of different methods has been shown to increase the chance to find mutations leading to more thermostable proteins [119, 121]. Thus, RMSF analysis extracted from MD simulations can be used as an alternative to the B-factor analysis to find the most flexible residues, probing at the same time the effect of multiple temperatures and regions that are more thermo- sensitive [18, 22]. Here, we used this strategy applied to *LigTh1519*, and found that the more thermo-sensitive region is located on region L300-E330 of the NBD (Figures 3-2 and 3-3).

To make thermo-sensitive regions resistant against harsh conditions nature uses different strategies. Several studies have addressed the differences between mesophilic and thermophilic homolog proteins [5, 6, 32-34]. Based on a number of reports one of the most common avenues taken by nature consists in altering charged residues on the protein surface to increase the number of salt bridges [5, 6, 32-34]. Our study based on multiple sequence alignments confirmed that in *LigTh1519* this network is significantly more populated than in mesophilic counterparts. In particular, there is a strong salt bridge network that preserves the connection of the thermo-sensitive region from residues L300 to E330 (Figure 3-3) at the NBD, presumably favoring the integrity of the domain at more elevated temperatures. Specifically, E280 and R318, which are conserved in all the mesophilic sequences, form a salt bridge that defines a minimal scaffold that helps preserving the connection of this domain to the rest of the enzyme. Then, addition of more salt bridges in thermophilic homologs as *LigTh1519* to this minimum scaffold can enhance the strength of this network preserving the integrity of protein subdomains against thermal perturbations. In fact, in *LigTh1519* a series of additional salt bridges (i.e. E272-R320, E305-K331, E305-R310, E327-R321, E327-K331, and E394-R308) are observed between this subdomain and the rest of the NBD, keeping it tight at the location (Table 3-1, Figure 3-3).

In particular, MD simulations of the individual mutants and *LigTh1519mut* system revealed that G304D in this region of the NBD showed a significant change in flexibility (Figure S3-3), likely due to the production of a tighter salt bridge network (e.g., D304 interacts with R310 and R308 with 53% occupancy). This is also consistent with the fact that this substitution is readily picked-up by our bioinformatics analysis and that this mutant is the one that seems to individually affect more significantly the half-life time at 94°C (Figure 3-5A). For the other two substitutions of this kind at positions 287 and 387, MD analysis of the single and combined mutations showed a less pronounced variation of flexibility (if not an increase as for K387, Figure S3-3), likely reflecting the fact that the newly formed salt bridge network is not as tight as in the G304D mutant (35% occupancy for K287, and 35% for K387) and that they are located on areas of the enzyme that are more solvent exposed and less connected with the rest of the enzyme in comparison to G304D (Figure 3-3). This feature is qualitatively consistent with the fact that these 2 mutations seem to individually account to a lesser extent to the overall thermostabilization of LigTh1519 (Figure 3-5A).

The identification of the thermo-sensitive regions and their stabilizing factors could be used as a lead, however cannot provide exhaustive information for proposing mutations to enhance thermostability. Although MD simulation is useful to locate candidate positions on the structure that are more temperature sensitive and can be targeted for thermostability engineering (as for position 304 in our case), it cannot directly propose appropriate substitutions to enhance thermostability. To predict suitable substitutions more general strategies based on comparison with sequences from higher thermostable proteins are useful.

Comparison with sequences from higher thermostable proteins also reduces the probability to perturb key conserved residues, as for instance those directly involved in the enzymatic catalytic activity. In *LigTh1519* five different highly conserved motifs are in fact located on the NBD that form the catalytic site. These motifs are essential for the three consecutive steps of DNA ligation (Figure 3-1) and substitutions in these regions can largely impact the enzymatic activity. This could explain the fact that, while enhancing thermostability to a greater extent among the single mutants tested (Figure 3-5A), G304D is also affecting the most the ligase enzymatic activity, likely due to its closer proximity with the active site (Figures 3-3 and 3-5B).

Specifically, thermostability analysis of *LigTh1519* by MD simulations at different temperatures showed that the most thermal sensitive regions are only located on NBD (Figure 3-2), while on DBD and OBD RMSF peaks are mainly located on coils and turns. It is also noteworthy how multiple sequence alignments confirm this point by suggesting mutations that are mostly located on NBD (residues M225-K420, Table 3-1) indicating that the detected thermo-sensitive region has been targeted for thermostabilization by nature. However, as the mutation candidates are based on major differences between amino acid content, there is little chance of proposing any mutation in more conserved motifs, as the catalytic site (Figure 3-4). Such a safe selection of mutation sites can be the main reason why our engineered DNA ligases did not show significant decrease of catalytic activity. In addition, using MSA we could detect and propose mutations that were impossible to detect only by using RMSF analysis. Among the proposed mutations (Table 3-2), G304 is immediately picked up by the MD based analysis (Figure 3-2), while on the other hand, S364 emerged only from multiple sequence alignments and was almost impossible to be detected only by using RMSF analysis. While sequence comparison could propose mutations that were able to optimize the hydrophobic core, thermo-sensitivity analysis using MD simulations were not much informative regarding this kind of interactions. All together these specific results show that MD simulations and sequence comparisons have to be used in synergy to have deeper insights into enzyme stability and thermostability engineering.

In addition, using sequence comparison approach does not always propose a single mutation for each candidate position on the sequence. Here, we used FoldX to rank the proposed mutations, strategy that turned out to give reasonable estimations given the intrinsic error associated with the calculations [61]. In the case energy calculations could not univocally select a suitable mutation, like for A387 in our case, MD simulations for both the candidate mutants, K387 and E387, revealed a very similar flexibility behavior, resulting from the multiple and equal possibilities of charged residues in this region to form salt bridges with the adjacent residues from the same helix (Figure 3-3). In particular, while in MD K387 can form salt bridges with E383 and E384, the E387 mutant can do the same with the basic residues K390 and R391. Based on this evidence, we predict that K387 and E387 can have an equal role for increasing thermostability. Seen the thermostability assessment of individual mutations (Figure 3-5A), both mutants are expected to have a similar, moderate effect (~15 min) on the global enhancement of half-inactivation time at 94°C.

In summary, the comparative in silico and in vitro analysis of the individual and combined impact of A287K, G304D S364I and A387K mutations allowed us to better dissect their contribution for LigTh1519 thermostability. We found that each of them represents relevant sites, invariantly producing a non-negligible, although moderate, contribution to thermostability. While the wild-type ligase has an half-life time at 94°C of 8 min, the mutant A387K reaches 15 min, S364I and A287K 20 min, and G304D 25 min (Figure 3-5A). A precise analysis to disentangle the contribution of each single mutation remains difficult, mainly because their non trivial, reciprocal spatial and dynamic relationship, however we can conclude that each mutation seems to have a quasi-additive effect on the overall half-life time at 94°C, resulting in a total increase of 41 min for the mutant carrying the 4 combined mutations (LigTh1519mut).

3.5 Conclusions

In conclusion, in this study we used computational techniques to deeper understand the origin of thermostability in *LigTh1519* DNA ligase as a thermostable enzyme. We engineered it to achieve higher thermal stability using rational design together with what we learned from the origin of thermostability of the wild type. Our results showed that an extended salt bridge network in the NBD plays a crucial role in stability preservation of *LigTh1519* at high temperatures over around 70 °C by keeping the thermo-sensitive region connected to the enzyme body. Finally, we suggested and tested four mutations, A287K, G304D, S364I and A387K to enhance the thermostability of the enzyme. Structural and dynamic analyses showed that A287K, G304D, A387K mutations play a role by electrostatic optimization of the surface, while S364I contributes to optimize the hydrophobic core of the enzyme. Experimental results proved that the proposed substitutions enhance thermostability of the protein by increasing in an additive fashion the half-inactivation time at 94°C from 8 to 41 minutes without significant reduction of specific catalytic activity.

The finding reported here emphasize on the importance and capabilities of molecular simulations combined with bioinformatics analyses in revealing stability/instability sources and predicting mutations able to enhance thermostability without affecting enzymatic activity. Therefore, the engineered hyperthermostable DNA ligase derived in this work could be used to optimize further PCR protocols [55, 117]. Furthermore, the approach presented here and applied to *LigTh1519* could be similarly adopted for other enzymes and can contribute to systematically produce more hyperthermostable enzymes for the industrial setting.

Note: The Supplementary Material of this work is reported in Appendix A1.

Chapter 4 Structure-Based Analyses for Protein Thermostability Engineering

Although proteins can function efficiently in the cells, reports indicate their stability is not completely optimized and they just satisfy minimum demands of stability to allow them to operate normally in the cell [1-4]. Therefore, at least in vitro, there are opportunities for enhancement of protein thermostability to satisfy industrial and biotechnological demands [1, 124, 125]. Using sequence and structure comparison, a number of factors have been reported to contribute for thermostability of proteins [83, 126-132]. In chapter 3, we used sequence and structure analysis to propose four thermo-stabilizing mutations for *LigTh1519* and experimental evaluations proved the efficiency of our predicted mutations. In this and the next two chapters, we will use our knowledge gained from protein thermostability engineering of *LigTh1519* and focus on the two major approaches that are commonly used for protein thermostability engineering; namely, *structure-* and *sequence-based* engineering methods. Our main aim here is to perform a systematic assessment of different popular means for protein thermostability engineering to better understand the accuracy of each one, and eventually find alternative approaches to time-consuming methods like MD simulations, developing a fast and accurate framework by integration of different methods and experimental data.

This chapter will be devoted to address structure based protein thermostability engineering techniques and in the next chapter, we will address strategies that can find thermo-stabilizing mutations based on analysis of homologous sequences without any structural information. Finally, through chapter 6, we will introduce a general protocol for protein thermostability engineering.

4.1 Introduction

From a structural point of view, increasing hydrogen bonding, promoting burial of hydrophobic surface area, producing better protein packing, modifying surface charges, increasing salt-bridges and disulfide bonds, producing more favorable helical dipolar interactions, improving secondary structure propensity, increasing rigidity and entropic stabilization are strategies all reported to be effective to enhance thermostability of proteins [5, 6, 32-34]. Although the mentioned factors are valuable strategies that can be used for thermo-stabilizing engineering of proteins [41-44, 46], none of the proposed thermostability enhancing factors could be used as a universal strategy for thermostability enhancement [126, 131, 132]. In fact, for different protein families the affecting factors could be different. More generally, studies suggested that stabilizing factors are even different between archaeal and bacterial proteins [133, 134]. Salt bridge formation is suggested to be the main stabilizing factor for bacterial proteins, while improved hydrophobic interactions for archaeal proteins are reported as key factors for stability [133, 134].

Therefore, to cover the thermostability enhancing factors based on structural data in this chapter we focused on two main strategies: (i) finding thermo-sensitive regions (so-called weak points or hot spots) that are potential targets for thermostability engineering, and (ii) strengthening hydrophobic (HB) interactions.

4.1.1 Finding Thermo-Sensitive Regions

Finding and optimizing thermo-sensitive regions on protein structures have been used extensively for protein thermostability engineering [18-26]. Thermo-sensitive regions are potential locations for initiation of unfolding at high temperatures. Optimization of thermo-sensitive regions results in a more stable structure at elevated temperatures [18, 19, 21]. To find the thermo-sensitive regions on a protein structure two methods have been mainly used, namely B-factor analysis and molecular dynamics (MD) simulation [18, 19, 21]. B-factor analysis of protein crystal structures could provide useful information about flexible regions on a protein structure [19, 118-123]. Reports have shown successful protein thermostability engineering studies by focusing on residues with higher B-factors compared to other residues in the protein and applying mutations that can stabilize them [118-123]. However, B-factors do not necessarily represent the flexibility state of residues in solution and they may be affected by crystal packing. To get a more realistic insight into protein dynamics, MD simulation has been used as the main tool to find the hot spots on protein structures [18, 21, 22, 24, 26]. In order to find hot spots with this approach, several MD simulations are usually run at different temperatures. Therefore, analysis of the per-residue root mean square fluctuations (RMSF) allows to find the most flexible fragments of the protein that are expected to be sensitive to temperature [18, 21, 22, 24, 26]. We used this strategy in chapter 3 and we successfully predicted some thermo-stabilizing mutations. However, such MD simulations are computationally expensive. Thus, it is also important to devise faster approaches to find hot spots on a protein structure. Here, our aim is to find protein hot spots using a simplified and fast technique like elastic network model (ENM). ENM is a topology-based coarse grained model at single-residue resolution (i.e., only C_α positions are most of the times taken into account), independent on the amino-acid sequence [66]. In such a model, details of residue interactions including salt bridges, hydrogen bonds and water solvation are not taken into account [66-68]. ENM has been used in a number of reports to study mechanics of protein structures (flexibility/rigidity) [67, 68]. Gaussian network model (GNM) and anisotropic network model (ANM) are two well established classes of ENMs [67, 68]. GNM has been mainly used to study fluctuations of amino acids in protein structures, as GNM can calculate the atomic fluctuations more accurately compared to ANM [67, 68]. On the other hand, ANM can provide directional information about atomic displacements, while GNM cannot [67, 68].

In this chapter we will explore the ability of a simple and fast method to find thermo-sensitive regions on protein structures. Next, we will examine its ability to locate candidate positions on protein structures to enhance thermostability by mutation. The detected thermo-sensitive regions can be stabilized via different strategies like introducing salt bridges.

In this study we used iterative GNM, as introduced by Su et al. [66] to originally study protein unfolding behavior. This method has been shown to be useful to study the unfolding priority for proteins and can show which residue has the tendency to unfolded first [66, 70, 135]. Finding unfolding nucleating centers can guide us to also reveal weak points on protein structure, which would be unfolding initiators at elevated temperatures, and subsequently suitable candidates for thermostability engineering. Although this method is well suited and validated for applications like ligand binding analyses [136], a systematic analysis of its applicability in the domain of thermostabilization is poorly explored. In this method, breaking non-covalent bonds between residues in close proximity simulates the unfolding of a protein. However, iterative GNM does not explore the conformational evolution during the unfolding explicitly and it can only give indication about the unfolding priority. Iterative GNM method has been used for CI2, barnase, and zinc finger [66, 135] structures to predict their unfolding priority. The sequence of unfolding events found by this method was reported to be consistent with thermal MD and MC simulation results [66, 137-147]. Although GNM is an amino-acid independent approach, using iterative GNM Su et al. concluded that the protein topology is an important factor for protein unfolding. Here, we studied the applicability of iterative GNM method to predict regions with high potential to enhance thermostability if they are mutated appropriately. In other words, we have studied how probable is finding sites on protein structures (as weak points) using iterative GNM that could significantly enhance the thermostability (i.e., increase ΔT_m by more than 5 °C). Using this method, we can only predict potential hot spots and no substitutions are suggested. Other methods (e.g. rational or consensus concept engineering strategies, see Chapter 5) should be used to find appropriate substitutions on the detected thermosensitive region by iterative GNM.

4.1.2 Improving Hydrophobic Interactions

Iterative GNM can only find hot spots, which are usually located on the surface of the protein, and cannot give useful information about clusters of HB residues that are usually buried inside the protein structure. These HB residues are usually keys to enable protein folding and general stability. Therefore, as a complementary strategy, we will explore and examine strategies that can enhance thermostability via engineering interactions among HB residues. The contribution of HB interactions on protein stability, by escaping from water media and forming and preserving HB-HB interactions, was first shown by Kauzmann [148]. HB cores have not only suggested to be essential for protein general stability [149-152] but also a number of studies suggested the HB environment/interaction modification as a principal stabilizing strategy for protein thermostability engineering [148, 149, 153]. In addition, the hydrophobicity content has been proposed to be informative to discriminate between mesophilic and thermophilic proteins [154]. Besides comparisons between mesophilic and thermophilic proteins, a number of mutagenesis studies have proven that protein stability could be altered significantly by mutations in HB cores [155]. However, the mechanism of stabilization/destabilization of such mutations in HB cores is not always well understood [155]. Some studies have proposed that change in stability is raised by factors like changes in transfer free energy and neighboring residues in structure [155, 156]. Specifically, for destabilizing mutations in HB cores, loss of hydrophobicity and disturbing well packed core residue side chains have been suggested to be at the origin of destabilization [149]. In addition, a number of experimental and theoretical works have addressed the effect of mutation size, in terms of small to large (and vice versa) substitution, for protein stability [151, 157-166]. Reports indicated that both large to small (i.e., by introducing cavities) [155, 159, 162, 163, 165, 167] and small to large (i.e., by introducing unfavorable contacts) [155, 168-173] mutations could result in protein instability. Therefore, mutations should make specific interactions and optimize HB cores and the “small to large mutations” strategy cannot be used as a universal rule for this kind of approach. However, L to I mutations have shown no change in protein stability [174] that could reveal the importance of side-chain hydrophobicity versus packing [174]. To gain more thermodynamics insights, different studies have correlated ΔΔG of substitution of a HB core member to different features of HB cores, like local packing density [151, 160], structural perturbations in neighboring atoms [175], and number of neighboring methyl and methylene groups [176]. On the other hand, while this approach is working for some specific protein, it has also been shown that generalization of these correlations is difficult [151, 165]. Even, somewhat simpler, correlations between ΔΔG and features of HB amino acids and consequent cavities and interactions [162, 165] could not be generalized to all protein families [151, 177].

Besides experimental studies, computational techniques have also been used to address the effect of HB core mutations on protein thermostability [155, 178]. However, the proposed computational techniques are usually costly and as discussed above, it is usually difficult to use them as general concepts for different proteins. Therefore, implementing faster, simpler, and less computationally costly methods, which can be generally applicable to different protein families, is still a pressing demand in the field of protein thermostability engineering [155]. Here, aiming at providing a fast and accurate protein thermostability engineering approach based on HB core optimization, we studied HB interactions and examined some simple rules that can be used for protein thermostability engineering. In other words, instead of prediction of all possible substitutions for HB residues, we just considered subclasses of HB substitution based on the available knowledge gained by structural comparisons and experimentally validated mutations.

4.2 Materials and Methods

4.2.1 Finding hot spots using a Gaussian Network Model (GNM)

The main assumption behind GNM is that inter-residue distances (fluctuations) around the equilibrium coordinates are following a Gaussian distribution [67-69]. In this model, C_α atoms are used as representative of residues in a protein and their coordinates in PDB files are considered as the equilibrium coordinates [68]. C_α atoms of i^{th} and j^{th} residues are connected by a classical spring with force constant of γ_{ij} if their distance is less than a cut-off radius of r_c.

For a protein, the internal Hamiltonian of the system can therefore be written as [66, 70]:

$$H = \frac{1}{2}\gamma[\Delta R^T(\Gamma \otimes E)\Delta R] \qquad \text{Eq. 1}$$

where ΔR represents residue fluctuations represented by the C_α atoms; the ΔR^T is the transpose of ΔR; E is the third-order identity matrix; $\otimes$ is the direct product; and Γ is the Kirchhoff matrix that is defined as follows:

$$\Gamma_{ij} = \begin{cases} -c & if|i-j| = 1 \\ -\gamma & if|i-j| > 1 \; and \; R_{ij} < r_c \\ 0 & if|i-j| > 1 \; and \; R_{ij} > r_c \\ -\sum_{i,j\neq i}\gamma_{ij} & if \; i = j \end{cases} \qquad \text{Eq. 2}$$

where R_{ij} is the distance between the i^{th} and j^{th} C_α atoms and r_c is the cutoff distance.

The N×N Kirchhoff matrix (Γ) can describe residue fluctuations and their cross correlations. It is noteworthy that the absolute value of γ does not affect the eigenvectors but uniformly scales eigenvalues [68]. The inverse of the Kirchhoff matrix can be decomposed as:

$$\Gamma^{-1} = U\Lambda^{-1}U^T \qquad \text{Eq. 3}$$

where columns of the orthogonal matrix U, u_i ($1 < i \leq N$), are the eigenvectors of Γ, and elements of the diagonal matrix Λ are eigenvalues of Γ, λ_i.

The cross-correlation fluctuations between the i^{th} and j^{th} residues can be calculated by:

$$< \Delta R_i . \Delta R_j > = \frac{3k_B T}{\gamma} [\Gamma^{-1}]_{ij} \qquad \text{Eq. 4}$$

where k_B is the Boltzmann constant and T is the absolute temperature. By assuming $i = j$, ($[\Gamma^{-1}]_{ii}$) using the equation 4 one can calculate the mean square fluctuation (MSF) of the i^{th} residue. Then, the Debye-Waller or B-factor of a residue can be calculated by:

$$B_i = \frac{8\pi^2}{3} < \Delta R_i . \Delta R_i > \qquad \text{Eq. 5}$$

To calculate the MSF in the distance vector R_{ij} between the residues i and j one can use:

$$\begin{aligned} \langle (\Delta R_{ij})^2 \rangle = \langle (R_{ij} - R_{ij}^0)^2 \rangle = \langle (\Delta R_i - \Delta R_j)^2 \rangle = \langle \Delta R_i . \Delta R_i \rangle + \langle \Delta R_j . \Delta R_j \rangle - 2 \langle \Delta R_i . \Delta R_j \rangle \\ = \frac{3k_B T}{\gamma} ([\Gamma^{-1}]_{ii} + [\Gamma^{-1}]_{jj} - 2[\Gamma^{-1}]_{ij}) \end{aligned}$$

Eq. 6

where R^0_{ij} and R_{ij} denote the equilibrium and instantaneous separation vectors between residues i and j, respectively.

Parameterization of the networks. B-factors in PDB files have been used to parameterize GNMs in terms of cutoff distance and spring constant [66-68]. In some studies for the covalent bonds, only for the links between two successive residues in the protein sequence, a different spring constant has been assigned to mimic the nature of its stronger bonding [66, 179, 180]. In such an approach, the spring constant is multiplied by a constant number that is assumed to be the same for all of covalent bonds.

Cutoff values are usually set in the range from 6.0 Å to 8.0 Å in different studies [67, 181-184]. Some studies have reported 7.3 Å on average by exploring the range 6 Å to 8.0 Å for cutoff values and residue interactions have been shown to be extended effectively up to 8 Å [67, 181-184]. Then, we explored the same region (6 Å to 8.0 Å) for cutoff distance to find the best cutoff that maximized the correlation coefficient between calculated B-factors (using Eq. 5) and experimental ones in PDB files [67, 181, 182] (Table 4-1)**.** To parameterize cutoff and the covalent bond constants a correlation coefficient was used as presented below:

$$\rho = \frac{\sum_{j=1}^{n}(x_j - x)(y_j - y)}{\left[\sum_{j=1}^{n}(x_j - x)^2 \sum_{j=1}^{n}(y_j - y)^2\right]^{1/2}} \qquad \text{Eq. 7}$$

where x_j and y_j represent experimental and computed B-factor values for j^{th} residue (C_α) respectively and x and y represent the average over all n residues in a protein. Relative rise and fall of the two B-factor curves can be calculated using a correlation coefficient that is independent of the scaling. Value of +1 indicates a perfect correlation while -1 represents a perfect anti-correlation [180].

As mentioned, spring constants have no effect on the correlation coefficient and are only a scaling factor. To optimize the force constant the least square of difference between the experimental and calculated B-factors was used [66, 179, 180].

Iterative GNM. Iterative GNM was implemented as introduced by Su et al. [66] in Python programming language [115] and a Prody [114] module was used for GNM eigenvalue/vector calculations. In iterative GNM, first MSFs are calculated for all residue pairs using Eq. 6. Then, the spring connecting the two residues with highest fluctuation is removed. This represents the breaking of non-bonded interaction between two residues. Finally, for the updated network the MSFs are calculated again and the above steps are repeated [66]. The algorithm flow is illustrated in Figure 4-1. In our study, we recorded the first 25% broken links in the network during the iterative GNM analysis. In order to evaluate the exact location of the mutations in the dataset (see next section) against the calculated hot spots, the closest atomic distance (D) between the mutated residues and the closest residue on the region with broken links was calculated. Then, the Loss-Number-of-Native-Contact (LNNC) with D < 4.5 Å was recorded (Table S4-1). The recorded LNNC and D for each mutation indicate how the mutated residue is located in respect to the hot spot region during the unfolding simulation.

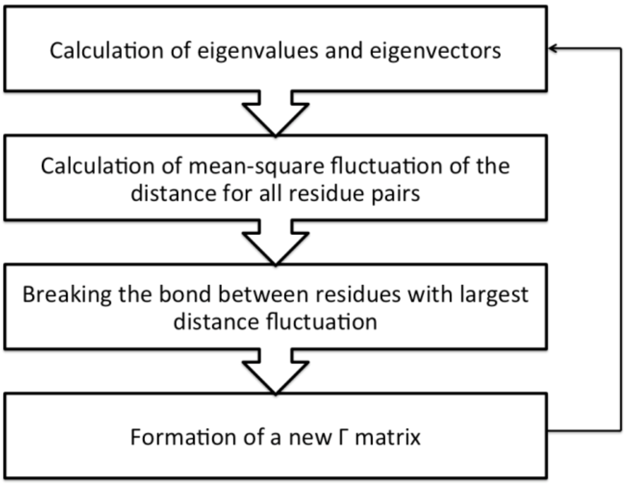

Calculation of eigenvalues and eigenvectors
Calculation of mean-square fluctuation of the distance for all residue pairs
Breaking the bond between residues with largest distance fluctuation
Formation of a new Γ matrix

Figure 4-1. Scheme for the algorithm of iterative GNM

Dataset. Mutations with $\Delta T_m > 5°C$ were extracted from ProTherm database [51]. T_m denotes the midpoint temperature at which thermal unfolding occurs and ΔT_m is defined as $T_m^{mutant} - T_m^{wild\text{-}type}$ [51]. Only mutations with reported X-ray crystal structures were used here, because B-factors are needed for parameterization of the elastic network. Taking these conditions into account, the total number of the mutated residues that we could find to define our dataset was 294 spread among 44 protein structures. The list of these structures and their PDB codes are reported in Table 4-1. Only mutations with ΔT_m higher than 5°C were considered to make sure that we focus our attention on residues with higher potential of increasing thermostability (i.e., to find potential hot spots). In fact, if a residue is located on weaker points its optimization could provide higher thermostability comparing with residues on rather stable regions. In other words, strongly stabilizing mutations are likely located at a position where the protein starts first to unfold [30, 185]. These mutations with their corresponding increase in T_m are shown in Table S4-1.

Depth Calculation. To have an estimation of the residue locations on a protein structure the depth of each residue was calculated using the DEPTH server [186]. DEPTH algorithm solvates a protein structure in a pre-equilibrated bath of water molecules and uses the distance between residues and the closest water molecules as a measure of the depth of residues on the protein [186].

4.2.2 Characterizing HB Interactions

We considered a group of amino acids as an HB core if there are at least 4 HB amino acids with closest atomic distance less than 4.5 Å and they were not located on long coils (length of coil should be shorter than 5 amino acids). For each residue, the secondary structure was assigned using the information on the header of the PDB files.

Single mutations with $|\Delta T_m|>1°C$, $5.5<pH<8.5$ in ProTherm database [51] that were studied by thermal denaturation and located on HB cores were extracted. The total number of the extracted mutations was 154 spread among 28 protein structures (Tables S4-3 and S4-4).

Using this dataset we followed simple rules (based on the review of the existing literature on the subject, see the Introduction section) for prediction of stabilizing/destabilizing effects of a mutation in a HB core:

1. Replacement of a HB core residue with non-HB amino acid results in destabilization.
2. The following order: G < A < V < I, L for amino acid substitution increases the protein stability, whereas the opposite order will result in decrease in stability. I and L have the highest probability to stabilize the protein and mutating them to any other amino acid will decrease the stability.
3. Replacement of any of the large HB amino acids in the HB core with G or A results in destabilization of the protein

4.3 Results and Discussion

4.3.1 Detection of Thermo-Sensitive Regions

By gradual increase of the temperature, fluctuation of residues leads to break of non-bonded native contacts between residues in protein structures, resulting eventually in un-folding. The weakest interactions, on hot spots here, are expected to be among the initial broken links [70]. Interestingly, even if two protein sequences vary significantly, but are sharing similar folding, they present similar unfolding rates [66, 187, 188]. Furthermore, experimental and theoretical studies have shown that the native topology of proteins strongly shapes the folding energy landscape and changes in sequences by evolution do not touch the native topology significantly [66, 189-192]. In addition, reports have indicated that the protein unfolding pathway does not change by increasing temperature [66, 193]. Putting all together, study of unfolding based on the native structures taken from PDB structures will be very useful to find weak points that nucleate the unfolding at elevated temperatures –and can thus targeted for protein thermostability engineering– in a protein family, even with significant variations in sequences of the family members [66].

However, it is noteworthy that from a theoretical point of view, GNM model is used to study fluctuations of the system around a single minimum while free energy landscape of protein unfolding/folding has many local minima. In an iterative GNM simulation, updating the normal modes after each link breaking is used as a possible strategy to overcome this problem [66, 179].

Elastic Network Parameterization. Elastic network parameters were calculated for the 44 protein structures in the dataset using the procedure described in methods section (Table 4-1). The calculated cutoff and bond constants for covalent bonds are shown along with their correlation coefficients. The number of links was calculated based on the corresponding calculated cutoff for each protein structure. This number was used to calculate the percentage of broken links that hold hot spot residues with favorable mutations. The average size of the dataset of our studied proteins was 156, ranging from 53 to 570 amino acids with distribution that is shown in Figure 4-2. Figure 4-2 demonstrates that the majority of structures (75%) have between 50 to 200 residues. In addition, for the studied structures, the average number of links in the protein networks was calculated as 701, ranging from 146 to 2804 links. The number of links is not only dependent on the number of residues and the protein 3D structure, but it also depends on the used R_c to construct the network. For the same structure, the higher the R_c the higher the number of links in the corresponding network. The average calculated R_c for the structures is 7.2± 0.56 Å, which is very close to the previously reported value (7.3 Å) [180]. The results shown in Table 4-1 indicated that, at least for this dataset, the parameter of the covalent bond did not play an important role. In other words, one can use uniform force constant without losing significant accuracy for the correlation coefficient between the calculated B-factors and the experimental ones (less than 3%). This fact is very useful because it allows us to decrease the number of parameters needed in our approach.

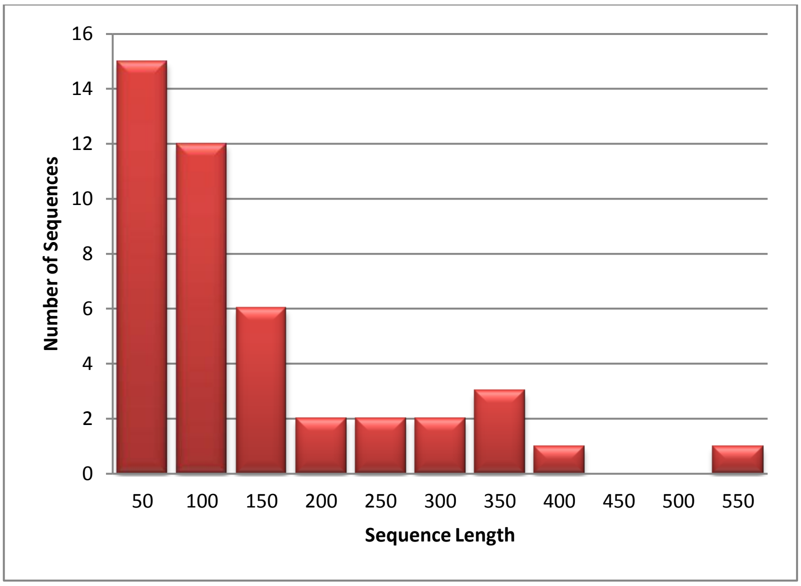


Figure 4-2. Size distribution of the studied proteins. 33 out of 44 structures (75%) have 50 to 200 residues.

Table 4-1. Parameterization and general information of the studied proteins. Cor. Coeff. shows correlation coefficient and Ratio shows the calculated optimum ratio for covalent bonds (see methods section). The reported correlation coefficient for R_c is calculated with homogenous spring constants for covalent and non-covalent links. Number of links shows the total number of links between all particles using the calculated Rc.

No.	**PDB ID**	**Length**	**Number of links**	**R_c (Å) (Cor. Coeff.)**	**γ/k_BT**	**Ratio (Corr. Coeff.)**
1	1JK9	153	486	6.3 (0.62)	0.9	2.6 (0.62)
2	451C	82	371	8.0 (0.22)	0.8	12.8 (0.27)
3	1HFZ	123	344	6.1 (0.80)	1	1 (0.80)
4	1CSP	67	284	7.5 (0.50)	0.63	1 (0.50)
5	1RGG	96	325	6.8 (0.47)	1.4	1 (0.47)
6	1SSO	62	244	7.8 (0.95)	3.16	3.0 (0.96)
7	1YCC	102	332	6.4 (0.52)	1.3	1.9 (0.53)
8	1PGA	56	202	6.9 (0.24)	1.4	1 (0.24)
9	1SHG	57	146	6.1 (0.88)	0.9	2.3 (0.89)
10	1N0J	198	845	7.3 (0.63)	0.84	3.6 (0.65)
11	3SSI	108	478	7.7(0.48)	0.44	10(0.51)
12	2RN2	155	581	6.7(0.77)	1.2	1.2(0.77)
13	2CBR	136	580	7.2 (0.68)	0.57	4.4 (0.70)
14	1KEV	351	1637	7.5 (0.40)	0.48	5 (0.40)
15	2HPR	87	390	7.5 (0.55)	0.73	6.1 (0.60)
16	2TRX	108	462	7.6 (0.73)	0.92	1.8 (0.74)
17	1TEN	89	252	6.4 (0.46)	1.3	1 (0.46)
18	1FTG	168	721	7.3 (0.53)	0.98	7.3 (0.54)
19	1guy	296	1063	6.5 (0.56)	0.89	11.2 (0.62)
20	1IRO	53	199	7.3 (0.66)	1.1	1.1 (0.66)
21	1BCX	185	933	8.0 (0.70)	1.1	1.1 (0.70)
22	1JU3	570	2804	7.7 (0.71)	0.68	7.0 (0.75)
23	1YNR	80	308	7.1 (0.23)	0.96	21.0 (0.26)
24	1B5M	84	269	6.5 (0.74)	0.98	1.1 (0.74)
25	1AZP	66	166	6.1 (0.73)	1.1	1.2 (0.73)
26	1H8V	117	431	6.9 (0.66)	1.02	4.9 (0.70)
27	1QLP	372	1368	6.8 (0.52)	0.85	1.4 (0.52)
28	1DIL	381	2104	8.0 (0.67)	0.6	21.0 (0.74)
29	2CI2	65	258	7.6 (0.90)	0.62	11.1 (0.93)
30	2AFG	129	641	7.9 (0.56)	0.77	1.1 (0.56)
31	1LZ1	130	626	8.0 (0.62)	0.62	10.5 (0.68)
32	1RN1	103	402	7.3 (0.78)	0.74	5.1 (0.79)
33	1SHF	59	171	6.4 (0.91)	1.09	1.3 (0.91)
34	4BLM	256	1039	6.9 (0.78)	1.1	1.9 (0.79)
35	2OCJ	194	818	7.4 (0.70)	0.73	3 (0.71)
36	1P3J	212	775	6.6 (0.77)	0.93	2.4 (0.77)
37	2ACY	98	443	7.5 (0.59)	0.51	19.8 (0.75)
38	1AYE	307	1227	6.9 (0.76)	0.97	1.9 (0.77)
39	1OSI	345	1650	7.5 (0.43)	0.88	1.1 (0.43)
40	1AQH	448	1966	7.3 (0.68)	0.89	3 (0.68)
41	1IO2	213	931	7.3 (0.75)	0.62	21 (0.85)
42	1RTP	109	411	7.0 (0.38)	1.06	12.1 (0.44)
43	1RTB	124	551	7.6 (0.38)	1.1	2.7 (0.39)
44	1DIV	149	617	7.3 (0.84)	0.48	20 (0.92)

Thermo-Sensitive Regions Analysis. To identify thermo-sensitive regions, we defined a region of 4.5 Å around any broken link and considered it as a hot-spot. For the mutations in the dataset we recorded the LNNC that a mutation was located in this region. Then, by dividing the LNNC by the total number of links within the corresponding elastic network model (see Table 4-1), we estimate the LNNC%. The LNNC% is reported instead of LNNC to normalize results relative to structures with different number of residues and links within their network models. Results of iterative GNM are summarized in Table S4-1 for 294 mutations. From the 294 mutations in the dataset, 281 (96%) mutated residues were located on the hot spot with LNNC% < 25%. As our aim is finding residues located on regions that get unfolded with higher priority, we only recorded LNNC% for mutations with a LNNC% less than 25% (Table S4-1).

Therefore, for example, for H93 of the protein with PDB ID: 2AFG (reference number 8299 in Table S4-1) a LNNC% of 0.9 means that residue H93 is located within 4.5 Å of the first 0.9% of broken links in the elastic model. For this residue the distance to the broken link is 0, which means that the residue is actually involved in the breaking of a link within the elastic network at that step. But, a residue like D75 on protein structure with PDB ID: 1FTG (reference number 20006 in Table S4-1) is located within 4.5 Å of the link broken at LNNC% = 1.2% with distance of 1.3 Å. This residue could be the residue with broken link during the next steps of the unfolding (that for this residue happens within LNNC% = 4%). Figure 4-3 illustrates the mentioned sample of the detected hot spot for residue D75 on PDB ID: 1FTG on its corresponding structure along with the built elastic network model. Specifically, for residue D75 on PDB ID: 1FTG (Figure 4-3A) detected as hot spot, mutation D75K increases the ΔT_m by 7.8 °C. Analysis of the corresponding structure shows local charge of -3 within 8 Å of D75, which clarifies the reason why D75K is able to enhance thermostability. In addition to the possibility of forming salt-bridges, this mutation can modify the local charge distribution of the region to decrease repulsive electrostatic interactions.

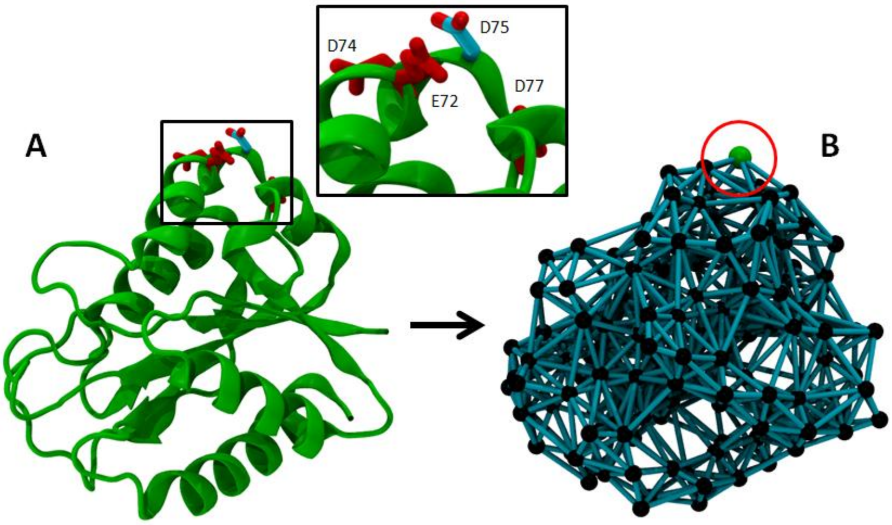


Figure 4-3. Residue D75 is located on one of the detected hot spots on protein structure with PDB ID:1FTG. A) Three negatively charged residues within 8 Å of D75 (D74, E72, D77) make local charge of -3 around it. B) The built elastic network for protein structure with PDB ID: 1FTG. C_α atoms are shown in black in the elastic network representation and D75 residue is highlighted in green.

Figure 4-4 shows the distribution of LNNC% for the 281 mutations: 73% of the 281 residues are found within the first 4% of LNNC% and this increases only by 9% by increasing the LNNC% up to 6% (Figure 4-4).

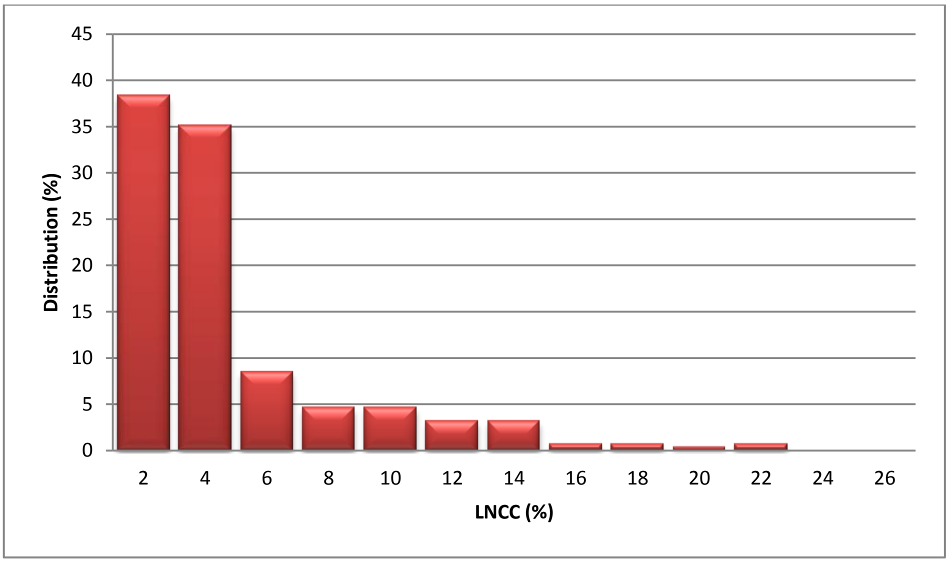


Figure 4-4. Distribution of LNNC% (LNNC number/total number of links in the protein) of mutations.

Thus to keep the list of candidate residues for thermostability engineering as short as possible we can conclude that it is better to focus on hot spots during the very early steps of unfolding. Therefore, for application purposes, one can use a 2% threshold to define a first candidate list of hot spot positions and, if more candidates are needed, the threshold can be increased considering a LNNC% cutoff up to 4% or 6% in the following steps. Figure 4-5 shows the distribution of the percentage of residues that loose their links during the first 2%, 4%, and 6% LNNC% in the studied 44 structures. The average percentage of residues that loose their links through the first 2%, 4%, and 6% LNNC% is 11% (± 2.1) (ranging from 7% to 15%), 19% (± 2.7) (ranging from 13% to 25%), and 25% (± 3.9) (ranging from 16% to 32%), respectively.

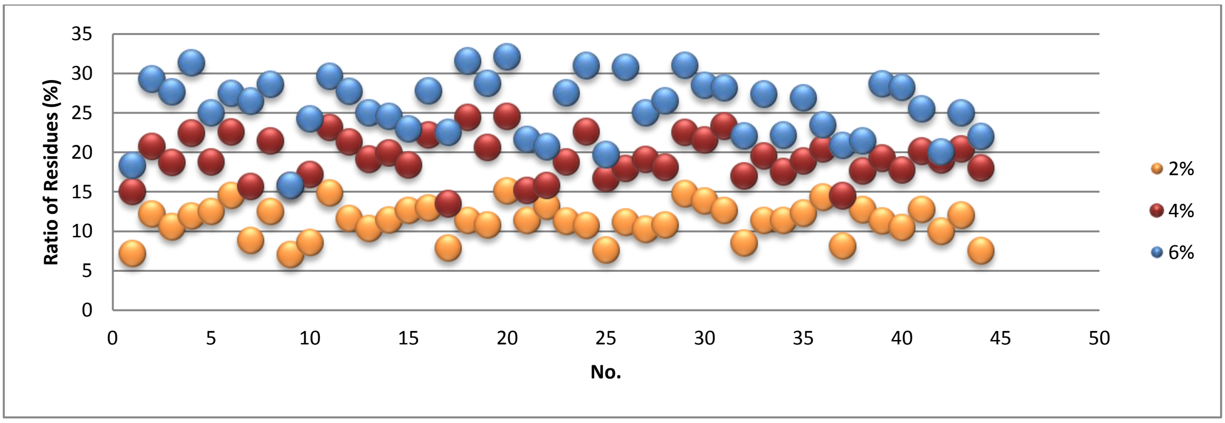


Figure 4-5. Ratio of residues loosing links during the first 2%, 4% and 6% of LNNC for the 44 structures used in this study.

To study the location of the calculated hot spots and mutations on the protein structures the depth of residues was calculated (Table S4-1). Figure 4-6 shows the distribution of depth of residues loosing links during the first 6% LNNC% on the protein structures. Depth analysis of the residues identified above showed that the majority (60%) of residues with LNNC% <6% were located on the surface of the protein (depth between 2 and 4 Å) (Figure 4-6). In addition, residues with higher depth that got unfolded within the first 6% LNNC% were very close to those located on the surface and unfolded after them. These results show that after unfolding the non-stable residues located at the surface, the local density of links decreases and can make the neighboring residues with higher depth unstable; this might induce adjacent links to break at subsequent steps. These residues can be even HB residues forming HB interactions. If these are residues forming HB interactions this indicates that engineering of these residues via modification of HB-HB interaction, could be of major interest to improve thermostability of the protein; because they are more prone to get exposed to water media and therefore destabilizing the protein. By optimizing those regions, the strength of the connected regions on the surface can be affected as well. (Due to the method used in the DEPTH server [186], all water molecules closer than 2.6 Å to the protein are removed first and the depth of each atom is given by calculation of its distance to the closest water molecule. This is why a few residues with depth less than 3 Å are reported in Figure 4-6).

Altogether this analysis shows that the hot-spots detected by iterative GNM approach are mainly located on the surface of protein structures. This fact discloses one of the limitations of iterative GNM method that is able to detect only residues located on protein surface and cannot detect potential candidate locations among buried residues inside the protein core, which is usually composed by HB amino acids. On the other hand, as discussed in the introduction, HB residues play important roles in stability of proteins. Therefore, alternative approaches should be used to find potential thermostability engineering targets among HB residues, topic that will be discussed in the next section.

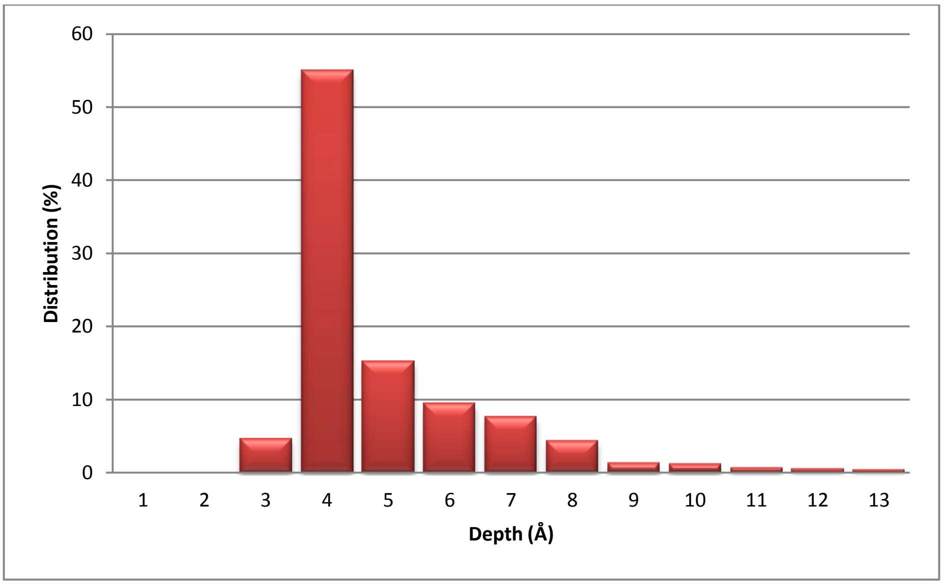


Figure 4-6. Distribution of residue depth for the residues located on the first 6% LNNC%.

We next studied the relationship between the LNNC% and the experimental ΔT_m by plotting ΔT_m vs. LNNC% (i.e., 69 mutations in our dataset from Table S4-1, see Table S4-2). Only single mutations were used here because we had no estimation of the contribution of each individual mutation in multiple mutations context. The plot of ΔT_m vs. LNNC% is illustrated in Figure 4-7 and raw data are reported in Table S4-2. Although no apparent correlation was observed between ΔT_m and LNNC% (Figure 4-7), for single mutations with very strong effect on thermostability (ΔT_m > 15 °C), 11 mutations out of 15 had LNNC% less than 2%. However, the mutation (i.e., G85R in PDB ID: 1JK9) with the highest increase in thermostability (ΔT_m = 45 °C) does not have such a low LNNC% (~18%). In addition, this residue, with calculated depth of 6.4 Å, is not located on surface (Table S4-1). Furthermore, there are other residues with LNNC% > 10% and significant effect on thermostability (ΔT_m > 10 °C) that are located on different locations corresponding to the surface. T53 on PDB ID: 1PGA, G93 on PDB ID: 1JK9, and D79 on PDB ID: 1RGG have calculated depth of 3.2 Å, 4.7 Å, and 5.4 Å respectively. Thus, although appropriate mutations on positions with LNNC% < 2% can have strong effects on thermostability enhancement, this criterion seems to be not exclusive and one should consider the hot spot optimization as one piece to solve the puzzle of thermostability engineering. For example, HB interactions, which are usually far from surface hot spots, are other important players in the thermostability of proteins.

However, besides the ability in finding hot spots on a protein structure, the main limitation of this method is that a 3D structure is needed. In addition to a 3D structure, B-factors should be assigned to atoms within the PDB file, meaning that this method is limited to X-ray crystal structures. For proteins sequences lacking a 3D structure, one can use homology modeling, if any template is available in the family, and assign B-factor using machine learning-based predictors [194]. In addition, as this method is sequence free, even if one does not find any structure for the target sequence, using an accurate alignment one can use results from currently available 3D structures from the same family and map the residues between sequences.

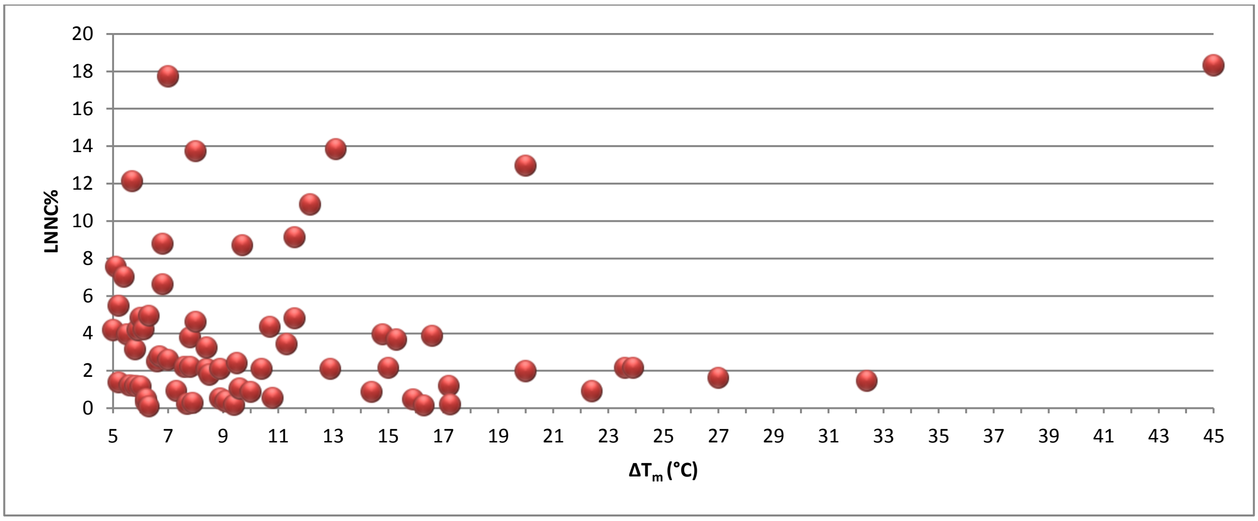


Figure 4-7. ΔT_m vs. LNCC% for single mutations.

4.3.2 *Improving HB Interactions*

Following the simple rules defined in the methods section, we studied the effect of mutations on thermostability for 154 mutations within the 28 protein structures in our dataset (Table S4-3). Table 4-2 shows the list of studied structures for HB mutation predictions. The average length of the sequences is 124 ranging from 51 to 381 amino acids. Tables S4-3 and S4-4 show prediction results for the 154 mutations.

Table 4-2. PDB IDs of the studied structures.

No	PDB	Length	No	PDB	Length
1	451C	82	15	1CYO	88
2	1HFZ	123	16	1DIL	381
3	1PGA	56	17	1H8V	117
4	1N0J	198	18	1HME	77
5	3SSI	108	19	1IOJ	57
6	2RN2	155	20	1PIN	157
7	2TRX	108	21	1QU7	226
8	1FTG	168	22	1ROP	56
9	1MBG	51	23	1SHF	59
10	1RTB	124	24	1STN	135
11	1RN1	103	25	1WQ5	268
12	1ARR	53	26	2CI2	65
13	1AZP	66	27	2LZM	164
14	1B5M	84	28	4LYZ	129

In Table S4-3 the mutations with correctly predicted effects on stability of proteins, using rules introduced in methods section, are shown. The total number of thermo-stabilizing and destabilizing mutations in the dataset was 27 and 127, respectively; while the effect of 146 mutations (95%) were successfully predicted by the rules. The number of true positive (TP) and true negative (TN) results is 22 and 124, respectively. On the other hand, Table S4-4 shows wrong predictions that are composed of 5 false negative (FN) and 3 false positive (FP) predictions. Results of statistical analysis for performance estimation, shown in Table 4-3, indicated a very good performance of this protocol, producing results with very high accuracy (95%), with the capability to distinguish between thermo-stabilizing and destabilizing mutations. The high TP rate and TN rate (0.81 and 0.98, respectively) indicate high efficiency for prediction of both thermo-stabilizing and destabilizing mutations. While the portion of thermo-stabilizing mutations is very low (18%) within the dataset, TP rate of 0.81 indicates robustness of predictor and shows that prediction results are far from random expectation. The estimated Matthew coefficient (MCC) of 0.82 confirms this outcome, as this quantity evaluates the correlation coefficient between the observations and the prediction results taking into account the TP, TN, FP, and FN (for more details see Chapter 2, section Statistical Analysis of Predictions). The calculated MCC is very close to 1, which indicates a quasi-optimal prediction, and far from 0, which indicates not better than a random prediction.

Table 4-3. Statistical analysis of HB-core mutation prediction results. TP, TN, FN, FP, ACC, TPR, TNR, MCC, and F1 represent true positive, true negative, false negative, false positive, accuracy, true positive rate, true negative rate, Mattews correlation coefficient, and F1 score respectively.

TP	TN	FN	FP	P	N	ACC	TPR	TNR	**MCC**	**F1**
22.00	124.00	5.00	3.00	27.00	127.00	0.95	0.81	0.98	0.82	0.85

Figure 4-8 shows the distribution of the applied rules for 146 mutations presented in Table S4-3. Interestingly, 73 of the correct predictions (almost 50%) are based on the G < A < V < I, L replacement rule that provides a specific way to find thermo-stabilizing/destabilizing mutation candidates.

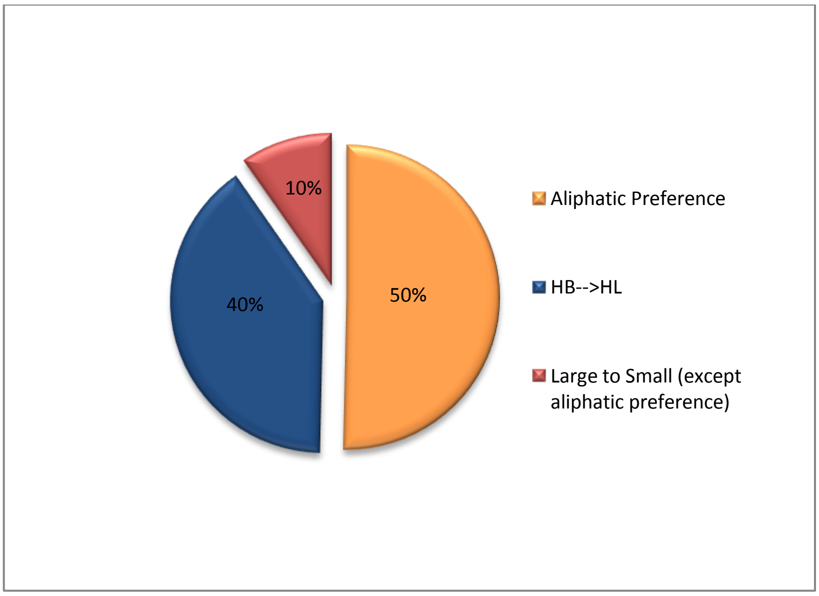


Figure 4-8. Distribution of the applied rules for 146 mutations that are presented in Table S4-3. Large to small mutations (shown in red) shows mutation of non-aliphatic large HB residues to A or G.

Fortunately, there are data (Table S4-3) that prove this affinity order for some individual points on protein structures. Such data are valuable to prove the validity of this rule for protein thermostability engineering. For example mutations of V74 to A, L, and I in the 2RN2 structure, V103 to A, L, and I in the 1MBG structure, A31 to V, L, and I in the 4LYZ structure, L41 to A and V in the 1ROP structure, I133 to A and V on 1SHF structure, V66 to A and L on 1STN structure, and I3 to G, A, and V on 2LZM structure shows the consistency of this rule. Figures 4-9 and 4-10 show the mentioned mutation sets on PDB IDs: 2RN2 and 4LYZ structures. Among the data shown in Table S4-3, 40% corresponds to HB to HL mutations that resulted in a decrease in the protein thermostability (Figure 4-8). The remaining 10%, which is represented as large to small mutations in the Figure 4-8, shows mutation of non-aliphatic large HB residues to A or G.

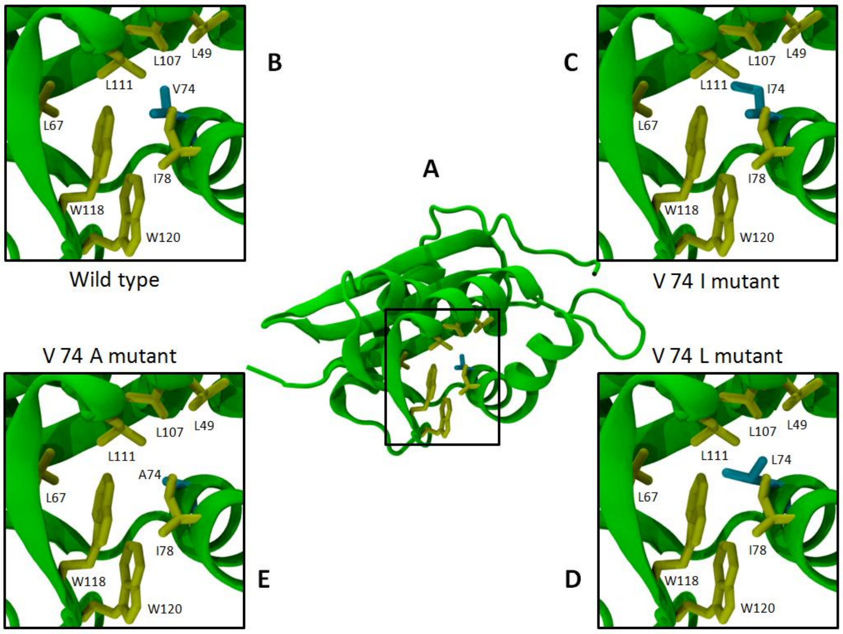


Figure 4-9. Structure of protein with PDB ID: 2RN2. The mutated HB residue (V74) and its interacting HB residues within the HB-core are illustrated in cyan and yellow respectively (A). B, C, D, and E show the magnified view of the same region for the wild type, Ile, Leu, and Ala substitutions, respectively. The mutants are built using Chimera package.

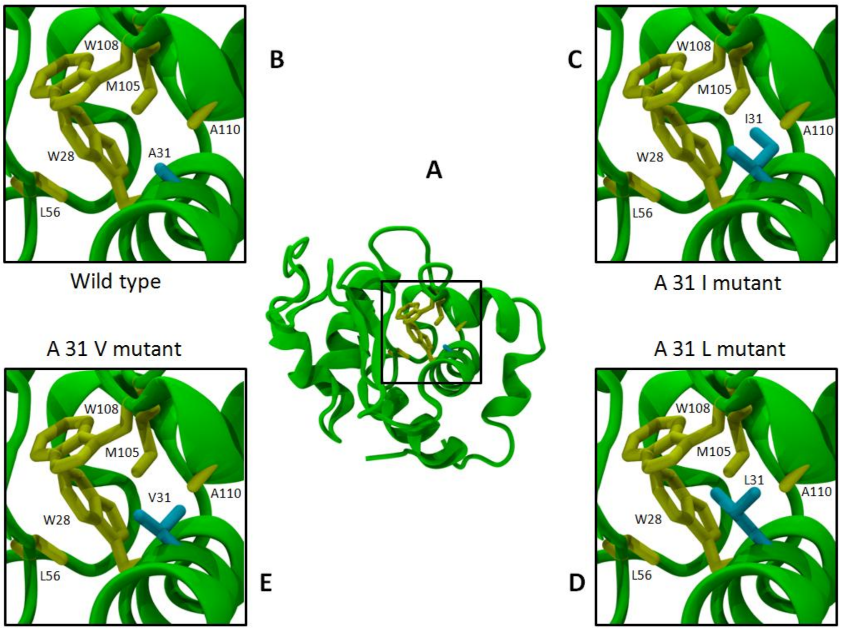


Figure 4-10. Structure of protein with PDB ID: 4LYZ. The mutated HB residue (A31) and its interacting HB residues within the HB-core are illustrated in cyan and yellow respectively (A). B, C, D, and E show the magnified view of the same region for the wild type, Ile, Leu, and Val substitutions respectively. The mutants are built using Chimera package [195].

Furthermore, distribution of ΔT_m of mutations in Table S4-3 is illustrated in Figure 4-11. Interestingly, the majority of mutations (85%), which were correctly predicted by the rules, were destabilizing mutations. This observation is reasonable because HB interactions are principal factors for stability of the protein and any disruption of this kind of interactions has drastic effect on the protein stability [148, 149, 153]. This observation also showed the power of these rules to filter out potential destabilizing mutations from mutation libraries.

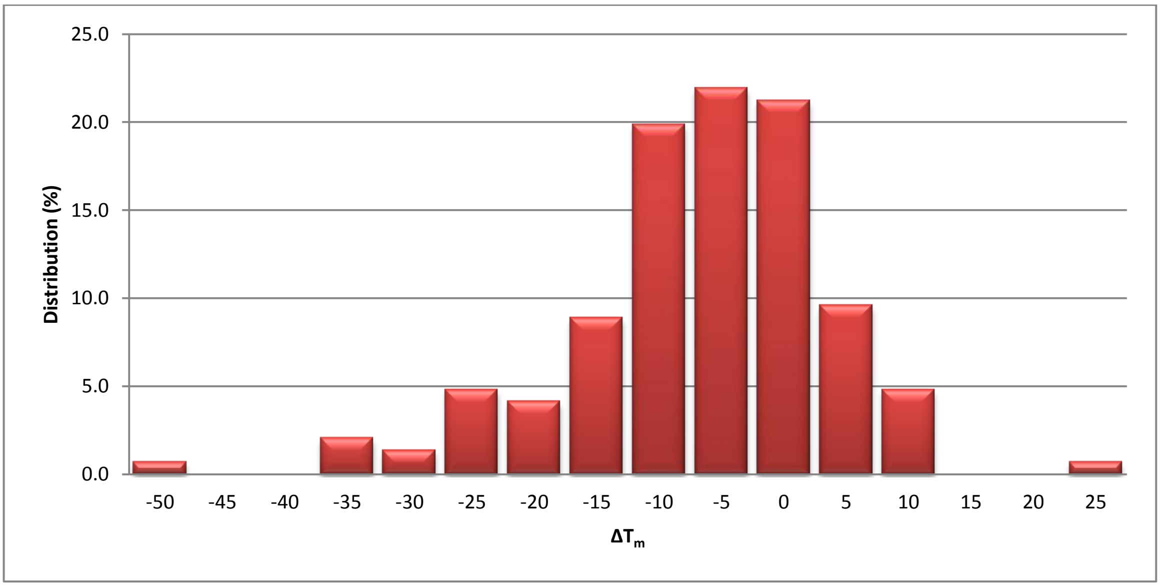


Figure 4-11. Distribution of change in thermostability of the studied mutations

Finally, it is noteworthy that small to large HB mutations may result in small structural rearrangements and consequently may damage the protein activities. Although mutations on HB cores may perturb it, reports indicate that some proteins can tolerate small perturbations in HB packing by small structural readjustments [1, 150, 151, 165, 196]. But, these mutations may lead to drastic energetic changes in protein stability [150, 151]. For an enzyme like T4 lysozyme the displacement of backbone atoms due to mutation of some HB core residues was reported to be as high as 2 Å [165]. In another study, HB core residues were substituted with large residues and the structure accommodated the potentially disruptive substitutes by a rearrangement of its α-helices. The resulted engineered protein was more stable than the wild type because of higher packing density and little geometric strain [197]. However, a significant reduction in the protein activity (in this case DNA binding) was observed due to displacement of functional residues [197]. Such observation showed that for some substitutions a protein can tolerate different degrees of flexibility in its main chain to gain higher stability if the substitution is not fit to the current HB core topology, but, on the other hand, it could result in disruptions in protein main function because of the structural rearrangement [197]. Therefore, a specific group of amino acids in the HB core should be selected not only because of optimal stability, but also because they can provide precise structural requirements for optimal activity [197]. Then, In case of using aliphatic affinity rule, special cares should be taken to prevent any unfavorable side effects on protein function. Besides experimental tests, it is strongly recommended to use such small to large mutations on HB cores close to cavities or on the larger HB cores, thus that residues could tolerate changes by side chains without altering the backbone.

All together, these simple rules seem to be very powerful to predict spot for protein thermostability engineering of the HB core of a protein. Specifically, using the Gly < Ala < Val < Ile, Leu replacement rule one can detect candidate thermo-stabilizing mutations on protein structures. On the other hand, the thermo-destabilization prediction power of rules can be very useful if coupled with other prediction methods to filter-out risky thermo-destabilizing mutations. While on one side this protocol is extremely fast, its main limitation is that is limited to targeting proteins of known 3D structure. However, as the high accuracy of this method in detection of both thermo-stabilizing and destabilizing mutations is tempting, on can use homology modeling or evolutionary coupling method to construct models, even at intermediate resolution, to detect HB interactions and apply the studied rules for the target protein sequence. Another limitation of these rules is that it is not possible to estimate the ΔT_m of the predicted mutations. To gain knowledge about ΔT_m one alternative is coupling these rules with machine learning techniques, training a model for ΔT_m prediction. However, even using machine-learning techniques can be challenging because the amount of data, that specifically satisfies these rules, is still limited.

4.4 Conclusions

In conclusion, two strategies were addressed in this chapter to find hot spots on protein structure and to optimize HB interactions. Iterative GNM method was shown to be powerful in finding hot spots on protein structure as targets for protein thermostability engineering. The detected hot spots can be optimized by different rational engineering methods like introducing salt bridges. We also showed that Iterative GNM is efficient to find hot spots only located on the protein surface and cannot give useful information for thermostability engineering of HB residues that are usually buried inside the protein. Therefore, we introduced simple, but very efficient, rules for thermostability engineering of the HB core. These rules were shown to be powerful in predicting thermo-stabilizing and destabilizing mutations. These strategies can be utilized together on a target protein structure to engineer weak points at the protein surface and strengthening more the HB core members to gain higher thermostability. However, both methods are suffering from a common weakness: the need for a 3D structure of the target protein. To solve this problem, alternative solutions using homology modeling or evolutionary coupling techniques were also discussed.

Chapter 5 Sequence-Based Protein Thermostability Engineering

5.1 Introduction

All the analyses presented in chapter 4 were based on the structural knowledge of the protein target. But, for a large part of (interesting) protein sequences there is no structure available. On the other hand, there are plenty of characterized and annotated protein sequences available in public databases. Therefore, any analysis tool that can provide hints towards enhancement of thermostability of proteins out of such a big treasury would be of great interest for scientists and biotechnologists. As an alternative to structure-based engineering, sequence-based engineering, which is also known as data driven engineering, has been recently attracted attentions. Specifically, advances in DNA sequencing technologies and increasing number of characterized protein sequences have made this approach more attractive in the field [47]. This approach uses all the available homologous sequences to propose thermo-stabilizing mutations [47]. To extract thermo-stabilizing mutations out of homologous sequences, the consensus concept (CC) is introduced that has been used in a number of studies as the main sequence-based protein thermostability engineering technique [48, 50, 96, 121, 198-201]. The logic behind this method is simple: Using a multiple sequence alignment (MSA) non consensus residues are substituted by consensus ones [47]. This method was first introduced by Pantoliano in 1989 [202] and a few years later Steipe, in 1994 [203], provided a statistical explanation for it by making an analogy to the thermodynamic canonical ensemble [203, 204]. The number of sequences is not important in this method, rather the ratio of frequencies of amino acids at each position on protein sequence [204]. Since the functional residues important for protein folding and enzyme activity belong to the consensus pool [205], using this method does not compromise stability and catalytic activity [201, 205, 206]. Furthermore, thermostabilized proteins that are engineered by this approach also have shown enhanced stability against water-miscible organic solvents and high concentration of kosmotropic and chaotropic salts [96]. However, using this method special care should be taken to detect the best thermostabilizing mutations. Actually, this does not guarantee that all the proposed individual mutations can increase thermostability. Reports indicate that proposed mutations by CC are usually composed of stabilizing, neutral, and destabilizing mutations that eventually counterbalance each other producing an overall stabilizing effect [48, 201, 206, 207]. On the other hand, removing destabilizing mutations has been shown to increase the thermostability [47, 48, 206]. Therefore, it is important to formulate strategies to refine CC results and more precisely detect stabilizing and destabilizing candidates among the proposed set of mutations. However, to the knowledge of the author, no systematic study on the accuracy and validity of CC method in this respect has been reported to date over an extended dataset. For further refinement of CC results and increasing the chance of selecting thermo-stabilizing mutations the rational analysis of the protein structure is one of the most popular approach [50, 198], while there are also few reports taking advantage for this purpose of comparison with thermophilic homologues [208].

Here, we used CC method to find out how accurate it actually is to predict the effect of mutations on stability of proteins. Then, with the help of additional structural information, we studied how the results of traditional CC method can be improved. In addition to structural analyses, we focused only on the subset of thermophilic homologous sequences to investigate if this could increase the accuracy and robustness of predictions done by traditional CC method.

5.2 Materials and Methods

Mutation data were extracted from ProTherm database [51]. All the single mutations with $|\Delta T_m| > 1°C$ and $5.5 < pH < 8.5$ which were studied by thermal denaturation were used. Finally, 363 experimentally studied mutations were extracted, which belong to 11 protein sequences with available structure. The average length of sequences was 118 ranging from 53 to 168 residues. The number of mutations studied in this work is much more extended than previously published reports on CC analysis (which were usually limited to ~10 mutations) [48, 201, 206, 207] and the sequences containing these mutations are somehow well distributed. Information about the studied protein sequences is summarized in Table 5-1.

Table 5-1. Sequence information. Second column shows the number of mutations extracted from ProTherm for the structure and the third column shows the number of wrongly predicted mutation effects by CC method. Column 4 shows the number of sequences used from Pfam. Columns fourth and fifth show the number of thermophilic sequences found from NCBI database respectively.

PDB ID	No. Res.	No. Mut.	No. no	No. Seq. Pfam	No. Thermo.
2LZM	164	46	12	865	1
1CSP	67	13	6	4873	74
1RGG	96	28	10	409	5
1FTG	168	30	15	1834	44
1HFZ	123	29	10	514	81
3SSI	108	13	5	96	0
2RN2	155	31	24	6022	48
1RTB	124	12	2	585	77
1RN1	103	29	3	213	3
1ARR	53	39	8	190	0
1STN	135	96	13	1640	53

5.2.1 Consensus Concept

In order to apply CC method for each protein sequence, first the corresponding family for each sequence was found from the Pfam-A database [209] using the default setting. Pfam-A is a database of protein families that also contains an accurate, manually curated MSA for each family [209]. Using an available and accurate MSA will increase the quality and speed of the analyses rather than performing MSA separately. Then the Prody Evol module [210] was used to fetch and analyze the MSAs. This module can fetch MSA for a protein family from Pfam database and apply different analysis like refinement on the MSA. Then, the MSA for each family was refined as follows: first, all the columns with gaps on the target sequence were removed. Next, sequences with more than

20% gap content were removed from the alignment. In addition, sequences with sequence identity more than 98% were considered as identical sequences. Finally, the distribution of amino acids in each column was calculated and the thermo-stabilizing and destabilizing mutations were predicted following the traditional method as reported in literature [204].

Denoting the amino acid frequency on MSA for the mutant amino acid as f_M and that of the target residue on the sequence as f_T, if $f_M > f_T$ we predicted the mutation to have a stabilizing effect and if $f_M < f_T$ we predicted it to have a destabilizing effect. Then, using the mutations, extracted from ProTherm database, we assessed our prediction results by comparing them with thermal denaturation data (i.e., considering negative ΔT_m as destabilizing and positive ΔT_m as stabilizing mutations). The described workflow for the analysis is illustrated in the right branch of Figure 5-1. Using this method we examined 363 mutations for 11 proteins for which we also have available structural data (see Tables 5-1 and 5-2).

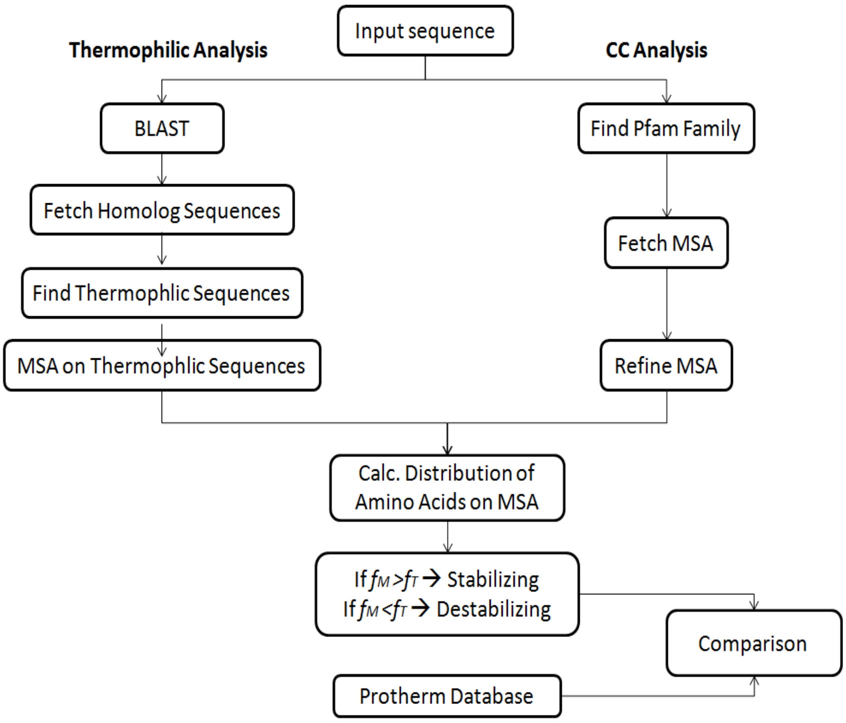

Input sequence
Thermophilic Analysis
CC Analysis
BLAST
Fetch Homolog Sequences
Find Thermophlic Sequences
MSA on Thermophlic Sequences
Find Pfam Family
Fetch MSA
Refine MSA
Calc. Distribution of Amino Acids on MSA
If $f_M > f_T$ → Stabilizing
If $f_M < f_T$ → Destabilizing
Comparison
Protherm Database

Figure 5-1. Process of prediction of thermo-stabilizing and destabilizing mutations by CC analysis and analysis of thermophilic sequences. The traditional CC analysis uses all available homologous sequences in the protein family database while comparative method just analyzes the thermophilic homologous sequences to the target sequence.

5.2.2 Analysis of the Thermophilic Sequences

Using Biopython modules [78], we first performed BLAST [97] with threshold e-value 10^{-4} and then chose homologous sequences from the non-redundant database of NCBI [73]. Next, the sequences that were recognized as homologous were fetched and then, using data implemented in Bioproject section of NCBI, the thermophilic sequences were detected [73]. Then, using the ClustalW2 package [75] MSA was carried out over homologous thermophilic sequences and distribution of amino acid on each column was calculated. Finally, we used the same process as explained above for the CC method, to predict the effect of mutations on stability. The process is illustrated in the left branch of Figure 5-1.

5.3 Results and Discussion

5.3.1 Consensus Concept

Although consensus method has been used in a number of studies [48, 50, 96, 121, 198-201], to the knowledge of the author this study is the first to address a systematic evaluation of this method in addition to providing some additional complementary protocols to improve its performance.

Among the 363 studied mutations, 252 were destabilizing with an average ΔT_m= -7.5 ± 6.3 °C ranging from -38 °C to -1 °C, and 111 were stabilizing mutations with average ΔT_m= 5.2 ±4.2 °C, ranging from 1 °C to 22 °C. Results for the analysis using CC method are represented in Table S5-1. Using the CC method, the effect of 253 (70%) mutations was predicted correctly including 226 true negative (TN) and 27 true positive (TP) predictions and the effect of 110 mutations were predicted wrongly including 26 false positive (FP) and 84 false negative (FN) predictions. Based on this outcome we computed a number of quantities, such as the accuracy, TP rate, TN rate, F1 score, and Mattews correlation coefficient (Table 5-2). Although the predictor showed accuracy as high as 0.7, the TP rate of 0.24 indicates the low efficiency of this method to predict thermostabilizing mutations in the pool of different mutations. This conclusion is supported by a F1 score, which gives higher score to positive predictions (for more details see at Chapter 2), of 0.33, almost half of the calculated accuracy. In addition, the low value of the Mattews correlation coefficient (0.18) points out that this method cannot distinguish between thermo-stabilizing and destabilizing mutations very well. It is noteworthy that value of zero for the Mattews correlation coefficient indicates that the predictor is not better than random. Although the accuracy of the method is somehow high (70%), the composition of the data set that contains 70% destabilizing mutations makes the high accuracy with low Mattews correlation coefficient reasonable. Therefore, we can conclude that the traditional CC method cannot be applied in protein thermostability engineering with high confidence and some further refining step should be added to increase its performance. Although addition of more mutations to the present dataset extracted from the ProTherm database is still possible, the dataset studied here is almost two order of magnitude bigger than the published reports on individual proteins, where the accuracy was calculated based on a limited number of mutations (normally less than 10) on single proteins, situation that precludes to collect enough statistics to probe the robustness of the method [48, 50, 96, 121, 198-201].

Table 5-2. Statistical analysis of different methods. Addition of restraints to the traditional CC method can considerably increase its performance. TP, TN, FN, FP, ACC, TPR, TNR, MCC, and F1 represent true positive, true negative, false negative, false positive, accuracy, true positive rate, true negative rate, Mattews correlation coefficient, and F1 score, respectively. CC, CC_str, CC_therm, and CC_s_therm represent consensus method, CC with structural analysis restraints, CC with thermophilic sequences, and CC dataset for sequences with available thermophilic homologues. CC_s_therm is presented here to make CC method comparable with its extension, as in thermophilic CC only 6 sequences within CC data set had thermophilic homologues.

Dataset	TP	TN	FN	FP	ACC	TPR	TNR	MCC	F1
CC	27	226	84	26	0.70	0.24	0.90	0.18	0.32
CC_str	76	238	35	14	0.87	0.68	0.94	0.67	0.76
CC_therm	28	136	42	2	0.79	0.40	0.99	0.51	0.56
CC_s_therm	16	122	54	16	0.66	0.23	0.88	0.15	0.31

However, to make sure that our engineering protocol does not affect any other functions of the protein/enzyme those residues that are highly conserved in the sequence should not be touched. Thus, a threshold (e.g. 50%) should be set for conserved residues to protect the potential functionally important residues. Interestingly, CC results (Table S5-1) indicate that the majority of highly conserved residues are stabilizing residues as well. But, on the other hand, there are cases, like D134 in PDB ID: 2RN2 (ribonuclease HI from *Escherichia coli*), that are the dominant amino acid in that position (i.e., 98% conserved) but they are the source of instability on the protein.

Accordingly, to increase the predictive power of the traditional CC method and rationalize its failures (like what we reported above for D134 in PDB ID: 2RN2), complementary analyses should be applied. Our studies showed that it is possible to increase the chance of successful predictions using rational concepts, like those coming from (i) structural analysis (Table S5-2) and (ii) homologous sequence analysis of thermophilic species (Table S5-3).

Using the help of structural information, two simple concepts, namely, hydrophobic cores (HB cores) and local charge distribution can explain failures by CC predictions for 61 out of 110 wrong predictions. Charged residue effects can be classified into two forms: salt bridges formation/breaking and accumulation of residues with the same charge on the structure. For example, 24 of 108 wrong predictions belonged to a single protein with PDB ID: 2RN2, and 10 out of these 24 belonged to mutations for a single position, namely D134, the residue that was mentioned above as the origin of instability of the protein, as substitution of it with several other amino acids made it more thermo-stable (see table S5-2), although it was highly conserved (98%). Interestingly, this residue is reported to lie close to the active site of the protein and the site-directed random mutagenesis confirmed that any mutation on this site, but D134H, disturbs the activity of the ribonuclease HI protein [211] (Table S5-1). Within a cutoff distance (R_c) of 8 Å from D134 residue, there are seven negatively charged residues while there are only two positively charged residues (Figure 5-2A). This shows that the region is negatively charged and any replacement with a non charged residue or a positively charged residue can alter the charge distribution and likely increase the stability. Results shown in Table S5-2 confirm this conclusion. Figure 5-3 illustrates this effect for the 37 mutations shown in Table S5-2 with local charge modification effect.

Replacing D134 with positively charged residues could modify the charge distribution and simultaneously could make salt bridges with neighboring negatively charged resides (Figure 5-2-A). However, such a highly conserved residue cannot be a target for mutation as it is a potential functionally important residue, like what we reported here for D134 on ribonuclease HI, except in the case there are enough evidences for a specific purpose like industrial or lab applications where this conserved residue is not playing a major role.

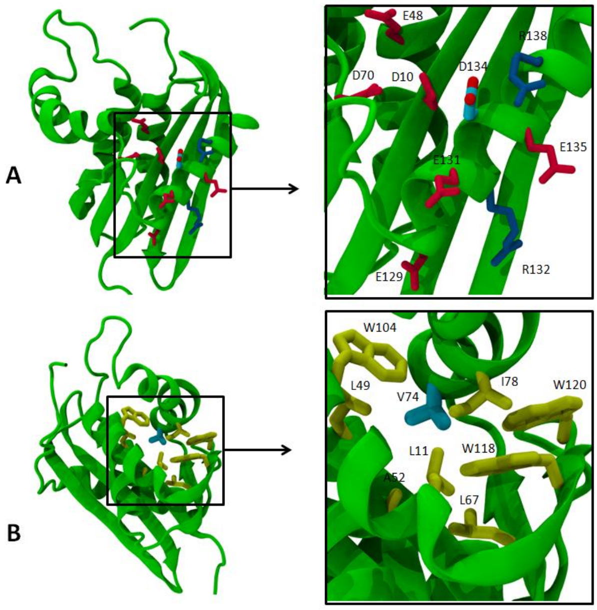


Figure 5-2. Two important residues on protein with PDB ID:2RN2. A) Reside D134 is surrounded by mainly negatively charged residues. B) V74 belongs to a HB core.

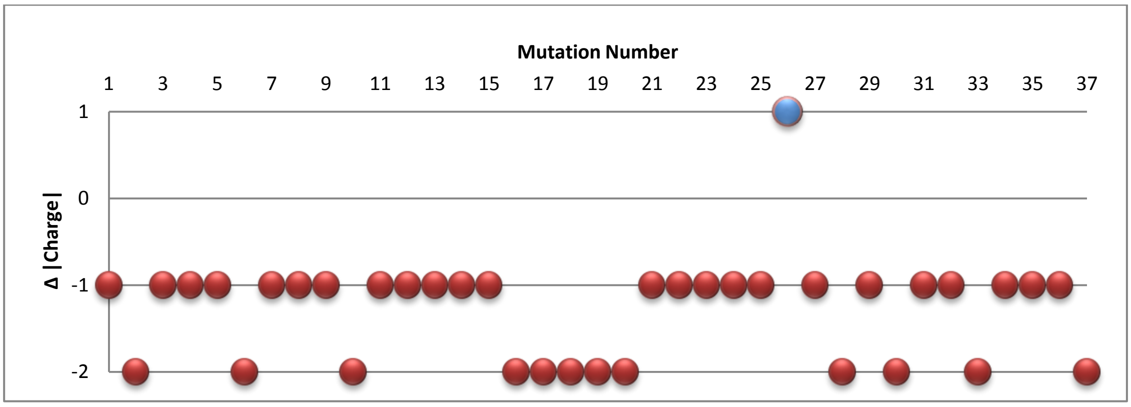


Figure 5-3. Illustration of structure base analysis summarized in Table S5-2 for charge modification effect of mutations on thermostability (data numbers from 1 to 37 in Table S5-2). This figure only illustrates the data that shows the effect of local charge distribution change of mutations on thermostability. For each mutation the net change in absolute local charge is calculated. Red and blue show thermo-stabilizing and destabilizing effect of the mutations respectively. Mutations that reduce the charge accumulation on protein structure increase its thermostability. For example for mutation number 2 (D134H), the absolute net charge decreased from 5 to 3 and for mutation number 1 (D134Q) or 3 (D134A) the absolute net charge decreases from 5 to 4.

Besides D134, three wrong predictions belong to V74 on 2RN2 structure. All three mutations are elements of a HB core for which the rules discussed for HB-cores in chapter 4 (that follow G<A<V<L,I order of mutation to enhance thermostability) should be taken into account (Figure 5-2-B). For the G<A<V<L,I affinity, the corresponding mutations in Table S5-2 are illustrated in Figure 5-4 (mutation numbers from 38 to 53 in Table S5-2). Thus, using the described structural analysis not only accuracy increased from 0.7 to 0.87, but also TP rate increased significantly up to 0.68 that together with increased F1 score (0.76 compared to 0.32 for traditional CC method) represents significant improvement of its thermo-stabilizing prediction power (Table 5-2). In addition, Mattews correlation coefficient of 0.67 indicates the enhanced power of distinguishing between thermo-stabilizing mutations compared to that of traditional CC method (which had Mattews correlation coefficient of 0.18).

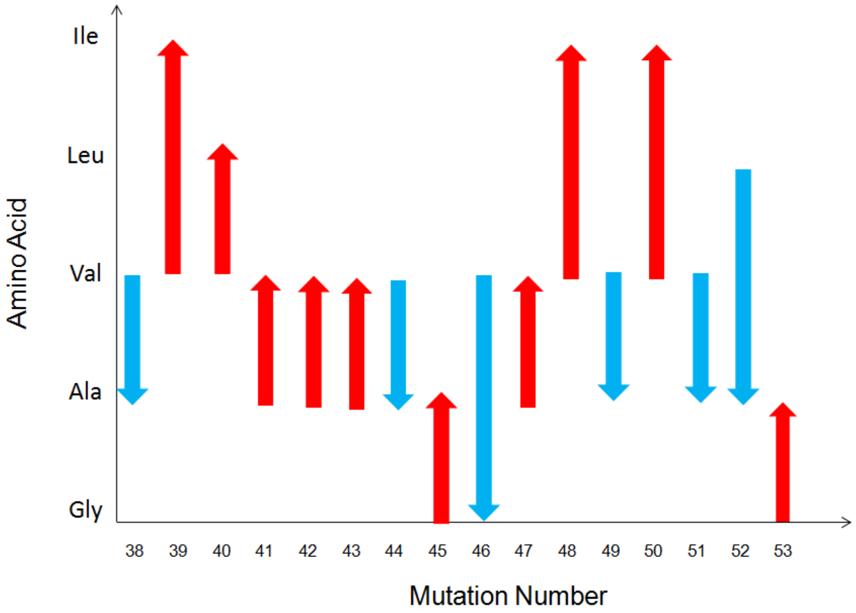


Figure 5-4. Illustration of structure base analysis summarized in Table S5-2 for the G<A<V<L, I rule (data numbers from 38 to 53 in Table S5-2). The directions of arrows show the mutation direction from the wild type to the mutant. Red and blue indicated thermo-stabilizing and destabilizing effect of the mutations, respectively. The numbering of mutations follows the Table S5-2.

However, keeping in mind that for the vast majority of interesting sequences for thermostability engineering no 3D structure is available, we need to devise protocols that can extract the used structural data, HB interactions and local distribution of charged residues, only from the sequence data. To take advantage of what we learnt, homology modeling and evolutionary coupling can be used to build accurate enough models. Specifically, as HB core members are fundamental for protein folding and stabilizing factors, they are expected to be conserved during the evolution in the protein families and then a good accuracy of detection of HB interactions is expected by these methods.

5.3.2 *Analysis of Thermophilic Sequences*

The sequences that are used in traditional CC method belong to organisms with different thermostability criteria and the only obligation is that they belong to the same protein family. Such diversity in sequences causes diversity and noise in amino acid distribution on MSA with respect thermostability properties. Furthermore, the vast majority of sequences that one can fetch for MSA are likely mesophilic sequences, while the aim of thermostability engineering is producing thermophilic or hyperthermophilic proteins. Therefore, limiting the analysis of the distribution of amino acid on a MSA to only thermophilic/hyperthermophilic sequences should provide more accurate and reliable results by revealing sites and substitutions that are chosen by nature during the evolution to boost thermostability. The main limitation of this approach is the limited number of sequenced thermophilic genomes.

Therefore, as an extension of CC method, we repeated the same procedure as before but only using thermophilic sequences to study the effect of this filtering on the accuracy of the predictions. Table 5-1 shows the number of the found thermophilic sequences for each sequence. Among them, there were only 6 sequences with more than 10 homologous thermophilic sequences. Focusing only on thermophilic sequences not only covered correctly predicted mutation effects on thermostability by traditional CC method (except three mutations, namely: L110R on PDB ID: 1HFZ, A46K on PDB ID: 1CSP, and G68A on PDB ID: 1FTG), but also it corrected the effect of mutations for 26 of the wrong predictions, without using any structural information (Table S5-3). Interestingly, among the 26 mutations, 12 were thermostabilizing (TP) that shows the power of using thermophilic sequences for the aim of thermostability engineering. Overall, using thermophilic sequences to aid traditional CC, 208 mutations were studies and the effect of 164 mutations was predicted correctly. Among them, 28 were TP and 136 were TN. For the wrong predictions, 2 were FP and 42 were FN (Table 5-2). Therefore, using thermophilic sequences analysis although the accuracy of prediction only increased by 20% (from 0.66 to 0.79), TP rate and F1 score, which show the prediction power of finding thermo-stabilizing mutations, increased by 74% (from 0.23 to 0.40) and 81% (from 0.31 to 0.56), respectively, compared to that of traditional CC method (Table 5-2). These changes are consistent with the estimated increase in Mattews correlation coefficient, from 0.15 to 0.51 (Table 5-2).

Overall, while improving the CC performance, this approach might suffer from shortage of thermophilic homologous sequences. Even for protein families with available thermophilic sequences, the number of available thermophilic sequences would still limited. Nevertheless, the principle behind this analysis is still promising: sequences belonging to species with higher thermostability have evolved to have higher thermostability. Therefore, comparing a sequence with its more temperature resistant homologues can be very useful to learn about the strategies taken by nature to optimize the protein even if only a few sequences are available.

5.4 Conclusions

In conclusion, we studied the performance of traditional CC method over a big dataset for the first time and clarified its weakness in detection of thermostabilizing mutations. Then, taking advantage of structural data and analysis of homologous thermophilic sequences we suggested two strategies to cover the weakness of the CC method and showed that these strategies work quite accurately. All together, although the traditional CC method, as it is, is weak for the purpose of protein thermostability engineering, its power in preserving functional residues, that cannot be detected using structure based methods, still keeps it as an attractive method. Specifically, if no structural data are available for the protein, sequence data would be the most reliable data, or even the only source for the aim of thermostability engineering. Therefore, appropriate strategies should be designed to aid this method as much as possible. Our analyses showed that beyond the limited availability of thermophilic sequences, they can help enhancing the predictive power of traditional CC. In addition, what we learned from coupling structural data with CC method can be applied on traditional CC as constraints even if there is no complete structure available for a target sequence. In fact, homology modeling and MSA analysis methods, specifically recently developed evolutionary coupling tools [212], can partially provide the needed structural information.

Chapter 6 A General Framework for Protein Thermostability Engineering

6.1 Introduction

Through chapters 3, 4, and 5 I showed that using protein sequence and structure analyses we can predict the effect of mutations on protein thermostability and find appropriate mutations that can lead to enhancement of protein thermostability. In addition to the specific enzyme (*LigTh1519*) on which we successfully applied our initial thermostability engineering protocols (see Chapter 3) different methods were used and evaluated based on published experimental results, as reported in Chapters 4 and 5. Specifically, in Chapter 5 I showed that a combination of sequence and structure analyses, even using simple concepts, can improve the prediction power significantly. Here, following what we learned from a case study, and a combination of the methods previously explained and evaluated, I propose a general rational engineering protocol for protein thermostability engineering. The aim of the protocol is to propose a minimum number of thermo-stabilizing mutations which would maximize ΔT_m with low risk to compromise stability and activity.

6.2 Materials and Methods

6.2.1 Details of the Protocol

This protocol is mainly designed based on a combination of different strategies discussed in Chapters 4 and 5, namely iterative GNM unfolding simulation, HB interactions, consensus concept, and comparative sequence analysis.

By applying this protocol, one of the benefits is that one can identify *forbidden mutations* and *beneficial mutations*. Forbidden mutations are those changes on the sequence, which are highly risky to engineer, because of their possible negative impact on protein thermostability and function. On the other hand, beneficial mutations on the sequence are expected to increase the protein thermostability. In other words, in this protocol, we aim at filtering out first risky substitutions and among the remaining positions on the protein, we aim at finding the best candidates for thermostability enhancement.

Here we call the positions on protein that should not be mutated as "Do Not Touch (DNT)" residues. Thus, the DNT list for each protein is composed of three set of residues: (i) residues that are already forming a salt-bridge, (ii) I or L residues involved in HB interactions, and (iii) residues that are highly conserved on the MSA with occupancy higher than a critical value (i.e., default = 50%). Mutation of such residues may result in instability of the protein by disturbing salt-bridges and HB interactions, or may jeopardize protein function provided by highly conserved residues (see chapters 4 and 5).

In addition, some mutations are detected as forbidden mutations. This means that the residue is not forbidden to be mutated (i.e., it does not belong to the DNT list), but there are certain substitutions that should be removed from the mutation library. This forbidden list of mutations is mainly built following HB rules, as shown in Chapter 4, and analysis of the global electrostatic charge of neighboring residues, as discussed in Chapter 5, following the simple rule that if the net charge of neighboring residues is not zero, mutations to residues with the same charge sign are considered as forbidden mutations, because they can create unfavorable electrostatic interactions. Obviously, if the wild type residue has the unfavorable charge, its mutation to an opposite or uncharged amino acid will have a thermostabilizing effect (see results and discussion in Chapter 5).

On the other hand, the protocol also suggests potential thermo-stabilizing candidates (beneficial mutations) for a set of positions on the protein. Here, the thermo-stabilizing mutations are classified as (1) HB mutations, (2) hydrophobic-hydrophilic (HB-HL) switch mutations, (3) charged mutations, (4) hot-spot mutations, (5) salt bridge forming mutations, and (6) consensus mutations.

1. *HB mutations* are proposed based on three HB interaction rules as outlined in Chapter 4, namely the following order of substitutions: G < A < V < I, L is expected to increase protein thermostability.

2. *HB-HL switch mutations* are proposed if the target residue is HB and the majority (more than a critical proportion, here 50% as default) of amino acids on the thermophilic MSA are HL, or vice versa. The mutant can be chosen among one of the most frequent amino acids on the thermophilic MSA. This analysis is very important because it targets one of the most difficult mutation effect. Because, removing, for example, a HB residue and replacing it with a HL amino acid will have a stabilizing effect only if the new HL residue will form more energetically favorable interactions with its neighboring residues compared to the native HB interactions. It is noteworthy that this strategy has been initially used successfully for a DNA ligase (*LigTh1519)* as discussed in Chapter 3, and here we aim at implementing it in a general protocol. In addition, we showed there that even using energy force fields, like FoldX [60], it is very difficult to find these mutations. However, bioinformatics analysis, which compares the sequence with its homologues having higher thermostability, could easily find the appropriate candidates for this type of substitution (see Chapter 3).

3*. Charged mutations* are proposed to improve the charge distribution on the protein structure. If the net charge of the neighboring residues for the target residue is not zero, the target residue is not one on the DNT list, and both consensus MSA and thermophilic MSA contain a favorable charged amino acid with higher frequency than the target residue, a favorable charged amino acid will be suggested for mutation. For example, a basic amino acid for negatively charged neighborhood will be proposed. The charge of the neighborhood is calculated by summation of charge of all residues with Cα closer than a critical distance (here we used 8 Å as default) to Cα of the target residue.

4. Hot-spot mutations are proposed for non DNT residues located on weak points (residues located on the first X% Loss-Number-of-Native-Contact (LNNC), where for this case X is set to 6% as default) with non zero neighboring net charge and at least one counterpart with favorable charge on thermophilic MSA with frequency higher than that of the target residue. The weak point analysis is performed as illustrated in Chapter 4 using Iterative GNM. One of these thermophilic MSA amino acids is proposed as a mutation that can fix the thermo-sensitive position on the structure.

5. Salt bridge forming mutations are proposed for two adjacent (Cα- Cα distance less than 8 Å) non-charged and non-DNT residues *i* and *j* if |*i-j*|>*5* and at least one of the residues are located on weak points with net charge of zero for its neighborhood. They should be more than 5 residues apart to increase the probability of intermediate or long range interaction to keep structural elements tightly interacting together. Even if one of the two residues is not located on flexible region, the formation of salt-bridge can fix the flexible point.

6. Consensus mutations are amino acids with frequency higher than that of the target residue in MSA of proteins in the family (for more details see Chapter 5) and not belonging to the DNT list. As shown in Chapter 5, the performance of this method to find stabilizing mutations is not high. Then, it is reasonable to use these lists only if the other methods could not find the needed number of mutations. Please note that these lists are used in combination with other structural analyses like in case of "charged mutations".

The work flow of the protocol is shown in Figure 6-1.

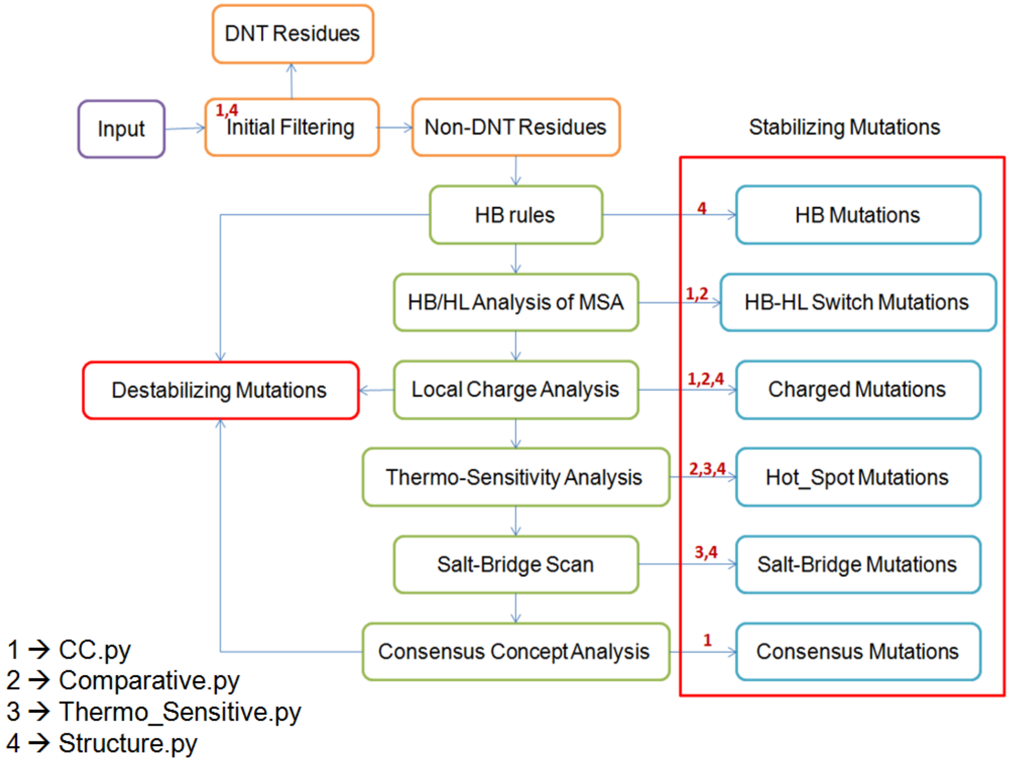


Figure 6-1. Workflow of the thermostability engineering protocol. For an input protein, the protocol filters-out first some residues as do not touch residues (DNT) and for the rest of residues predicts a set of thermo-stabilizing or destabilizing mutations. Four Python modules were developed and used within this workflow: *1) CC.py*, *2) Comparative.py*, *3)Thermo_Sensitive.py,* and *4) Structure.py, see text for details.*

6.2.2 Dataset and Protein Sequences

Proteins with available crystal structure in RCSB protein data bank [213], thermophilic homologues from NCBI database [73], MSA from Pfam database [209], and experimental data on ProTherm database [51] were chosen in this study to assess the accuracy and performance of the protocol. Thus, 19 protein structures (Table 6-1) were selected and analyzed for protein thermostability engineering, for which information was available on each of these databases. The average sequence length of the pool was 171, ranging from 62 to 570, residues. Then, to validate the results of the protein thermostability protocol, ProTherm database [51] was used. The stabilizing and destabilizing mutations detected by the protocol were searched within the ProTherm database to find corresponding experimental values for ΔT_m. Finally, we compared our protocol with fast and popular predictors like I-Mutant2.0 [79] and AUTO-MUTE [80] servers to predict the effect of the detected thermo-stabilizing and destabilizing mutations on the stability of protein structures. They can both conduct structure-based prediction for thermostability.

The described pipeline is implemented in a Python library that gets as input a protein sequence (in FASTA format) for sequence based analyses, or a structure (in PDB format) for structure based analyses. The sequence based analyses toolkit is composed of two main modules called *CC.py*, for consensus concept analysis and *Comparative.py*, for thermophilic sequence analysis. On the other hand, the structure analyses toolkit is composed of the *Thermo_Sensitive.py* and *Structure.py* modules. *Thermo_Sensitive.py* module applies the thermo-sensitive detection analyses, as discussed in Chapter 4, using elastic networks models and the *Structure.py* module performs the structure analyses including detection of HB interactions, finding salt-bridges and other general structural features.

Table 6-1. Results for the 19 studied structures. Each column represents data as follows: L: length of the sequence, DNT: number of DNTs, HB_M: number of HB mutations, HB_F: number of HB forbidden mutations, HB_HL: number of HB_HL switch mutations, CH: number of charged mutations, Hot: number of hot-spot mutations, CH_F: number of charged forbidden mutations, CC: number of consensus mutations, Salt: number of proposed salt-bridge making pairs, Ncc: number of sequences in the consensus analysis, Nth: number of thermophilic sequences. These results are obtained using parameters described in the methods section (described as default values in the text).

PDB_I	L	DN	HB_	HB_	HB_H	CH_	CH_	Ho	CC	Sal	Ncc	Nth
1AZP	66	54	14	18	0	0	7	0	8	0	10	4
1CSP	67	34	17	21	6	3	26	1	20	1	487	63
1DIL	38	92	81	87	85	1	228	3	3	40	307	32
1FTG	16	59	35	31	68	4	87	11	99	3	183	44
1H8V	11	52	56	56	45	1	95	2	69	62	336	7
1HFZ	12	53	16	19	63	3	53	14	59	7	514	59
1JU3	57	164	160	188	120	6	290	13	14	17	846	12
1N0J	19	71	45	50	27	5	95	3	36	7	265	12
1RGG	96	38	19	20	19	6	50	2	34	7	409	4
1RN1	10	26	25	23	22	0	62	0	52	5	213	2
1RTB	12	50	20	29	49	1	60	5	33	3	585	59
1SSO	62	53	14	14	4	1	7	1	7	0	10	4
1STN	13	47	30	35	16	8	75	2	66	0	164	57
1WQ5	25	116	71	87	46	6	105	4	13	3	206	71
2AFG	12	60	21	21	27	8	66	8	62	2	566	11
2CBR	13	43	26	34	20	3	69	2	73	18	134	86
2RN2	15	62	35	37	17	6	76	3	10	4	602	44
3SSI	10	49	39	40	19	0	38	0	28	1	96	4
4BLM	25	92	53	63	22	6	126	4	21	4	444	48

6.3 Results and Discussion

The number of DNT residues, forbidden and proposed mutations for each protein are listed in the Table 6-1. The predicted thermo-stabilizing and destabilizing mutations were compared with the data in the ProTherm database and comparative analysis is reported in Table S6-1. Thus, conserved DNT residues were filtered out at the initial step and no thermostbility analysis was applied for them because they are supposed to be functionally important residues and should not be subjected to mutation. However, the rest of DNTs, salt-bridge forming residues and I/L residues on HB interactions were considered as destabilizing mutations and reported in Table S6-1.

On average, 43% of residues on the sequences in our benchmark set were detected as DNT residues. There are two extreme cases for proteins 1AZP and 1SSO, which have more than 80% residues in the DNT list. For both these structures, the majority of DNTs are highly conserved residues. This is likely due to the fact that when the MSA contains only a few sequences (for example 10 sequences) with limited diversity of species, the distribution of amino acids will not be diverse, and thus the majority of residues on the sequence will be recognized as conserved residues. Therefore, for proteins with only few sequences available in the MSA, the conserved DNTs should be considered carefully and more information about functionally important residues should be obtained from literature. Another way to decrease the number of DNTs for such proteins is to increase the threshold value of acceptance (here 50% is used) for considering a residue as a potentially functional residue.

On the other hand, for the remaining residues on the target proteins a significant number of thermo-stabilizing mutations are proposed. The average percentage of residues proposed to be mutated based on HB rules was 24%. All the HB thermo-stabilizing mutations were detected based on G < A < V < I, L substitution rule, while the other rules can only detect thermo-destabilizing mutations. The rest of the suggested mutations were produced, respectively, by rules concerning HB_HL switch mutations (20%), charged mutations (2%), hot-spot mutations (2%), salt-bridge forming mutations (6%), and consensus mutations (40%). Such low prediction rate for charged mutations, hot-spot mutations, and salt-bridge forming pairs has to be accounted for the several constraints applied to the protocol (see Methods section). For the results based on consensus rules, the number of proposed mutations is strongly dependent on the diversity of the sequences used in the family. If a family contains members with a very diverse range of sequence identities then a large number of proposed mutations are expected. In addition, It is noteworthy that residues on the protein sequences are shown to co-vary during evolution in the protein family [214] . This kind of co-variation cannot be detected using traditional consensus concept techniques. Then for proteins with high diversity in homologues sequences one can use co-variation analysis to make the proposed mutation list via consensus concept shorter [214].

To assess the performance of this protocol, the predicted stabilizing and destabilizing mutations were compared with the data present in the ProTherm database, namely the effect of experimental mutations (i.e., ΔT_m values) were compared with results of our protocol. For the 19 structures in our dataset, we could find experimental reports for 146 mutations, among which the effects of 130 mutations (89%) were predicted by our protocol correctly. Comparing with experimental results from the ProTherm database the accuracy and F1 score were calculated as high as 0.89 and 0.84, respectively. Interestingly, both TP and TN rates (~0.9, Table 6-2) showed very high accuracy for prediction of thermo-stabilizing and destabilizing mutations, reflected also by the high value of the Mattews correlation coefficient (0.76). (For more information about the meanings of statistical analyses see Chapter 2).

In Table S6-1, the predictions done using other software, like I-Mutant2.0, and AUTO-MUTE, are reported along with the comparison with experimental values found in ProTherm data base. For both the tools the protein 3D structures were used as input to be consistent with our protocol. In particular, AUTO-MUTE could not perform the calculations for structures with PDB ID: 1HFZ, 1RN1, 1WQ5, and 4BLM (the reason was unknown to us). For consistency, all statistical analyses were repeated over predictions made by our protocol for the same data set as for AUTO-MUTE (without considering mutations corresponding to structures with PDB ID: 1HFZ, 1RN1, 1WQ5, and 4BLM). The correspondent statistical analyses for I-Mutant2.0 and AUTO-MUTE prediction with the same experimental values are also summarized in Table 6-2.

I-Mutant2.0 predictions were considerably weaker than our protocol. The low Mattews correlation coefficient and TP rate indicated that although I-Mutant2.0 showed an accuracy as high as 73%, it was not successful in distinguishing between thermo-stabilizing and destabilizing mutations very well, similarly to what we discussed for consensus method in Chapter 5 (see section Results and Discussion). Specifically, the TP rate and F1 score were considerably lower than that of our protocol (0.33 and 0.43 vs. 0.89 and 0.84) confirming its low efficiency in detection of thermo-stabilizing mutations. On the other hand, the AUTO_MUTE predictor showed better results compared to I-Mutant2.0, but still not as good as our protocol. Our protocol can predict thermo-stabilizing mutations as shown by the F1 score (0.85 vs. 0.73) and TP rate (0.86 vs. 0.67), and also showed an improved ability to distinguish between thermo-stabilizing and de-stabilizing mutations as confirmed by the Mattews correlation coefficient (0.77 vs. 0.62).

In addition to a better statistical performance, it should be highlighted that our protocol provides a more rational based approach, thus that for each mutation it gives an explanation, like the origin of stabilizing/destabilizing effect, that makes it more valuable for engineering purpose. For example, the user can see if a specific mutation is enhancing thermostability because of the optimization of HB interactions, or if this is thermo-destabilizing because it causes salt-bridge breaking, etc. However, no matter how accurate machine learning-based tools like I-Mutant2.0 and AUTO_MUTE can be, they do not provide any rational details about the origin of the stability/instability of the predicted mutations.

Table 6-2. Statistical analysis of different methods. TP= true positive, TN = true negative and so on , FN= false negative, FP= false positive, ACC= accuracy, TPR= true positive rate, TNR= true negative rate, MCC= Mattews correlation coefficient, and F1= F1 score. As the AUTO-MUTE did not work for structures with PDB ID: 1HFZ, 1RN1, 1WQ5, and 4BLM, for consistency, all statistical analyses of our pipeline , shown in first row, was repeated without considering results of the mentioned structures and shown in Data_automut_sub row.

Predictor	TP	TN	FN	FP	ACC	TPR	TNR	MCC	F1
Our protocol	41	89	5	11	0.89	0.89	0.89	0.76	0.84
I-Mutatnt2.0	15	91	31	9	0.73	0.33	0.91	0.30	0.43
AUTO_MUT	24	68	12	6	0.84	0.67	0.92	0.62	0.73
Data_automut_sub	31	68	5	6	0.90	0.86	0.92	0.77	0.85

Along with the statistical analysis of the results of our framework, it is also very interesting to look at the relative ΔT_m distribution of thermo-stabilizing mutations suggested by our protocol. Within the large number of random mutagenesis experiments for protein thermostability studies, the majority of the reported single mutations are thermo-destabilizing and the thermo-stabilizing ones usually increase the T_m less than 2°C, while single mutations resulting in ΔT_m higher than 4°C are not commonly observed [208]. Analyzing the available thermo-stabilizing single mutations in the current version of ProTherm database with

5.5 <pH< 8.5, we found that the total number of thermo-stabilizing mutations was 614 with an average ΔT_m=4.4 ± 6.5℃. 46% of them had ΔT_m < 2℃ and less than 35% of them had ΔT_m >4℃. It is noteworthy that 67% of the single mutations in ProTherm database had ΔT_m < 0℃ (Figure 6-2).

Strikingly, the average ΔT_m of the correctly suggested thermostabilizing mutations predicted by our protocol was 5.5 ± 0.9 ℃ and for the destabilizing mutations it was -8.3 ± 5.2 ℃. Figure 6-3 illustrate the distribution of ΔT_m of the detected thermo-stabilizing and destabilizing mutations by our protocol. The majority of the thermo-stabilizing mutations are in the range of 2 and 8℃ (63%) with highest pick between 4 and 6℃ (24%) (see Figure 6-3), indicating that not only the accuracy of our method is high, but it provides higher changes to predict mutations that can more significantly improve the thermostability of the target protein. On the other hand, for the destabilizing mutations, the highest ΔT_m pick occurs between -6 and -4℃ (20%) and the majority (54%) of the detected destabilizing mutations had ΔT_m less than -6℃ (Figure 6-3), showing that the protocol can predict the most important unfavorable mutations and subsequently filters them out from the mutation library.

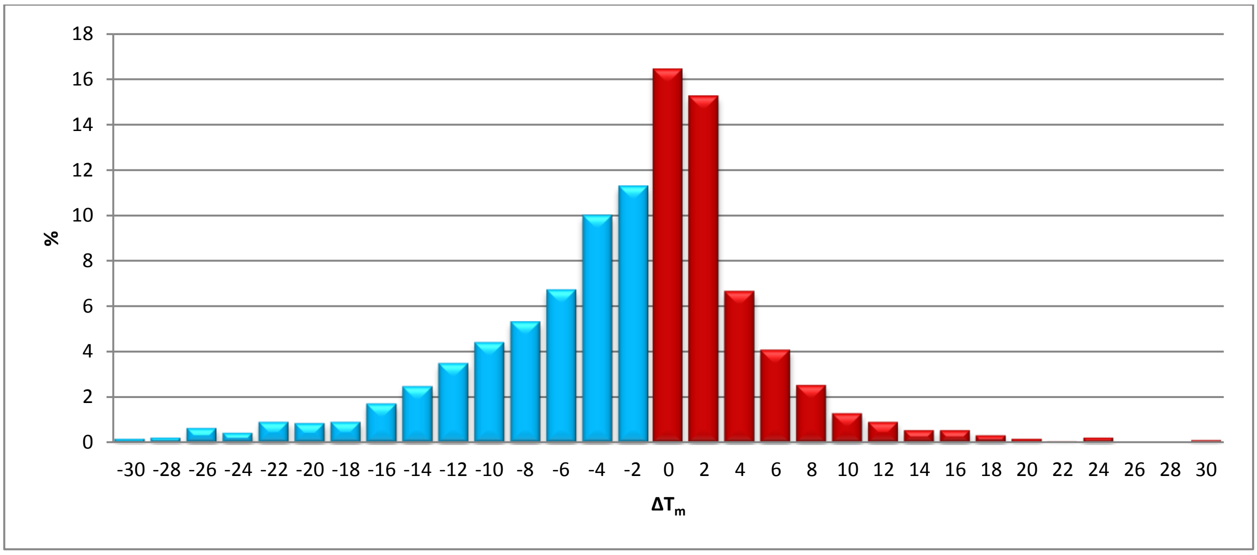


Figure 6-2. Distribution of the thermo-stabilizing (red)/destabilizing (blue) single mutations reported in ProTherm database. The shown distribution for ΔT_m contains 99% of the data.

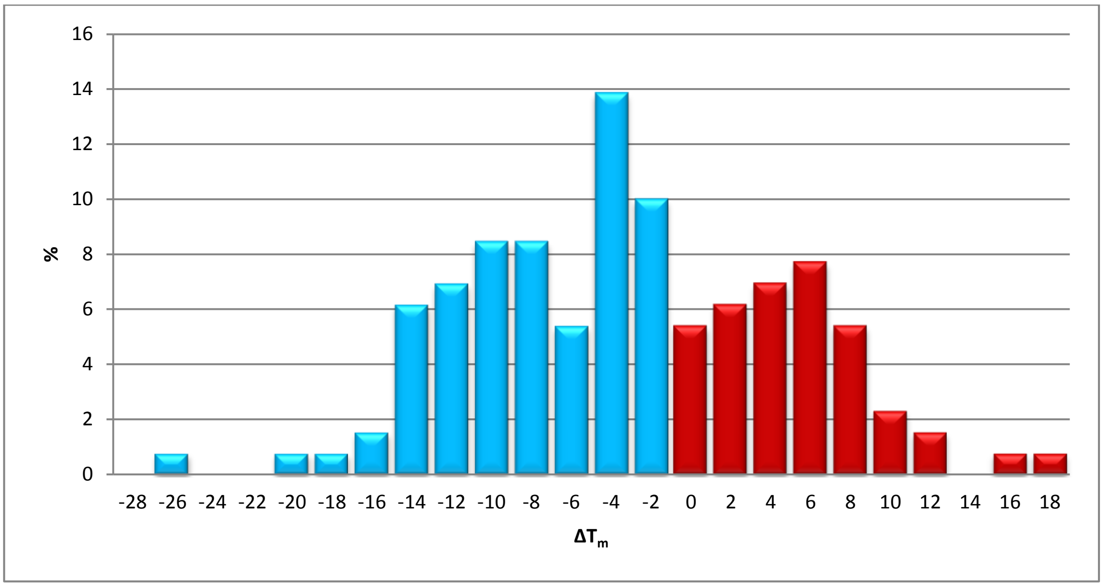


Figure 6-3. Distribution of ΔT_m of the detected thermo-stabilizing (red) and destabilizing (blue) mutations as detected by our protocol.

6.4 Conclusions

In conclusion, the protocol that I developed was able to predict mutations with significant thermo-stabilizing effect in addition to filtering out unfavorable destabilizing mutations. The prediction power of this protocol is not limited to specific protein families and it showed very promising results for 19 proteins with high diversity in length and coming from different protein families. This protocol also showed greater accuracy in selecting thermo-stabilizing mutations compared to other popular methods. In addition, as the point mutation effects on protein thermostability are shown to be independent and additive [52], combination of single mutations proposed by this framework can result in a synergistic increment in ΔT_m. Finally, it should be considered that even for proteins without available 3D structures, the majority of structural data, which are needed for the protocol, like HB interactions and salt-bridge networks, could be in principle estimated using homology modeling or recently developed evolutionary coupling methods [212].

Chapter 7 Conclusions and Perspectives

Using proteins isolated from microorganisms living in non-standard temperature conditions has become nowadays an attractive strategy for industrial and biotechnological applications, in order to achieve better yield in chemical reactions and developing new technologies. However, the optimal conditions for industrial applications of proteins, especially enzymes, still differ significantly from the cellular conditions. Specifically, the favourable conditions for industrial applications are usually considered quite harsh conditions for any living cells, like elevated temperature, presence of organic solvents and high concentration of substrate/product. Therefore, the task of protein engineering techniques is quite challenging aiming at optimizing proteins for adverse conditions with the lowest number of mutations, while preserving their original function. For this purpose, *in silico* methods have provided a significant contribution in designing efficient mutation libraries by rationalizing strategies taken by nature to enhance thermostability of proteins during evolution. Although several reports have addressed the problem using *in silico* analyses of protein sequence and structure, providing fast and accurate protocols that can enhance thermostability remains still challenging.

In this thesis, I first used computational techniques to deeper understand the origin of thermostability in the *LigTh1519* DNA ligase, a thermophilic enzyme. I engineered it to achieve higher thermal stability using a rational design approach. Results showed that an extended salt bridge network plays a crucial role in stability preservation of *LigTh1519* at high temperatures, i.e. over around 70 °C, by keeping the thermo-sensitive region tightly connected to the enzyme body. Based on this careful analysis, I tested four mutations, namely A287K, G304D, S364I and A387K, predicted to enhance the thermostability of the enzyme. Structural and dynamic analyses showed that A287K, G304D, A387K mutations play a role in the electrostatic optimization of the surface, while S364I contributes to optimize the hydrophobic core of the enzyme. Experimental results proved that the proposed substitutions enhanced indeed thermostability of the protein by increasing in an additive fashion the half-inactivation time at 94°C from 8 to 41 minutes without significant reduction of specific catalytic activity. The findings reported here emphasize on the importance and capabilities of molecular modelling and simulations combined with bioinformatics analyses in revealing stability/instability sources and predicting mutations able to enhance thermostability without affecting enzymatic activity. Therefore, the engineered hyperthermostable DNA ligase derived in this work could be used to optimize further PCR protocols.

It is important to stress the fact that this was a particular harsh case for computational methods, since I was able to further stabilize and significantly increase the half-inactivation time of an enzyme that was already thermophilic. This is hinting to promising future applications of a similar protocol to produce thermostable enzymes at the hyperthermophilic level. The approach presented here and applied to *LigTh1519* could be in fact similarly adopted for other enzymes and can contribute to systematically produce more hyperthermostable enzymes for the industrial setting.

Following what I learned from the study on *LigTh1519*, I decided to implement this knowledge in a protein thermostability engineering protocol and a corresponding toolbox that can be used for a broad array of proteins. I first devised a strategy that is alternative to MD simulations to detect thermo-sensitive regions. Thermo-sensitive regions are flexible/weak regions on protein structures that will likely initiate the denaturation upon temperature increase. I showed that this approach could detect regions on the protein structure that have high potential to enhance thermostability if they are mutated appropriately. In addition, following structure-based methods, I introduced simple rules that can predict thermo-stabilizing and destabilizing mutations accurately for hydrophobic interacting residues. I showed that the selection rules following the Gly<Ala<Val<Leu,Ile scale for HB mutations is very accurate to find thermo-stabilizing mutations and, following the opposite order, disruption of hydrophobic interactions by hydrophilic mutations can destabilize the protein. Then, I examined the most popular sequence-based protein thermostability engineering method, *consensus con-*

cept, and, testing it over a large dataset, I showed its weakness in prediction of thermo-stabilizing mutations, and suggested efficient strategies to cure these drawbacks. Finally, I put all these pieces together and built a thermostability engineering protocol with its corresponding toolbox, and showed that it is accurate in prediction of thermo-stabilizing and destabilizing mutations for different protein families. As the point mutation contribution on protein thermostability has been shown to be independent and additive [52], combination of single mutations proposed by this framework can result in promising thermostability enhancement of target proteins.

Although the framework presented in this thesis is promising in terms of robustness and accuracy, it has still obvious limitations, which provide large room for improvements. The future development can be conducted in different ways: First, additional methods can be added to the protocol to improve the performance. For example, despite all advantages, my proposed method for finding thermo-sensitive regions on protein structures is unable to propose the most energetically favourable mutation to secure the detected thermo-sensitive regions. To solve this problem, force field based methods can be added to the framework to have energetic estimations upon mutations. But, as they are generally still computationally costly, fast methods like FoldX [60] can be a valid alternative. On the other hand, although FoldX can predict changes in protein stability for single and multiple mutations quickly, its accuracy appears to be still limited [61].

In addition, although HB interaction engineering was shown to be very accurate, the lack of 3D structures for some proteins still represents a limitation. To take advantage of my findings about HB interaction engineering, one can use recent evolutionary coupling analyses [212] to find additional structural information from sequence. However, for such approach to be applied a comprehensive analysis is needed to address the robustness of evolutionary coupling methods in detection of HB interactions. The same strategy can be applied to find cluster of residues producing salt-bridge networks.

Finally, another limitation of the present protocol is that it can only predict potential thermo-stabilizing or destabilizing mutations but cannot predict the resultant change in T_m upon the suggested mutations. Coupling my protocol with machine learning techniques can be a step forward to have an accurate tool that can predict the most probable thermo-stabilizing mutations, having maximum increase in T_m, with the lowest risk to affect function. In addition, it is noteworthy that all the described methods and the protocols are implemented in a Python library that could manage all the computational steps automatically.

In conclusion, in this thesis I focused on the most popular sequence and structure based thermostability engineering methods and examined their efficiency. I developed faster alternatives to time-consuming methods and suggested complementary analyses to enhance their robustness. Finally, I proposed and tested an accurate protein thermostability engineering protocol that can be used for different protein families. In addition to its application to thermostability engineering, it could also be used for mutagenesis experiments to predict if the target mutation will have a stabilizing or destabilizing effect, helping to rationally design new experiments and produce more advanced working hypotheses. As a complementary application that I did not address in this thesis, the framework can also be used to work in the opposite direction, namely by engineering mesophilic variants of thermophilic protein targets. In fact, examples of thermo-destabilization have been recently shown for some proteins, like I-DmoI from archaeon *Desulfurococcus mobilis* [215]. However, as the number of destabilizing mutations can be significantly higher than stabilizing ones and they could cause more serious, uncontrolled damages on protein structure and function, adapting the protocol for applications of this kind would require the development of new strategies for the selection of appropriate mutations.

References

1. Ventura, S. and L. Serrano, *Designing proteins from the inside out.* Proteins-Structure Function and Bioinformatics, 2004. **56**(1): p. 1-10.
2. Adams, M.W.W. and R.M. Kelly, *Biocatalysis at Extreme Temperatures - Enzyme-Systems near and above 100-Degrees-C - Preface.* Biocatalysis at Extreme Temperatures, 1992. **498**: p. R7-R8.
3. Demetrius, L., *Thermodynamics and kinetics of protein folding: An evolutionary perspective.* Journal of Theoretical Biology, 2002. **217**(3): p. 397-411.
4. Kim, D.E., H.D. Gu, and D. Baker, *The sequences of small proteins are not extensively optimized for rapid folding by natural selection.* Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(9): p. 4982-4986.
5. Bouzas, T.D., J. Barros-Velazquez, and T.G. Villa, *Industrial applications of hyperthermophilic enzymes: A review.* Protein and Peptide Letters, 2006. **13**(7): p. 645-651.
6. Vieille, C. and G.J. Zeikus, *Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability.* Microbiol Mol Biol Rev, 2001. **65**(1): p. 1-43.
7. Vieille, C. and G.J. Zeikus, *Hyperthermophilic enzymes: Sources, uses, and molecular mechanisms for thermostability.* Microbiology and Molecular Biology Reviews, 2001. **65**(1): p. 1-+.
8. Widmann, M., J. Pleiss, and A.K. Samland, *Computational tools for rational protein engineering of aldolases.* Comput Struct Biotechnol J, 2012. **2**: p. e201209016.
9. Rubingh, D.N., *Protein engineering from a bioindustrial point of view.* Current Opinion in Biotechnology, 1997. **8**(4): p. 417-422.
10. Moore, J.C., et al., *Strategies for the in vitro evolution of protein function: Enzyme evolution by random recombination of improved sequences.* Journal of Molecular Biology, 1997. **272**(3): p. 336-347.
11. Kumamaru, T., et al., *Enhanced degradation of polychlorinated biphenyls by directed evolution of biphenyl dioxygenase.* Nature Biotechnology, 1998. **16**(7): p. 663-666.
12. Miesenbock, G., D.A. De Angelis, and J.E. Rothman, *Visualizing secretion and synaptic transmission with pH-sensitive green fluorescent proteins.* Nature, 1998. **394**(6689): p. 192-195.
13. Luetz, S., L. Giver, and J. Lalonde, *Engineered enzymes for chemical production.* Biotechnol Bioeng, 2008. **101**(4): p. 647-53.
14. Cowan, D.A., *Thermophilic proteins: Stability and function in aqueous and organic solvents.* Comparative Biochemistry and Physiology a-Molecular & Integrative Physiology, 1997. **118**(3): p. 429-438.
15. Sterner, R. and W. Liebl, *Thermophilic adaptation of proteins.* Critical Reviews in Biochemistry and Molecular Biology, 2001. **36**(1): p. 39-106.
16. Steipe, B., *Evolutionary approaches to protein engineering.* Combinatorial Chemistry in Biology, 1999. **243**: p. 55-86.

17. Barrozo, A., et al., *Computational Protein Engineering: Bridging the Gap between Rational Design and Laboratory Evolution.* International Journal of Molecular Sciences, 2012. **13**(10): p. 12428-12460.
18. Huang, X., D. Gao, and C.G. Zhan, *Computational design of a thermostable mutant of cocaine esterase via molecular dynamics simulations.* Organic & Biomolecular Chemistry, 2011. **9**(11): p. 4138-4143.
19. Joo, J.C., et al., *Thermostabilization of Bacillus circulans xylanase: Computational optimization of unstable residues based on thermal fluctuation analysis.* Journal of Biotechnology, 2011. **151**(1): p. 56-65.
20. Bradley, E.A., et al., *Investigations of the Thermostability of Rubredoxin Models Using Molecular-Dynamics Simulations.* Protein Science, 1993. **2**(4): p. 650-665.
21. Meharenna, Y.T. and T.L. Poulos, *Using Molecular Dynamics To Probe the Structural Basis for Enhanced Stability in Thermal Stable Cytochromes P450.* Biochemistry, 2010. **49**(31): p. 6680-6686.
22. Kundu, S. and D. Roy, *Structural study of carboxylesterase from hyperthermophilic bacteria Geobacillus stearothermophilus by molecular dynamics simulation.* Journal of Molecular Graphics & Modelling, 2010. **28**(8): p. 820-827.
23. Kundu, S. and D. Roy, *Comparative structural studies of psychrophilic and mesophilic protein homologues by molecular dynamics simulation.* Journal of Molecular Graphics & Modelling, 2009. **27**(8): p. 871-880.
24. Rahman, M.B.A., et al., *Deciphering the Flexibility and Dynamics of Geobacillus zalihae Strain T1 Lipase at High Temperatures by Molecular Dynamics Simulation.* Protein and Peptide Letters, 2009. **16**(11): p. 1360-1370.
25. Spiwok, V., et al., *Cold-active enzymes studied by comparative molecular dynamics simulation.* Journal of Molecular Modeling, 2007. **13**(4): p. 485-497.
26. Purmonen, M., et al., *Molecular dynamics studies on the thermostability of family 11 xylanases.* Protein Engineering Design & Selection, 2007. **20**(11): p. 551-559.
27. Goldstein, R.A., *Amino-acid interactions in psychrophiles, mesophiles, thermophiles, and hyperthermophiles: insights from the quasi-chemical approximation.* Protein Sci, 2007. **16**(9): p. 1887-95.
28. Perl, D. and F.X. Schmid, *Electrostatic stabilization of a thermophilic cold shock protein.* J Mol Biol, 2001. **313**(2): p. 343-57.
29. Schweiker, K.L., et al., *Computational design of the Fyn SH3 domain with increased stability through optimization of surface charge-charge interactions.* Protein Science, 2007. **16**(12): p. 2694-2702.
30. Eijsink, V.G.H., et al., *Rational engineering of enzyme stability.* Journal of Biotechnology, 2004. **113**(1-3): p. 105-120.
31. Reetz, M.T., J. D Carballeira, and A. Vogel, *Iterative saturation mutagenesis on the basis of B factors as a strategy for increasing protein thermostability.* Angewandte Chemie-International Edition, 2006. **45**(46): p. 7745-7751.
32. Haki, G.D. and S.K. Rakshit, *Developments in industrially important thermostable enzymes: a review.* Bioresource Technology, 2003. **89**(1): p. 17-34.
33. Sawle, L. and K. Ghosh, *How Do Thermophilic Proteins and Proteomes Withstand High Temperature?* Biophysical Journal, 2011. **101**(1): p. 217-227.
34. Petsko, G.A., *Structural basis of thermostability in hyperthermophilic proteins, or "there's more than one way to skin a cat".* Hyperthermophilic Enzymes, Pt C, 2001. **334**: p. 469-478.
35. Jaenicke, R., *Do ultrastable proteins from hyperthermophiles have high or low conformational rigidity?* Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(7): p. 2962-2964.
36. Zavodszky, P., et al., *Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins.* Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(13): p. 7406-7411.

37. Merkley, E.D., W.W. Parson, and V. Daggett, *Temperature dependence of the flexibility of thermophilic and mesophilic flavoenzymes of the nitroreductase fold.* Protein Engineering Design & Selection, 2010. **23**(5): p. 327-336.
38. Radestock, S. and H. Gohlke, *Protein rigidity and thermophilic adaptation.* Proteins-Structure Function and Bioinformatics, 2011. **79**(4): p. 1089-1108.
39. Radestock, S. and H. Gohlke, *Exploiting the Link between Protein Rigidity and Thermostability for Data-Driven Protein Engineering (vol 8, pg 507, 2008).* Engineering in Life Sciences, 2008. **8**(6): p. 657-657.
40. LeMaster, D.M., et al., *Enhanced thermal stability achieved without increased conformational rigidity at physiological temperatures: Spatial propagation of differential flexibility in rubredoxin hybrids.* Proteins-Structure Function and Bioinformatics, 2005. **61**(3): p. 608-616.
41. Chen, J.M., et al., *Increasing the thermostability of staphylococcal nuclease: Implications for the origin of protein thermostability.* Journal of Molecular Biology, 2000. **303**(2): p. 125-130.
42. Dumon, C., et al., *Engineering hyperthermostability into a GH11 xylanase is mediated by subtle changes to protein structure.* Journal of Biological Chemistry, 2008. **283**(33): p. 22557-22564.
43. Tian, J.A., et al., *Enhanced thermostability of methyl parathion hydrolase from Ochrobactrum sp. M231 by rational engineering of a glycine to proline mutation.* Febs Journal, 2010. **277**(23): p. 4901-4908.
44. Watanabe, K., et al., *Proline Residues Responsible for Thermostability Occur with High-Frequency in the Loop Regions of an Extremely Thermostable Oligo-1,6-Glucosidase from Bacillus-Thermoglucosidasius Kp1006.* Journal of Biological Chemistry, 1991. **266**(36): p. 24287-24294.
45. Gu, J. and V.J. Hilser, *Sequence-Based Analysis of Protein Energy Landscapes Reveals Nonuniform Thermal Adaptation within the Proteome.* Molecular Biology and Evolution, 2009. **26**(10): p. 2217-2227.
46. Pace, C.N., *Single surface stabilizer.* Nature Structural Biology, 2000. **7**(5): p. 345-346.
47. Chaparro-Riggers, J.F., K.M. Polizzi, and A.S. Bommarius, *Better library design: data-driven protein engineering.* Biotechnol J, 2007. **2**(2): p. 180-91.
48. Lehmann, M., et al., *The consensus concept for thermostability engineering of proteins.* Biochim Biophys Acta, 2000. **1543**(2): p. 408-415.
49. Shoichet, B.K., et al., *A Relationship between Protein Stability and Protein Function.* Proceedings of the National Academy of Sciences of the United States of America, 1995. **92**(2): p. 452-456.
50. Polizzi, K.M., et al., *Structure-guided consensus approach to create a more thermostable penicillin G acylase.* Biotechnol J, 2006. **1**(5): p. 531-6.
51. Kumar, M.D.S., et al., *ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions.* Nucleic Acids Research, 2006. **34**: p. D204-D206.
52. Zhang, X.J., et al., *Enhancement of protein stability by the combination of point mutations in T4 lysozyme is additive.* Protein Engineering, 1995. **8**(10): p. 1017-1022.
53. Song, J.K. and J.S. Rhee, *Simultaneous enhancement of thermostability and catalytic activity of phospholipase A(1) by evolutionary molecular engineering.* Appl Environ Microbiol, 2000. **66**(3): p. 890-4.
54. Wilkinson, A., J. Day, and R. Bowater, *Bacterial DNA ligases.* Molecular Microbiology, 2001. **40**(6): p. 1241-1248.
55. Le, Y.L., et al., *Properties of an NAD(+)-dependent DNA ligase from the hyperthermophile Thermotoga maritima and its application in PCR amplification of long DNA fragments.* Enzyme and Microbial Technology, 2010. **46**(2): p. 113-117.

56. Seitz, P., et al., *ComEA Is Essential for the Transfer of External DNA into the Periplasm in Naturally Transformable Vibrio cholerae Cells.* Plos Genetics, 2014. **10**(1).
57. Cornell, W.D., et al., *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995).* Journal of the American Chemical Society, 1996. **118**(9): p. 2309-2309.
58. Fox, T. and P.A. Kollman, *Application of the RESP methodology in the parametrization of organic solvents.* Journal of Physical Chemistry B, 1998. **102**(41): p. 8070-8079.
59. Darden, T., D. York, and L. Pedersen, *Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems.* Journal of Chemical Physics, 1993. **98**(12): p. 10089-10092.
60. Schymkowitz, J., et al., *The FoldX web server: an online force field.* Nucleic Acids Research, 2005. **33**: p. W382-W388.
61. Christensen, N.J. and K.P. Kepp, *Accurate Stabilities of Laccase Mutants Predicted with a Modified FoldX Protocol.* Journal of Chemical Information and Modeling, 2012. **52**(11): p. 3028-3042.
62. Frenkel, D. and B. Smit, *Understanding molecular simulation: from algorithms to applications*. Vol. 1. 2001: Academic press.
63. Nose, S., *A Unified Formulation of the Constant Temperature Molecular-Dynamics Methods.* Journal of Chemical Physics, 1984. **81**(1): p. 511-519.
64. Andersen, H.C., *Molecular-Dynamics Simulations at Constant Pressure and-or Temperature.* Journal of Chemical Physics, 1980. **72**(4): p. 2384-2393.
65. Neumann, M. and O. Steinhauser, *Influence of Boundary-Conditions Used in Machine Simulations on the Structure of Polar Systems.* Molecular Physics, 1980. **39**(2): p. 437-454.
66. Su, J.G., et al., *Protein unfolding behavior studied by elastic network model.* Biophysical Journal, 2008. **94**(12): p. 4586-4596.
67. Bahar, I., A.R. Atilgan, and B. Erman, *Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential.* Folding & Design, 1997. **2**(3): p. 173-181.
68. Bahar, I., et al., *Normal Mode Analysis of Biomolecular Structures: Functional Mechanisms of Membrane Proteins.* Chemical Reviews, 2010. **110**(3): p. 1463-1497.
69. Haliloglu, T., I. Bahar, and B. Erman, *Gaussian dynamics of folded proteins.* Physical Review Letters, 1997. **79**(16): p. 3090-3093.
70. Liu, M., et al., *Insight into the Structure, Dynamics and the Unfolding Property of Amylosucrases: Implications of Rational Engineering on Thermostability.* Plos One, 2012. **7**(7).
71. Popov, I., et al., *Bioinformatics in Proteomics: A Review on Methods and Algorithms.* Biotechnology & Biotechnological Equipment, 2009. **23**(1): p. 1115-1120.
72. Pavlopoulou, A. and I. Michalopoulos, *State-of-the-art bioinformatics protein structure prediction tools (Review).* International Journal of Molecular Medicine, 2011. **28**(3): p. 295-310.
73. Geer, L.Y., et al., *The NCBI BioSystems database.* Nucleic Acids Research, 2010. **38**: p. D492-D496.
74. Feng, D.F. and R.F. Doolittle, *Progressive Sequence Alignment as a Prerequisite to Correct Phylogenetic Trees.* Journal of Molecular Evolution, 1987. **25**(4): p. 351-360.
75. Larkin, M.A., et al., *Clustal W and clustal X version 2.0.* Bioinformatics, 2007. **23**(21): p. 2947-2948.
76. Finn, R.D., et al., *Pfam: the protein families database.* Nucleic Acids Research, 2014. **42**(D1): p. D222-D230.
77. Bava, K.A., et al., *ProTherm, version 4.0: thermodynamic database for proteins and mutants.* Nucleic Acids Research, 2004. **32**: p. D120-D121.
78. Cock, P.J.A., et al., *Biopython: freely available Python tools for computational molecular biology and bioinformatics.* Bioinformatics, 2009. **25**(11): p. 1422-1423.

79. Capriotti, E., P. Fariselli, and R. Casadio, *I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure.* Nucleic Acids Research, 2005. **33**: p. W306-W310.
80. Masso, M. and I.I. Vaisman, *AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements.* Protein Engineering Design & Selection, 2010. **23**(8): p. 683-687.
81. Gill, P., T.T. Moghadam, and B. Ranjbar, *Differential scanning calorimetry techniques: applications in biology and nanoscience.* J Biomol Tech, 2010. **21**(4): p. 167-93.
82. Fawcett, T., *An introduction to ROC analysis.* Pattern Recognition Letters, 2006. **27**(8): p. 861-874.
83. Demirjian, D.C., F. Moris-Varas, and C.S. Cassidy, *Enzymes from extremophiles.* Current Opinion in Chemical Biology, 2001. **5**(2): p. 144-151.
84. Zamost, B.L., H.K. Nielsen, and R.L. Starnes, *Thermostable Enzymes for Industrial Applications.* Journal of Industrial Microbiology, 1991. **8**(2): p. 71-81.
85. Kumar, S. and R. Nussinov, *How do thermophilic proteins deal with heat?* Cellular and Molecular Life Sciences, 2001. **58**(9): p. 1216-1233.
86. Razvi, A. and J.M. Scholtz, *Lessons in stability from thermophilic proteins.* Protein Science, 2006. **15**(7): p. 1569-1578.
87. Zhou, H.X., *Toward the physical basis of thermophilic proteins: Linking of enriched polar interactions and reduced heat capacity of unfolding.* Biophysical Journal, 2002. **83**(6): p. 3126-3133.
88. Das, R. and M. Gerstein, *The stability of thermophilic proteins: a study based on comprehensive genome comparison.* Funct Integr Genomics, 2000. **1**(1): p. 76-88.
89. Gianese, G., F. Bossa, and S. Pascarella, *Comparative structural analysis of psychrophilic and meso- and thermophilic enzymes.* Proteins-Structure Function and Bioinformatics, 2002. **47**(2): p. 236-249.
90. Zeldovich, K.B., I.N. Berezovsky, and E.I. Shakhnovich, *Protein and DNA sequence determinants of thermophilic adaptation.* Plos Computational Biology, 2007. **3**(1): p. 62-72.
91. Ladenstein, R. and G. Antranikian, *Proteins from hyperthermophiles: stability and enzymatic catalysis close to the boiling point of water.* Adv Biochem Eng Biotechnol, 1998. **61**: p. 37-85.
92. Adams, M.W.W., *Enzymes and Proteins from Organisms That Grow near and above 100-Degrees-C.* Annual Review of Microbiology, 1993. **47**: p. 627-658.
93. Paiardini, A., et al., *"Hot cores" in proteins: Comparative analysis of the apolar contact area in structures from hyper/thermophilic and mesophilic organisms.* Bmc Structural Biology, 2008. **8**: p. 8-14.
94. Greaves, R.B. and J. Warwicker, *Mechanisms for stabilisation and the maintenance of solubility in proteins from thermophiles.* Bmc Structural Biology, 2007. **7**: p. 7-18.
95. Chakravarty, S. and R. Varadarajan, *Elucidation of factors responsible for enhanced thermal stability of proteins: A structural genomics based study.* Biochemistry, 2002. **41**(25): p. 8152-8161.
96. Vazquez-Figueroa, E., et al., *Thermostable variants constructed via the structure-guided consensus method also show increased stability in salts solutions and homogeneous aqueous-organic media.* Protein Engineering Design & Selection, 2008. **21**(11): p. 673-680.
97. Altschul, S.F., et al., *Basic Local Alignment Search Tool.* Journal of Molecular Biology, 1990. **215**(3): p. 403-410.
98. Maunders, M.J., *DNA and RNA ligases (EC 6.5.1.1, EC 6.5.1.2, and EC 6.5.1.3).* Methods Mol Biol, 1993. **16**: p. 213-230.
99. Smagin, V.A., et al., *[Isolation and characteristics of new thermostable DNA ligase from archaea of the genus Thermococcus].* Prikl Biokhim Mikrobiol, 2008. **44**(5): p. 523-528.

100. Petrova, T., et al., *ATP-dependent DNA ligase from Thermococcus sp 1519 displays a new arrangement of the OB-fold domain.* Acta Crystallographica Section F-Structural Biology and Crystallization Communications, 2012. **68**: p. 1440-1447.
101. Tomkinson, A.E., et al., *DNA ligases: Structure, reaction mechanism, and function.* Chemical Reviews, 2006. **106**(2): p. 687-699.
102. Doherty, A.J. and S.W. Suh, *Structural and mechanistic conservation in DNA ligases.* Nucleic Acids Research, 2000. **28**(21): p. 4051-4058.
103. Ellenberger, T. and A.E. Tomkinson, *Eukaryotic DNA ligases: Structural and functional insights.* Annual Review of Biochemistry, 2008. **77**: p. 313-338.
104. Martin, I.V. and S.A. MacNeill, *ATP-dependent DNA ligases.* Genome Biology, 2002. **3**(4): p. reviews3005.1–reviews3005.7.
105. Eswar, N., et al., *Comparative protein structure modeling using Modeller.* Curr Protoc Bioinformatics, 2006. **Chapter 5**: p. Unit 5 6.
106. Phillips, J.C., et al., *Scalable molecular dynamics with NAMD.* J Comput Chem, 2005. **26**(16): p. 1781-802.
107. Hornak, V., et al., *Comparison of multiple amber force fields and development of improved protein backbone parameters.* Proteins-Structure Function and Bioinformatics, 2006. **65**(3): p. 712-726.
108. Jorgensen, W.L., et al., *Comparison of Simple Potential Functions for Simulating Liquid Water.* Journal of Chemical Physics, 1983. **79**(2): p. 926-935.
109. Martyna, G.J., D.J. Tobias, and M.L. Klein, *Constant-Pressure Molecular-Dynamics Algorithms.* Journal of Chemical Physics, 1994. **101**(5): p. 4177-4189.
110. Feller, S.E., et al., *Constant-Pressure Molecular-Dynamics Simulation - the Langevin Piston Method.* Journal of Chemical Physics, 1995. **103**(11): p. 4613-4621.
111. Jean-Paul Ryckaert, G.C., Herman J.C Berendsen, *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes.* Journal of Computational Physics, 1977. **23**(3): p. 327–341.
112. Darden, T., et al., *New tricks for modelers from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations.* Structure (London, England : 1993), 1999. **7**(3): p. R55-60.
113. Humphrey, W., A. Dalke, and K. Schulten, *VMD: Visual molecular dynamics.* Journal of Molecular Graphics & Modelling, 1996. **14**(1): p. 33-38.
114. Bakan, A., L.M. Meireles, and I. Bahar, *ProDy: Protein Dynamics Inferred from Theory and Experiments.* Bioinformatics, 2011. **27**(11): p. 1575-1577.
115. Van Rossum, G. and F.L. Drake, *Python language reference manual*. 2003, Bristol: Network Theory Limited. ii, 112 p.
116. Lauer, G., et al., *Cloning, Nucleotide-Sequence, and Engineered Expression of Thermus-Thermophilus DNA-Ligase, a Homolog of Escherichia-Coli DNA-Ligase.* Journal of Bacteriology, 1991. **173**(16): p. 5047-5053.
117. Yilin, L., et al., *Thermostable DNA Ligase-Mediated PCR Production of Circular Plasmid (PPCP) and Its Application in Directed Evolution via In situ Error-Prone PCR.* DNA Research, 2013. **20**(4): p. 375-382.
118. Kim, S.J., et al., *The Development of a Thermostable CiP (Coprinus cinereus peroxidase) Through in silico Design.* Biotechnology Progress, 2010. **26**(4): p. 1038-1046.
119. Kim, H.S., A.T.L. Quang, and Y.H. Kim, *Development of thermostable lipase B from Candida antarctica (CalB) through in silico design employing B-factor and RosettaDesign.* Enzyme and Microbial Technology, 2010. **47**(1-2): p. 1-5.
120. Parthasarathy, S. and M.R.N. Murthy, *Protein thermal stability: insights from atomic displacement parameters (B values).* Protein Engineering, 2000. **13**(1): p. 9-13.
121. Blum, J.K., M.D. Ricketts, and A.S. Bommarius, *Improved thermostability of AEH by combining B-FIT analysis and structure-guided consensus method.* Journal of Biotechnology, 2012. **160**(3-4): p. 214-221.

122. Zhang, J.H., et al., *High-throughput screening of B factor saturation mutated Rhizomucor miehei lipase thermostability based on synthetic reaction.* Enzyme and Microbial Technology, 2012. **50**(6-7): p. 325-330.
123. Siglioccolo, A., R. Gerace, and S. Pascarella, *"Cold spots" in protein cold adaptation: Insights from normalized atomic displacement parameters (B '-factors).* Biophysical Chemistry, 2010. **153**(1): p. 104-114.
124. van den Burg, B. and V.G.H. Eijsink, *Selection of mutations for increased protein stability.* Current Opinion in Biotechnology, 2002. **13**(4): p. 333-337.
125. Villegas, V., et al., *Stabilization of proteins by rational design of alpha-helix stability using helix/coil transition theory.* Folding & Design, 1996. **1**(1): p. 29-34.
126. Clark, A.T., et al., *Thermodynamics of core hydrophobicity and packing in the hyperthermophile proteins Sac7d and Sso7d.* Biochemistry, 2004. **43**(10): p. 2840-2853.
127. Niehaus, F., et al., *Extremophiles as a source of novel enzymes for industrial application.* Applied Microbiology and Biotechnology, 1999. **51**(6): p. 711-729.
128. Egorova, K. and G. Antranikian, *Industrial relevance of thermophilic Archaea.* Current Opinion in Microbiology, 2005. **8**(6): p. 649-655.
129. Ravot, G., J.M. Masson, and F. Lefevre, *Applications of extremophiles: The industrial screening of extremophiles for valuable biomolecules.* Extremophiles, 2006. **35**: p. 785-+.
130. Willies, S., M. Isupov, and J. Littlechild, *Thermophilic enzymes and their applications in biocatalysis: a robust aldo-keto reductase.* Environmental Technology, 2010. **31**(10): p. 1159-1167.
131. Szilagyi, A. and P. Zavodszky, *Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey.* Structure, 2000. **8**(5): p. 493-504.
132. Kumwenda, B., et al., *Analysis of Protein Thermostability Enhancing Factors in Industrially Important Thermus Bacteria Species.* Evolutionary Bioinformatics, 2013. **9**: p. 327-342.
133. Takano, K., et al., *Evolvability of Thermophilic Proteins from Archaea and Bacteria.* Biochemistry, 2013. **52**(28): p. 4774-4780.
134. Ding, Y.R., et al., *Comparison of the structural basis for thermal stability between archaeal and bacterial proteins.* Extremophiles, 2012. **16**(1): p. 67-78.
135. Chang, S., et al., *Stability and Folding Behavior Analysis of Zinc-Finger Using Simple Models.* International Journal of Molecular Sciences, 2010. **11**(10): p. 4014-4034.
136. Erman, B., *The Gaussian network model: Precise prediction of residue fluctuations and application to binding problems.* Biophysical Journal, 2006. **91**(10): p. 3589-3599.
137. Reich, L. and T.R. Weikl, *Substructural cooperativity and parallel versus sequential events during protein unfolding.* Proteins-Structure Function and Bioinformatics, 2006. **63**(4): p. 1052-1058.
138. Lazaridis, T. and M. Karplus, *"New view" of protein folding reconciled with the old through multiple unfolding simulations.* Science, 1997. **278**(5345): p. 1928-1931.
139. Li, A.J. and V. Daggett, *Characterization of the Transition-State of Protein Unfolding by Use of Molecular-Dynamics - Chymotrypsin Inhibitor-2.* Proceedings of the National Academy of Sciences of the United States of America, 1994. **91**(22): p. 10430-10434.
140. Kazmirski, S.L., et al., *Protein folding from a highly disordered denatured state: The folding pathway of chymotrypsin inhibitor 2 at atomic resolution.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(8): p. 4349-4354.
141. Li, L. and E.I. Shakhnovich, *Constructing, verifying, and dissecting the folding transition state of chymotrypsin inhibitor 2 with all-atom simulations.*

Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(23): p. 13014-13018.

142. Li, A.J. and V. Daggett, *Identification and characterization of the unfolding transition state of chymotrypsin inhibitor 2 by molecular dynamics simulations.* Journal of Molecular Biology, 1996. **257**(2): p. 412-429.
143. Day, R. and V. Daggett, *Sensitivity of the folding/unfolding transition state ensemble of chymotrypsin inhibitor 2 to changes in temperature and solvent.* Protein Science, 2005. **14**(5): p. 1242-1252.
144. Ferrara, P., J. Apostolakis, and A. Caflisch, *Targeted molecular dynamics simulations of protein unfolding.* Journal of Physical Chemistry B, 2000. **104**(18): p. 4511-4518.
145. Ozkan, S.B., G.S. Dalgyn, and T. Haliloglu, *Unfolding events of Chymotrypsin Inhibitor 2 (CI2) revealed by Monte Carlo (MC) simulations and their consistency from structure-based analysis of conformations.* Polymer, 2004. **45**(2): p. 581-595.
146. Fersht, A.R., *Protein-Folding and Stability - the Pathway of Folding of Barnase.* Febs Letters, 1993. **325**(1-2): p. 5-16.
147. Kmiecik, S. and A. Kolinski, *Characterization of protein-folding pathways by reduced-space modeling.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(30): p. 12330-12335.
148. Chirakkal, H., G.C. Ford, and A. Moir, *Analysis of a conserved hydrophobic pocket important for the thermostability of Bacillus pumilus chloramphenicol acetyltransferase (CAT-86).* Protein Engineering, 2001. **14**(3): p. 161-166.
149. Northey, J.G.B., A.A. Di Nardo, and A.R. Davidson, *Hydrophobic core packing in the SH3 domain folding transition state.* Nature Structural Biology, 2002. **9**(2): p. 126-130.
150. Dong, H.J., et al., *Hydrophobic effect on the stability and folding of a hyperthermophilic protein.* Journal of Molecular Biology, 2008. **378**(1): p. 264-272.
151. Bueno, M., et al., *Energetics of aliphatic deletions in protein cores.* Protein Science, 2006. **15**(8): p. 1858-1872.
152. Krittanai, C., et al., *Mutation of the hydrophobic residue on helix alpha 5 of the Bacillus thuringiensis Cry4B affects structural stability.* Protein and Peptide Letters, 2003. **10**(4): p. 361-368.
153. Gromiha, M.M., et al., *Hydrophobic environment is a key factor for the stability of thermophilic proteins.* Proteins-Structure Function and Bioinformatics, 2013. **81**(4): p. 715-721.
154. Banerji, A. and I. Ghosh, *A new computational model to study mass inhomogeneity and hydrophobicity inhomogeneity in proteins.* European Biophysics Journal with Biophysics Letters, 2009. **38**(5): p. 577-587.
155. Kono, H., et al., *Designing the hydrophobic core of Thermus flavus malate dehydrogenase based on side-chain packing.* Protein Engineering, 1998. **11**(1): p. 47-52.
156. Daopin, S., et al., *Structural and Thermodynamic Analysis of the Packing of 2 Alpha-Helices in Bacteriophage-T4 Lysozyme.* Journal of Molecular Biology, 1991. **221**(2): p. 647-667.
157. Yutani, K., et al., *Dependence of Conformational Stability on Hydrophobicity of the Amino-Acid Residue in a Series of Variant Proteins Substituted at a Unique Position of Tryptophan Synthase Alpha-Subunit.* Proceedings of the National Academy of Sciences of the United States of America, 1987. **84**(13): p. 4441-4444.
158. Matsumura, M., W.J. Becktel, and B.W. Matthews, *Hydrophobic Stabilization in T4 Lysozyme Determined Directly by Multiple Substitutions of Ile-3.* Nature, 1988. **334**(6181): p. 406-410.
159. Kellis, J.T., K. Nyberg, and A.R. Fersht, *Energetics of Complementary Side-Chain Packing in a Protein Hydrophobic Core.* Biochemistry, 1989. **28**(11): p. 4914-4922.

160. Shortle, D., W.E. Stites, and A.K. Meeker, *Contributions of the Large Hydrophobic Amino-Acids to the Stability of Staphylococcal Nuclease.* Biochemistry, 1990. **29**(35): p. 8033-8041.
161. Baase, W.A., et al., *Dissection of Protein-Structure and Folding by Directed Mutagenesis.* Faraday Discussions, 1992. **93**: p. 173-181.
162. Eriksson, A.E., et al., *Response of a Protein-Structure to Cavity-Creating Mutations and Its Relation to the Hydrophobic Effect.* Science, 1992. **255**(5041): p. 178-183.
163. Serrano, L., et al., *The Folding of an Enzyme .2. Substructure of Barnase and the Contribution of Different Interactions to Protein Stability.* Journal of Molecular Biology, 1992. **224**(3): p. 783-804.
164. Takano, K., et al., *Contribution of the hydrophobic effect to the stability of human lysozyme: Calorimetric studies and X-ray structural analyses of the nine Valine to Alanine mutants.* Biochemistry, 1997. **36**(4): p. 688-698.
165. Xu, J.A., et al., *The response of T4 lysozyme to large-to-small substitutions within the core and its relation to the hydrophobic effect.* Protein Science, 1998. **7**(1): p. 158-177.
166. Chen, J.M. and W.E. Stites, *Energetics of side chain packing in staphylococcal nuclease assessed by systematic double mutant cycles.* Biochemistry, 2001. **40**(46): p. 14004-14011.
167. Jackson, S.E., et al., *Effect of Cavity-Creating Mutations in the Hydrophobic Core of Chymotrypsin Inhibitor-2.* Biochemistry, 1993. **32**(42): p. 11259-11269.
168. Karpusas, M., et al., *Hydrophobic Packing in T4 Lysozyme Probed by Cavity-Filling Mutants.* Proceedings of the National Academy of Sciences of the United States of America, 1989. **86**(21): p. 8237-8241.
169. Sandberg, W.S. and T.C. Terwilliger, *Influence of Interior Packing and Hydrophobicity on the Stability of a Protein.* Science, 1989. **245**(4913): p. 54-57.
170. Lim, W.A. and R.T. Sauer, *The Role of Internal Packing Interactions in Determining the Structure and Stability of a Protein.* Journal of Molecular Biology, 1991. **219**(2): p. 359-376.
171. Hurley, J.H., W.A. Baase, and B.W. Matthews, *Design and Structural-Analysis of Alternative Hydrophobic Core Packing Arrangements in Bacteriophage-T4 Lysozyme.* Journal of Molecular Biology, 1992. **224**(4): p. 1143-1159.
172. Lim, W.A., D.C. Farruggio, and R.T. Sauer, *Structural and Energetic Consequences of Disruptive Mutations in a Protein Core.* Biochemistry, 1992. **31**(17): p. 4324-4333.
173. Baldwin, E., et al., *Thermodynamic and structural compensation in "size-switch" core repacking variants of bacteriophage T4 lysozyme.* Journal of Molecular Biology, 1996. **259**(3): p. 542-559.
174. Anil, B., et al., *Fine structure analysis of a protein folding transition state; distinguishing between hydrophobic stabilization and specific packing.* Journal of Molecular Biology, 2005. **354**(3): p. 693-705.
175. Buckle, A.M., P. Cramer, and A.R. Fersht, *Structural and energetic responses to cavity-creating mutations in hydrophobic cores: Observation of a buried water molecule and the hydrophilic nature of such hydrophobic cavities.* Biochemistry, 1996. **35**(14): p. 4298-4305.
176. Otzen, D.E., M. Rheinnecker, and A.R. Fersht, *Structural Factors Contributing to the Hydrophobic Effect - the Partly Exposed Hydrophobic Minicore in Chymotrypsin Inhibitor-2.* Biochemistry, 1995. **34**(40): p. 13051-13058.
177. Vlassi, M., G. Cesareni, and M. Kokkinidis, *A correlation between the loss of hydrophobic core packing interactions and protein stability.* Journal of Molecular Biology, 1999. **285**(2): p. 817-827.
178. Priyakumar, U.D., *Role of Hydrophobic Core on the Thermal Stability of Proteins - Molecular Dynamics Simulations on a Single Point Mutant of Sso7d.* Journal of Biomolecular Structure & Dynamics, 2012. **29**(5): p. 961-971.

179. Tian, X.H., et al., *Computational model for protein unfolding simulation.* Physical Review E, 2011. **83**(6).
180. Kundu, S., et al., *Dynamics of proteins in crystals: Comparison of experiment with simple models.* Biophysical Journal, 2002. **83**(2): p. 723-732.
181. Miyazawa, S. and R.L. Jernigan, *Estimation of Effective Interresidue Contact Energies from Protein Crystal-Structures - Quasi-Chemical Approximation.* Macromolecules, 1985. **18**(3): p. 534-552.
182. Gromiha, M.M., et al., *Role of structural and sequence information in the prediction of protein stability changes: comparison between buried and partially buried mutations.* Protein Engineering, 1999. **12**(7): p. 549-555.
183. Manavalan, P. and P.K. Ponnuswamy, *Study of Preferred Environment of Amino-Acid Residues in Globular Proteins.* Archives of Biochemistry and Biophysics, 1977. **184**(2): p. 476-487.
184. Manavalan, P. and P.K. Ponnuswamy, *Hydrophobic Character of Amino-Acid Residues in Globular Proteins.* Nature, 1978. **275**(5681): p. 673-674.
185. Wijma, H.J., R.J. Floor, and D.B. Janssen, *Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability.* Curr Opin Struct Biol, 2013.
186. Tan, K.P., et al., *Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pK(a) of ionizable residues in proteins.* Nucleic Acids Research, 2013. **41**(W1): p. W314-W321.
187. Alm, E. and D. Baker, *Matching theory and experiment in protein folding.* Current Opinion in Structural Biology, 1999. **9**(2): p. 189-196.
188. Perl, D., et al., *Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins.* Nature Structural Biology, 1998. **5**(3): p. 229-235.
189. Alm, E. and D. Baker, *Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures.* Proceedings of the National Academy of Sciences of the United States of America, 1999. **96**(20): p. 11305-11310.
190. Clementi, C., P.A. Jennings, and J.N. Onuchic, *How native-state topology affects the folding of dihydrofolate reductase and interleukin-1 beta.* Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(11): p. 5871-5876.
191. Koga, N. and S. Takada, *Roles of native topology and chain-length scaling in protein folding: A simulation study with a Go-like model.* Journal of Molecular Biology, 2001. **313**(1): p. 171-180.
192. Galzitskaya, O.V. and A.V. Finkelstein, *A theoretical search for folding/unfolding nuclei in three-dimensional protein structures.* Proceedings of the National Academy of Sciences of the United States of America, 1999. **96**(20): p. 11299-11304.
193. Day, R., et al., *Increasing temperature accelerates protein unfolding without changing the pathway of unfolding.* Journal of Molecular Biology, 2002. **322**(1): p. 189-203.
194. Yuan, Z., T.L. Bailey, and R.D. Teasdale, *Prediction of protein B-factor profiles.* Proteins-Structure Function and Bioinformatics, 2005. **58**(4): p. 905-912.
195. Pettersen, E.F., et al., *UCSF chimera - A visualization system for exploratory research and analysis.* Journal of Computational Chemistry, 2004. **25**(13): p. 1605-1612.
196. Johnson, E.C., et al., *Solution structure and dynamics of a designed hydrophobic core variant of ubiquitin.* Structure with Folding & Design, 1999. **7**(8): p. 967-976.
197. Lim, W.A., et al., *The Crystal-Structure of a Mutant Protein with Altered but Improved Hydrophobic Core Packing.* Proceedings of the National Academy of Sciences of the United States of America, 1994. **91**(1): p. 423-427.
198. Vazquez-Figueroa, E., J. Chaparro-Riggers, and A.S. Bommarius, *Development of a thermostable glucose dehydrogenase by a structure-guided consensus concept.* Chembiochem, 2007. **8**(18): p. 2295-301.

199. Lehmann, M., et al., *The consensus concept for thermostability engineering of proteins: further proof of concept.* Protein Eng, 2002. **15**(5): p. 403-11.
200. Anbar, M., et al., *Improved Thermostability of Clostridium thermocellum Endoglucanase Cel8A by Using Consensus-Guided Mutagenesis.* Applied and Environmental Microbiology, 2012. **78**(9): p. 3458-3464.
201. Lehmann, M. and M. Wyss, *Engineering proteins for thermostability: the use of sequence alignments versus rational design and directed evolution.* Current Opinion in Biotechnology, 2001. **12**(4): p. 371-375.
202. Pantoliano, M.W., et al., *Large Increases in General Stability for Subtilisin Bpn' through Incremental Changes in the Free-Energy of Unfolding.* Biochemistry, 1989. **28**(18): p. 7205-7213.
203. Steipe, B., et al., *Sequence Statistics Reliably Predict Stabilizing Mutations in a Protein Domain.* Journal of Molecular Biology, 1994. **240**(3): p. 188-192.
204. Steipe, B., *Consensus-based engineering of protein stability: From intrabodies to thermostable enzymes.* Protein Engineering, 2004. **388**: p. 176-186.
205. Lehmann, M., et al., *The consensus concept for thermostability engineering of proteins.* Biochimica Et Biophysica Acta-Protein Structure and Molecular Enzymology, 2000. **1543**(2): p. 408-415.
206. Lehmann, M., et al., *The consensus concept for thermostability engineering of proteins: further proof of concept.* Protein Engineering, 2002. **15**(5): p. 403-411.
207. Ohage, E.C., et al., *beta-Turn propensities as paradigms for the analysis of structural motifs to engineer protein stability.* Protein Science, 1997. **6**(1): p. 233-241.
208. Xiao, Z.H., et al., *Improvement of the thermostability and activity of a pectate lyase by single amino acid substitutions, using a strategy based on melting-temperature-guided sequence alignment.* Applied and Environmental Microbiology, 2008. **74**(4): p. 1183-1189.
209. Bateman, A., et al., *The Pfam Protein Families Database.* Nucleic Acids Research, 2002. **30**(1): p. 276-280.
210. Liu, Y. and I. Bahar, *Sequence Evolution Correlates with Structural Dynamics.* Molecular Biology and Evolution, 2012. **29**(9): p. 2253-2263.
211. Haruki, M., et al., *Investigating the Role of Conserved Residue Asp134 in Escherichia-Coli Ribonuclease Hi by Site-Directed Random Mutagenesis.* European Journal of Biochemistry, 1994. **220**(2): p. 623-631.
212. Hopf, T.A., et al., *Sequence co-evolution gives 3D contacts and structures of protein complexes.* Elife, 2014. **3**.
213. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Research, 2000. **28**(1): p. 235-242.
214. Lee, Y., et al., *A Coevolutionary Residue Network at the Site of a Functionally Important Conformational Change in a Phosphohexomutase Enzyme Family.* Plos One, 2012. **7**(6).
215. Prieto, J., et al., *Generation and analysis of mesophilic variants of the thermostable archaeal I-DmoI homing endonuclease.* Journal of Biological Chemistry, 2008. **283**(7): p. 4364-4374.

Appendix A1: Supplementary Information for Chapter 3

In this appendix I report the supplementary material adapted from the paper submitted to *Biochemistry* and currently under second revision:

"Understanding and Engineering Thermostability in the DNA Ligase from Thermococcus sp. 1519", Hassan P. Modarres, Boris D. Dorokhov, Vladimir O. Popov, Nikolai V. Ravin, Konstantin G. Skryabin, Matteo Dal Peraro.

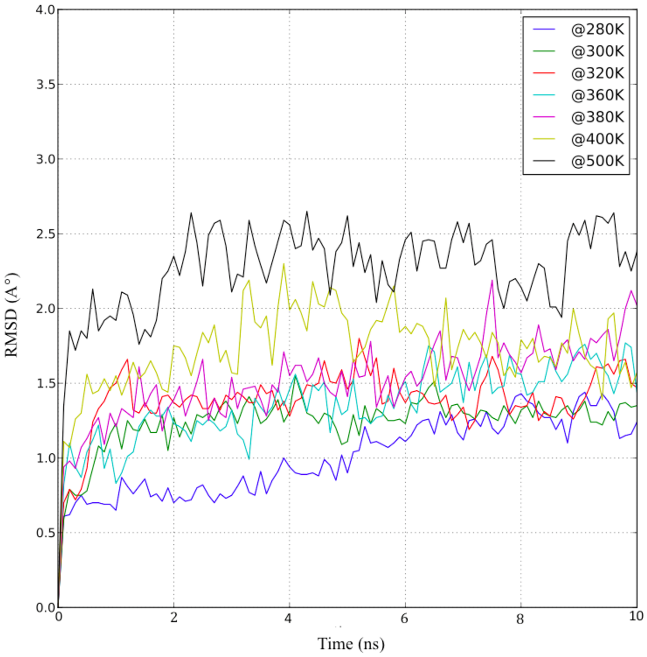


Figure S3-1. RMSD of the NBD at different temperatures during MD simulations.

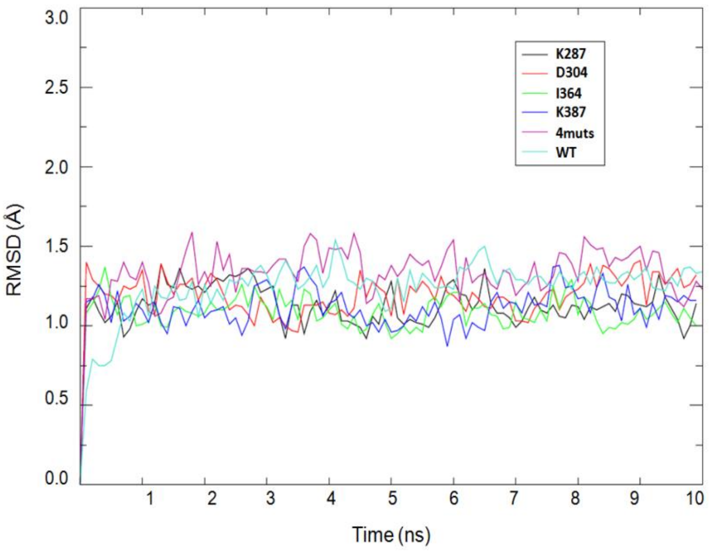


Figure S3-2. RMSD of the NBD for MD simulations of the wild-type (WT), the 4 predicted mutations alone and combined (4muts).

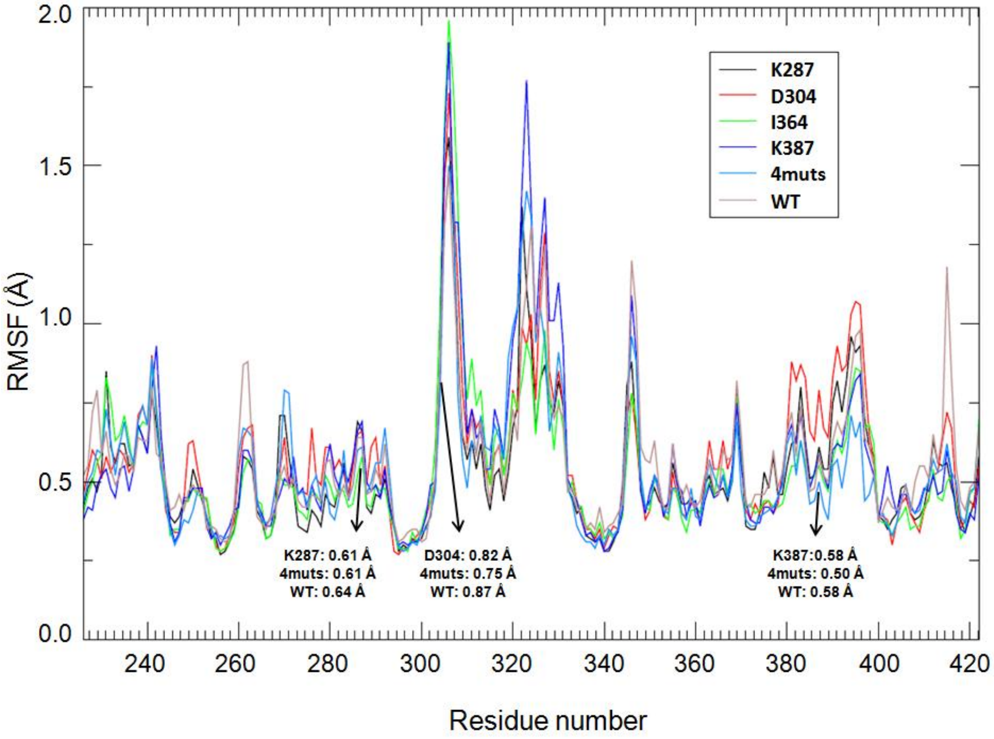


Figure S3-3. RMSF of the NBD for MD simulations of the wild-type (WT), the 4 predicted mutations alone and combined (4muts).

Table S3-1. List of accession code of mesophilic homologues. These sequences are chosen based on the lowest E-value, least gaps, highest alignment length, highest number of identities and temperature as far as possible from the thermophilic region.

Accession codes
NP_279843, NP_217578, NP_625491, NP_962051, YP_001827832, ZP_05975074, NP_988090, NP_828312

Table S3-2. List of accession codes of sequences belonging to hyperthermophilic mesophilic species.

Accession codes
YP_919839, Q9HHC4, NP_579364, YP_003400285, YP_003247998, YP_001012632, NP_213963, NP_614282, AAD00532, YP_003477672, NP_147713, YP_003420222, YP_001054945, YP_001041317, NP_126234, YP_004782082, YP_358898, YP_002994293, NP_558882, YP_003457537, YP_003323128, YP_002307902, YP_004341316, NP_963789, YP_002960084, YP_003649444, YP_002427972, YP_004176453, YP_002838243, NP_376074, YP_003128064, YP_004101968, ZP_09026322, YP_004340961, YP_003901955, YP_184553, YP_002829987, YP_002839895, YP_003669131, NP_247139, YP_001152343, YP_002582768, NP_341745, NP_070553, YP_001793472, YP_001435529, NP_069457, O67398, NP_143476, YP_003858975, O29632, YP_001540942, YP_920155, YP_003816621, YP_003616430

Table S3-3. Non-bonded interaction energy for residues at position 287 and 387 for the wild-type (denoted by WT) and mutant (denoted by mut) systems.

System	Average (kcal/mol)	Standard deviation
WT_A287	-55	4
mut_K287	-91	20
mut_E287	-39	10
WT_A387	-62	2
mut_K387	-109	39
mut_E387	-99	29

Appendix A2: Supplementary Information for Chapter 4

Table S4-1. Result of iterative GNM calculations. The Entry code is the code used in Protherm database. Mutations are shown in form of: W-type residue number Mutant. For multiple mutations, all individual mutations are listed below it individually. LNNC means Loss-number-of-native-contact. LNNC% represents percentage of LNNC as divided by the total number of links in the corresponding proteins (see Table 1). The reportecd LNNC% in the forth column shows the first LNNC% that the mutated residue is located at distance (D) less than 4.5 Å. residues that are not located in the hot spot during the first 25% LNNC are marked as C in this column. Depth shows the distance to water molecules interacting to the surface of the protein.

Entry	PDB	Mutation	ΔT_m	LNNC	Depth
1784	1N0J	Y 34 F	15.9	0.5:1.3	4.49
1785	1N0J	Q 143 N	15.3	3.7:4.5	6.43
1230	2CBR	R 111 Q, R 131 Q	14.5		
		131		9.1:3.7	6.22
		111		0.2:4.1	8.97
1230	2CBR	R 131 Q	11.6	9.1:3.7	6.22
1230	2CBR	R 111 Q	9.4	0.2:4.1	8.97
4070	3SSI	D 83 C	17.25	0.2:4.3	3.37
9773	451C	F 7 A, V 13 M, F 34 Y, E 43 Y, V 78 I	32.9		
		34		0.3:4.2	3.97
		7		2.4:1.3	4.15
		13		4.0:3.2	3.35
		43		7.5:3.2	3.05
		78		3.2:1.3	6.8
6405	451C	F 34 Y, E 43 Y	20.3		
		34		0.3:4.2	3.97
		43		7.5:3.2	3.05

6407	451C	F 34 Y, Q 37 R, E 43 Y	17.5		
		34		0.3:4.2	3.97
		37		5.1:3.7	3.22
		43		7.5:3.2	3.05
6401	451C	F 34 Y	16	0.3:4.2	3.97
	451C	F 7 A, V 13 M, F 34 Y, E 43 Y	14.8		
		13		4.0:3.2	3.35
		7		2.4:1.3	4.15
		34		0.3:4.2	3.97
		43		7.5:3.2	3.05
1518	451C	F 7 A, V 13 M, V 78 I	13		
		7		2.4:1.3	4.15
		13		4.0:3.2	3.35
		78		3.2:1.3	6.8
6404	451C	F 34 Y, Q 37 R	12.5		
		34		0.3:4.2	3.97
		37		5.1:3.7	3.22
6400	451C	F 7 A, V 13 M	12		
		7		2.4:1.3	4.15
		13		4.0:3.2	3.35
1518	451C	F 34 Y, E 43 Y, V 78 I	10.7		
		34		0.3:4.2	3.97
		43		7.5:3.2	3.05
		78		3.2:1.3	6.8
6398	451C	F 7 A	9.5	2.4:1.3	4.15
2015	451C	Y 27 F	8.9	0.5:3.6	7.09
6408	451C	V 78 I	8.4	3.2:1.3	6.8
6403	451C	E 43 Y	5.1	7.5:3.2	3.05
1996	2HPR	A 2 R, N 71 E	10.1		
		2		0.8:0	3.93
		71		0.8:0	3.87
8930	1KEV	S 24 P, L 316 P	11.1		
		24		C	3.07
		316		0.5:3.0	6.59
8927	1KEV	L 316 P	10.8	0.5:3.0	6.59
1755	1KEV	V 224 E, S 254 K, Q 165 E, M 304 R	8.6		
		254		2.4:3.2	4.05
		224		2.6:0	3.93
		165		C	3.07
		304		5.1:1.3	3.0
1755	1KEV	V 224 E, S 254 K, Q 165 E, M 304 R, R 238 A	8.5		
		254		2.4:3.2	4.05
		224		2.6:0	3.93
		165		C	3.07
		304		5.1:1.3	3.0
		238		7.2:3.6	4.54
8929	1KEV	A 22 P, S 24 P	6.1		
		22		C	3.46
		24		C	3.07
1755	1KEV	V 224 E, S 254 K	5.4		
		254		2.4:3.2	4.05
		224		2.6:0	3.93

1755	1KEV	Q 165 E, M 304 R	5.3		
		165		C	3.07
		304		5.1:1.3	3.0
1270	1HFZ	D 87 N	32.4	1.5:1.3	3.67
1971	1HFZ	D 37 N	14.4	0.9:3	3.33
6611	1HFZ	K 114 N	10.7	4.4:1.3	2.98
1970	1HFZ	E 1 M	10	0.9:0	3.04
1971	1HFZ	E 11 L	9.7	8.7:2.8	3.67
1971	1HFZ	E 7 Q	7.9	0.3:1.3	3.1
6601	1HFZ	L 110 H	6.3	4.9:1.3	3.21
1854	1CSP	M 1 R, E 3 K, K 65 I, E 66 L	31.2		
		1		2.1:3.8	3.07
		3		3.9:4.2	3.17
		65		1.1:3	3.32
		66		2.1:32.	3.53
2001	1FTG	E 20 K, E 72 K	9.5		
		20		1.4:3.4	3.15
		72		2.2:4.4	3.1
2000	1FTG	Q 99 A	8.5	1.8:3.2	10.38
1798	1FTG	D 126 K	8	13.7:2.	3.19
2000	1FTG	D 75 K	7.8	2.2:1.3	3.19
1798	1FTG	E 72 K	7.6	2.2:4.4	3.1
2000	1FTG	G 68 A	6.7	2.8:1.3	3.29
1798	1FTG	E 40 K	6.6	2.5:1.3	2.99
1804	1FTG	T 122 S	5.9	C	3.68
2000	1FTG	A 101 V	5.4	C	8.75
1798	1FTG	E 20 K	5.2	1.4:3.4	3.15
1853	1CSP	M 1 R, E 3 K, K 65 I	29.9		
		1		2.1:3.8	3.07
		3		3.9:4.2	3.17
		65		1.1:3	3.32
1213	1CSP	E 3 R, T 64 V, E 66 L	23.2		
		3		3.9:4.2	3.17
		64		0.7:4.1	3.18
		66		2.1:32.	3.53
1854	1CSP	A 46 K, S 48 R, E 66 L	23.2		
		46		2.1:4.3	4.3
		48		2.1:3.1	3.24
		66		2.1:32.	3.53
1853	1CSP	M 1 R, E 3 K	23.2		
		1		2.1:3.8	3.07
		3		3.9:4.2	3.17
1853	1CSP	M 1 R, K 65 I, E 66 K	21.9		
		1		2.1:3.8	3.07
		65		1.1:3	3.32
		66		2.1:32.	3.53
1213	1CSP	E 3 R, E 66 L	21		
		3		3.9:4.2	3.17
		66		2.1:32.	3.53
1854	1CSP	E 3 R, A 46 K, S 48 R	20.7		

		3		3.9:4.2	3.17
		46		2.1:4.3	4.3
		48		2.1:3.1	3.24
1854	1CSP	E 3 K, K 65 I, E 66 K	18		
		3		3.9:4.2	3.17
		65		1.1:3	3.32
		66		2.1:32.	3.53
1852	1CSP	E 43 K, A 46 K, S 48 K	17.4		
		43		6.0:4.3	3.01
		46		2.1:4.3	4.3
		48		2.1:3.1	3.24
1854	1CSP	E 3 R, A 46 K, S 48 R, E 66 L	17		
		3		3.9:4.2	3.17
		46		2.1:4.3	4.3
		48		2.1:3.1	3.24
		66		2.1:2.9	3.53
1852	1CSP	E 3 K	16.6	3.9:4.2	3.17
1852	1CSP	A 46 K, S 48 R	15.9		
		46		2.1:4.3	4.3
		48		2.1:3.1	3.24
1853	1CSP	M 1 R, E 66 K	15.7		
		1		2.1:3.8	3.07
		66		2.1:2.9	3.53
1853	1CSP	E 66 L, A 67 P	13		
		66		2.1:2.9	3.53
		67		16.9:0	3.06
1853	1CSP	E 66 K	12.9	2.1:2.9	3.53
1852	1CSP	M 1 R	10.4	2.1:3.8	3.07
1853	1CSP	K 65 I	9.6	1.1:3	3.32
1854	1CSP	S 48 R	8.9	2.1:3.1	3.24
1854	1CSP	A 46 K	8.4	2.1:4.3	4.3
1853	1CSP	M 1 R, E 3 K, E 66 K	7.3		
		1		2.1:3.8	3.07
		3		3.9:4.2	3.17
		66		2.1:32.	3.53
5898	1SSO	F 31 A	27	1.6:2.3	6.15
2585	1JK9	G 85 R	45	18.3:4.	6.43
2585	1JK9	G 93 A	20	13.0:3.	4.74
7022	2RN2	D 10 A	14.8	4.0:3.9	4.86
2166	2RN2	A 52 V, V 74 L	8.07		
		52		3.8:3.3	9.31
		74		3.8:0	7.65
753	2RN2	A 52 V	7.8	3.8:3.3	9.31
681	2RN2	D 134 H	7	17.7:3.	3.46
12	2RN2	K 95 G	6.8	8.8:4.3	3.13
2148	2RN2	K 91 R, K 95 G	5.8		
		91		8.8:4.4	3.08
		95		8.8:4.3	3.13
2535	2RN2	K 91 R, D 94 E, K 95 G	5.6		
		91		8.8:4.4	3.08
		95		8.8:4.3	3.13
		94		10.7:0	3.15

152	2RN2	D 70 N	5.5	4.0:	3.83
2146	2RN2	D 94 R, K 95 G	5.5		
		95		8.8:4.3	3.13
		94		10.7:0	3.15
709	2RN2	H 62 A, H 83 A, H 124 A, H 127 A	5.5		
		62		1.7:1.3	3.1
		83		C	3.09
		124		15.8:1.	3.14
		127		13.4:3.	3.14
2563	1RGG	D 25 K, E 74 K, S 31 P, S 42 G, S 48 P, T 76	27.9		
		25		0.6:4	3.07
		74		4.0:2.4	3.02
		31		0.6:4.2	3.1
		42		0.9:2.7	3.34
		48		1.2:1.3	3.09
		76		13.5:0	3.03
		79		13.8:3.	5.41
		77		14.5:1.	3.1
1905	1RGG	D 79 F	13.1	13.8:3.	5.41
1876	1PGA	T 16 I, T 18 I, T 25 E, V 29 F	24.7		
		16		5.4:4.3	3.3
		18		0.5:3.8	3.62
		25		0.5:3.6	3.34
		29		1.0:3.2	3.59
1876	1PGA	T 16 I, T 18 I, V 29 W	22.2		
		16		5.4:4.3	3.3
		18		0.5:3.8	3.62
		29		1.0:3.2	3.59
3006	1PGA	T 53 Y	12.17	10.9:4.	3.82
2196	1PGA	V 21 P, A 23 P	6.9		
		21		0.5:0	2.99
		23		1.0:3.2	3.64
2196	1PGA	V 21 P	6.2	0.5:0	2.99
1644	1SHG	A 11 V, V 23 L, M 25 V, V 44 I, V 58 L	17.1		
		11		2.1:4	4.83
		23		3.4:3.8	5.72
		25		3.4:3.4	5.55
		44		2.1:4.3	7.18
		58		2.1:3.9	3.68
1644	1SHG	D 48 G	11.3	3.4:2.5	3.12
1885	2TRX	L 79 C, T 89 C	9.8		
		79		0.6:3.8	6.23
		89		0.6:1.3	3.24
1885	2TRX	T 77 C, V 91 C	7		
		77		3.0:3.7	5.63
		91		1.7:1.3	3.32
1989	1TEN	Q 808 K, L 820 K, D 850 K, T 890 K	9.6		
		808		4.0:1.3	3.03
		850		4.8:4.3	3.05
		820		4.0:3.7	3.67

		890		7.1:3.4	3.35
2968	1YCC	N 52 I, C 102 A	25		
		52		1.2:2.8	5.96
		102		4.8:1.3	4.87
2963	1YCC	N 52 I	17.2	1.2:2.8	5.96
2965	1YCC	C 102 A	11.6	4.8:1.3	4.87
4783	1YCC	V 20 C	6.8	6.6:0	4.48
3563	1YCC	K 73 V	6.1	4.2:3.1	2.99
1598	1DIV	K 12 M	16.3	0.2:3.1	3.03
1643	1DIV	D 23 A	7	2.6:0	3.0
1663	1RTB	F 46 A	22.4	0.9:3.9	7.89
2013	1RTB	S 123 A	5.9	4.2:1.3	3.58
1669	1RTP	H 26 P	5.6	1.2:1.3	4.15
1669	1RTP	A 21 P	5.8	3.2:0	3.12
2027	1IO2	E 8 A	8	4.6:3.6	5.59
2027	1IO2	D 135 A	5.2	5.5:0	5.1
1678	1AQH	K 300 R, N 150 D, V 196 F, Q 164 I, T 232 V	7.8		
		232		12.9:4	6.18
		164		0.3:2.8	3.69
		196		1.1:0	3.47
		150		1.3:2.1	3.13
		300		0.5:3.9	10.31
2291	1OSI	L 134 N, V 181 T, P 324 T, A 335 E	5.5		
		324		21.6:3.	6.18
		181		7.9:4.6	6.94
		134		2.4:3.8	5.08
		335		C	3.18
		324		21.6:3.	
1989	1AYE	Q28A E, Q 32A K, K41E, E48K,	9.8		
		32		1.7:2.9	4.27
		69		8.4:3.1	7.29
		41		2.8:1.3	4.24
		78		0.2:1.3	10.74
		48		11.9:3.	4.41
		73		0.2:3.3	3.75
		28		12.4:2.	3.09
1990	2ACY	K 24 E, E 63 K, N 82 K, Q 95 K	5.4		
		82		10.2:3.	3.06
		63		0.5:3.2	3.15
		60		0.5:1.3	3.13
		95		1.1:1.3	3.15
2467	1P3J	L 3 I, G 17 A, D 23 K, K 69 R, G 73 S, D 75 S,	11.6		
		3		1.7:3.4	6.57
		69		0.5:0	3.04
		103		0.5:2.7	3.31
		73		0.5:3.5	3.39
		75		2.1:3.1	2.97
		114		0.4:2	3.23
		17		0.1:1.3	6.69
		118		0.2:10	3.12
		105		2.1:3.1	3.84
		23		0.1:2.5	3.06

2467	1P3J	L 3 I, G 17 A, D 23 K, K 69 R, G 73 S, D 75 S, I	12.5		
		3		1.7:3.4	6.57
		69		0.5:0	3.04
		103		0.5:2.7	3.31
		73		0.5:3.5	3.39
		75		2.1:3.1	2.97
		114		0.4:2	3.23
		17		0.1:1.3	6.69
		99		0.5:0	4.6
		105		2.1:3.1	3.84
		23		0.1:2.5	3.06
2467	1P3J	D 23 K, Y 103 M, E 114 Q, S 169 T, Q 180 A,	5		
		169		1.0:3.3	3.16
		193		1.8:4.1	4.22
		103		0.5:2.7	3.31
		208		0.1:3.2	5.96
		210		1.5:1.3	3.11
		114		0.4:2	3.23
		180		5.9:1.3	3.13
		187		11.7:4.	3.0
		23		0.1:2.5	3.06
2490	2OCJ	M 133 L, V 203 A, N 239 Y, N 268 D	5.7		
		133		C	11.12
		203		0.4:20	5.54
		239		6.5:4.5	5.15
		268		10.8:0	6.52
2490	2OCJ	M 133 L, V 203 A, N 239 Y, N 268 D, Y 236 F,	6		
		133		C	11.12
		203		0.4:20	5.54
		239		6.5:4.5	5.15
		268		10.8:0	6.52
		236		9.4:3.5	10.99
		253		10.5:1.	10.85
3062	4BLM	K 234 A	5.2	C	8.44
3333	1SHF	S 124 K	5.4	7.0:4.5	4.69
5539	1RN1	W 59 Y	20	2.0:3.3	5.73
6683	1LZ1	Q 58 G	5.7	12.1:3.	4.8
8299	2AFG	H 93 G	7.3	0.9:1.3	3.13
1825	2AFG	H 21 Y, L 44 F	5.2		
			21	2.3:1.3	4.54
			44	0.3:1.3	6.71
1825	2AFG	H 21 Y, H 102 Y	5		
			21	2.3:1.3	4.54
			102	1.2:0	4.59
1826	2AFG	H 21 Y, L 44 F, H 102 Y	6.7		
		21		2.3:1.3	4.54
		44		0.3:1.3	6.71
		102		1.2:0	4.59
1826	2AFG	H 21 Y, L 44 F, F 108 Y	6		
		21		2.3:1.3	4.54

		44		0.3:1.3	6.71
**		108		1.4:3.2	5.64
1826	2AFG	H 21 Y, H 102 Y, F 108 Y	6.5		
		21		2.3:1.3	4.54
		102		1.2:0	4.59
		108		1.4:3.2	5.64
1826	2AFG	H 21 Y, L 44 F, H 102 Y, F 108 Y	7.8		
		21		2.3:1.3	4.54
		102		1.2:0	4.59
		108		1.4:3.2	5.64
		44		0.3:1.3	6.71
1057	2CI2	P 44 A	9.1	0.4:3.6	3.05
1269	1DIL	A 53 L	5	4.2:3.8	12.86
1580	1QLP	K 331 F	6	1.2:3	5.08
1639	1H8V	A 35 V	7.7	0.2:3.9	3.51
1698	1AZP	V 30 I	5.8	1.2:3.8	6.47
1720	1B5M	D 60 R	6	4.8:1.3	3.0
2368	1B5M	R 15 H, E 20 S	14.8		
		15		3.3:3.3	3.16
		20		3.3:1.3	2.98
2015	1YNR	Y 25 F	6.2	0.3:3.5	7.48
2571	1JU3	T 172 R, G 173 Q	6.41		
		172		0.1:1.3	3.5
		173		0.1:3.3	5.83
2571	1JU3	L 169 K	6.3	0.1:3.3	3.7
2232	1BCX	S 100 C, N 148 C	7.5		
		100		9.9:3	4.0
		148		8.4:3.2	3.04
2237	1IRO	V 24 I, I 33 L	8.2		
		24		2.0:3	4.88
		33		2.0:1.3	4.36
2237	1IRO	Y 4 W, V 24 I, I 33 L	6.7		
		24		2.0:3	4.88
		33		2.0:1.3	4.36
		4		0.5:2.7	3.58
1678	1GUY	T 187 C	15	2.2:3.5	3.73
1725	1GUY	E 165 Q	23.6	2.2:3.4	3.03
1726	1GUY	E 165 K	23.9	2.2:3.4	3.03

Table S4-2. Single mutation data extracted from Table S4-1.

Entry	**PDB**	**Mutation**	**ΔTm**	**%**	**Entry**	**PDB**	**Mutation**	**ΔTm**	**%**
25715	1JU3	L 169 K	6.3	0.11	6398	451C	F 7 A	9.5	2.43
15985	1DIV	K 12 M	16.3	0.16	17986	1FTG	E 40 K	6.6	2.5
12301	2CBR	R 111 Q	9.4	0.17	16436	1DIV	D 23 A	7	2.59
4070	3SSI	D 83 C	17.25	0.21	20004	1FTG	G 68 A	6.7	2.77
16399	1H8V	A 35 V	7.7	0.23	16695	1RTP	A 21 P	5.8	3.16
19710	1HFZ	E 7 Q	7.9	0.29	6408	451C	V 78 I	8.4	3.23
20154	1YNR	Y 25 F	6.2	0.32	16448	1SHG	D 48 G	11.3	3.42
10572	2CI2	P 44 A	9.1	0.39	17855	1N0J	Q 143 N	15.3	3.67
17847	1N0J	Y 34 F	15.9	0.47	753	2RN2	A 52 V	7.8	3.79
21961	1PGA	V 21 P	6.2	0.5	18529	1CSP	E 3 K	16.6	3.87
20156	451C	Y 27 F	8.9	0.54	152	2RN2	D 70 N	5.5	3.96
8927	1KEV	L 316 P	10.8	0.55	7022	2RN2	D 10 A	14.8	3.96
19709	1HFZ	E 1 M	10	0.87	20130	1RTB	S 123 A	5.9	4.17
19714	1HFZ	D 37 N	14.4	0.87	12699	1DIL	A 53 L	5	4.18
16630	1RTB	F 46 A	22.4	0.91	3563	1YCC	K 73 V	6.1	4.22
8299	2AFG	H 93 G	7.3	0.94	6611	1HFZ	K 114 N	10.7	4.36
18530	1CSP	K 65 I	9.6	1.06	20274	1IO2	E 8 A	8	4.62
15801	1QLP	K 331 F	6	1.17	2965	1YCC	C 102 A	11.6	4.82
16988	1AZP	V 30 I	5.8	1.2	17202	1B5M	D 60 R	6	4.83
2963	1YCC	N 52 I	17.2	1.2	6601	1HFZ	L 110 H	6.3	4.94
16691	1RTP	H 26 P	5.6	1.22	20277	1IO2	D 135 A	5.2	5.48
17985	1FTG	E 20 K	5.2	1.39	4783	1YCC	V 20 C	6.8	6.63
12705	1HFZ	D 87 N	32.4	1.45	3333	1SHF	S 124 K	5.4	7.02
5898	1SSO	F 31 A	27	1.64	6403	451C	E 43 Y	5.1	7.55
20008	1FTG	Q 99 A	8.5	1.8	19711	1HFZ	E 11 L	9.7	8.72
5539	1RN1	W 59 Y	20	1.99	12	2RN2	K 95 G	6.8	8.78
18546	1CSP	A 46 K	8.4	2.11	12302	2CBR	R 131 Q	11.6	9.14
18548	1CSP	S 48 R	8.9	2.11	3006	1PGA	T 53 Y	12.17	10.89
18528	1CSP	M 1 R	10.4	2.11	6683	1LZ1	Q 58 G	5.7	12.14
18531	1CSP	E 66 K	12.9	2.11	25853	1JK9	G 93 A	20	12.96
16788	1GUY	T 187 C	15	2.16	17982	1FTG	D 126 K	8	13.73
17259	1GUY	E 165 Q	23.6	2.16	19053	1RGG	D 79 F	13.1	13.85
17260	1GUY	E 165 K	23.9	2.16	681	2RN2	D 134 H	7	17.73
17987	1FTG	E 72 K	7.6	2.22	25855	1JK9	G 85 R	45	18.31
20006	1FTG	D 75 K	7.8	2.22					

Table S4-3. HB analysis results. Successful stabilizing/destabilizing mutation effect predictions using rules introduced in methods section.

No	Entry	PDB	Mutation	ΔT_m	No	Entry	PDB	Mutation	ΔT_m
1	6408	451C	V 78 I	8.4	74	17377	1SHF	I 133 A	-35.9
2	6615	1HFZ	V 42 A	-4.5	75	17378	1SHF	I 133 V	-2.2
3	6616	1HFZ	V 42 G	-5.2	76	17362	1SHF	I 111 S	-33.5
4	6621	1HFZ	W 104 Y	-12.5	77	17341	1SHF	F 87 S	-27.2
5	6607	1HFZ	A 106 S	-6.1	78	17363	1SHF	I 111 A	-27.1
6	7903	1PGA	F 30 H	-53	79	17372	1SHF	A 122 S	-26.9
7	3099	1N0J	I 58 T	-13.6	80	17367	1SHF	W 120 S	-25.6
8	2586	3SSI	V 13 G	-27.7	81	17356	1SHF	F 103 A	-22.3
9	2370	3SSI	M 103 G	-13.57	82	17368	1SHF	W 120 A	-21.6
10	675	2RN2	G 23 A	1.8	83	17346	1SHF	A 89 G	-18.6
11	752	2RN2	A 24 V	3.2	84	17382	1SHF	V 138 A	-17.1
12	753	2RN2	A 52 V	7.8	85	17373	1SHF	A 122 G	-12
13	7020	2RN2	I 53 A	-12.1	86	24666	1SHF	L 3 A	-10.5
14	749	2RN2	V 74 A	-12.7	87	17342	1SHF	F 87 A	-8.9
15	747	2RN2	V 74 L	3.3	88	17364	1SHF	I 111 V	-2.2
16	748	2RN2	V 74 I	2.1	89	22606	1SHF	V 74 T	-16.9
17	3097	2TRX	L 78 R	-16.1	90	22509	1SHF	V 99 T	-15.2
18	3096	2TRX	L 78 K	-13.7	91	22504	1STN	V 23 T	-14.7
19	18029	1FTG	L 6 A	-11.8	92	22602	1STN	V 23 T	-14
20	19998	1FTG	V 18 I	4.9	93	22515	1STN	V 66 S	-13.9
21	18030	1FTG	I 22 V	-6.3	94	2868	1STN	A 69 T	-13.7
22	18052	1FTG	I 51 V	-4.7	95	22613	1STN	V 66 S	-12.9
23	18053	1FTG	I 52 V	-4.1	96	2866	1STN	V 23 A	-12.5
24	18050	1FTG	V 31 A	-4.8	97	2948	1STN	V 66 A	-12.2
25	20004	1FTG	G 68 A	6.7	98	2865	1STN	L 7 A	-3.2
26	20007	1FTG	V 83 I	1	99	109	1STN	L 25 A	-11.5
27	18055	1FTG	A 84 G	-5.4	100	2952	1STN	V 66 A	-11.1
28	18044	1FTG	L 143 A	-1.5	101	22510	1STN	V 104 T	-10.3
29	18046	1FTG	V 160 A	-8.2	102	22507	1STN	V 66 T	-6.3
30	6151	1MBG	V 103 L	22.4	103	110	1STN	V 66 L	3.9
31	6150	1MBG	V 103 I	7	104	1000	1WQ5	P 28 S	-6.4
32	6148	1MBG	V 103 A	-36.8	105	996	1WQ5	G 51 D	-4.5
33	20133	1RTB	A 4 S	-1.54	106	995	1WQ5	P 21 S	-4.3
34	20135	1RTB	A 5 S	-1.11	107	984	1WQ5	P 78 S	-1.6
35	3222	1RTB	P 114 G	-9.5	108	4691	2CI2	L 27 A	-22.5
36	11641	1RN1	V 16 C	-9.6	109	4697	2CI2	V 70 A	-16.8
37	11640	1RN1	V 16 T	-11.3	110	4693	2CI2	I 39 V	-14.8
38	11639	1RN1	V 16 S	-15.6	111	4702	2CI2	L 51 A	-9.1
39	11638	1RN1	V 16 A	-8.2	112	4701	2CI2	V 53 T	-8.3
40	2575	1RN1	W 59 Y	-4.2	113	4694	2CI2	I 48 V	-7.2
41	11653	1RN1	V 78 T	-11.3	114	4692	2CI2	V 38 A	-4.7
42	11642	1RN1	V 78 A	-7.9	115	4700	2CI2	V 53 A	-2.8
43	11643	1RN1	V 78 S	-16.6	116	4703	2CI2	L 51 V	-1
44	11645	1RN1	V 89 S	-21.6	117	7370	2LZM	L 66 P	-38.2
45	11646	1RN1	V 89 T	-13.2	118	7369	2LZM	L 91 P	-34.3
46	11647	1RN1	V 89 C	-4.9	119	1524	2LZM	L 99 A, F 153 A	-22.8
47	661	1ARR	L 12 A	-15.6	120	1253	2LZM	L 133 D	-17.9
48	667	1ARR	W 14 A	-26.4	121	1514	2LZM	L 99 A	-11.4

49	637	1ARR	V 18 A	-2.1	122	1519	2LZM	F 153 A	-9.3
50	849	1ARR	L 19 Q	-19	123	1327	2LZM	I 3 D	-8.5
51	662	1ARR	L 21 A	-18.3	124	1314	2LZM	I 3 W	-8
52	659	1ARR	V 33 A	-13.8	125	1070	2LZM	L 46 A	-6.4
53	664	1ARR	M 42 A	-22.3	126	1323	2LZM	I 3 T	-6
54	16988	1AZP	V 30 I	5.8	127	1332	2LZM	A 146 T	-6
55	18642	1AZP	W 24 A	-9.6	128	1320	2LZM	I 3 C	-3.7
56	15591	1B5M	L 71 S	-10.2	129	1315	2LZM	I 3 Y	-5.9
57	3110	1CYO	F 35 H	-11.3	130	1325	2LZM	I 3 G	-5.8
58	9184	1CYO	V 45 E	-9.9	131	1326	2LZM	I 3 E	-5.7
59	6558	1CYO	V 61 K	-8.7	132	1515	2LZM	L 99 V	-5.2
60	12698	1DIL	A 69 V	3	133	1324	2LZM	I 3 S	-4.6
61	12699	1DIL	A 53 L	5	134	1356	2LZM	P 86 H	-4
62	16402	1H8V	A 35 S	-4	135	1354	2LZM	P 86 R	-3
63	16399	1H8V	A 35 V	7.7	136	1322	2LZM	I 3 A	-1.8
64	16404	1H8V	G 41 A	2.5	137	1318	2LZM	I 3 V	-1.2
65	17701	1HME	I 34 H	-4.4	138	1355	2LZM	P 86 D	-1
66	17702	1HME	G 35 H	-1.1	139	1818	4LYZ	I 55 T	-13.1
67	6819	1IOJ	G 15 A	7	140	1817	4LYZ	I 55 A	-11.6
68	25134	1PIN	F 25 A	-26.1	141	1814	4LYZ	I 55 F	-6.5
69	25116	1PIN	M 15 A	-6.8	142	1812	4LYZ	I 55 M	-6
70	25128	1PIN	V 22 A	-4.4	143	1770	4LYZ	F 3 Y	-1.2
71	4726	1QU7	A 438 V	3.1	144	1783	4LYZ	A 31 V	3.1
72	3106	1ROP	L 41 A	-18.1	145	1784	4LYZ	A 31 I	3.6
73	3105	1ROP	L 41 V	-5.7	146	1785	4LYZ	A 31 L	4.7

Table S4-4. This table shows those mutations that rules presented in methods section fails to predict the effect of corresponding mutation on thermostability.

No.	Reference	PDB ID	Mutation	dT_m
1	17375	1SHF	A 39 L	-31.7
2	17381	1SHF	V 55 S	1.5
3	935	1WQ5	G 211 E	1.8
4	4696	2CI2	I 49 V	1.4
5	6601	1HFZ	L 110 H	6.3
6	2588	3SSI	V 13 I	-2.4
7	20001	1FTG	I 21 G	3.8
8	653	1ARR	G 30 A	-10

Appendix A3: Supplementary Information for Chapter 5

Table S5-1. Protein structures and their corresponding mutations extracted from Protherm database. The second column shows the reference number of mutation in Protherm database and the last column indicates whether the prediction using CC is in agreement with experiments or not. If it is yes it means that the consensus method has correctly predicted if the thermostability change by mutation is stabilizing ($\Delta T_m>0$) or destabilizing ($\Delta T_m<0$).

No	Ref.	PDB ID	Mutation	ΔT_m	Acceptance
1	1091	2LZM	T 26 S	1.3	No
2	3709	2LZM	H 31 N	-11	Yes
3	346	2LZM	S 38 D	1.5	Yes
4	1092	2LZM	N 40 D	1.28	Yes
5	1093	2LZM	A 41 D	1.1	Yes
6	1067	2LZM	K 43 A	-3.08	Yes
7	1070	2LZM	L 46 A	-6.4	Yes
8	1071	2LZM	D 47 A	-2.75	No
9	1072	2LZM	K 48 A	-1.68	Yes
10	1329	2LZM	C 54 V	-2	No
11	1330	2LZM	C 54 T	1	Yes
12	1290	2LZM	N 55 G	-1.6	Yes
13	1538	2LZM	T 59 A	-4	Yes
14	1537	2LZM	T 59 V	-4	Yes
15	1536	2LZM	T 59 G	-4.1	Yes
16	1535	2LZM	T 59 D	-3.1	Yes
17	1533	2LZM	T 59 N	-2.8	Yes
18	7370	2LZM	L 66 P	-38.2	Yes
19	1175	2LZM	Q 69 P	-7.6	Yes
20	1176	2LZM	D 72 P	-6.9	Yes
21	1183	2LZM	A 74 P	-12.4	Yes
22	200	2LZM	A 82 P	2.1	No

23	377	2LZM	K 83 H	-1	Yes
24	1354	2LZM	P 86 R	-3	Yes
25	374	2LZM	S 90 H	-2.7	Yes
26	7369	2LZM	L 91 P	-34.3	Yes
27	2206	2LZM	D 92 N	-3.7	No
28	2931	2LZM	R 96 H	-7.8	Yes
29	1515	2LZM	L 99 V	-5.2	Yes
30	1514	2LZM	L 99 A	-11.4	Yes
31	1516	2LZM	L 99 I	-3.7	Yes
32	1517	2LZM	L 99 M	-1.5	Yes
33	1198	2LZM	M 102 L	-2.31	No
34	1142	2LZM	Q 105 A	-1.6	Yes
35	1143	2LZM	Q 105 E	-3	Yes
36	1144	2LZM	Q 105 G	-3.9	Yes
37	2207	2LZM	T 109 D	1.5	No
38	1199	2LZM	V 111 F	-4.63	Yes
39	1200	2LZM	V 111 I	-2.6	Yes
40	1097	2LZM	G 113 E	1.03	No
41	348	2LZM	N 116 D	1.6	No
42	367	2LZM	Q 123 E	1.2	No
43	1253	2LZM	L 133 D	-17.9	Yes
44	370	2LZM	N 144 E	1.5	No
45	350	2LZM	N 144 D	1.4	No
46	1332	2LZM	A 146 T	-6	Yes
47	18529	1CSP	E 3 K	16.6	Yes
48	12131	1CSP	E 3 L	9.1	No
49	12130	1CSP	E 3 R	16	No
50	3525	1CSP	F 17 A	-11.3	Yes
51	3524	1CSP	F 27 A	-6	Yes
52	3527	1CSP	F 38 A	2.2	No
53	12132	1CSP	A 46 E	-5	No
54	18546	1CSP	A 46 K	8.4	Yes
55	18548	1CSP	S 48 R	8.9	No
56	18530	1CSP	K 65 I	9.6	No
57	12133	1CSP	E 66 L	12.8	Yes
58	18531	1CSP	E 66 K	12.9	Yes
59	16643	1RGG	T 16 V	1	Yes
60	16638	1RGG	T 18 V	-4.7	Yes
61	19020	1RGG	D 33 A	-16	Yes
62	16647	1RGG	V 36 T	-4.6	Yes
63	16728	1RGG	Q 38 A	3.5	No
64	3409	1RGG	N 39 D	-5.2	Yes
65	3411	1RGG	N 39 A	-7.6	Yes
66	17862	1RGG	N 39 S	-8	Yes
67	17863	1RGG	N 39 A	-7.6	Yes
68	16729	1RGG	E 41 K	-2.5	Yes
69	16648	1RGG	V 43 T	-1.6	Yes
70	16737	1RGG	E 54 Q	-7.9	Yes
71	16639	1RGG	T 56 V	-6.3	Yes
72	16649	1RGG	V 57 T	-15	Yes
73	16644	1RGG	T 59 V	-5.6	Yes
74	16731	1RGG	R 65 A	-3.4	Yes

75	16732	1RGG	E 74 K	3.1	No
76	19053	1RGG	D 79 F	13.1	No
77	19052	1RGG	D 79 A	11.7	No
78	19051	1RGG	D 79 N	7.6	No
79	19014	1RGG	D 79 K	7.6	No
80	19013	1RGG	D 79 L	8.7	No
81	19012	1RGG	D 79 R	9	No
82	19011	1RGG	D 79 I	9.6	Yes
83	19009	1RGG	D 79 Y	9.6	No
84	19015	1RGG	D 79 W	7.6	Yes
85	19016	1RGG	D 79 H	5.6	No
86	16641	1RGG	T 82 V	-5.7	Yes
87	19998	1FTG	V 18 I	4.9	No
88	17936	1FTG	E 20 K	-1.5	Yes
89	20001	1FTG	I 21 G	3.8	No
90	18030	1FTG	I 22 V	-6.3	Yes
91	18031	1FTG	V 31 A	-4.7	No
92	17979	1FTG	E 40 K	4.5	No
93	18033	1FTG	I 51 V	-4.1	Yes
94	18034	1FTG	I 52 V	-1.4	Yes
95	17947	1FTG	E 61 K	2.7	No
96	17939	1FTG	D 65 K	-2.1	Yes
97	20004	1FTG	G 68 A	6.7	Yes
98	18035	1FTG	S 71 A	-1	Yes
99	17940	1FTG	E 72 K	1.2	No
100	17988	1FTG	D 75 K	4	No
101	20007	1FTG	V 83 I	1	No
102	18036	1FTG	A 84 G	-4.3	Yes
103	18037	1FTG	N 97 A	-2.4	Yes
104	20008	1FTG	Q 99 A	8.5	No
105	20009	1FTG	A 101 V	5.4	No
106	18039	1FTG	I 104 V	-4	Yes
107	20010	1FTG	E 107 A	2.7	No
108	18040	1FTG	S 110 A	-1.7	No
109	20011	1FTG	Q 111 G	3.7	No
110	18060	1FTG	V 117 A	-6.4	Yes
111	18042	1FTG	T 122 S	5.9	No
112	17942	1FTG	D 126 K	4.6	No
113	18062	1FTG	V 139 A	-1.5	Yes
114	18044	1FTG	L 143 A	-1.5	Yes
115	17983	1FTG	D 150 K	-6.2	Yes
116	18064	1FTG	I 156 V	-8.7	Yes
117	19710	1HFZ	E 7 Q	7.9	No
118	19711	1HFZ	E 11 L	9.7	No
119	19712	1HFZ	D 14 N	-3	Yes
120	19713	1HFZ	E 25 A	1.1	No
121	6600	1HFZ	H 32 A	-11.6	Yes
122	19714	1HFZ	D 37 N	14.4	Yes
123	6615	1HFZ	V 42 A	-4.5	Yes
124	6616	1HFZ	V 42 G	-5.2	No

125	6614	1HFZ	V 42 N	-1.1	No
126	6617	1HFZ	Q 54 A	-1.9	Yes
127	6618	1HFZ	I 59 W	-4.5	No
128	12704	1HFZ	D 87 N	6.8	No
129	6619	1HFZ	Y 103 P	-1.1	Yes
130	6620	1HFZ	Y 103 A	-10.7	No
131	6621	1HFZ	W 104 Y	-12.5	Yes
132	6607	1HFZ	A 106 S	-6.1	Yes
133	6609	1HFZ	H 107 A	-4.1	Yes
134	6608	1HFZ	H 107 Y	-1	No
135	6610	1HFZ	H 107 W	-8.7	No
136	6601	1HFZ	L 110 H	6.3	Yes
137	6603	1HFZ	L 110 E	-1.1	Yes
138	6602	1HFZ	L 110 R	1.9	Yes
139	6613	1HFZ	K 114 E	-3.2	Yes
140	6611	1HFZ	K 114 N	10.7	Yes
141	6612	1HFZ	K 114 Q	-2.7	Yes
142	6604	1HFZ	Q 117 A	-4.1	Yes
143	6606	1HFZ	W 118 H	-3.3	Yes
144	6605	1HFZ	W 118 Y	-5.4	Yes
145	2590	3SSI	V 13 M	-12.6	Yes
146	2589	3SSI	V 13 L	-4.5	Yes
147	2588	3SSI	V 13 I	-2.4	Yes
148	2587	3SSI	V 13 F	-13.5	Yes
149	2586	3SSI	V 13 G	-27.7	Yes
150	2585	3SSI	V 13 A	-13.2	Yes
151	3081	3SSI	M 73 I	-1.78	Yes
152	3079	3SSI	M 73 V	-1.17	Yes
153	3075	3SSI	M 73 D	3.08	No
154	3076	3SSI	M 73 E	1.99	No
155	3078	3SSI	M 73 A	1.16	No
156	4076	3SSI	D 83 C	13.3	No
157	4090	3SSI	D 83 N	1.07	No
158	7022	2RN2	D 10 A	14.8	No
159	675	2RN2	G 23 A	1.8	No
160	752	2RN2	A 24 V	3.2	No
161	7023	2RN2	R 27 A	-5.6	Yes
162	765	2RN2	R 41 C	1.6	No
163	753	2RN2	A 52 V	7.8	No
164	7020	2RN2	I 53 A	-12.1	Yes
165	755	2RN2	H 62 P	4.1	Yes
166	754	2RN2	H 62 R	1.3	No
167	715	2RN2	H 62 A	1	No
168	749	2RN2	V 74 A	-12.7	No
169	748	2RN2	V 74 I	2.1	No
170	677	2RN2	V 74 L	3.3	No
171	758	2RN2	Q 76 L	1.1	No
172	10	2RN2	D 94 E	-1.6	Yes
173	766	2RN2	K 95 N	3.2	No
174	12	2RN2	K 95 G	6.8	No
175	7021	2RN2	Q 105 G	-9.6	Yes
176	739	2RN2	Q 113 P	-2.1	Yes

177	717	2RN2	H 114 A	-7.7	Yes
178	762	2RN2	E 119 V	2.7	No
179	683	2RN2	D 134 Q	4.8	No
180	678	2RN2	D 134 H	7	No
181	689	2RN2	D 134 A	5.5	No
182	688	2RN2	D 134 L	5.5	No
183	687	2RN2	D 134 I	4.6	No
184	682	2RN2	D 134 E	3.1	No
185	685	2RN2	D 134 T	3.9	No
186	684	2RN2	D 134 S	3.9	No
187	686	2RN2	D 134 V	4.1	No
188	681	2RN2	D 134 H	7	No
189	20133	1RTB	A 4 S	-1.54	No
190	20142	1RTB	A 5 S	-1.42	Yes
191	16628	1RTB	F 46 L	-11.1	Yes
192	16629	1RTB	F 46 V	-12.1	Yes
193	16630	1RTB	F 46 A	22.4	No
194	3541	1RTB	N 67 D	-1.4	Yes
195	23702	1RTB	S 75 T	-8.9	Yes
196	23701	1RTB	S 75 A	-7.9	Yes
197	23703	1RTB	S 75 C	-17.9	Yes
198	23704	1RTB	S 75 R	-22.9	Yes
199	3221	1RTB	P 93 G	-6.2	Yes
200	3222	1RTB	P 114 G	-9.5	Yes
201	11638	1RN1	V 16 A	-8.2	No
202	11651	1RN1	V 16 T	-9.4	Yes
203	11639	1RN1	V 16 S	-15.6	Yes
204	11641	1RN1	V 16 C	-9.6	Yes
205	2333	1RN1	S 17 A	1.7	Yes
206	10041	1RN1	Y 24 W	1.6	No
207	2884	1RN1	Q 25 K	3.3	No
208	2328	1RN1	Y 42 F	3.4	Yes
209	3414	1RN1	N 44 S	-5	Yes
210	2337	1RN1	N 44 A	-5.5	Yes
211	17865	1RN1	N 44 D	-5.5	Yes
212	2576	1RN1	Y 45 W	-1.6	Yes
213	2329	1RN1	Y 56 F	-2.1	Yes
214	2330	1RN1	Y 57 F	-1.3	Yes
215	2298	1RN1	E 58 A	-2.3	Yes
216	2575	1RN1	W 59 Y	-4.2	No
217	2334	1RN1	S 64 A	-4.6	Yes
218	2331	1RN1	Y 68 F	-4	Yes
219	5720	1RN1	D 76 A	-15.2	Yes
220	5719	1RN1	D 76 S	-13.6	Yes
221	5718	1RN1	D 76 N	-13.8	Yes
222	11653	1RN1	V 78 T	-11.3	Yes
223	11643	1RN1	V 78 S	-16.6	Yes
224	11642	1RN1	V 78 A	-7.9	Yes
225	2338	1RN1	N 81 A	-8.6	Yes
226	11646	1RN1	V 89 T	-13.2	Yes

227	11647	1RN1	V 89 C	-4.9	Yes
228	11645	1RN1	V 89 S	-21.6	Yes
229	10049	1RN1	H 92 A	-1.3	Yes
230	634	1ARR	S 5 A	-1.5	Yes
231	639	1ARR	M 7 A	-3.5	No
232	622	1ARR	P 8 A	15.1	No
233	890	1ARR	P 8 L	13.9	No
234	642	1ARR	F 10 A	-19.6	Yes
235	624	1ARR	N 11 A	3.1	No
236	661	1ARR	L 12 A	-15.6	Yes
237	635	1ARR	R 13 A	-1.7	Yes
238	667	1ARR	W 14 A	-26.4	Yes
239	656	1ARR	P 15 A	-11.3	Yes
240	636	1ARR	E 17 A	-2	Yes
241	637	1ARR	V 18 A	-2.1	Yes
242	849	1ARR	L 19 Q	-19	Yes
243	641	1ARR	L 19 A	-11.7	Yes
244	640	1ARR	D 20 A	-3.7	Yes
245	662	1ARR	L 21 A	-18.3	No
246	648	1ARR	R 23 A	-1.2	Yes
247	846	1ARR	K 24 T	-10	Yes
248	649	1ARR	K 24 A	-1.6	Yes
249	650	1ARR	E 28 A	-2.2	No
250	658	1ARR	N 29 A	-12.6	Yes
251	653	1ARR	G 30 A	-10	Yes
252	663	1ARR	R 31 A	-20.8	Yes
253	665	1ARR	S 32 A	-24.4	Yes
254	659	1ARR	V 33 A	-13.8	Yes
255	644	1ARR	N 34 A	5.1	No
256	623	1ARR	S 35 A	4.4	Yes
257	666	1ARR	Y 38 A	-24.9	Yes
258	645	1ARR	Q 39 A	3.5	Yes
259	668	1ARR	R 40 A	-26.7	Yes
260	664	1ARR	M 42 A	-22.3	No
261	638	1ARR	E 43 A	-2.9	Yes
262	657	1ARR	S 44 A	-11.6	Yes
263	655	1ARR	K 47 A	-10.7	Yes
264	660	1ARR	E 48 A	-14.7	Yes
265	652	1ARR	G 49 A	-9.2	Yes
266	654	1ARR	R 50 A	-10	Yes
267	651	1ARR	I 51 A	-7	Yes
268	646	1ARR	G 52 A	3	No
269	22534	1STN	T 33 S	-5.5	Yes
270	22550	1STN	T 33 C	-3.6	Yes
271	22557	1STN	T 33 I	1.1	No
272	22640	1STN	T 33 V	2.3	No
273	25490	1STN	F 34 C	-13	Yes
274	25491	1STN	L 36 C	-14	Yes
275	22603	1STN	V 39 T	-6	Yes
276	22513	1STN	V 39 S	-11	Yes
277	22656	1STN	T 41 I	4.2	No
278	22551	1STN	T 41 C	1.3	No

279	22535	1STN	T 41 S	-3.1	Yes
280	22543	1STN	T 41 V	2.7	No
281	22642	1STN	T 44 V	1	No
282	22657	1STN	T 44 I	-3.3	Yes
283	22568	1STN	K 49 F	-1.2	Yes
284	9219	1STN	G 50 C	-2.6	Yes
285	9228	1STN	G 50 F	-2.5	Yes
286	22506	1STN	V 51 T	-1	Yes
287	22524	1STN	Y 54 L	-12.4	Yes
288	22518	1STN	Y 54 F	-1.8	Yes
289	22570	1STN	E 57 F	-1.1	Yes
290	9221	1STN	A 60 C	-3.6	Yes
291	22600	1STN	T 62 V	2.3	No
292	22560	1STN	T 62 I	-3.9	Yes
293	22599	1STN	T 62 S	-9	Yes
294	22687	1STN	T 62 C	-3.5	Yes
295	22686	1STN	T 62 A	-8.1	Yes
296	22690	1STN	T 62 H	-8	Yes
297	22689	1STN	T 62 G	-15.9	Yes
298	22594	1STN	T 62 K	-17.5	Yes
299	22595	1STN	T 62 L	-4.4	No
300	22694	1STN	T 62 M	-2.3	Yes
301	22695	1STN	T 62 N	-15.9	Yes
302	22598	1STN	T 62 Q	-13.5	Yes
303	22669	1STN	M 65 F	-5.7	Yes
304	2780	1STN	V 66 A	-11.9	Yes
305	22515	1STN	V 66 S	-13.9	Yes
306	2935	1STN	V 66 L	5.7	Yes
307	22507	1STN	V 66 T	-6.3	Yes
308	22572	1STN	E 67 F	-6.6	Yes
309	2868	1STN	A 69 T	-13.7	Yes
310	9222	1STN	K 70 C	-2.8	Yes
311	3102	1STN	K 70 W	-3.4	Yes
312	12690	1STN	E 73 G	-15.2	Yes
313	22573	1STN	E 73 F	-5.3	No
314	22606	1STN	V 74 T	-16.9	No
315	2943	1STN	E 75 V	-6.6	No
316	2870	1STN	E 75 G	-15.2	Yes
317	2869	1STN	E 75 A	-5.6	Yes
318	2941	1STN	D 77 A	-10.7	Yes
319	12691	1STN	D 77 G	-8.1	Yes
320	2871	1STN	G 79 S	-4.4	Yes
321	22574	1STN	Q 80 F	-1.6	Yes
322	22652	1STN	T 82 C	-1.2	Yes
323	22538	1STN	T 82 S	-3.5	Yes
324	22519	1STN	Y 85 F	-1.3	Yes
325	22576	1STN	G 86 F	-9.1	Yes
326	2872	1STN	G 88 W	4.6	Yes
327	2937	1STN	G 88 V	4.3	Yes
328	113	1STN	A 90 S	-10.7	Yes

329	22624	1STN	Y 91 L	-13.8	Yes
330	22520	1STN	Y 91 F	-7.2	Yes
331	22521	1STN	Y 93 F	-7.8	Yes
332	22675	1STN	G 96 F	-11.3	Yes
333	22509	1STN	V 99 T	-15.2	Yes
334	22676	1STN	E 101 F	-13.3	Yes
335	22510	1STN	V 104 T	-10.3	Yes
336	9224	1STN	R 105 C	-11.1	Yes
337	9232	1STN	R 105 F	-15.1	Yes
338	22609	1STN	V 111 T	-10.2	Yes
339	22579	1STN	A 112 F	-6.9	Yes
340	9225	1STN	A 112 C	-2.8	Yes
341	22512	1STN	V 114 T	-1	Yes
342	22528	1STN	Y 115 L	-1.4	Yes
343	382	1STN	K 116 G	3.8	Yes
344	379	1STN	P 117 G	5	No
345	22637	1STN	T 120 S	-1.2	No
346	22653	1STN	T 120 C	-5.6	Yes
347	22660	1STN	T 120 I	-4.6	Yes
348	22645	1STN	T 120 V	-4.8	Yes
349	22581	1STN	E 122 F	-1	Yes
350	22582	1STN	Q 123 F	-2	Yes
351	114	1STN	H 124 L	5.1	Yes
352	22583	1STN	K 127 F	-1.2	Yes
353	22682	1STN	S 128 F	-3.3	Yes
354	22627	1STN	S 128 A	4.9	Yes
355	9226	1STN	K 134 C	-2.9	Yes
356	9234	1STN	K 134 F	-1.7	Yes
357	22585	1STN	E 135 F	-3.5	Yes
358	22684	1STN	K 136 F	-5.5	Yes
359	2873	1STN	L 137 A	-8.7	Yes
360	18696	1STN	W 140 F	-11.6	Yes
361	18697	1STN	W 140 Y	-12.1	Yes
362	18698	1STN	W 140 H	-8.9	Yes
363	22530	1STN	S 141 A	-1.3	Yes

Table S5-2. Wrongly thermostability effect prediction results by normal consensus that could be explained using structural information. The last column shows the structural information that can explain that why mutation has a stabilizing or destabilizing effect. HB indicates that using HB core rules (G<A<V<L,I) we could find the explanation. CH means that by charge distribution around the target residue (Rc=8Å) we can find the explanation. CH (a- vs. b+) shows that there are “a” negatively charged amino acids around it vs. “b” positively charged amino acids. CH (salt-bridge break) shows that the mutation leads to break of a salt-bridge. CH (salt-bridge) shows that the mutation can result in salt-bridge formation on the structure.

No.	PDB ID	Mutation	Explanation
1	2RN2	D 134 Q	CH (7- vs. 2+)
2	2RN2	D 134 H	CH (7- vs. 2+)
3	2RN2	D 134 A	CH (7- vs. 2+)
4	2RN2	D 134 L	CH (7- vs. 2+)
5	2RN2	D 134 I	CH (7- vs. 2+)
6	2RN2	D 134 K	CH (7- vs. 2+)
7	2RN2	D 134 T	CH (7- vs. 2+)
8	2RN2	D 134 S	CH (7- vs. 2+)
9	2RN2	D 134 V	CH (7- vs. 2+)
10	2RN2	D 134 H	CH (7- vs. 2+)
11	2RN2	D10A	CH (5- vs. 1+)
12	2RN2	K 95 N	CH (1- vs. 4+)
13	2RN2	K 95 G	CH (1- vs. 4+)
14	1HFZ	E 11 L	CH (5- vs. 2+)
15	1HFZ	D 87 N	CH (4- vs. 1+)
16	3SSI	D 126 K	CH (6- vs. 1+)
17	3SSI	D 75 K	CH(4- vs. 1+)
18	3SSI	E 72 K	CH (4- vs. 2+)
19	3SSI	E 61 K	CH(3- vs. 1+)
20	3SSI	E 40 K	CH (3- vs. 1+)
21	2LZM	Q 123 E	CH (1- vs. 3+)
22	2LZM	N 144 E	CH (1- vs. 3+)
23	2LZM	N 144 D	CH (1- vs. 3+)
24	1CSP	E 3 L	CH (5- vs. 1+)
25	1CSP	E 3 R	CH (5- vs. 1+)
26	1CSP	A 46 E	CH (4- vs. 2+) and HB
27	1CSP	S 48 R	CH (3- vs. 1+)
28	1RGG	D 79 H	CH (4- vs. 1+)
29	1RGG	D 79 Y	CH (4- vs. 1+)
30	1RGG	D 79 R	CH (4- vs. 1+)
31	1RGG	D 79 L	CH (4- vs. 1+)
32	1RN1	Q 25 K	CH (3- vs. 0+)
33	1RGG	D 79 K	CH (4- vs. 1+)
34	1RGG	D 79 N	CH (4- vs. 1+)
35	1RGG	D 79 A	CH (4- vs. 1+)
36	1RGG	D 79 F	CH (4- vs. 1+)
37	1RGG	E 74 K	CH (4- vs. 1+)
38	2RN2	V 74 A	HB

39	2RN2	V 74 I	HB
40	2RN2	V 74 L	HB
41	2RN2	A 52 V	HB
42	1HFZ	A 52 V	HB
43	2RN2	A 24 V	HB
44	2RN2	V 74 A	HB
45	2RN2	G 23 A	HB
46	1HFZ	V 42 G	HB
47	3SSI	A 101 V	HB
48	3SSI	V 83 I	HB
49	3SSI	V 31 A	HB
50	3SSI	V 18 I	HB
51	1RN1	V 16 A	HB
52	1ARR	L 21 A	HB
53	1ARR	G 52 A	HB
54	1HFZ	V 42 N	HB
55	1RTB	A 4 S	HB
56	1STN	V 74 T	HB
57	1HFZ	I 59 W	HB
58	2LZM	D 47 A	CH (salt-bridge break)
59	2LZM	D 92 N	CH (salt-bridge break)
60	2LZM	G 113 E	CH (salt-bridge)
61	2LZM	N 116 D	CH (salt-bridge)

Table S5-3. Results for thermophilic sequences. The effect of mutations presented here could not be predicted correctly using normal CC method.

No.	Ref.	PDB ID	Mutation
1	12131	1CSP	E 3 L
2	12130	1CSP	E 3 R
3	12132	1CSP	A 46 E
4	19998	1FTG	V 18 I
5	18031	1FTG	V 31 A
6	17988	1FTG	D 75 K
7	20008	1FTG	Q 99 A
8	20010	1FTG	E 107 A
9	20011	1FTG	Q 111 G
10	17942	1FTG	D 126 K
11	19711	1HFZ	E 11 L
12	6616	1HFZ	V 42 G
13	6608	1HFZ	H 107 Y
14	6605	1HFZ	W 118 Y
15	715	2RN2	H 62 A
16	749	2RN2	V 74 A
17	12	2RN2	K 95 G
18	762	2RN2	E 119 V
19	20133	1RTB	A 4 S
20	9219	1STN	G 50 C
21	22595	1STN	T 62 L
22	22695	1STN	T 62 N
23	22573	1STN	E 73 F
24	22606	1STN	V 74 T
25	2943	1STN	E 75 V
26	22637	1STN	T 120 S

Appendix A4: Supplementary Information for Chapter 6

Table S6_1. Comparison between our protocol's result with experimental data from Protherm database. The sixth column (Code) represents the code of the method that detected the mutation effect as follows: 1 for HB DNT, 2 for Salt-bridge DNT, 3for HB mutation, 4 for HB_HL switch mutation, 5 for HB forbidden mutation, 6 for consensus mutation, 7 for charged mutation, and 8 for charge forbidden mutation. Proth., Imut., and Auto. indicate Protherm, I-Mutant2.0, and AUTO-MUTE results. + shows agreement and - shows disagreement with experimental values found in Protherm database. For Auto., "0" points out those mutations that AUTO-MUTE cannot make any prediction for.

No.	**Ref.**	**PDB_ID**	Mutation	dT_m	Code	Proth.	Imut.	Auto.
1	18029	1FTG	L 6 A	-11.8	1	+	+	+
2	19998	1FTG	V 18 I	4.9	3	+	-	+
3	17978	1FTG	E 20 K	-2.7	2	+	+	+
4	20001	1FTG	I 21 G	3.8	1	_	-	-
5	18030	1FTG	I 22 V	-6.3	1	+	+	+
6	18050	1FTG	V 31 A	-4.8	5	+	+	+
7	18052	1FTG	I 51 V	-4.7	1	+	+	+
8	18053	1FTG	I 52 V	-4.1	1	+	+	+
9	20004	1FTG	G 68 A	6.7	3	+	-	+
10	20007	1FTG	V 83 I	1	3	+	-	-
11	18055	1FTG	A 84 G	-5.4	5	+	+	+
12	20009	1FTG	A 101 V	5.4	3	+	+	+
13	18039	1FTG	I 104 V	-4	1	+	+	+
14	20010	1FTG	E 107 A	2.7	2, 4	_	-	-
15	18041	1FTG	V 117 A	-8	5	+	+	+
16	18043	1FTG	V 139 A	-6.4	5	+	+	+
17	18044	1FTG	L 143 A	-1.5	1	+	+	+
18	18064	1FTG	I 156 V	-8.7	1	+	+	+
19	18046	1FTG	V 160 A	-8.2	5	+	+	+
20	2865	1STN	L 7 A	-3.2	1	+	+	+
21	22504	1STN	V 23 T	-14.7	3	+	+	+
22	109	1STN	L 25 A	-11.5	1	+	+	+

23	22564	1STN	K 28 F	-4.2	2	+	-	+
24	25491	1STN	L 36 C	-14	1	+	+	+
25	22513	1STN	V 39 S	-11	3	+	+	+
26	22568	1STN	K 49 F	-1.2	2	+	-	+
27	22506	1STN	V 51 T	-1	3	+	+	+
28	9221	1STN	A 60 C	-3.6	3	+	+	+
29	22595	1STN	T 62 L	-4.4	6	_	+	+
30	2950	1STN	V 66 A	-11.9	5	+	+	+
31	22515	1STN	V 66 S	-13.9	3	+	+	+
32	22605	1STN	V 66 T	-5.7	3	+	+	+
33	2779	1STN	V 66 L	5.7	3	+	+	+
34	22572	1STN	E 67 F	-6.6	2	+	-	-
35	2868	1STN	A 69 T	-13.7	3	+	+	+
36	22508	1STN	V 74 T	-18.3	3	+	+	+
37	2869	1STN	E 75 A	-5.6	2	+	-	+
38	2870	1STN	E 75 G	-15.2	2	+	+	+
39	2943	1STN	E 75 V	-6.6	2	+	+	+
40	111	1STN	G 79 S	-7.6	3	+	+	+
41	22576	1STN	G 86 F	-9.1	3	+	+	+
42	112	1STN	G 88 V	2.9	3	+	-	+
43	3103	1STN	G 88 W	4.8	6	+	-	+
44	113	1STN	A 90 S	-10.7	3	+	+	+
45	22509	1STN	V 99 T	-15.2	3	+	+	+
46	22510	1STN	V 104 T	-10.3	3	+	+	+
47	9232	1STN	R 105 F	-15.1	2	+	+	+
48	9224	1STN	R 105 C	-11.1	2	+	+	+
49	22609	1STN	V 111 T	-10.2	3	+	+	+
50	9225	1STN	A 112 C	-2.8	3	+	+	+
51	22512	1STN	V 114 T	-1	3	+	+	+
52	382	1STN	K 116 G	3.8	4	+	-	+
53	22581	1STN	E 122 F	-1	2	+	-	-
54	381	1STN	H 124 L	5.8	6	+	+	+
55	22627	1STN	S 128 A	4.9	4	+	-	-
56	22585	1STN	E 135 F	-3.5	2	+	-	+
57	2873	1STN	L 137 A	-8.7	1	+	+	+
58	18698	1STN	W 140 H	-8.9	3	+	+	+
59	18697	1STN	W 140 Y	-12.1	3	+	+	+
60	675	2RN2	G 23 A	1.8	3	+	+	+
61	752	2RN2	A 24 V	3.2	3	+	+	+
62	7023	2RN2	R 27 A	-5.6	2	+	+	+
63	753	2RN2	A 52 V	7.8	3	+	+	+
64	7020	2RN2	I 53 A	-12.1	1	+	+	+
65	738	2RN2	H 62 P	4.1	6	+	+	+
66	748	2RN2	V 74 I	2.1	3	+	+	+
67	749	2RN2	V 74 A	-12.7	5	+	+	+
68	747	2RN2	V 74 L	3.3	3	+	+	+
69	739	2RN2	Q 113 P	-2.1	6	_	+	+
70	18529	1CSP	E 3 K	16.6	7	+	-	+
71	3526	1CSP	F 15 A	-15.8	5	+	+	+
72	3525	1CSP	F 17 A	-11.3	5	+	+	+
73	20133	1RTB	A 4 S	-1.54	6	_	+	+
74	16630	1RTB	F 46 A	22.4	5	_	-	-

75	3225	1RTB	P 114 G	-10.2	5	+	+	+
76	19711	1HFZ	E 11 L	9.7	2, 4	+	-	0
77	19712	1HFZ	D 14 N	-3	2	+	+	0
78	19714	1HFZ	D 37 N	14.4	6	+	-	0
79	6615	1HFZ	V 42 A	-4.5	3	+	+	0
80	6616	1HFZ	V 42 G	-5.2	3	+	+	0
81	6614	1HFZ	V 42 N	-1.1	3	+	+	0
82	6618	1HFZ	I 59 W	-4.5	1	+	+	0
83	6620	1HFZ	Y 103 A	-10.7	6	_	+	0
84	6607	1HFZ	A 106 S	-6.1	3	+	+	0
85	6609	1HFZ	H 107 A	-4.1	4	_	+	0
86	6608	1HFZ	H 107 Y	-1	6	_	-	0
87	6610	1HFZ	H 107 W	-8.7	6	_	-	0
88	6601	1HFZ	L 110 H	6.3	1,6	+	-	0
89	6602	1HFZ	L 110 R	1.9	1,6	+	+	0
90	6613	1HFZ	K 114 E	-3.2	8	+	-	0
91	6611	1HFZ	K 114 N	10.7	6	+	+	0
92	6606	1HFZ	W 118 H	-3.3	8	+	+	0
93	18642	1AZP	W 24 A	-9.6	5	+	+	-
94	16988	1AZP	V 30 I	5.8	3	+	-	-
95	12699	1DIL	A 53 L	5	3	+	-	+
96	12698	1DIL	A 69 V	3	3	+	-	+
97	16399	1H8V	A 35 V	7.7	3	+	-	+
98	16402	1H8V	A 35 S	-4	3	+	+	-
99	16404	1H8V	G 41 A	2.5	3	+	+	+
100	25715	1JU3	L 169 K	6.3	1	_	-	-
101	3099	1N0J	I 58 T	-13.6	1	+	+	+
102	16646	1RGG	V 2 T	-3.2	3	+	+	+
103	16643	1RGG	T 16 V	1	6	+	-	-
104	19020	1RGG	D 33 A	-16	2	+	+	-
105	16647	1RGG	V 36 T	-4.6	3	+	+	+
106	16648	1RGG	V 43 T	-1.6	3	+	+	+
107	16730	1RGG	E 54 Q	-5.9	2	+	+	-
108	16649	1RGG	V 57 T	-15	3	+	+	+
109	16731	1RGG	R 65 A	-3.4	2	+	+	+
110	19015	1RGG	D 79 W	7.6	4	+	-	-
111	19011	1RGG	D 79 I	9.6	6	+	-	+
112	2327	1RN1	Y 11 F	-6	4	_	+	0
113	11638	1RN1	V 16 A	-8.2	5	+	+	0
114	11639	1RN1	V 16 S	-15.6	3	+	+	0
115	11641	1RN1	V 16 C	-9.6	3	+	+	0
116	11651	1RN1	V 16 T	-9.4	3	+	+	0
117	2333	1RN1	S 17 A	1.7	6	+	+	0
118	2328	1RN1	Y 42 F	3.4	4	+	-	0
119	2298	1RN1	E 58 A	-2.3	2	+	+	0
120	11653	1RN1	V 78 T	-11.3	3	+	+	0
121	11643	1RN1	V 78 S	-16.6	3	+	+	0
122	11642	1RN1	V 78 A	-7.9	5	+	+	0
123	11645	1RN1	V 89 S	-21.6	3	+	+	0
124	11647	1RN1	V 89 C	-4.9	3	+	+	0

125	11646	1RN1	V 89 T	-13.2	3	+	+	0
126	16990	1SSO	I 29 V	-3.9	1	+	+	+
127	996	1WQ5	G 51 D	-4.5	3	+	+	0
128	975	1WQ5	M 101 T	7.6	4	+	-	0
129	18246	2AFG	H 21 Y	3.6	4, 6	+	-	+
130	18280	2AFG	L 44 F	2.2	1	_	-	-
131	12727	2AFG	L 73 V	-8.5	1	+	+	+
132	8299	2AFG	H 93 G	7.3	6	+	-	+
133	18248	2AFG	H 102 Y	1.9	6	+	+	-
134	18249	2AFG	F 108 Y	1.6	4, 6	+	-	-
135	12728	2AFG	V 109 L	-3.3	3	_	+	+
136	12302	2CBR	R 131 Q	11.6	8	+	-	+
137	12301	2CBR	R 111 Q	9.4	8	+	-	+
138	2589	3SSI	V 13 L	-4.5	3	_	+	+
139	2588	3SSI	V 13 I	-2.4	3	_	+	+
140	2586	3SSI	V 13 G	-27.7	5	+	+	+
141	2585	3SSI	V 13 A	-13.2	5	+	+	+
142	3074	3SSI	M 73 K	0.65	6	+	+	+
143	2371	3SSI	M 103 A	-4.2	5	+	+	+
144	2370	3SSI	M 103 G	-13.57	5	+	+	+
145	3059	4BLM	K 234 A	4.2	6	+	-	0
146	3058	4BLM	K 234 E	-2.8	8	+	+	0

Appendix A5

In this appendix I report a published work on a subject that differs from the main objective of my thesis. In this work, I conducted protein modeling and simulation for better understanding the biological function of the competence protein ComEA from *Vibrio cholera*. My contribution, within the collaboration led Prof. M. Blokesch at Laboratory of Molecular Microbiology – EPFL, was key for the finding of the DNA binding mode of ComEA and for proposing a possible DNA uptake mechanism:

> *ComEA Is Essential for the Transfer of External DNA into the Periplasm in Naturally Transformable Vibrio cholerae Cells*
>
> Patrick Seitz, Hassan Pezeshgi Modarres, Sandrine Borgeaud, Roman D. Bulushev, Lorenz J. Steinbock, Aleksandra Radenovic, Matteo Dal Peraro, Melanie Blokesch.

ComEA Is Essential for the Transfer of External DNA into the Periplasm in Naturally Transformable *Vibrio cholerae* Cells

Patrick Seitz[1], Hassan Pezeshgi Modarres[2,3], Sandrine Borgeaud[1], Roman D. Bulushev[4], Lorenz J. Steinbock[4], Aleksandra Radenovic[4], Matteo Dal Peraro[2,3], Melanie Blokesch[1]*

1 Laboratory of Molecular Microbiology, Global Health Institute, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, **2** Laboratory for Biomolecular Modeling, Institute of Bioengineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, **3** Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland, **4** Laboratory of Nanoscale Biology, Institute of Bioengineering, School of Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Abstract

The DNA uptake of naturally competent bacteria has been attributed to the action of DNA uptake machineries resembling type IV pilus complexes. However, the protein(s) for pulling the DNA across the outer membrane of Gram-negative bacteria remain speculative. Here we show that the competence protein ComEA binds incoming DNA in the periplasm of naturally competent *Vibrio cholerae* cells thereby promoting DNA uptake, possibly through ratcheting and entropic forces associated with ComEA binding. Using comparative modeling and molecular simulations, we projected the 3D structure and DNA-binding site of ComEA. These *in silico* predictions, combined with *in vivo* and *in vitro* validations of wild-type and site-directed modified variants of ComEA, suggested that ComEA is not solely a DNA receptor protein but plays a direct role in the DNA uptake process. Furthermore, we uncovered that ComEA homologs of other bacteria (both Gram-positive and Gram-negative) efficiently compensated for the absence of ComEA in *V. cholerae*, suggesting that the contribution of ComEA in the DNA uptake process might be conserved among naturally competent bacteria.

Citation: Seitz P, Pezeshgi Modarres H, Borgeaud S, Bulushev RD, Steinbock LJ, et al. (2014) ComEA Is Essential for the Transfer of External DNA into the Periplasm in Naturally Transformable *Vibrio cholerae* Cells. PLoS Genet 10(1): e1004066. doi:10.1371/journal.pgen.1004066

Editor: William F. Burkholder, Institute of Molecular and Cell Biology (IMCB), A*STAR, Singapore

Received July 19, 2013; **Accepted** November 12, 2013; **Published** January 2, 2014

Copyright: © 2014 Seitz et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the Swiss National Science Foundation (http://www.snf.ch) grants 200020_138013 (to MDP) and grants 31003A_127029 and 31003A_143356 (to MB) and by the European Research Council (http://erc.europa.eu) grants 259398-PorABEL (to AR) and 309064-VIR4ENV (to MB). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: melanie.blokesch@epfl.ch

Introduction

Recombination between the bacterial chromosome and DNA fragments that enter the cell through horizontal gene transfer (HGT) either replace damaged or mutated alleles with the original alleles, thereby repairing the gene, or transfer mutated alleles or new genes to naïve strains. Thus, HGT plays a key role in transferring genetic information from one bacterium to another and maintaining the balance between genome maintenance and evolution. Natural competence for transformation is one of three modes of HGT in bacteria and promotes the uptake of free DNA from the environment (for recent reviews see [1–6]).

Many naturally transformable bacteria have been described [7], including the pathogenic bacterium *Vibrio cholerae* [6,8]. The physiological state of natural competence of this Gram-negative bacterium is associated with its primary niche, the aquatic environment. Within this habitat, *V. cholerae* attaches to the exoskeleton of zooplankton or zooplankton molts [9]. Those exoskeletons comprise the polymer chitin, which is the natural inducer of competence in *V. cholerae* [6,8,10]. Whereas the regulatory network driving competence has been well investigated (reviewed by Seitz and Blokesch [6]), so far very little is known about the DNA uptake complex of *V. cholerae* [11]. With respect to the DNA uptake machinery of naturally transformable bacteria it has been suggested that a (pseudo-)pilus [1,2], similar to type IV pili (Tfp) [12], represents a core element of the DNA import machinery. However, it is still unclear how the proteins interact to pull the transforming DNA through the cell envelope [3]. A proposed mechanism for DNA uptake involves repeating cycles of pilus extension and retraction [1,2,4,13] although recent review articles suggested that other competence proteins, such as ComEA, might be involved in pulling the DNA into the cell [4,14] (though without experimental evidence). The present study reinforces those ideas and shows that ComEA is a prerequisite for DNA uptake in naturally competent *V. cholerae*. Furthermore, based on an earlier study on DNA ejection from bacteriophages [15] we propose a model suggesting that the DNA translocation across the outer membrane is possibly accomplished by ratcheting and entropic forces associated with the binding of ComEA to the incoming DNA.

Currently, the majority of studies on the cellular localization of competence proteins were performed on Gram-positive bacteria [16–19], whereas far less is known about competence protein localization in Gram-negative bacteria. We recently identified the minimal competence gene set of *V. cholerae* and provided first insight into the DNA uptake machinery of this organism [11].

Author Summary

Horizontal gene transfer (HGT) plays a key role in transferring genetic information from one organism to another. Natural competence for transformation is one of three modes of HGT used by bacteria to promote the uptake of free DNA from the surrounding. The human pathogen *Vibrio cholerae* enters such a competence state upon growth on chitinous surfaces, which represent its natural niche in the aquatic environment. Whereas we have gained a reasonable understanding on how the competence phenotype is regulated in *V. cholerae* we are only at the beginning of deciphering the mechanistic aspects of the DNA uptake process. In this study, we characterize the competence protein ComEA. We show that ComEA is transported into the periplasm of *V. cholerae* and that it is required for the uptake of DNA across the outer membrane. We demonstrate that ComEA aggregates around incoming DNA *in vivo* and that the binding of DNA is dependent on specific residues within a conserved helix-hairpin-helix motif. We propose a model indicating that the DNA uptake process across the outer membrane might be driven through ratcheting and entropic forces associated with ComEA binding.

Notably, through the analysis of knockout strains lacking specific components of the DNA uptake complex we demonstrated that natural transformation still occurred in the absence of the proteins involved in the Tfp structure and biosynthesis though at very low frequencies. Such rare transformants were never detectable for *comEA*$^{-}$ strains [11], suggesting that ComEA plays an important role in the DNA uptake process, the focus of this work.

In studies on *B. subtilis* and *S. pneumoniae* it was reported that binding of transforming DNA to those Gram-positive cells is at least partially mediated by ComEA and that ComEA is "absolutely required" for DNA uptake and transformation [20–22]. Likewise, ComE (ComEA homolog)-negative strains of *Neisseria gonorrhoeae* [23] and *V. cholerae* [8,24] were severely or completely impaired for natural transformability, indicating that ComEA might also play an important role in Gram-negative bacteria. A recent study by Lo Scrudato and Blokesch indicated that *comEA* and the gene encoding the inner membrane transporter *comEC* were differentially regulated from the Tfp-like components of the DNA uptake machinery [25,26], which, together with our study on the DNA uptake machinery, suggest that DNA transport might be a multi-step process in *V. cholerae* (as previously proposed for *Helicobacter pylori* [14,27], which does not contain a *bona fide* Tfp-based DNA uptake machinery). Here, we show that the Tfp-like elements of the DNA uptake machinery of *V. cholerae* are not sufficient to translocate DNA across the outer membrane and that the competence protein ComEA plays an essential role in this process.

Results and Discussion

ComEA localizes to the periplasm in naturally competent *V. cholerae* cells

In a previous study by Chen and Gotschlich the authors predicted a 19-residues signal sequence for sec-dependent transport of the ComEA-homolog of *Neisseria gonorrhoeae* (ComE) into the periplasm [23]. Such a predictable signal sequence (amino acid residues 1–25) is also present in ComEA of *V. cholerae*. To experimentally address the localization of the ComEA protein we aimed at visualizing it *in vivo* by constructing a functional translational fusion between ComEA and mCherry. Using this construct we observed a uniform localization pattern of ComEA (Fig. 1A), which is consistent with the presence of such an N-terminal signal sequence and the transport of ComEA to the periplasm. To validate this microscopical observation, we generated a translational fusion between *comEA* and the gene encoding beta-lactamase (*bla*; without the region encoding the signal sequence), which replaced the wild-type *comEA* allele on the *V. cholerae* large chromosome. The resulting strain retained natural transformability at a frequency of $2.5\times10^{-5}\pm3.0\times10^{-5}$ compared with $7.9\times10^{-5}\pm2.5\times10^{-5}$ for the parental wild-type strain (average of four biological replicates ± SD) indicating the functionality of the fusion construct. Most importantly, the construct conferred full resistance to ampicillin, which provides further evidence for the periplasmic localization of ComEA-bla as beta-lactamase can only exert activity against beta-lactam antibiotics in the periplasm of Gram-negative bacteria (Fig. S1).

Next, we aimed to investigate whether the ComEA protein is motile within the periplasm. To this extent we used a fluorescence loss in photobleaching (FLIP; Fig. 1B) approach because photobleaching can reveal protein dynamics in live cells [28]. In

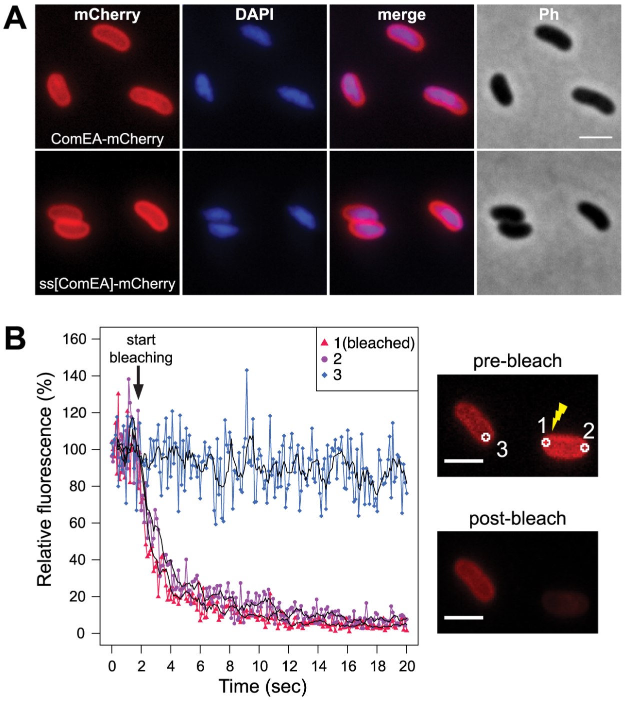


Figure 1. Localization of the ComEA protein in naturally competent *V. cholerae* cells. (A) Expression and distribution of ComEA-mCherry (upper row) or signal sequence[ComEA] (amino acids 1–25)-mCherry fusion proteins (lower row) within competent *V. cholerae* cells. Fluorescent signals for mCherry or DAPI-stained genomic DNA were visualized and compared with each other (merge) and the corresponding phase contrast image (Ph). (B) Representative fluorescence loss in photobleaching (FLIP) experiment to demonstrate the degree of mobility of ComEA-mCherry in live bacteria. Bleaching of the region-of-interest (ROI) 1 (indicated as 1 in the images on the right) was initiated after the acquisition of 20 frames and repeated after every frame. The fluorescence intensities of ROIs 1–3 were measured for a total of 20 sec and normalized to the average fluorescence intensity of the first 10 frames. The moving averages (period n = 5) are indicated with black lines. The average fluorescence intensity projections before (pre-bleach) and after bleaching (post-bleach) are shown on the right. Scale bars, 2 µm.

doi:10.1371/journal.pgen.1004066.g001

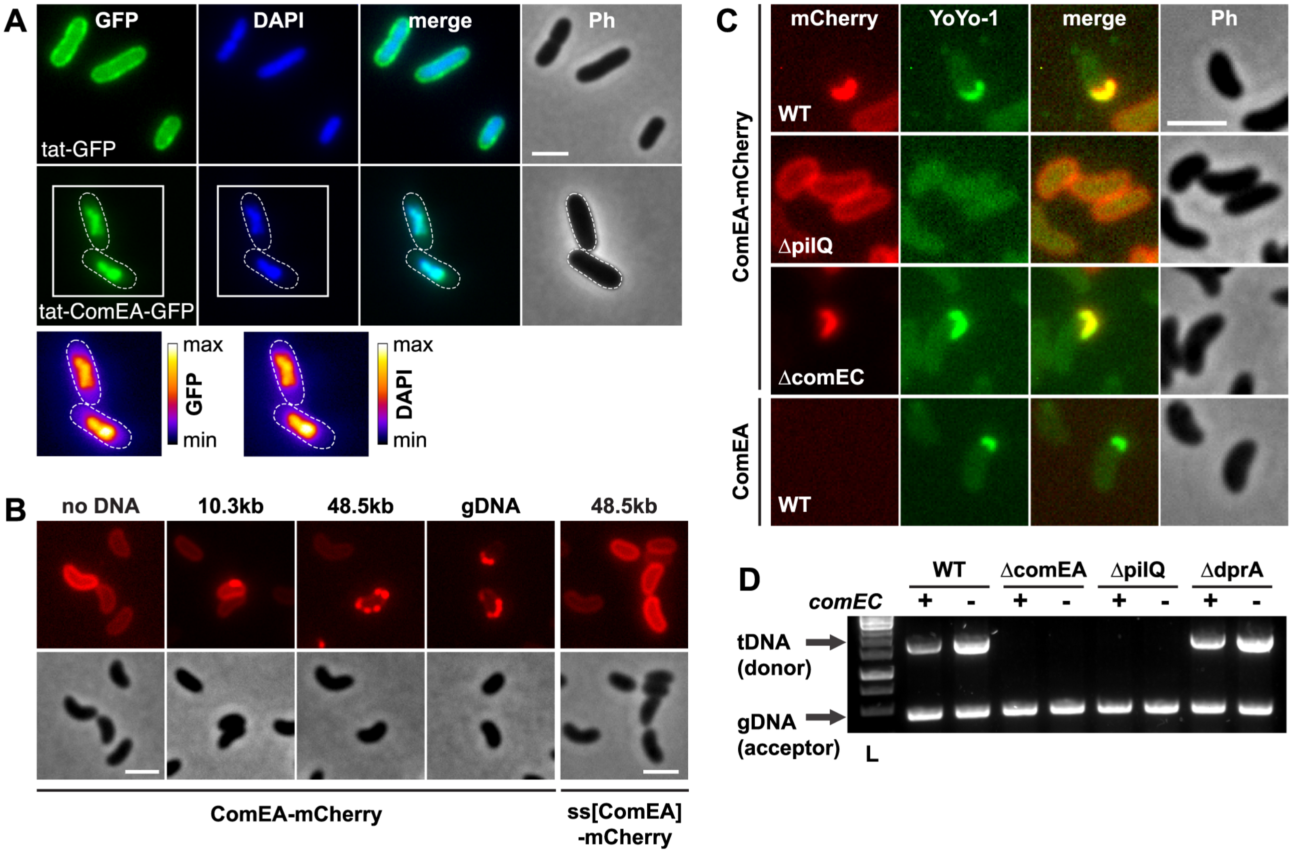


Figure 2. ComEA binds to DNA *in vivo*. (A) Plasmid-encoded *gfp* (tat-GFP) or *comEA-gfp* (tat-ComEA-GFP), both preceded by a tat-signal sequence, were expressed in *E. coli*. The images shown correspond to the GFP channel, DAPI channel (to visualize DAPI-stained DNA), merged fluorescent images (merge), and phase contrast (Ph). The cells are outlined with dashed lines for tat-ComEA-GFP. Heat-maps showing the fluorescence intensities of the GFP and DAPI signals are depicted for the *tat-comEA-gfp* expressing cells below the images. (B) ComEA-mCherry aggregation and foci formation after the addition of external DNA. Competence-induced cells without (no DNA) or with external DNA were imaged in the red (mCherry; upper row) or the phase contrast channel (lower row) to visualize ComEA-mCherry localization. The DNA fragments differed in lengths (PCR fragment, 10.3 kb; λDNA, 48.5 kb; gDNA, various lengths). Transforming DNA did not lead to foci formation of periplasmic mCherry alone (preceded by the ComEA signal sequence; ss[ComEA]-mCherry). (C) Colocalization (merged image) of ComEA-mCherry (red channel) and YoYo-1-stained transforming DNA (green channel). The outline of the cells is shown in the phase contrast image (Ph). Scale bars in all images, 2 μm. (D) DNA uptake requires ComEA. DNA uptake of competent *V. cholerae* cells was tested using a whole-cell duplex PCR assay. All mutant strains were tested in a *comEC* positive (+) and negative (−) background. The lower PCR fragments indicate acceptor strain DNA (gDNA, acceptor); the upper band indicates internalized transforming DNA (tDNA). L, ladder.
doi:10.1371/journal.pgen.1004066.g002

contrast to fluorescence recovery after photobleaching (FRAP), where fluorescent proteins within a small area of the cell are bleached and the back-diffusion of the surrounding non-bleached proteins into this region is recorded, FLIP consist of repetitive bleaching of the same region (e.g. region of interest 1 in Fig. 1B), thereby preventing fluorescence recovery in that region. Moreover, any mobile protein from elsewhere in the same compartment (e.g. region of interest 2 in Fig. 1B) will also enter this continuously photo-bleached area, eventually resulting in a complete loss of fluorescence in the compartment. In contrast, any not connected compartment will be spared from bleaching (e.g. region of interest 3 in Fig. 1B). Therefore, FLIP is often used to reveal the mobility of proteins within certain compartments of the cell [29], which is what we were aiming for. Indeed, our FLIP experiments indicated that ComEA was highly motile within the periplasm (Fig. 1B). Likewise, a translational fusion between the signal sequence of ComEA (amino acid residues 1–25; ss[ComEA]) alone and mCherry resulted in a similar localization (Fig. 1A) and mobility pattern (Fig. S2). This uniform localization pattern differed from that obtained from previous studies on *B. subtilis*, where Hahn *et al.* used immunofluorescence microscopy to show that ComEA localizes in a non-uniform punctate manner [16]. Kaufenstein *et al.* confirmed those data and concluded that the distinct assemblies of ComEA were mobile [19].

ComEA binds to DNA *in vivo*

Studies using purified tagged ComEA/ComE homologs demonstrated that the protein binds DNA *in vitro*; thus, ComEA was considered as a DNA receptor protein [21,23,30,31]. DNA binding could be attributed to a conserved helix-hairpin-helix (HhH) motif [32]. Notably and in contrast to helix-turn-helix or helix-loop-helix motifs, which are widespread in proteins that interact with DNA in a sequence-dependent manner, HhH motifs bind DNA in a non-sequence-specific manner. Such binding is based on hydrogen bonding between the protein and the DNA phosphate groups [32] and HhH motifs have been described in various protein classes, including DNA polymerases, DNA ligases or DNA glycosylases [32,33]. However, the *in vivo* binding of DNA through ComEA has never been demonstrated. We genetically engineered a fusion protein between ComEA and GFP, which was

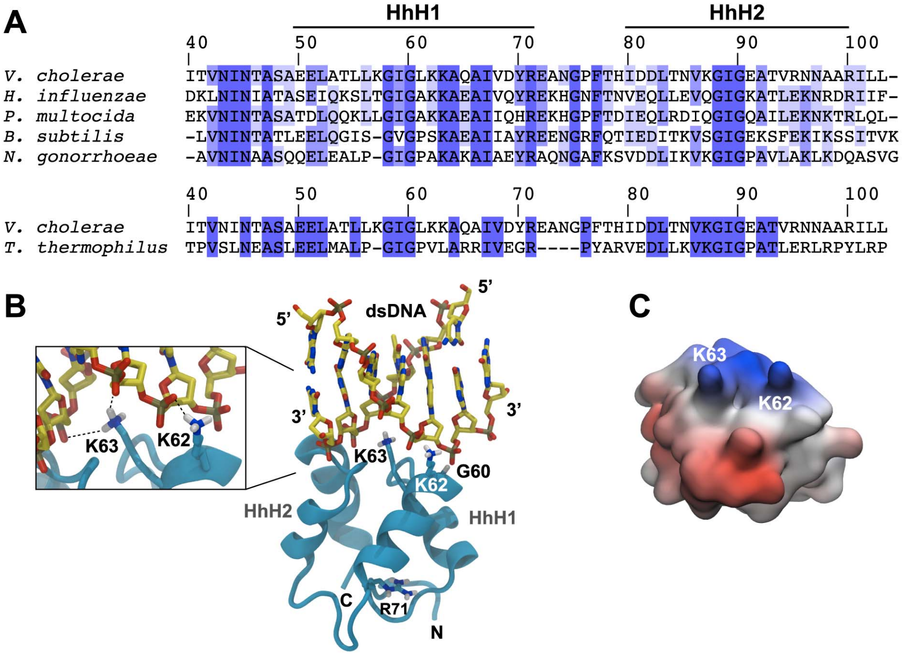

A
HhH1
HhH2
40 50 60 70 80 90 100
V. cholerae ITVNINTASAEELATLLKGIGLKKAQAIVDYREANGPFTHIDDLTNVKGIGEATVRNNAARILL-
H. influenzae DKLNINIATASEIQKSLTGIGAKKAEAIVQYREKHGNFTNVEQLLEVQGIGKATLEKNRDRIIF-
P. multocida EKVNINTASATDLQQKLLGIGAKKAEAIIQHREKHGPFTDIEQLRDIQGIGQAILEKNKTRLQL-
B. subtilis -LVNINTATLEELQGIS-GVGPSKAEAIIAYREENGRFQTIEDITKVSGIGEKSFEKIKSSITVK
N. gonorrhoeae -AVNINAASQQELEALP-GIGPAKAKAIAEYRAQNGAFKSVDDLIKVKGIGPAVLAKLKDQASVG
40 50 60 70 80 90 100
V. cholerae ITVNINTASAEELATLLKGIGLKKAQAIVDYREANGPFTHIDDLTNVKGIGEATVRNNAARILL
T. thermophilus TPVSLNEASLEELMALP-GIGPVLARRIVEGR----PYARVEDLLKVKGIGPATLERLRPYLRP
B
5'
dsDNA
5'
3'
3'
K63
K62
K63
K62
G60
HhH2
HhH1
C
R71
N
C
K63
K62

Figure 3. *In silico*-prediction of the ComEA-DNA complex. (A) Protein sequence alignment of the two helix-hairpin-helix motifs (HhH1-2) of ComEA/ComE homologs from the indicated organisms (using the ComEA residue numbering). The sequence conservation is shown in tones of blue (dark blue = highly conserved). (B) 3D model of ComEA and its predicted DNA binding mode based on comparative modeling using the ComEA-related protein of *T. thermophilus* HB8 (PDB ID: 2DUY) as a template (see also movie S1). The non-sequence-specific DNA backbone phosphate interactions with K62 and K63 are shown (inset). (C) The electrostatic potential at the molecular surface of ComEA is reported within a ±160 k_BT/e range (negative values in red, positive values in blue).

doi:10.1371/journal.pgen.1004066.g003

transported across the inner membrane via the Tat-transport machinery in a folded state (as GFP is improperly folded when translocated to the periplasm in a sec pathway dependent manner [34]). Interestingly, the protein failed to translocate in *Escherichia coli*; instead, ComEA was tightly bound to the bacterial chromosome, which appeared as a highly compacted structure (Fig. 2A). The increased protein expression levels resulted in cell death, indicating that the strong binding of ComEA to the DNA *in vivo* interfered with cellular processes. Due to this lack of translocation of ComEA-GFP into the periplasm and the *in vivo* binding to the chromosome we conducted further experiments using the ComEA-mCherry fusion despite the lower signal intensity of mCherry compared with GFP.

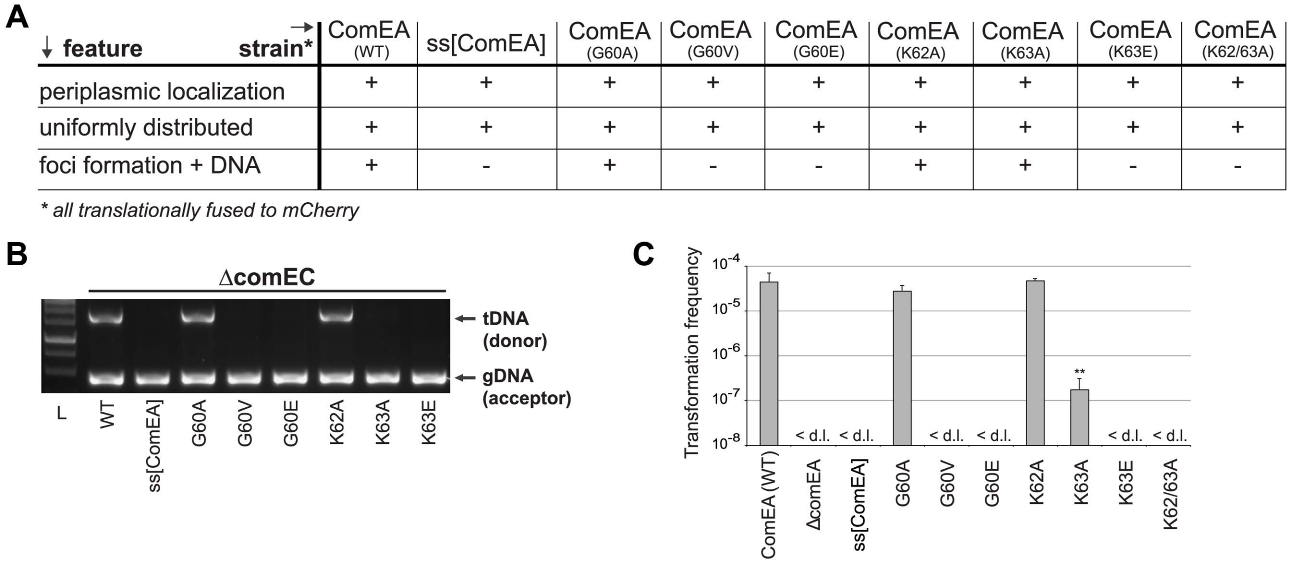
ComEA is required for uptake of transforming DNA into the periplasm

To investigate the function of ComEA *in vivo*, we excluded artifacts caused by artificial (over-)expression as those have been recognized as having detrimental effects on subcellular localization [35]. Thus, all *V. cholerae* strains used in these experiments were generated through the substitution of chromosomal *comEA* with diverse *comEA-mCherry* alleles. In these strains, the expression of *comEA-mCherry* was driven through its native promoter and consequently co-regulated with other competence genes. The functionality of the chromosomally encoded ComEA fusion protein was confirmed using a transformation assay, and the chromosomally-encoded fusion protein was uniformly localized within the periplasm (Fig. 2B and Fig. S3). Importantly, the addition of external transforming DNA (tDNA) led to the formation of distinctive ComEA-mCherry foci (Fig. 2B). The size and numbers of these protein aggregates was dependent on the length of the supplemented tDNA. Periplasmic mCherry alone did not aggregate (ss[ComEA]-mCherry; Fig. 2B). A similar relocalization pattern after the addition of external DNA was also observed when the cells were grown on chitin surfaces mimicking the natural reservoir *V. cholerae* (Fig. S4). This observation suggested that ComEA binds transforming DNA in the periplasm thereby potentially contributing to DNA translocation across the outer membrane.

To test this hypothesis, we repeated the experiments using YoYo-1-labeled DNA. Indeed, a perfect colocalization pattern was observed when the fluorescent signals of ComEA-mCherry and DNA were compared (Fig. 2C). Foci formation through ComEA and colocalization with YoYo-1-labeled DNA were absent in a strain lacking the outer membrane pore PilQ [2,8,11], whereas the absence of the inner membrane transporter ComEC did not interfere with ComEA-DNA colocalization (Fig. 2C). Similar foci formation of YoYo-1-labeled DNA was also observed in a strain carrying wild-type ComEA, excluding a translational artifact resulting from the mCherry-fusion (Fig. 2C). Notably, YoYo-1 foci were absent in a *comEA*-negative strain, which was also the case for a strain lacking the major Tfp subunit PilA (Fig. S5A).

Using a whole-cell duplex PCR-based DNA uptake assay [24,11] that aims at detecting DNA strands, which have either entered the periplasm or have already reached the cytoplasm of the competent bacteria (thereby becoming resistant against externally applied DNase), we confirmed that tDNA (both unlabeled or YoYo-1-labeled) was undetectable in *comEA*-negative strains even though it was readily detectable in the wild-type strain and in *comEC* negative derivatives (Fig. 2D and Fig. S5).

Whereas the absence of YoYo-1 labeled DNA foci and PCR-amplifiable DNA in *comEA* negative strains is indicative of a failure



↓ feature \ strain* →	ComEA (WT)	ss[ComEA]	ComEA (G60A)	ComEA (G60V)	ComEA (G60E)	ComEA (K62A)	ComEA (K63A)	ComEA (K63E)	ComEA (K62/63A)
periplasmic localization	+	+	+	+	+	+	+	+	+
uniformly distributed	+	+	+	+	+	+	+	+	+
foci formation + DNA	+	-	+	-	-	+	+	-	-

* all translationally fused to mCherry

Figure 4. *In vivo* validation of ComEA-DNA interaction sites. Strains carrying *comEA-mCherry* or variants thereof on the chromosome were tested for uniform periplasmic localization of the fusion protein and foci formation after the addition of external transforming DNA (A), for DNA uptake (B) as described for Fig. 2, or for natural transformation (C). The transformation frequencies shown on the Y-axis are averaged from at least three independent replicates (± SD). <d.l., below detection limit (on average 3.4×10^{-8} ± SD of 1.1×10^{-8}). Statistically significant differences were determined using Student's t-test on log-transformed values. $^{**}P<0.01$.

doi:10.1371/journal.pgen.1004066.g004

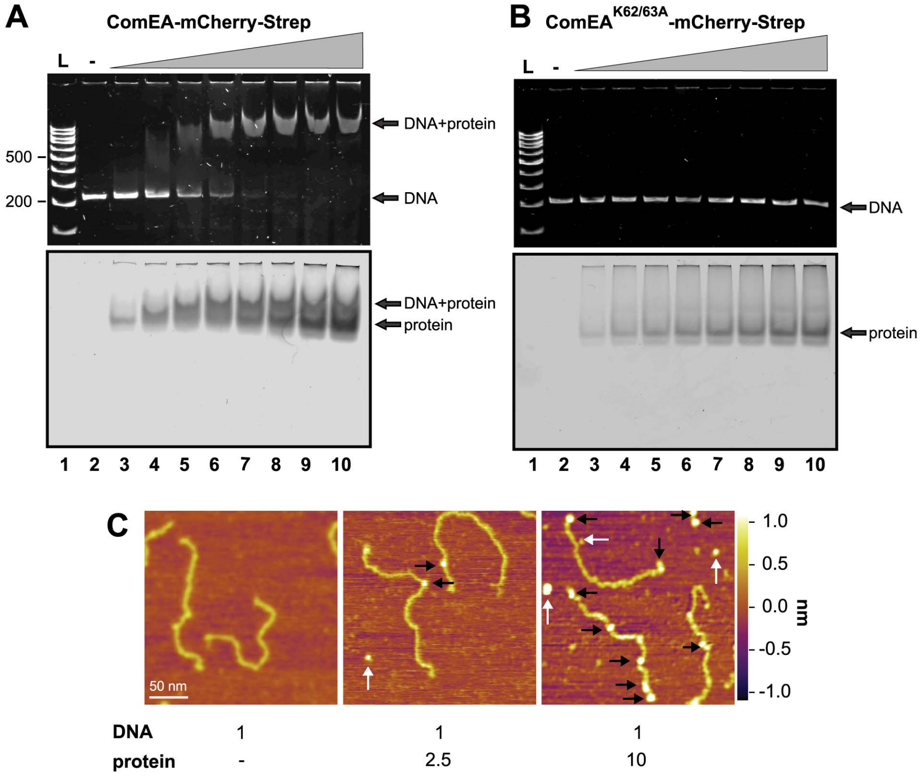


Figure 5. ComEA but not ComEA$^{K62/63A}$ binds to DNA *in vitro*. EMSA using the 200 bp upstream region of the *comEA* gene as a probe (panels A and B). A total of 0.4 pmol of DNA was incubated without or with increasing amounts of the ComEA-mCherry-Strep (A) and ComEA$^{K62/63A}$-mCherry-Strep (B) protein (lanes 2 to 10: 0, 2, 4, 6, 8, 10, 12, 16, and 20 pmol of protein). Free DNA, free protein, and the DNA/protein complex are indicated by the arrows. L: DNA ladder (representative bp are indicated on the left). Panel C: AFM images of DNA, proteins, and DNA/protein complexes absorbed on mica. AFM images from left to right: bare DNA fragments (DNA to protein ratio 1:0); DNA/protein complex at a molecular ratio of 1:2.5; DNA/protein complex at a molecular ratio of 1:10. The proteins bound to the DNA are marked with black arrows; unbound proteins are labeled by white arrows. The height or Z scale is shown on the right and is the same for all three panels displaying 270 nm×270 nm scan areas.

doi:10.1371/journal.pgen.1004066.g005

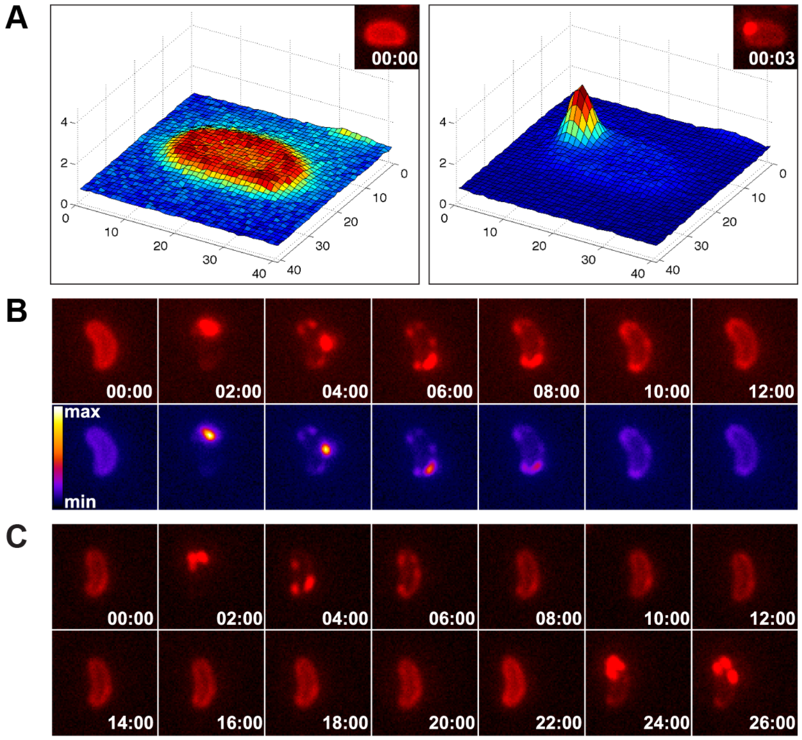


Figure 6. Time-lapse microscopy series of bacteria exposed to exogenous DNA. The images were captured in the red channel to visualize ComEA-mCherry at intervals of 3 sec (A) or 2 min (B). Matlab-computed maximal fluorescence intensity-plots are shown in A (corresponding fluorescent image in inset). Heat-maps showing the fluorescence intensities of the mCherry signal are depicted in the lower row of panel B. (C) Time-lapse microscopy series as in (B), but in a *comEC* minus background. The corresponding movies are available online (movies S2, S3, S4).
doi:10.1371/journal.pgen.1004066.g006

to transport tDNA across the outer membrane, such results would also be consistent with ComEA's main function being to protect and stabilize incoming tDNA against potential nucleases. Indeed, two nucleases have been described for *V. cholerae*, Dns and Xds, which are solely responsible for extracellular nuclease activity in this organism [36]. Interestingly, Focareta and Manning demonstrated that even though Dns can be recovered from culture supernatants, it was also detectable in the periplasmic space of *V. cholerae* [37]. We recently confirmed the extracellular localization of Dns [38] but also its at least partial association with the bacterial cells (through western blot analysis; [25]). Moreover, Blokesch and Schoolnik showed that expression of *dns* has to be silenced in *V. cholerae* to allow natural transformation to occur at high cell density [26,38]. Thus, to rule out the possibility that ComEA might protect incoming tDNA against either of those two nucleases we tested *dns*, *xds*, and *comEA* single, double, and triple mutants for natural transformation and the recovery of DNase resistant tDNA in whole cells (Fig. S6). Notably, the absence of *dns* resulted in higher transformability (Fig. S6A), consistent with an early study [38], and in the detection of increased amounts of DNase-resistant tDNA within the bacteria (Fig. S6B). However, no transformants or translocated tDNA were detectable if *comEA* was concomitantly absent (Fig. S6). We therefore conclude that ComEA's main role is not to protect incoming tDNA against degradation by the nucleases Xds or Dns, though we cannot exclude the presence of any other hitherto unidentified nuclease in the periplasm of *V. cholerae*. Instead, we suggest that translocation of tDNA across the outer membrane is not solely driven through Tfp-like elements of the DNA uptake machinery but also requires ComEA.

In silico prediction and *in vivo* validation of a ComEA-DNA complex

To gain insights into the molecular mechanism through which ComEA binds dsDNA, we predicted the structure of ComEA and characterized the interactions of this protein with the transforming DNA. First, we used comparative modeling to create a 3D structure of ComEA using the X-ray structure of the ComEA-related protein HB8 from *Thermus thermophilus* (PDB ID: 2DUY, unpublished) as a template (Fig. 3, movie S1). Based on structural similarity with structures from the HhH family [39], we identified K62 and K63 as candidate residues for DNA binding interactions and could model the putative ComEA-DNA adduct (Fig. 3B). The electrostatic potential of the ComEA model is consistent with the identified DNA-binding region, showing positively charged regions corresponding to the lysine pair (Fig. 3C).

To validate this model, we used site-directed mutagenesis to create ComEA variants with single or double amino acid substitutions. All *comEA-mCherry* alleles were inserted into the chromosome, thereby replacing the wild-type *comEA* copy. The ComEA-mCherry variants were tested for expression and periplasmic localization, foci formation upon provision of tDNA, for their ability to induce DNA translocation into a DNase resistant state (using the DNA uptake assay) and to restore natural transformation (Fig. 4 and Fig. S7). Consistent with the *in silico* predictions, K63 was of major importance. ComEA$^{\mathrm{K63A}}$ was severely impaired for natural transformation (~250-fold reduction; Fig. 4C), resulting in DNA uptake levels below the limit of detection (Fig. 4B). The substitution of K63 with a negatively charged residue (ComEA$^{\mathrm{K63E}}$) or the concomitant exchange of K62 (ComEA$^{\mathrm{K62/63A}}$) completely abolished natural transformation (Fig. 4C). The ComEA-DNA model also explains why K63 has the major role in DNA binding: while K62 is engaged with a single backbone phosphate moiety, K63 is inserted into the DNA minor groove, chelating the backbone of both strands (Fig. 3B, inset). Moreover, a substitution of the nearby glycine residue at position 60 by alanine had no effect on DNA binding and transformation, whereas strains producing ComEA^{G60V} and ComEA^{G60E} were impaired in DNA uptake and were non-transformable (Fig. 4). We suggest that the combined effect of impairing the interactions of K62 and K63 with the dsDNA (as in the case of ComEA^{G60E}) and perturbing the HhH1 GIG hairpin motif (Fig. 3A) has a major impact on the ability of ComEA to bind DNA.

To further investigate whether the lysine pair is indeed involved in DNA binding we heterologously expressed those variants as tat-ComEA-GFP fusions in *E. coli* (Fig. S8; as for wild-type ComEA in Fig. 2A). Using this approach we showed that the ComEA^{K63E} and ComEA$^{K62/63A}$ variants behaved differently from WT ComEA in that they localized evenly within the cytoplasm. In addition, most of the *E. coli* cells did not show any compaction of the chromosome (and if so the variant did not co-localize with the compacted chromosome). The same phenotype was observable for variants that lacked either of the two HhH motifs (Fig. S8), suggesting that those variants had lost their ability to bind DNA. In contrast, a K63A variant showed an intermediate phenotype (Fig. S8) consistent with the ~250-fold decreased transformation frequency observed for the ComEA^{K63A}-mCherry variant in *V. cholerae* (Fig. 4).

Apart from this patch at HhH1, the only other amino acid important for the *in vivo* functionality of ComEA was the conserved arginine residue at position 71 (Fig. 3A). The DNA uptake ability of ComEA^{R71A} was slightly reduced, and less DNA-protein foci were observed for this variant (Fig. S7). However, the strain containing ComEA^{R71A} remained naturally transformable, a feature that was completely abolished for the ComEA^{R71D} variant. The latter mutant protein was also unable to bind DNA within the periplasmic space and did not foster the uptake of transforming DNA (Fig. S7). Based on our ComEA model structure, R71 is located in a position not particularly favorable for DNA binding (Fig. 3B); therefore, it is likely that R71 might be important for the structural stability of ComEA.

In vitro binding of ComEA to DNA only occurs in the presence of the lysine pair K62/K63

To unambiguously show that the lysine residues are required for DNA binding we purified a tagged (*Strep*-tag II) version of ComEA, ComEA$^{K62/63A}$, ComEA-mCherry, and ComEA$^{K62/63A}$-mCherry (Fig. S9). The purified ComEA protein showed an unexpected UV-Vis spectrum, which was consistent with bound DNA (due to an absorption peak around 260 nm; Fig. S9A). Interestingly, if we compared purified ComEA-mCherry with the ComEA$^{K62/63A}$-mCherry, we observed that the peak at 260 nm was absent in this variant, indicating that the protein was indeed no longer able to bind DNA.

To remove any pre-bound DNA from the ComEA protein we included a DNase treatment step prior to the elution of the protein from the affinity column (see Material and Methods; Fig. S9C and D). All four proteins were tested for *in vitro* binding to DNA using an electrophoretic mobility shift assay (EMSA). Notably, ComEA-mCherry and ComEA bound to DNA in a concentration dependent manner as visualized by the retarded migration of the DNA probe (Fig. 5A and Fig. S10A) and the likewise changed migration of the protein (visualized by the fluorescence of mCherry; Fig. 5A). Notably, the K62/63A variants of ComEA did not change the migration behavior of the DNA probe (Fig. 5B and Fig. S10B), again confirming that the protein had lost the ability for DNA binding.

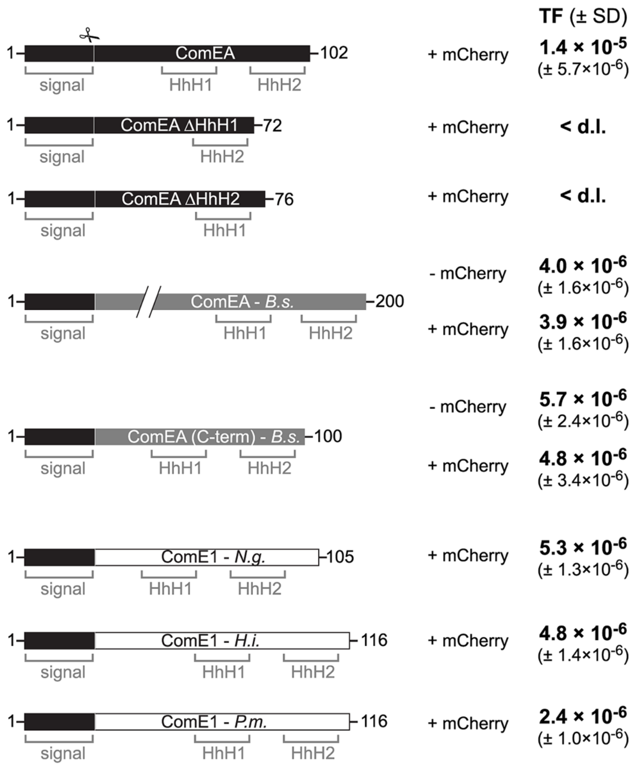

TF (± SD)
1– ComEA –102 + mCherry 1.4 × 10^-5 (± 5.7×10^-6)
signal HhH1 HhH2
1– ComEA ΔHhH1 –72 + mCherry < d.l.
signal HhH2
1– ComEA ΔHhH2 –76 + mCherry < d.l.
signal HhH1
- mCherry 4.0 × 10^-6 (± 1.6×10^-6)
1– ComEA - B.s. –200
signal HhH1 HhH2
+ mCherry 3.9 × 10^-6 (± 1.6×10^-6)
- mCherry 5.7 × 10^-6 (± 2.4×10^-6)
1– ComEA (C-term) - B.s. –100
signal HhH1 HhH2
+ mCherry 4.8 × 10^-6 (± 3.4×10^-6)
1– ComE1 - N.g. –105 + mCherry 5.3 × 10^-6 (± 1.3×10^-6)
signal HhH1 HhH2
1– ComE1 - H.i. – 116 + mCherry 4.8 × 10^-6 (± 1.4×10^-6)
signal HhH1 HhH2
1– ComE1 - P.m. – 116 + mCherry 2.4 × 10^-6 (± 1.0×10^-6)
signal HhH1 HhH2

Figure 7. ComEA homologs from different naturally competent bacteria compensate for the absence of ComEA in *V. cholerae*. First row: Schematic representation of the ComEA protein of *V. cholerae* from amino acid 1 to the end. Domains such as the helix-hairpin-helix motifs (HhH) and the signal sequence (signal) are also depicted. The signal sequence cleavage site (scissor) was predicted through the SignalP 4.1 server [73] for ComEA$^{V.c.}$ and for its homologs. Designed constructs replacing wild-type ComEA on the chromosome of the respective *V. cholerae* strains are indicated below the WT ComEA scheme. First, the two ComEA mutants lacking either of both HhH motifs are indicated. In gray: the ComEA protein of *B. subtilis* (*B.s.*; not to scale compared to the other constructs and as indicated by the two diagonal lines). The transmembrane domain of ComEA$^{B.s.}$ was removed to avoid toxicity/insolubility problems [21]. ComEA(C-term) of *B. subtilis* refers to the C-terminal part of the protein, which still allowed DNA binding *in vitro* [21]. Both ComEA$^{B.s.}$ and ComEA(C-term)$^{B.s.}$ were tested without and with mCherry fused to the C-terminus (− mCherry/ + mCherry). In white: ComEA homologs (ComE1) of the Gram-negative bacteria *Neisseria gonorrhoeae* (*N.g.*), *Haemophilus influenzae* (*H.i.*), and *Pasteurella multocida* (*P.m.*). The original signal sequence of each of those proteins was removed. All constructs were expressed from the native *comEA* promoter and encoded the *V. cholerae* ComEA-specific signal sequence (residues 1–25) to allow proper translocation across the inner membrane. Natural transformation was tested for all strains and the transformation frequencies (TF) are indicated on the right. The average of at least three independent biological replicates is shown (± SD). <d.l., below detection limit.
doi:10.1371/journal.pgen.1004066.g007

It should be noted that the shifted DNA signal was detectable at DNA to protein ratios as low as 1:10 and the probe seemed completely shifted at a ratio of 1:25–30 (Fig. 5A and Fig. S10A), which was significantly lower than what has been described for the *B. subtilis* ComEA homolog (98% of the DNA probe was shifted when 5.5×10^{-11} M of DNA was incubated with 1.6 μM of purified protein; [21]) or for the neisserial ComE ortholog [23]. A possible explanation for this difference could be that the ComEA/ComE proteins investigated in those earlier studies were pre-occupied by DNA as we observed for ComEA of *V. cholerae* in the absence of DNase treatment.

V. cholerae ComEA does not show cooperativity for DNA binding

Provvedi and Dubnau suggested that the *in vitro* DNA binding behavior of the ComEA protein of *B. subtilis* was indicative of cooperative binding [21]. To test whether any cooperative binding was observable for ComEA of *V. cholerae* we used Atomic Force Microscopy (AFM). AFM allows investigating the extent of ComEA-mCherry binding to a DNA fragment and to also determine where on the DNA the protein is bound (e.g. fractional occupancies at any specific site, binding to the ends, or to nonspecific sites). To minimize overestimation of the binding affinity that can occur in the case when coverage of protein on the surface is too high, such that the protein coincidently lands on DNA, we kept the DNA-protein molecular ratio low by not exceeding a ratio of 1:10 (DNA to protein). Prior to AFM imaging, we pre-incubated the ComEA-mCherry protein with a random PCR fragment (809 bp) at a molecular ratio of 1:2.5 or 1:10. As illustrated in Fig. 5 we observed a mixture of bare DNA molecules, free protein molecules, and protein/DNA complexes. To identify the ComEA-mCherry protein in topographic AFM images we used height and width criteria (height >2 nm, width from 10 to 20 nm). Using an approach reported by Yang *et al.* [40] we found that the probability of protein molecules located on DNA was 5 times higher than it would be for stochastically binding of the protein to the mica surface. Moreover, in the case of a DNA to protein ratio of 1:10 we observed 2.5-fold higher affinity of the protein to the free ends of DNA than to random sites on the DNA strand. These AFM data indicate that, at least at the measured concentrations, no cooperative binding of the ComEA protein to DNA occurred and again contradicts the hypothesis that binding of ComEA might primarily protect the tDNA from degradation. Such protective effect has been demonstrated for the competence protein DprA of *Streptococcus pneumoniae* [41], which binds the single-stranded tDNA after its translocation into the cytoplasm. Indeed, Mortier-Barrière *et al.*, described in their study that DprA binding to DNA appeared to be cooperative since fully covered protein-DNA complexes were observed next to free ssDNA molecules at a protein to nucleotide ratio of 1:20. We never observed such scenario for ComEA's binding to dsDNA using AFM (though we used a ~4-fold lower protein to nucleotide ratio).

Entry of DNA into the periplasm of naturally competent *V. cholerae* cells occurs at one distinct location

Interestingly, a passive DNA uptake mechanism has recently been proposed for single-stranded T-DNA translocation into plant cells involving the VirE2 protein of *Agrobacterium tumefaciens* [42]. We reasoned that if a similar mechanism is responsible for DNA uptake in competent *V. cholerae* cells, although dsDNA is involved and ComEA shows no similarity to VirE2, then the aggregation of ComEA should occur at one distinct DNA entry point (most likely next to the PilQ secretin). To test this hypothesis, we performed time-lapse microscopy experiments using ComEA-mCherry-expressing *V. cholerae* strains in the presence of external DNA (Fig. 6). We consistently observed the accumulation of ComEA as one large focus before smaller subclusters separated from the main ComEA focus and spread throughout the periplasm until the uniform localization of ComEA was restored (Fig. 6, movies S2, S3, S4).

The function of ComEA might be conserved among naturally competent bacteria

Based on the data presented above we hypothesize that ComEA might play a direct role in the translocation of DNA across the outer membrane solely based on its ability to bind to DNA. If this were the case then ComEA homologs of other naturally competent bacteria should be able to replace ComEA of *V. cholerae*. And indeed, ComEA of *B. subtilis* was able to efficiently compensate for the absence of ComEA of *V. cholerae* (Fig. 7). Moreover, even the C-terminal $(HhH)_2$ motif of ComEA of *B. subtilis* alone, which was shown to bind DNA *in vitro* [21], was sufficient to restore natural transformation of a *comEA* negative *V. cholerae* strain as were the ComEA homologs from *N. gonorrhoeae, Haemophilus influenzae*, and *Pasteurella multocida* (Fig. 7). It should be noted that Sinha *et al.* suggested that *H. influenzae* might contain an additional but so far unidentified paralog of *comE1* due to the modest effect observed for a *comE1* minus strain [43].

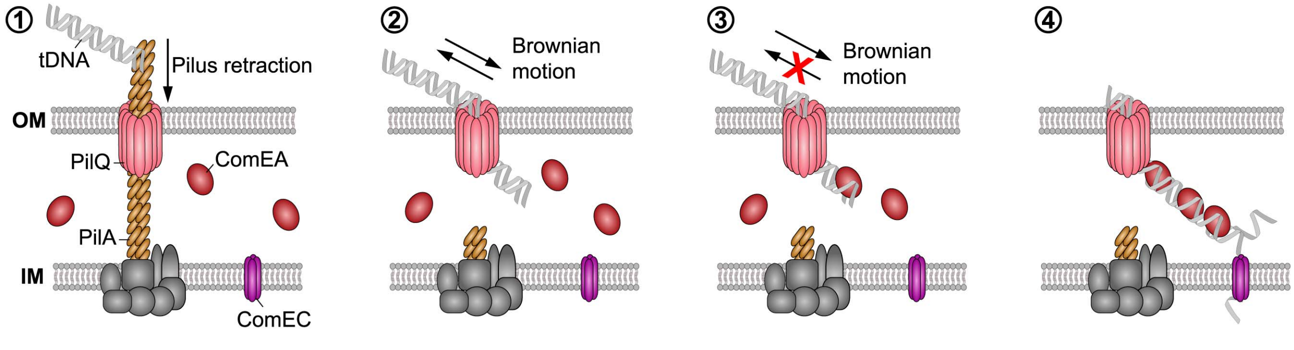


Figure 8. Working model of how DNA translocation across the outer membrane might occur in *V. cholerae* (based on the current study and [2–4,13,14]). The key components addressed in this study are indicated. It has been suggested that a (pseudo)pilus [1,2], which is similar to type 2 secretion systems (T2SS) and type IV pili (Tfp), represents a core element of the DNA uptake machinery [11]. It is assumed that the Tfp crosses the outer membrane through a secretin pore formed by PilQ [2,4,74,75]. This secretin could also provide the point of entry for incoming DNA. A single pilus retraction event might open the secretin pore so that short stretches of the tDNA can enter the periplasm by Brownian motion (or through partial binding to the pilus structure). ComEA would then bind to the tDNA via its HhH-associated lysine residues favoring translocation via a Brownian ratchet mechanism, which is modulated by the effective ComEA DNA-binding kinetics, binding spacing, and concentration. The ComEA-loaded tDNA might eventually interact with the inner membrane transporter ComEC [76], which would transport the DNA into the cytoplasm. Similar to Gram-positive bacteria, it is assumed that incoming DNA enters the cytoplasm of Gram-negative bacteria single-stranded.
doi:10.1371/journal.pgen.1004066.g008

It is tempting to speculate that ComEA might fulfill a similar role in Gram-positive bacteria. Indeed, the localization of ComEA has been previously described for *B. subtilis* [16,17,19] but those studies were either based on immunofluorescence microscopy [16], which does not allow following protein localization over time, or were done in the absence of tDNA [17,19]. Therefore, it was concluded by Kaufenstein *et al.* that ComEA localizes to many sites of the cell membrane and only occasionally co-localizes with the polar DNA uptake machinery, which was mainly achieved by changing the artifical inducer concentration [19]. However if the cell wall would be considered as a similar barrier in Gram-positive bacteria as the outer membrane is in Gram-negatives, creating a kind of periplasmic space between the cell wall and the (inner) membrane as suggested by Matias and Beveridge [44], then the binding of ComEA could also participate in the transport of DNA across the cell wall layer. However, in contrast to ComEA of Gram-negative bacteria, ComEA of Gram-positives is anchored to the membrane and therefore accumulation of ComEA can only occur in two dimensions, which might still be sufficient to prevent backward diffusion of the tDNA and contribute to DNA translocation across the cell wall. Notably, while this article was under revision Bergé *et al.* published a study on the nuclease EndA of naturally competent *Streptococcus pneumoniae* [45]. The authors demonstrated that EndA aggregates at midcell in this Gram-positive bacterium and that this recruitment is dependent on "the dsDNA receptor" ComEA [45]. Interestingly, ComEA also localized to the midcell and the authors speculated "a direct interaction of EndA and ComEA, an hypothesis which received indirect support" [45].

A working model for ComEA-dependent DNA translocation across the outer membrane

Our findings suggest that the ability of ComEA proteins to bind to dsDNA emerging from the PilQ pore can potentially prevent the retrograde movement of the substrate, and ComEA binding might contribute to pull DNA into the periplasm (Fig. 8). It has been suggested that ratcheting produced through binding proteins can significantly accelerate translocation events [46,47], as for the case of phage DNA injection into bacterial cells [15]. Based on our data, a similar mechanism can be envisioned for the ComEA-mediated transfer of DNA into the periplasm, with the rate of uptake depending on the specific binding kinetics and concentration of ComEA [15]. We hypothesize that ComEA-mediated DNA internalization might start occurring once short stretches of tDNA would enter the periplasm (most likely through the outer membrane secretin PilQ and potentially after a single Tfp retraction event). The ratio between the periplasmic ComEA protein and the incoming tDNA should be high at that stage thereby leading to an increased ComEA effective binding density, which, potentially together with the higher affinity of ComEA for DNA ends as observed by AFM (Fig. 5), would promote efficient DNA internalization. The absence of cooperative ComEA-DNA binding revealed by our AFM data (Fig. 5) is not an obstacle to a ComEA-mediated ratchet mechanism of internalization, as cooperativity would only contribute to increase the relative speed of the process [15,46,47]. The binding of proteins has undeniably been recognized as a driving force, both in the translocation of proteins as well as of DNA [48]. To this extent, Salman *et al.* investigated the translocation of double-stranded (ds) DNA through the nuclear pore complex using a combination of epifluorescence microscopy and single-molecule manipulation techniques [49]. They presented evidence that the DNA uptake process in their reconstituted system was based on a passive ratchet, directed by the retention of the already translocated segment of the DNA [49]. We suggest that ComEA might play a similar role in the DNA uptake process in naturally competent *V. cholerae* cells.

In summary, we used a cell biological approach to better understand DNA uptake in naturally competent *V. cholerae* cells. We visualized the competence protein ComEA and observed the *in vivo* binding of this protein to dsDNA in real time. Structural modeling and AFM experiments suggested that the binding of ComEA to DNA is primarily responsible for DNA translocation across the outer membrane. Consistent with this suggestion, ComEA variants unable to bind to DNA *in vivo* were also defective in promoting DNA uptake and natural transformation. We hypothesize that ComEA encounters incoming DNA immediately after short stretches of DNA have crossed the outer membrane (through the PilQ secretin or in exceptional cases also in a Tfp-independent manner [11]) and that ComEA subsequently promotes DNA translocation across the outer membrane without the need for any external energy source (Fig. 8). ComEA might therefore be more than a DNA receptor protein, but rather a crucial player for mediating DNA uptake in *V. cholerae* and potentially also other naturally competent bacteria.

Materials and Methods

Bacterial strains, plasmids, and culture conditions

Vibrio cholerae strains and plasmids used in this study are listed in Table S1. *Escherichia coli* strain DH5α [50] was used as host for cloning purposes and for heterologous expression of ComEA and its variants for protein purification. Genomic DNA (gDNA) extracted from *E. coli* BL21 (DE3) [51] was utilized to test DNA uptake by PCR as described [24]. *E. coli* S17-1λpir [52] served as donor strain for bacterial mating with *V. cholerae*.

All *V. cholerae* and *E. coli* strains were grown aerobically in Luria-Bertani (LB) medium at 30°C and 37°C, respectively. Solid LB plates contained 1.5% agar. For *tfoX* expression and induction of other constructs under control of the P_{BAD} promoter the LB medium was supplemented with 0.02% L-arabinose (L-ara). For expression of *tat-gfp*, *tat-comEA-gfp*, and its derivatives in *E. coli* DH5α (Fig. 2A and Fig. S8) L-ara concentrations were lowered to 0.002%. Thiosulfate Citrate Bile Salts Sucrose (TCBS) agar plates were prepared following the manufacturer's instructions (Fluka) and used to counterselect *E. coli* after bacterial mating. For sucrose-based counterselection, NaCl-free LB medium containing 6% sucrose was used. LB medium and LB agar plates were supplemented with antibiotics when required. Final concentrations of antibiotics were 50 μg/ml, 75 μg/ml and 100 μg/ml for gentamicin, kanamycin, and ampicillin, respectively. The ampicillin concentration was lowered to 50 μg/ml for *V. cholerae* strains induced for competence.

Recombinant DNA techniques

Standard molecular biology-based methods were used for DNA manipulations. Restriction enzymes and DNA modifying enzymes were obtained from New England Biolabs, Taq DNA polymerase (GoTaq) was obtained from Promega and used for colony PCR, and Pwo DNA Polymerase (Roche) was used for high-fidelity PCR amplifications. Modified DNA sequences were verified using Sanger sequencing (Microsynth, CH).

Plasmid construction

All plasmid constructs were based on pBAD/Myc-HisA (Invitrogen), which contains the *ara*BAD (P_{BAD}) promoter followed by a multiple cloning site (MCS) for dose-dependent protein expression. A derivative of pBAD/Myc-HisA, pBAD(kan), was created through substitution of the ampicillin resistance cassette (*bla*) with a kanamycin resistance cassette (*aph*). The genes and translational fusion constructs were PCR amplified and cloned into the MCS of pBAD/Myc-HisA or pBAD(kan). For the amplification of *V. cholerae* genes, the gDNA of strain A1552 [53] served as a template. The accuracy of the plasmids was verified through sequencing.

Strain constructions

Genes were deleted from the parental strain A1552, using either a gene disruption method based on the counter-selectable plasmid pGP704-Sac28 [54], or natural transformation and FLP recombination, as recently described (TransFLP method [55–57]).

Strains containing *comEA-mCherry* or site-directed variants thereof were constructed using the TransFLP method [55–57]. For the construction of ComEA site-directed variants, a silent 'watermark' restriction site was inserted close to or including the changed nucleotide sequence. This watermark simplified screening purposes after homologous recombination.

The *comEA*$^{B.s.}$ gene (or parts thereof) was amplified from gDNA derived from *B. subtilis* strain 168. The DNA fragment containing *comE1* from *Neisseria gonorrhoeae (N.g.*; *Neisseria gonorrhoeae* strain FA 1090, NCBI Reference Sequence: NC_002946.2; locus YP_208252), *Haemophilus influenzae* (*H.i.*; *Haemophilus influenzae* strain R2846, NCBI Reference Sequence: NC_017452.1; locus YP_005829750), and *Pasteurella multocida* (*P.m.*; *Pasteurella multocida subsp. multocida* str. Pm70, NCBI Reference Sequence: NC_002663.1; locus NP_246604, hypothetical protein PM1665) was synthesized using the GeneArt® Strings™ technology (Life technologies/Invitrogen) and served as PCR template for the TransFLP strain construction method [55–57]. The beta-lactamase gene (*bla*) was amplified from plasmid pBR-flp [55–57]. All strains were verified through colony PCR (in part followed by restriction enzyme digestion according to inserted watermarks) and confirmed through PCR amplification and sequencing.

Wide-field fluorescence microscope settings and image analysis

Microscopy images were obtained using a Zeiss Axio Imager M2 epifluorescence microscope. Details about the instrumentation and configurations are provided elsewhere [25]. All bacterial samples were mounted on 2% agarose/PBS pads. Image processing and annotation was done using ImageJ and Adobe Illustrator.

Microscopy of strains expressing fluorescent fusion proteins

Strains carrying fluorescent fusion constructs were grown aerobically for ~5 h in LB supplemented with the respective antibiotics and 0.02% L-arabinose (0.002% L-ara for *E. coli* experiments; Fig. 2 and Fig. S8). The strains carrying chromosomally encoded fluorescent fusion proteins were grown aerobically and at 30°C in LB supplemented with 0.02% L-ara for ~7 h (OD_{600} 2.5; [11]). The samples were washed once in PBS and immediately imaged.

The staining of chromosomal DNA was performed through the addition of 4′,6-diamidino-2-phenylindole (DAPI; final concentration 5 µg/ml) to the bacterial cultures for at least 5 min.

To characterize the ComEA-mCherry localization dynamics during DNA uptake, *comEA-mCherry*-expressing strains were grown as described above. A total of 50 µl of washed culture was mixed with 1 µg of either gDNA derived from *V. cholerae* strain A1552-lacZ-Kan [58], commercially available phage lambda DNA (Roche) or a 10.3 kb fragment amplified through PCR. After 5 min of incubation with the DNA the bacteria were mounted on agarose pads and imaged. To visualize the DNA during the relocalization of ComEA-mCherry, phage lambda DNA (Roche) was pre-stained with 10 µM YoYo-1 (Molecular Probes/Invitrogen) at 4°C corresponding to a base pair to dye ratio of 15:1. The bacterial culture was mixed with the pre-stained DNA and incubated for 20 min. The cells were washed in PBS, mounted on agarose pads and imaged.

For time-lapse microscopy, the samples were prepared as described above, but immediately imaged after the addition of DNA. The images were taken every 3 or 120 sec as indicated in the figure and movie legends. For time-lapse imaging, the agarose pads were sealed using a mixture of Vaseline, lanolin and paraffin (VALAP).

Fluorescence loss in photobleaching

Fluorescence loss in photobleaching (FLIP) experiments were performed on a Zeiss LSM710 microscope equipped with a 561 nm solid-state laser (20 mW). A Plan-Apochromat 63×/1.40 Oil objective was used. The microscope was controlled with the Zen 2009 software suite (Zeiss). Time intervals ranged from 104 to 120 ms/frame for live cells to max. 160 ms/frame for fixed cells. The maximum (100%) laser power was used for bleaching.

V. cholerae strains ΔcomEA-Tn*tfoX* harboring pBAD(kan)-*comEA*-mCherry or pBAD(kan)-ss[ComEA]-mCherry were grown aerobically for 5 h in LB supplemented with 0.02% arabinose and 75 µg/ml of kanamycin. After the cells were mounted, the slides were sealed and the bacteria were immediately imaged (live samples; Fig. 1 and Fig. S2A). Alternatively, the cells were fixed for 30 min (4% paraformaldehyde/150 mM phosphate buffer) before imaging (fixed samples; Fig. S2B).

For FLIP data acquisition a circular bleaching region of ~440 nm width was defined at one cell pole (region-of-interest (ROI) 1; labeled as 1 in Fig. 1). A circular ROI of the same size was defined at the opposite cell pole of the same bacterium (labeled as 2 in Fig. 1) and in an adjacent cell (labeled as 3 in Fig. 1). The average fluorescence intensity of all regions was recorded. Bleaching of ROI 1 was initiated after a lag of 20 frames and repeated after each frame. The acquired data were exported and processed in 'R' [59]. The recorded fluorescence intensities were normalized to the average fluorescence intensity of the first 10 frames. Moving averages were calculated using the SMA(x, n = 5) function from the 'TTR' package [59].

Natural transformation assays

Transformation assays were performed as previously described [25] with gDNA of strain A1552-lacZ-Kan [58] as transforming material. Transformation frequencies were calculated as the number of transformants divided by the total number of colony forming units (CFU). Differences in transformation frequencies were considered significant for *P*-values below 0.05 (*) or 0.01 (**) as determined by Student's t-test on log-transformed data.

Detection of DNA uptake by PCR

DNA uptake was verified using a whole-cell duplex PCR assay as described [24] with slight modifications. Briefly, competence-induced bacteria were grown aerobically until an OD_{600} of 1.0–1.5 before genomic DNA (gDNA) (2 µg/ml) of *E. coli* strain BL21 (DE3) was added for 2 h. For the uptake of YoYo-1-labeled DNA gDNA of *E. coli* strain BL21 (DE3) was pre-labeled as described for the microscopy experiments and YoYo-1 was maintained in the solution throughout the 2 h incubation period. Next, cells were harvested and treated with DNase I (Roche) for 15 min at 37°C. Excess nuclease was removed by washing and cells were resuspended in 100 µl PBS. ~3×10^6 bacteria were used as template in a whole-cell duplex PCR. Primer pairs were specific for the donor DNA derived from *E. coli* BL21 (DE3) and for gDNA of the *V. cholerae* acceptor strain (at a 10-fold lower concentration). The latter reaction served as control for the total number of acceptor bacteria [24].

3D modeling of ComEA, its interaction with dsDNA, and molecular simulations

A 3D model structure was produced for ComEA (truncating the first 37 residues including the 25 residue-containing signal peptide) using comparative modeling (MODELLER package [60]) on the *Thermus thermophilus* HB8 (PDB ID: 2DUY) template (with 43% sequence identity) (Fig. 3). The ComEA-DNA complex was modeled, to identify structurally similar DNA-binding proteins using the DALI server [39]. The DNA polymerase, PolC, from *Geobacillus kaustophilus* (PDB ID: 3F2D) [61] was selected as the best match, with 24% sequence identity and a root mean square deviation (RMSD) of 2.4 Å compared with the modeled ComEA of *V. cholerae*. The PolC X-ray structure complexed with DNA was used to identify potential DNA poses on the *V. cholerae* ComEA model using the Chimera MatchMaker tool [62]. This assessment led to the production of a DNA-ComEA model (Fig. 3B, movie S1), which was further refined and equilibrated using the minimization and molecular simulations detailed below. The estimated binding energy for the ComEA-DNA association is in the order of 29±8 kcal/mol, based on MM/PBSA calculation on the MD trajectory.

Molecular dynamics simulation was used to relax and study the dynamics and energetics of ComEA and the ComEA-DNA complex for 55 and 50 ns, respectively. The MD simulations were run using the NAMD simulation package [63] with Amber force field (with Barcelona modification for nucleic acids [64] and the TIP3P water model [65]. The systems were first energy minimized using constrained C-alpha atoms, followed by analysis without any constraint for 2000 steps. To equilibrate the system, the temperature was gradually increased up to 300 K in the NVT ensemble and maintained at 300 K for 100 ps with a 1 fs time step. Finally, an NPT simulation was run at 300 K for 500 ps with a 2 fs time step to complete the equilibration procedure. The equilibrated structure was used as starting point for production simulations. All production MD simulations were run at 1 bar with a time step of 2 fs, using SHAKE algorithm [66] on all bonds and PME [67] for treating electrostatic interactions. To control the temperature and the pressure, Langevin dynamics and the Nose-Hoover Langevin piston, respectively, were used [68,69]. The trajectories were saved every 500 steps in the production simulations. To characterize the binding affinity of different systems, the free binding energies were calculated using the MMPBSA.py package [70]. 100 frames were sampled from the trajectories for analysis using MMPBSA.py. The entropy portion of the free energy was not considered in the calculation. In addition, the PME module in VMD was used to estimate the electrostatics potential of the modeled ComEA monomers (Fig. 3C).

Purification of recombinant ComEA and its variants

ComEA, ComEA$^{K62/63A}$, ComEA-mCherry, and ComEA$^{K62/63A}$-mCherry (all containing the eight amino acid *Strep*-tag II sequence at the C-terminus) were purified as previously described [71] with minor modification. Briefly, *E. coli* cells containing the respective plasmids (Table S1) were grown aerobically at 37°C until an OD_{600} of 1.0. At that time expression was induced by the addition of 0.2% arabinose to the culture medium and the cells were further incubated for 2 hours before their harvest at 4°C and storage of the cell pellet at −80°C. The cells were lysed by sonication (Vibra-cell; 10 min. in total with 30 sec on and 30 sec off intervals and an amplitude of 80%) and the lysate was further processed as described [71]. Notably, after realization that the protein was pre-occupied by DNA (see results section), we included a on-column DNase treatment step (10 µg/ml of DNase I (Roche) in 100 mM Tris/HCl pH 8.0 buffer containing 20 mM $MgCl_2$ and 0.2 mM $CaCl_2$; 30 min. at 30°C) after the soluble protein fraction was loaded onto the streptactin resin and washed with 5 column volumes of washing buffer. The DNase I treatment step was followed by extensive washing of the column (10 to 30 volumes) before the respective protein was eluted as described [71]. The eluted proteins were concentrated using Amicon Ultra spin columns (with a MWCO of 3 kDa or 10 kDa; Millipore). For the AFM experiment, the protein was dialyzed against AFM buffer (5 mM Tris/HCl pH = 8.0 and 10 mM $MgCl_2$). The protein concentration was determined according to Bradford [72].

Electrophoretic Mobility Shift Assays (EMSAs)

Electrophoretic Mobility Shift Assays were basically performed as previously explained [71]. However, as preliminary experiments indicated that neither the absence of DTT nor the storage of the protein in the absence of glycerol and at 4°C did change the results of the experiments, the protocol was changed accordingly. The 200 bp DNA fragment was PCR-amplified using gDNA of strain A1552 as template and represented the upstream region of the *comEA* gene. Other DNA fragments (e.g. the *aphA* promoter region as previously tested [71]) were similarly shifted (data not shown). The protein/DNA mixture was incubated for 5 min at room temperature before electrophoretic separation on an 8% polyacrylamide gel. DNA was visualized by ethidium bromide staining [71] whereas the fusion proteins (ComEA-mCherry and ComEA$^{K62/63A}$-mCherry were detected using a Typhoon scanner (GE Healthcare; excitation at 532 nm (green) and emission detected with a 610 BP30 (red) filter).

Atomic Force Microscopy (AFM)

To prepare the protein/DNA complex we mixed 0.85 ng/µl of a PCR-amplified DNA fragment (809 bp) with the protein in the molecular ratios of 1:2.5 and 1:10 (DNA:protein) in buffer containing 5 mM Tris/HCl pH 8.0 and 10 mM $MgCl_2$. After incubation for 10 min at 37°C, 15 µl of the mixture was deposited on freshly cleaved mica and rinsed thoroughly with ddH_2O for two minutes. Preparation of the sample with bare DNA was done under the same conditions but in the absence of the protein. The AFM images were acquired in air and in tapping mode using an Asylum Research Cypher microscope. We used Olympus silicon cantilevers (Olympus OMCL-AC240TS-R3) with a spring constant of 1.7 N/m and a resonant frequency of 70 kHz. The typical scan rate was 2.0 Hz.

Supporting Information

Figure S1 Periplasmic localization of ComEA. *V. cholerae* wild-type strain (A1552-Tn*tfoX*) and strain ComEA-bla-Tn*tfoX* (encoding a translational fusion between ComEA and beta-lactamase) were grown for 3 h at 30°C in LB medium in the absence or presence of the competence inducer L-arabinose (0.02 or 0.2% as shown on the left). Ampicillin (50 µg/ml) was added to the indicated cultures (+AMP) and growth of all cultures was resumed for 3 h. The protective effect of the ComEA-bla fusion protein located in the periplasm was checked by phase contrast imaging (boxed region).
(PDF)

Figure S2 Fluorescence loss in photobleaching (FLIP) experiment of ss[ComEA]-mCherry and of ComEA-mCherry in fixed cells. (A) Live *V. cholerae* cells expressing *ss*[ComEA]-*mcherry* (mCherry preceded solely by the signal sequence of ComEA; residues 1 to 25) were tested for mCherry mobility within the periplasmic space using FLIP. (B) The same bacterial strain as in Fig. 1B was tested, but the cells were fixed before microscopy. The settings for (A) and (B) were as described for Fig. 1B. Scale bars, 2 µm.
(PDF)

Figure S3 Localization of chromosomally encoded ComEA. The *comEA* gene of *V. cholerae* was replaced with the *comEA-mCherry* or the *ss*[ComEA]-*mCherry* allele using bacterial genetics (TransFLP [55–57]). The DNA was stained with DAPI. The fusion proteins were localized as in Fig. 1A. Images from left to right: mCherry channel (red), DAPI-stained chromosomal DNA (blue), overlaid fluorescent channels (merge), and phase contrast channel (Ph). Scale bar, 2 µm.
(PDF)

Figure S4 Expression and localization of ComEA-mCherry under chitin-inducing competence conditions. The *V. cholerae* strain harboring the *comEA-mCherry* translational fusion on the chromosome was grown on chitin surfaces for ~24 h as described [25]. The bacteria were mounted for microscopy in the absence (left) or presence (right) of external gDNA. ComEA-mCherry was visualized in the red channel. The edge of the chitin bead is indicated with the dotted line. Scale bar, 2 µm.
(PDF)

Figure S5 ComEA is required for foci formation of YoYo-1-labeled DNA. (A) Visualization of YoYo-1-stained transforming DNA (green channel) in wild-type (WT), or in a *pilA* or *comEA* negative strain. The outline of the cells is shown in the phase contrast image (Ph). Scale bar, 2 µm. (B) DNA uptake assay using the indicated strains and YoYo-1-labeled tDNA as donor DNA. Details as in Fig. 2. L, ladder.
(PDF)

Figure S6 The absence of the nucleases Dns and Xds does not rescue the ΔcomEA phenotype. (A) Natural transformation of the indicated strains was scored using a chitin-independent transformation assay [25]. The transformation frequencies shown on the Y-axis are averaged from at least three independent replicates (± SD). <d.l., below detection limit. Statistically significant differences are indicated (**$P<0.01$); n.s. not statistically different. (B) DNA uptake assay as described for Fig. 2. The genotypes of the tested strains are indicated above the figure. L, ladder.
(PDF)

Figure S7 Localization and functionality of ComEA-mCherry variants. Additional ComEA-mCherry variants were tested for uniform periplasmic localization and the ability to aggregate after the addition of transforming DNA. (A) Table as in Fig. 4. (B) Representative images (for panel A and Fig. 4) showing uniform periplasmic localization, foci formation upon DNA binding, and DNA-independent aggregation (only observed for ComEA$^{N43I/N45A}$). (C) DNA uptake assay of selected variants as described for Fig. 2. (D) Natural transformation assay as described for Fig. 4. The average of at least three independent biological replicates is shown (± SD). <d.l., below detection limit.
(PDF)

Figure S8 Localization of tat-ComEA-GFP variants. Variants tested: tat-ComEA-GFP (WT; A), tat-ComEA^{K63A}-GFP (B), tat-ComEA^{K63E}-GFP (C), tat-ComEA$^{K62/63A}$-GFP (D), tat-ComEAΔHhH1-GFP (E), and tat-ComEAΔHhH2-GFP (F). Details are as in Fig. 2A. Scale bar, 2 µm.
(PDF)

Figure S9 Purification of ComEA, ComEA$^{K62/63A}$, ComEA-mCherry and ComEA$^{K62/63A}$-mCherry. ComEA and its variants (all containing a C-terminal *Strep*-tagII sequence) were purified by affinity chromatography. UV-Vis spectra of purified ComEA-Strep (A) or ComEA-mCherry-Strep (B, red line) and ComEA$^{K62/63A}$-mCherry-Strep (B, back line) were recorded. Panel C and D: Purification of the ComEA-mCherry-Strep (C) and ComEA$^{K62/63A}$-mCherry-Strep (D) protein followed by 11% SDS PAGE of the pooled fractions at each step. Lane 1, molecular mass (kDa) standard; lanes 2 to 10: cell lysate of the respective *E. coli* strains before and after induction, S17 extract, aliquots of the last two washing steps after on-column DNase treatment, and elution fractions 1 to 4. The gels were stained with Coomassie. The respective UV-Vis spectra are indicated below the gel images.
(PDF)

Figure S10 ComEA behaves similar as ComEA-mCherry in EMSA. EMSA were performed using purified ComEA-Strep (A) and ComEA$^{K62/63A}$-Strep (B). Details as described in Fig. 5.
(PDF)

Movie S1 3D model of ComEA based on comparative modeling using the ComEA-related protein of *T. thermophilus* HB8 (PDB ID: 2DUY) as template and its predicted DNA binding site. Residues K62, K63 and R71 are highlighted (as in Fig. 3).
(MPG)

Movie S2 Time-lapse microscopy series of *V. cholerae* exposed to exogenous DNA. Images were taken in the red channel to visualize ComEA-mCherry at intervals of 3 sec (as in Fig. 6A).
(MOV)

Movie S3 Time-lapse microscopy series in the presence of exogenous DNA. Images were taken in the red channel to visualize ComEA-mCherry at intervals of 2 min (as in Fig. 6B).
(MOV)

Movie S4 Time-lapse microscopy series of a *V. cholerae* strain lacking the inner membrane channel ComEC (ΔcomEC). Images were taken in the red channel to visualize ComEA-mCherry at intervals of 2 min (as Fig. 6C). Please note that two DNA uptake events occur within the total duration of 36 min.
(MOV)

Table S1 Bacterial strains and plasmids.
(DOCX)

Acknowledgments

We acknowledge Tiziana Scrignari and Clémentine Thévoz for technical assistance and Thierry Laroche and Arne Seitz from the Bioimaging facility (PT-BIOP) within the School of Life Sciences of EPFL for help with the FLIP experimentation. We also thank Prof. Andras Kis (Laboratory of Nanoscale Electronics and Structures, EPFL) for training and access to the atomic force microscope. Further acknowledgments go to the three anonymous reviewers as well as to a self-consulted reviewer for their very valuable input.

Author Contributions

Conceived and designed the experiments: PS HPM RDB LJS AR MDP MB. Performed the experiments: PS HPM SB RDB LJS MDP MB. Analyzed the data: PS HPM RDB LJS AR MDP MB. Contributed reagents/materials/analysis tools: PS HPM SB RDB LJS AR MDP MB.

Wrote the paper: MB. Conceived the project: PS MB. Performed the cellular biology-based experiments: PS MB. Designed and constructed the bacterial strains and plasmids: PS SB MB. Performed biochemical experiments/protein purification: SB MB. Performed the biomolecular modeling experiments: HPM MDP. Performed the atomic force microscopy experiments: RDB LJS AR. Contributed to the text: PS HPM RDB LJS AR MDP.

References

1. Chen I, Christie PJ, Dubnau D (2005) The ins and outs of DNA transfer in bacteria. Science 310: 1456–1460.
2. Chen I, Dubnau D (2004) DNA uptake during bacterial transformation. Nat Rev Microbiol 2: 241–249.
3. Allemand JF, Maier B (2009) Bacterial translocation motors investigated by single molecule techniques. FEMS Microbiol Rev 33: 593–610.
4. Burton B, Dubnau D (2010) Membrane-associated DNA transport machines. Cold Spring Harb Perspect Biol 2: a000406.
5. Claverys JP, Prudhomme M, Martin B (2006) Induction of competence regulons as a general response to stress in gram-positive bacteria. Annu Rev Microbiol 60: 451–475.
6. Seitz P, Blokesch M (2013) Cues and regulatory pathways involved in natural competence and transformation in pathogenic and environmental Gram-negative bacteria. FEMS Microbiol Rev 37: 336–363.
7. Lorenz MG, Wackernagel W (1994) Bacterial gene transfer by natural genetic transformation in the environment. Microbiol Rev 58: 563–602.
8. Meibom KL, Blokesch M, Dolganov NA, Wu C-Y, Schoolnik GK (2005) Chitin induces natural competence in *Vibrio cholerae*. Science 310: 1824–1827.
9. Lipp EK, Huq A, Colwell RR (2002) Effects of global climate on infectious disease: the cholera model. Clin Microbiol Rev 15: 757–770.
10. Blokesch M (2012) Chitin colonization, chitin degradation and chitin-induced natural competence of *Vibrio cholerae* are subject to catabolite repression. Environ Microbiol 14: 1898–1912.
11. Seitz P, Blokesch M (2013) DNA-uptake machinery of naturally competent *Vibrio cholerae*. Proc Natl Acad Sci USA 110: 17987–17992.
12. Pelicic V (2008) Type IV pili: *e pluribus unum*? Mol Microbiol 68: 827–837.
13. Claverys JP, Martin B, Polard P (2009) The genetic transformation machinery: composition, localization, and mechanism. FEMS Microbiol Rev 33: 643–656.
14. Krüger NJ, Stingl K (2011) Two steps away from novelty-principles of bacterial DNA uptake. Mol Microbiol 80: 860–867.
15. Inamdar MM, Gelbart WM, Phillips R (2006) Dynamics of DNA ejection from bacteriophage. Biophys J 91: 411–420.
16. Hahn J, Maier B, Haijema BJ, Sheetz M, Dubnau D (2005) Transformation proteins and DNA uptake localize to the cell poles in *Bacillus subtilis*. Cell 122: 59–71.
17. Kidane D, Graumann PL (2005) Intracellular protein and DNA dynamics in competent *Bacillus subtilis* cells. Cell 122: 73–84.
18. Kramer N, Hahn J, Dubnau D (2007) Multiple interactions among the competence proteins of *Bacillus subtilis*. Mol Microbiol 65: 454–464.
19. Kaufenstein M, van der Laan M, Graumann PL (2011) The three-layered DNA uptake machinery at the cell pole in competent *Bacillus subtilis* cells is a stable complex. J Bacteriol 193: 1633–1642.
20. Inamine GS, Dubnau D (1995) ComEA, a *Bacillus subtilis* integral membrane protein required for genetic transformation, is needed for both DNA binding and transport. J Bacteriol 177: 3045–3051.
21. Provvedi R, Dubnau D (1999) ComEA is a DNA receptor for transformation of competent *Bacillus subtilis*. Mol Microbiol 31: 271–280.
22. Berge M, Moscoso M, Prudhomme M, Martin B, Claverys JP (2002) Uptake of transforming DNA in Gram-positive bacteria: a view from *Streptococcus pneumoniae*. Mol Microbiol 45: 411–421.
23. Chen I, Gotschlich EC (2001) ComE, a competence protein from *Neisseria gonorrhoeae* with DNA-binding activity. J Bacteriol 183: 3160–3168.
24. Suckow G, Seitz P, Blokesch M (2011) Quorum sensing contributes to natural transformation of *Vibrio cholerae* in a species-specific manner. J Bacteriol 193: 4914–4924.
25. Lo Scrudato M, Blokesch M (2012) The regulatory network of natural competence and transformation of *Vibrio cholerae*. PLoS Genet 8: e1002778.
26. Blokesch M (2012) A quorum sensing-mediated switch contributes to natural transformation of *Vibrio cholerae*. Mob Genet Elements 2: 224–227.
27. Stingl K, Muller S, Scheidgen-Kleyboldt G, Clausen M, Maier B (2010) Composite system mediates two-step DNA uptake into *Helicobacter pylori*. Proc Natl Acad Sci USA 107: 1184–1189.
28. White J, Stelzer E (1999) Photobleaching GFP reveals protein dynamics inside live cells. Trends Cell Biol 9: 61–65.
29. Ishikawa-Ankerhold HC, Ankerhold R, Drummen GP (2012) Advanced fluorescence microscopy techniques–FRAP, FLIP, FLAP, FRET and FLIM. Molecules 17: 4047–4132.
30. Mullen LM, Bosse JT, Nair SP, Ward JM, Rycroft AN, et al. (2008) Pasteurellaceae ComE1 proteins combine the properties of fibronectin adhesins and DNA binding competence proteins. PLoS One 3: e3991.
31. Jeon B, Zhang Q (2007) Cj0011c, a periplasmic single- and double-stranded DNA-binding protein, contributes to natural transformation in *Campylobacter jejuni*. J Bacteriol 189: 7399–7407.
32. Doherty AJ, Serpell LC, Ponting CP (1996) The helix-hairpin-helix DNA-binding motif: a structural basis for non-sequence-specific recognition of DNA. Nucleic Acids Res 24: 2488–2497.
33. Shao X, Grishin NV (2000) Common fold in helix-hairpin-helix proteins. Nucleic Acids Res 28: 2643–2650.
34. Feilmeier BJ, Iseminger G, Schroeder D, Webber H, Phillips GJ (2000) Green fluorescent protein functions as a reporter for protein localization in *Escherichia coli*. J Bacteriol 182: 4068–4076.
35. Lybarger SR, Johnson TL, Gray MD, Sikora AE, Sandkvist M (2009) Docking and assembly of the type II secretion complex of *Vibrio cholerae*. J Bacteriol 191: 3149–3161.
36. Focareta T, Manning PA (1991) Distinguishing between the extracellular DNases of *Vibrio cholerae* and development of a transformation system. Mol Microbiol 5: 2547–2555.
37. Focareta T, Manning PA (1991) Genetic analysis of the export of an extracellular DNase of *Vibrio cholerae* using DNase-beta-lactamase fusions. Gene 108: 31–37.
38. Blokesch M, Schoolnik GK (2008) The extracellular nuclease Dns and its role in natural transformation of *Vibrio cholerae*. J Bacteriol 190: 7232–7240.
39. Holm L, Rosenstrom P (2010) Dali server: conservation mapping in 3D. Nucleic Acids Res 38: W545–549.
40. Yang Y, Sass LE, Du C, Hsieh P, Erie DA (2005) Determination of protein-DNA binding constants and specificities from statistical analyses of single molecules: MutS-DNA interactions. Nucleic Acids Res 33: 4322–4334.
41. Mortier-Barriere I, Velten M, Dupaigne P, Mirouze N, Pietrement O, et al. (2007) A key presynaptic role in transformation for a widespread bacterial protein: DprA conveys incoming ssDNA to RecA. Cell 130: 824–836.
42. Grange W, Duckely M, Husale S, Jacob S, Engel A, et al. (2008) VirE2: a unique ssDNA-compacting molecular machine. PLoS Biol 6: e44.
43. Sinha S, Mell JC, Redfield RJ (2012) Seventeen Sxy-dependent cyclic AMP receptor protein site-regulated genes are needed for natural transformation in *Haemophilus influenzae*. J Bacteriol 194: 5245–5254.
44. Matias VR, Beveridge TJ (2005) Cryo-electron microscopy reveals native polymeric cell wall structure in *Bacillus subtilis* 168 and the existence of a periplasmic space. Mol Microbiol 56: 240–251.
45. Berge MJ, Kamgoue A, Martin B, Polard P, Campo N, et al. (2013) Midcell recruitment of the DNA uptake and virulence nuclease, EndA, for pneumococcal transformation. PLoS Pathog 9: e1003596.
46. Simon SM, Peskin CS, Oster GF (1992) What drives the translocation of proteins? Proc Natl Acad Sci U S A 89: 3770–3774.
47. Peskin CS, Odell GM, Oster GF (1993) Cellular motions and thermal fluctuations: the Brownian ratchet. Biophys J 65: 316–324.
48. Ambjörnsson T, Metzler R (2004) Chaperone-assisted translocation. Phys Biol 1: 77–88.
49. Salman H, Zbaida D, Rabin Y, Chatenay D, Elbaum M (2001) Kinetics and mechanism of DNA uptake into the cell nucleus. Proc Natl Acad Sci U S A 98: 7247–7252.
50. Yanisch-Perron C, Vieira J, Messing J (1985) Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. Gene 33: 103–119.
51. Studier FW, Moffatt BA (1986) Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. J Mol Biol 189: 113–130.
52. Simon R, Priefer U, Pühler A (1983) A broad host range mobilization system for *in vivo* genetic engineering: transposon mutagenesis in Gram negative bacteria. Nat Biotechnol 1: 784–791.
53. Yildiz FH, Schoolnik GK (1998) Role of *rpoS* in stress survival and virulence of *Vibrio cholerae*. J Bacteriol 180: 773–784.
54. Meibom KL, Li XB, Nielsen AT, Wu CY, Roseman S, et al. (2004) The *Vibrio cholerae* chitin utilization program. Proc Natl Acad Sci USA 101: 2524–2529.
55. De Souza Silva O, Blokesch M (2010) Genetic manipulation of *Vibrio cholerae* by combining natural transformation with FLP recombination. Plasmid 64: 186–195.
56. Blokesch M (2012) TransFLP – a method to genetically modify *V. cholerae* based on natural transformation and FLP-recombination. J Vis Exp 68: e3761, doi:3710.3791/3761
57. Borgeaud S, Blokesch M (2013) Overexpression of the *tcp* gene cluster using the T7 RNA polymerase/promoter system and natural transformation-mediated genetic engineering of *Vibrio cholerae*. PLoS One 8: e53952.
58. Marvig RL, Blokesch M (2010) Natural transformation of *Vibrio cholerae* as a tool-optimizing the procedure. BMC Microbiol 10: 155.

59. R Development Core Team (2009) R: A language and environment for statistical computing. ViennaAustria: R Foundation for Statistical Computing. 409 p.
60. Sanchez R, Sali A (2000) Comparative protein structure modeling. Introduction and practical examples with modeller. Methods Mol Biol 143: 97–129.
61. Evans RJ, Davies DR, Bullard JM, Christensen J, Green LS, et al. (2008) Structure of PolC reveals unique DNA binding and fidelity determinants. Proc Natl Acad Sci USA 105: 20695–20700.
62. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, et al. (2004) UCSF Chimera–a visualization system for exploratory research and analysis. J Comput Chem 25: 1605–1612.
63. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, et al. (2005) Scalable molecular dynamics with NAMD. J Comput Chem 26: 1781–1802.
64. Perez A, Marchan I, Svozil D, Sponer J, Cheatham TE, 3rd, et al. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. Biophys J 92: 3817–3829.
65. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of Simple Potential Functions for Simulating Liquid Water. J Chem Phys 79: 926–935.
66. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys 23: 327–341.
67. Darden T, Perera L, Li L, Pedersen L (1999) New tricks for modelers from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations. Structure 7: R55–60.
68. Martyna GJ, Tobias DJ, Klein ML (1994) Constant-Pressure Molecular-Dynamics Algorithms. J Chem Phys 101: 4177–4189.
69. Feller SE, Zhang YH, Pastor RW, Brooks BR (1995) Constant-Pressure Molecular-Dynamics Simulation - the Langevin Piston Method. J Chem Phys 103: 4613–4621.
70. Miller BR, McGee TD, Swails JM, Homeyer N, Gohlke H, et al. (2012) MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. J Chem Theory Comput 8: 3314–3321.
71. Lo Scrudato M, Blokesch M (2013) A transcriptional regulator linking quorum sensing and chitin induction to render *Vibrio cholerae* naturally transformable. Nucleic Acids Res 41: 3644–3658.
72. Bradford MM (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem 72: 248–254.
73. Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8: 785–786.
74. Wolfgang M, van Putten JP, Hayes SF, Dorward D, Koomey M (2000) Components and dynamics of fiber formation define a ubiquitous biogenesis pathway for bacterial pili. EMBO J 19: 6408–6418.
75. Korotkov KV, Gonen T, Hol WG (2011) Secretins: dynamic channels for protein transport across membranes. Trends Biochem Sci 36: 433–443.
76. Draskovic I, Dubnau D (2005) Biogenesis of a putative channel protein, ComEC, required for DNA uptake: membrane topology, oligomerization and formation of disulphide bonds. Mol Microbiol 55: 881–896.

Hasan Pezeshgi Modarres

Curriculum Vitae

Education

- **PhD. in Biotechnology and Bioengineering**, *Ecole Polytechnique Fédérale de Lausanne (EPFL)*, Lausanne. 2011- 2015.
  Thesis title: "Modeling and Engineering Proteins Thermostability” .
- **Master of Science in Chemical Engineering**, Biotechnology Engineering, Sharif University of Technology, Tehran 2006- 2009.
  Thesis title: "Investigation of Surfactin Behavior at the Presence of Water, Oil, and Asphaltene Using Molecular Dynamics Simulation".
- **Bachelor of Science in Chemical Engineering**, Polymer Engineering, Isfahan University of Technology, Isfahan, 2002-2006.

Experiences

- Teacher Assistant, "Biomolecular structure and mechanics", School of life Science, EPFL, 2011- 2014.
- Teacher Assistant, "Bioprocess Engineering", School of life Science, EPFL, 2012.
- Teacher Assistant, "Cell Quantitative Physiology", Department of Mechanical Engineering, Sharif University of technology. 2009 and 2010.
- Visiting Scholar, Molecular Cell Biomechanics Laboratory, University of California, Berkeley, 2008-2010.
- Programming the Main Software for “Molecular Dynamics Simulation”, Data interpretation and Visualization, Research council of Isfahan University of Technology, 2004-2005.

Avenue de Morges, 76, Lausanne, 1004
hassan.pezeshki@gmail.com +41 78 700 86 82

Publications

- Patrick Seitz, Hassan Pezeshgi Modarres, Sandrine Borgeaud, Roman D. Bulushev, Lorenz J. Steinbock, Aleksandra Radenovic, Matteo Dal Peraro, Melanie Blokesch (2014). ComEA Is Essential for the Transfer of External DNA into the Periplasm in Naturally Transformable Vibrio cholerae Cells. PLoS Genet 10(1): e1004066.
- Hassan P. Modarres, Boris D. Dorokhov, Vladimir O. Popov, Nikolai V. Ravin, Konstantin G. Skryabin, Matteo Dal Peraro, “Understanding and Engineering Thermostability in the DNA Ligase from Thermococcus sp. 1519”, *Biochemistry*, DOI**:** 10.1021/bi501227b.

Poster Presentations

- Hasan P. Modarres and Matteo Dal Peraro, “Conformational Changes During DNA Ligation Studied by Molecular Dynamics”, Simulations towards in silico biological cell: Bridging experiments and simulations, CECAM HQ-EPFL, Lausanne, Switzerland, 2012.
- Hasan P. Modarres and Matteo Dal Peraro, “DNA Ligases: Structure, Mechanism, Cofactor Specificity, and Protein Thermostability Engineering”, Free energy calculations: From theory to applications, Ecole des Ponts, Champssur-Marne, France, 2012.

Research Interests

- Protein Engineering and Design, Bioinformatics, Molecular Dynamics Simulation.

Computer Skills

- Python, MATLAB, Linux

Interests

- Travel, Music, Sports.