Molecular Modeling and Artificial Intelligence for Computational Protein Design : conception of optimized enzymes and nanobodies

Jelena Vucinic

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

l'Institut National des Sciences Appliquées de Toulouse (INSA de Toulouse)

Présentée et soutenue le *05/03/2021* **par :**

Jelena VUCINIC

Molecular Modeling and Artificial Intelligence for Computational Protein Design: conception of optimized enzymes and nanobodies

---

JURY

Charlotte DEANE	Professeur d'Université	Rapporteur
Bernard OFFMANN	Professeur d'Université	Rapporteur
Claire DUMON	Directrice de Recherche	Présidente du Jury
Derek WOOLFSON	Professeur d'Université	Examinateur
Sophie BARBE	Directrice de Recherche	Directrice de thèse
Thomas SCHIEX	Directeur de Recherche	Directeur de thèse

École doctorale et spécialité :

SEVAB : Ingénieries microbienne et enzymatique

Unité de Recherche :

TBI - Toulouse Biotechnology Institute, Bio & Chemical Engineering

MIAT - Mathématiques et Informatique Appliquées de Toulouse

Directeur(s) de Thèse :

Sophie BARBE et *Thomas SCHIEX*

Rapporteurs :

Charlotte DEANE et *Bernard OFFMANN*

Acknowledgments

Je souhaite tout d'abord remercier mes directeurs de thèse, Sophie Barbe et Thomas Schiex pour leur soutien et leurs conseils qui m'ont aidé à suivre le bon chemin. Si j'ai pu m'épanouir scientifiquement durant cette thèse c'est certainement grâce à eux et à tout ce qu'ils ont pu m'apporter par leur connaissances, leur pédagogie et leur rigueur scientifique. Pendant cette aventure extraordinaire qui a durée un peu plus de trois ans, j'ai également beaucoup appris sur un plan moins scientifique et peut être plus humain. Cela m'a permis de grandir en tant que personne et je leur en suis réellement reconnaissante.

Je souhaite également remercier Charlotte Dean et Bernard Offmann pour avoir rapporté cette thèse, ainsi que Claire Dumon et Dek Woolfson d'avoir accepté de faire parti de mon jury. Je me sens extrêmement chanceuse et c'est un grand honneur pour moi d'avoir pu partager mon travail avec tous ces scientifiques extraordinaires.

Une bonne partie de mes travaux de thèse a été effectuée en collaboration avec des équipes d'expérimentateurs. Ils ont rendu possible la validation de nos méthodes computationnelles et ont fait de cette thèse une histoire encore plus intéressante. Pour cela je tiens à exprimer ma gratitude envers nos collègues expérimentateurs du CRCT: Claudine Tardy, Coralie Morand, Patrick Chinestra et Aurélien Olichon ainsi que nos collègues de TBI: Manon Darribère, Thomas Enjalbert, Sophie Bozonnet, Cédric Montanier, Gianluca Cioci et Claire Dumon. C'était un plaisir de travailler avec vous!

Je remercie la région Occitanie et l'INRAe pour avoir financé cette thèse.

Pendant ces trois ans j'ai eu la chance de passer mon temps dans deux laboratoires différents: TBI et MIAT. Cela a représenté pour moi une expérience très enrichissante et stimulante. Pour cela, côté TBI j'aimerais remercier Isabelle André et Jérémy Esque pour leur gentillesse, leur disponibilité et les discussion scientifiques très intéressantes. J'aimerais remercier tous les collègues qui ont fait de ce séjour à TBI un séjour très plaisant: Mounir, Julien, Manon M., Younes, Marie, Sabine, Maxant, Awilda, Emma, Alex, Alexandra, Dhoha, Gleb, Raj, Tarun, Vinciane. Merci pour votre gentillesse et votre bonne humeur! Côté MIAT je tiens à remercier Sylvain Jasson en premier lieu. Merci pour ton grand soutien...merci d'être le Gandalf des doctorants! :) Merci à Manon R., Lise, Jean, Paul, Damien, Emilien, Andrea, Clémence, Céline, Khaoula pour tous les moments passés ensemble que ce soit en conférence, dans la salle café MIAT, à faire du vélo ou autre. Merci à tous les autres collègues de MIAT et plus particulièrement Nathalie V, Matthias, Claire, Annick, Marie-Jo, Simon, David A., Régis, Fred, Ronan, Fabienne, Alain pour tous nos échanges.

Merci à mes top modèles'o: David S., Akli B, Benoit D. qui sont devenus ma famille, ici à Toulouse. Ma première année de thèse restera, grâce à vous, le plus beau souvenir que je garderai toujours dans mon cœur!

Merci à mes amis, éparpillés un peu par tout dans le monde, d'être toujours là pour moi, malgré la distance. Votre soutien inconditionnel et votre amour sont

précieux.

Enfin, je ne peux pas finir ces remerciements sans évoquer les personnes grâce auxquelles je suis là aujourd'hui. Je remercie infiniment mes parents et mon frère. Merci d'avoir toujours été là pour moi, merci d'être ma source d'inspiration, de motivation et de courage. Sans vous je n'en serais pas là aujourd'hui.

A mes parents

Contents

Thesis contributions

Publications

1. Positive Multistate Protein Design, **Vucinic, J.**, Simoncini, D., Ruffini, M., Barbe, S., Schiex, T., Bioinformatics, Volume 36, Issue 1, 1 January 2020, Pages 122–130, https://doi.org/10.1093/bioinformatics/btz497

2. Guaranteed Diversity & Quality for the Weighted CSP, Ruffini, M., **Vucinic, J.**, de Givry, S., Katsirelos, G., Barbe, S., & Schiex, T. (2019, November). In 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 18-25). IEEE. doi:10.1109/ICTAI.2019.00012

3. Guaranteed Diversity and Optimality in Cost Function Network Based Computational Protein Design Methods, Ruffini, M., **Vucinic, J.**, de Givry, S., Katsirelos, G., Barbe, S., Schiex, T., Algorithms 2021, 14, 168. https://doi.org/10.3390/a14060168

4. A Comparative Study to Decipher the Structural and Dynamics Determinants Underlying the Activity and Thermal Stability of GH-11 Xylanases, **Vucinic, J.**, Novikov, G., Montanier, C.Y., Dumon, C., Schiex, T., Barbe, S., International Journal of Molecular Sciences. 2021; 22(11):5961. https://doi.org/10.3390/ijms22115961

5. Improving activity and thermostability of a GH-11 xylanase by Computational Design, **Vucinic, J.** et al. (In preparation)

6. Nanobody scaffold design, **Vucinic, J.** et al. (Journal article and patent in preparation)

Oral communications

1. Pushing the computational frontiers of Multistate Protein Design, **Vucinic, J.**, Simoncini, D., Ruffini, M., Schiex T., Barbe, S. GGMM 2019, Nice, France, 2019

2. Constraint Programming and Graphical models - Pushing data into your models, The protein design case, Schiex, T., Barbe, S., Simoncini, D., **Vucinic, J.**, Rufini, M., Allouche, D. 23rd International Symposium on Mathematical Programming (ISMP-18)

3. Qualité et diversité garanties dans les réseaux de fonctions de coût, Ruffini M., **Vucinic J.**, de Givry S., Katsirelos G., Barbe S., Schiex T. 2019. JFPC 2019 - Journées Francophones de Programmation par Contraintes, Albi, France, Juin, 2019.

Posters

1. Combining structural, dynamic and evolutionary information for computational protein design, **Vucinic, J.**, Simoncini, D., Ruffini, M., Schiex T., Barbe, S., CECAM workshop, Stuttgart, Germany, 2018.

2. Computational strategy for protein design based on structure-dynamics-activity relationship insights: GH11 Xylanases as a case study, Novikov, G., **Vucinic, J.**, Montanier, C., Schiex, T., Dumon, C., Barbe, S. CBM13, Toulouse, France, 2019.

3. Positive Multistate Protein Design: Modeling Protein Flexibility, **Vucinic, J.**, Simoncini, D., Ruffini, M., Barbe, S., Schiex, T., JOBIM 2019, Nantes, France, 2019.

Introduction

Proteins are fundamental components of life. They are indispensable to the structure and function of living cells and viruses and are in charge of many essential processes in all living organisms. They can carry energy, transmit signals, provide structure to cells or promote particular chemical reactions. Over the billions of years of evolution, proteins have evolved to perform better and faster certain functions or to achieve new functions in order to pursue the biological needs under diverse and changing conditions.

Most proteins have a particular three dimensional structure which is directly related to its specific function. The structure and the function of a protein arise from a set of building blocks that compose a protein sequence, called amino acid residues. For a given sequence length, the protein sequence space describes an ensemble of possible combinations of amino acid residues at each sequential position. For example, for a 100 residue protein, the sequence space contains 20^{100} sequences. Naturally occurring proteins cover a very small amount of this space. A large portion of sequences is unexplored by Nature and many functional proteins are certainly yet to be discovered. In recent years, the interest for proteins with new or improved properties has increased in many domains. However, synthesizing all the possible sequences remains unimaginable. Despite the success of approaches such as directed evolution, crowned by Frances Arnold's Nobel prize in 2018, this method has limited sequence space exploration abilities. Therefore, the need for accurate computational methods is crucial in order to rationalize and speed-up the conception of new proteins.

The last decade has been marked by major scientific advances that allowed a deeper comprehension of proteins at different levels. Many biochemical and kinetic data allowed better comprehension of proteins structural and functional properties which in turn led to an extension of the structure-function paradigm to include protein structural dynamics. X-ray crystallography, Nuclear Magnetic Resonance spectroscopy and cryogenic electron microscopy have provided a huge number of protein structures. Computational methods completely revolutionized the domain of protein structure prediction [1]. Also, Molecular Dynamics simulations on proteins, acknowledged by Martin Karplus and Michael Levitt's Nobel prize in 2013, allowed the investigation of proteins at the atomic level. All these advances greatly contributed to refining our comprehension of proteins sequence-structure-function relationship. The amount of available protein structures and our understanding of their functions render structure-based Computational Protein Design (CPD) possible.

Because of the vastness of the sequence search space and the intractable combination of many degrees of freedom of a protein, the most usual CPD approaches model proteins as a single rigid protein backbone, and usually ignore protein flexibility. This traditional Single State Protein Design (SSD) contrasts with the increas-

ing evidence that proteins do not remain fixed in a unique conformational state but rather sample conformational ensembles. Furthermore, large-scale protein motions ranging from local flexibility to large conformational rearrangements are known to play key roles on protein properties and functions. The objective of this thesis is, first, to develop a new method that alleviates SSD limitations by considering several conformational states simultaneously, and second, to demonstrate the interest in applying this method on relevant protein design examples for applications in health and white biotechnology. The manuscript is structured in 3 parts and 8 chapters.

The first part provides background for this work. Chapter 1 introduces proteins with more general notions about protein's function, structure and flexibility. Chapter 2 provides some details about basic principles in Molecular Modeling and Design techniques. Computational Protein Design is then presented along with different state-of-the art approaches. Chapter 3 introduces Computational Protein Design methods based on the Cost Function Networks framework.

The second part of this thesis describes the development of new multistate design methods. Chapter 4 describes a MultiState Design (MSD) approach which allows taking into account multiple conformational protein states simultaneously. During this thesis, numerous interactions with our experimental collaborators allowed improving our method through the introduction of new functionalities which are presented in Chapter 5.

The third and the final part of this thesis presents the application of MSD to two different case studies in which the computational predictions were experimentally validated. The first case study focuses on the redesign of GH11 Xylanases, an enzyme widely used in industrial bio-refinery processes. Chapter 6 describes a Molecular Dynamics study that was conducted to gain deeper insights on the structure-dynamics-activity relationship of this class of enzymes and identify the molecular determinants governing their thermal stability and activity. In Chapter 7, characteristics revealed by the latter study were further used for designing new xylanases with improved thermostability and catalytic activity. Finally, Chapter 8 presents the second case study where the computational design strategies were used to redesign a synthetic humanized nanobody scaffold. Results showed that this new nanobody is highly expressed and possesses suitable affinity with different CDR loops.

At the end of this manuscript, a general conclusion provides a summary of different studies done during this PhD and gives some perspectives and future research directions.

Part I

Background

Chapter 1

Proteins

Contents

1.1 Definition and function

In all multicellular organisms, the smallest unit of life, the cell, can generate, from one initial cell, hundreds of different kinds of cells. Cells have diverse properties such as shape, size, color, surface composition and thus constitute our muscle, skin, bone, neuron or blood cells. Cell's structure and function are based on different kinds of molecules that are fundamental to life. Proteins carry energy, transmit signals, give cells structure and most essential function by performing most cellular tasks [2]. Proteins are large, complex macromolecules that represent the main cell's building blocks and are designed to work in particular places within the cell. By assuming a large variety of functions, proteins have been involved in a multitude of fundamental biological processes over the billions of years of evolution. Therefore, proteins can have many purposes ranging from biological sensors that can modify different cell properties to structural components of a cell. They can import and export substances across the membrane, bind to a specific gene in order to regulate its expression. They can also be extracellular signals that are released from one cell to communicate with other cells or intracellular signals carrying information within the cell. They can be enzymes catalysing chemical reactions or antibodies that defend against infections and foreign substances. All these are some of the examples of proteins and their crucial functions within an organism. A basic foundation in protein science states that a protein function is directly related to its structure [3]. In other words, understanding the role of proteins requires prior knowledge and understanding of their corresponding structure.

1.2 Structure representation

A protein is composed of one or several chains of amino acid residues. The association of different amino acid residues is controlled by the genetic code, which controls when and where proteins are created. The genes that are giving this information have a coding region which specifies the exact order of the amino acid sequence, and a regulatory region that tells when and in which cell or part of the cell this protein is being made. The concept of distinguishing proteins by their amino acid sequences was first introduced by Frederic Sanger in 1952 who described this sequential nature of proteins by studying and sequencing insulin. During protein synthesis, amino acids are being attached to one another by creation of a peptide bond. A particular order of amino acids in these polypeptide chains form an amino acid sequence which further folds in order to generate compact, functional three-dimensional protein structures. In the early 1970's, Anfinsen declared that the protein structure is only determined only by its sequence [4]. Know as the Anfinsen's dogma, this principle is the basis of what today is called the protein sequence-structure-function relationship. Under physiological conditions, a protein sequence does not remain in the form of a long unstructured filament. Instead, it has different levels of structure, ranging from the unfolded sequence to its three-dimensional (3D) structure. 3D protein structures are often relatively stable, well determined, and convey a particular biological function.

Four different levels of protein structures have been defined (Figure 1.1):

- Primary structure: the unfolded amino acid sequence
- Secondary structure: the arrangements of amino acid residues in a local three-dimensional structures
- Tertiary structure: three-dimensional structure, complete spatial organisation of local structures
- Quaternary structure: association of multiple chains and their relative organisation within the 3D structure

1.2.1 The primary structure of proteins

When referring to protein's primary structure, we refer to its sequence, or in other words its amino acid composition.

Amino Acids

There are 20 natural amino acids and each one of these small building blocks consists of two different chemical moieties: a common backbone or main-chain and a variable side-chain. The backbone is composed of an amino group (NH2) on one side, a carboxyl group (COOH) on the other, and a central C_α carbon connected to the two latter moieties, a hydrogen atom and the variable side-chain (Figure 1.2).

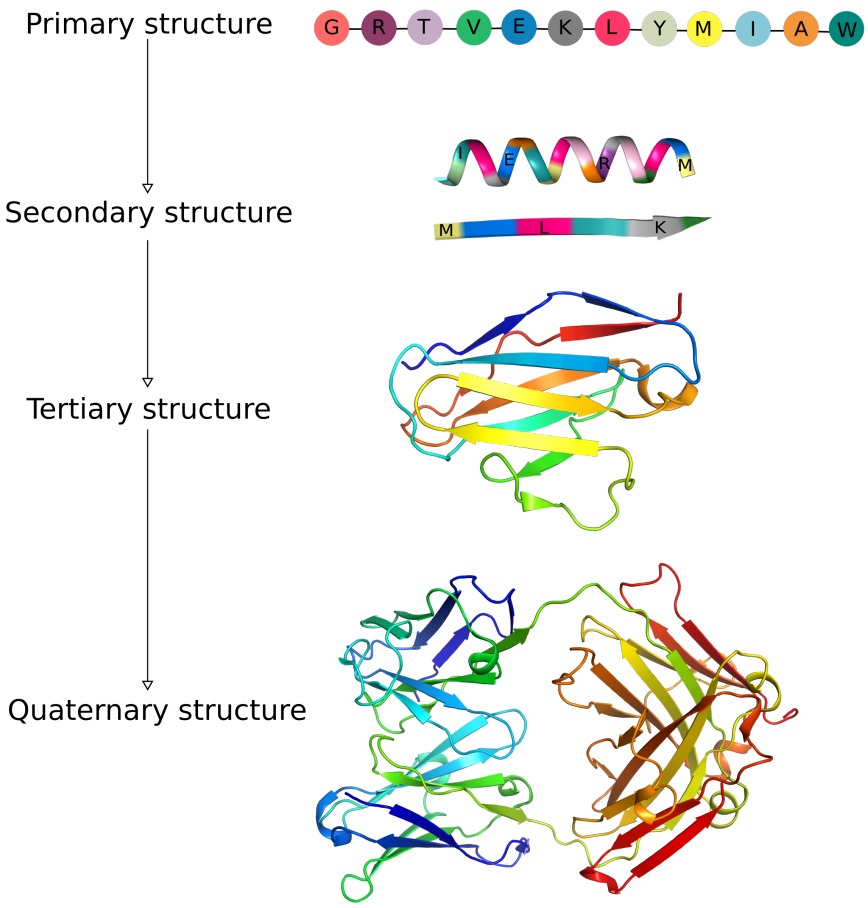

Figure 1.1: Schematic representation of the four levels used to describe the protein structure.

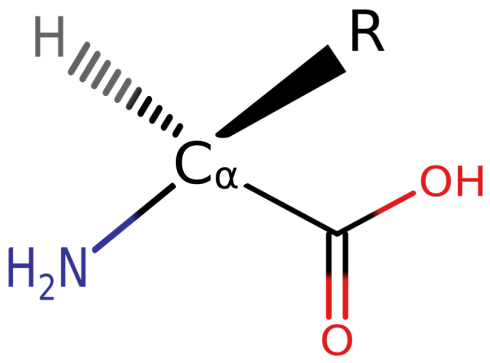

Figure 1.2: The chemical structure of an amino acid. The common backbone composed of the central α carbon ($C\alpha$), an amino group (NH2) and a carboxyl group (COOH). The R group (R) represents the variable side chain which is specific to every amino acid. This generic representation does not concern Proline which has a cyclic form.

This variable side chain is what differentiates one amino acid from another and what confers to each amino acid its specific physico-chemical properties. Based on the chemical properties conferred by their side chain, the 20 natural amino acids can be classified in different groups. The most common classification of amino acids divides them in three groups: hydrophobic, polar and charged. Hydrophobic amino acids are usually buried within the protein core and are known to contribute to the protein stabilization by participating in Van der Waals interactions.

Other sub-classifications have been proposed such as: large or small, aliphatic or aromatic, positively or negatively charged [5].

Because of their characteristics, each one of the 20 amino acids is located in a different chemical environment and has a particular role within the protein structure. This is why it is very difficult to classify all amino acids of the same type into the same group. This can be illustrated by some examples such as the case of Tyrosine which is amphiphile. This amino acid can be found in two different groups at the same time. Tyrosine can be considered as hydrophobic because of its aromatic cycle (the phenol group), but also polar because of the hydroxyl -OH on the phenol group. Histidine is another example which, depending on the environment and the pH of the solution, can be polar or charged. Another example that is worth mentioning is Cysteine and its two different oxidation states : $C_{S\text{-}S}$ and $C_{S\text{-}H}$. According to some classifications cysteine is considered to be hydrophobic while others consider it polar because of its usual presence at the protein surfaces and the relative polarity of its thiol moiety. $C_{S\text{-}S}$ indicates that two cysteines are connected, and form a disulphide bond. As the role of this amino acid is very dependent on the cellular location of the protein, the formation of the covalent bond between two cysteines is very rare within an intracellular environment. Thus, $C_{S\text{-}H}$ indicates its free, unbound form.

The peptide bond

The primary structure of a protein is simply a sequence of amino acids that compose it. The linkage between amino acid residues is ensured by the formation of the peptide bond between the carboxyl group of one amino acid with the amino group of the next consecutive amino acid in the sequence (Figure 1.3). As the process repeats, the polypeptide chain elongates, starting with an N-terminal end formed by the free amino group of the first amino acid in the sequence and ending with a C-terminal end formed by the free carboxyl group of the last amino acid in the sequence. The atoms of the amino acid residues that are involved in the peptide bond define the protein backbone. The creation of a carboxamide group upon the formation of the peptide bond locks it in a quite rigid planar conformation. Therefore, the degrees of freedom of the polypeptide chain exist mainly for the bonds formed by the Cα carbon (NH-Cα and Cα-CO). These two rotations are identified as ϕ and ψ dihedral angles and are shown in Figure 1.3. However, because of the steric hindrance, ϕ and ψ angles are constrained and thus not all of the conformations are possible. Allowed conformations defined for certain ranges of ϕ and ψ angles have been studied by Ramachandran and coworkers in 1968 whose results are presented in the famous Ramachandran plot that maps the entire conformational space of a polypeptide [6].

1.2.2 The secondary structure of proteins

We can think of the secondary structure of proteins as the local spatial rearrangements occurring during the folding of a polypeptide chain. These arrangements are called secondary structures and they represent the core elements of the protein. The most frequent and most stable secondary structure elements are called α helices and β sheets. These regular structures represent the majority of elements seen in proteins, but there are other regions of irregular structures which are called loops or coils.

α Helices

An α-helix is a secondary structure element created by the folding of the polypeptide backbone into a spiral (Figure 1.4). The structure of the α helix is stabilized by hydrogen bonding occurring in the core of the helix while the surface is covered in side-chain groups. This hydrogen interaction within the core involves the carbonyl oxygen of the peptide bond of the residue i and an amide hydrogen of the peptide bond of the residue i+4. The α helix has 3.6 residues per helical turn.

There are other types of helices that have been observed in proteins: 3_{10}helix and π helix. However these helices have an energetically less favorable geometry and are therefore rare.

β Sheets

This type of secondary structure motif is also formed by hydrogen interactions between amide hydrogens and carbonyl oxygens of the peptide backbone of regularly

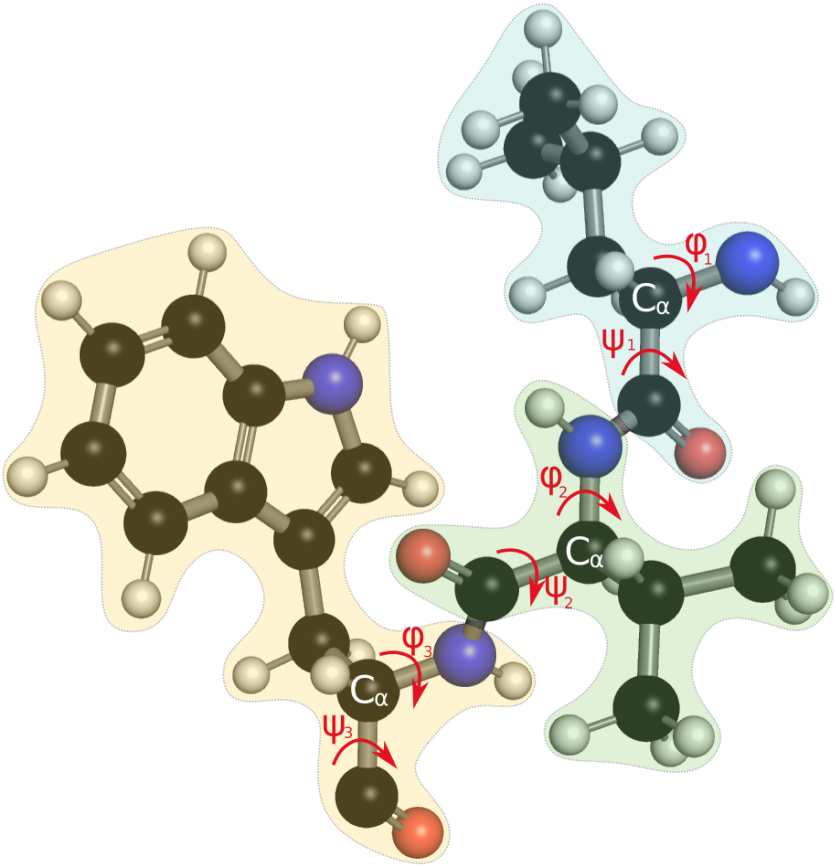

Figure 1.3: Polypeptide chain of 3 amino acid residues (L-blue,V-green,W-yellow). The main-chain atoms are linked through the Cα atoms. Each residue has two degrees of freedom and thus can rotate around two bonds. The ϕ angle represents the angle of rotation around the N-Cα bond and the ψ angle represents the angle of rotation around the Cα-C bond.

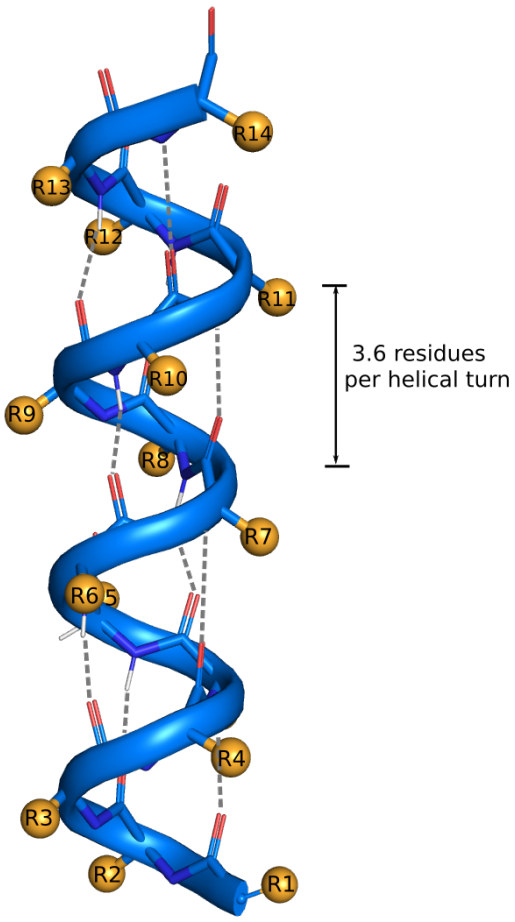

Figure 1.4: The α helix secondary structure element. The polypeptide backbone folded into a spiral is shown in blue. Different hydrogen bonds that stabilize the structure are shown in gray. The surface of the helix is covered with side chain R groups of different amino acid residues (represented in orange).

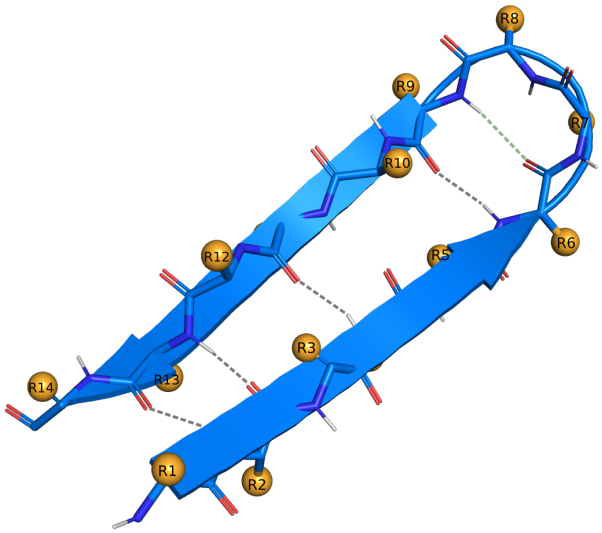

Figure 1.5: The β sheet secondary structure element. Representation of a simple two-stranded β-sheet with antiparellel β strands. Different hydrogen bonds that stabilize the structure are shown in gray. The side chain R groups of different amino acid residues are represented in orange. The short turn between the β strands is also stabilized by a hydrogen bonds (shown in green dashed lines).

arranged consecutive segments in the polypeptide chain. The part of the polypeptide chain that is engaged into a β sheet is called a β strand. These β strands can adopt two different configurations in order to form a β sheet : anti-parallel and parallel. In the parallel configuration, β strands are oriented in the same direction with reference to their N-terminal ends. In the anti-parallel configuration, the orientation of the residue side chains alternates between the two facets of the strand, so that N-terminal of one strand is adjacent to the C-terminal of the other. Hydrogen bondings between the two strands are planar which makes this configuration of beta sheets very stable (Figure 1.5). There is also another variant of the standard β sheet which is the β bulge. This short structure is observed in anti-parallel β sheets and can allow the polypeptide chain to change direction in space.

Turns

Turn represents another type of important secondary structure element present in proteins. Turns are composed of three or four residues and are located at the protein's surface. They are stabilized by a hydrogen bond as shown in Figure 1.5 and exhibit just a few well-defined structures. They allow the polypetide chain to be redirected and proteins to be folded into compact structures. The amino acids residues commonly present in turns are glycine and proline.

Irregular secondary structures

Contrary to an α helix and a β sheet, a loop or coil is an irregular secondary

structure and it represents the third most common secondary structure in proteins. They are usually formed of 2 to 16 residues and are usually found in solvent exposed areas such as the protein surface. They can be defined as the transitions connecting the regular secondary structure elements, but contrary to turns, loops can be formed in many different ways. They can be very flexible and their flexibility can have an important impact on proteins function. As a matter of fact, loop flexibility can play a key role in many protein-protein or protein-ligand interaction processes.

1.2.3 The tertiary and quaternary structure of proteins

The tertiary structure of a protein is the organization of the secondary structure elements into a stable and functional three-dimensional structure or domain. While backbone interactions are important for the formation of different secondary structure elements such as α helices or β sheets, the interactions between amino acids side chains mainly contribute to the stabilization of the final three-dimensional structure. Because of the diversity of chemical properties of the 20 amino acids, there are various types of interactions within the protein structure. The main molecular interaction that leads to protein folding is the hydrophobic effect [7]. Amino acids with a hydrophobic side chain are kept away from the water molecules that constitute the solvent and stay buried in the core of the protein, while other polar and charged residues are usually located in solvent exposed areas such as the protein surface. Along with hydrophobic interactions, other interactions such as hydrogen bonds, salt bridges or covalent bonds (disulphide bridges) are also critical for protein folding, for providing protein stability and flexibility.

The quaternary structure of the protein consists of an association of several polypeptide chains, where each chain is called a monomer and the ensemble of chains an oligomer. This association between the monomers can be formed by the same different types of interactions mentioned above. Antibodies are one of the examples of proteins that contain several domains and thus have a quaternary structure. It is important to point out that the final protein structure depends on the interactions that are made within the polypeptide chain but also between the different domains (if there is more than one).

The Protein Data Bank (PDB) [8] is a database that contains information about the experimentally determined 3D structures of proteins. In this database, each protein has a corresponding PDB file that describes the average protein conformational state “in solution” with the corresponding 3D Cartesian coordinates of its constitutive atoms. The PDB currently contains more than 170,000 protein structures mainly determined by X-ray crystallography, nuclear magnetic resonance (NMR) or transmission electron cryo-microscopy (cryoEM).

1.3 Protein flexibility

Understanding biological processes requires comprehension of protein function at the atomic level. For decades, X-ray crystallography has been used for the deter-

mination of protein structures and has become a powerful method for the study of the structure-function relationship. Each crystallographic structure is represented by a unique single conformation. In a given crystal structure, the relative vibrational dynamics of each atom is quantified by the Debye-Waller factor (also known as B-factor). However, this unique structure only reveals limited information on the protein dynamics. In the cellular environments, proteins are in motion: they fluctuate over a large number of conformational states. Thus the assumption by which the "native" state of the protein can be represented by a single conformation has been shown to be a considerable simplification [9]. Many three-dimensional protein structures are highly flexible and undergo conformational changes allowing proteins to adapt to environmental variations or respond to the presence of other molecules. Protein flexibility has been characterized in many studies [10, 11, 12, 13] and has shown to play a crucial role in the function of proteins [14]. Some experimental techniques such as cryoEM or NMR allow exploration of conformational fluctuations of proteins [15]. Computational methods can also contribute and help further understanding of the protein structure-function relationship through the prediction of protein flexibility [16, 17, 18, 19]. Molecular Dynamics simulation is one of the techniques which is widely used for studying protein dynamics in a simulated explicit environment based on a general physical model.

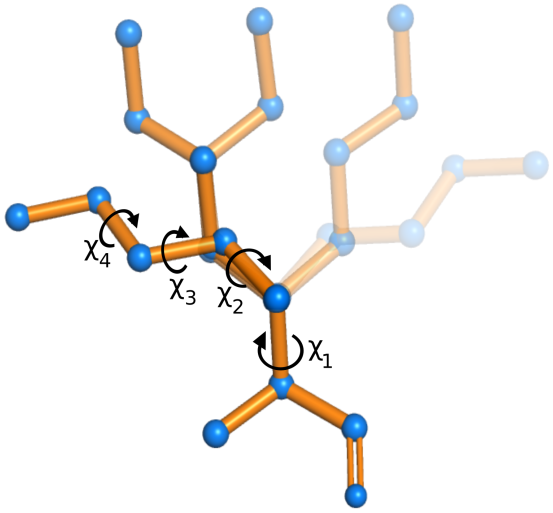

Figure 1.6: Dihedral χ angles in Lysine. Different fading conformations illustrate some of lysine rotamers.

At some level, protein flexibility can be described by different degrees of freedom that can be observed in the protein structure. As it was mentioned in the previous section, the polypeptide chain is defined by the rotational freedom of bonds formed by the α carbons, or more precisely the ϕ and ψ angles. However these are not the only degrees of freedom within the protein, considering dihedral angles of amino

acid side-chains which are referred to as χ angles (Figure 1.6). There are as many χ angles as there are dihedrals for a given amino acids. These χ angles describe the degrees of freedom of the amino acid side chains and were shown to adopt a finite set of favored average conformations, know as rotamers. Rotamer libraries, defined by a discrete set of conformations, contain statistically preferred χ angles observed in natural proteins and are greatly used for all molecular modeling and design methods.

1.4 Enzymes and Enzyme Engineering

Enzymes are proteins that have a unique ability to catalyse a wide range of biochemical reactions. They are called biological catalysts. The function of many proteins depends on their ability to bind other molecules or ligands. In the particular case of enzymes, the molecules upon which enzymes react are called substrates. Enzymes (E) bind specific substrates (S) to further convert them into different molecules called products (P) (Equation 1.1).

$$E + S \rightleftharpoons ES \rightarrow E + P \tag{1.1}$$

Likewise to any other chemical reaction, in which a given reactant is transformed into a given reaction product, a change in the free energy of the reaction pathway between the reactant and the product respective states is also observed in enzymatic reactions. For any chemical reaction to occur, the system must have a sufficient energy to be able to cross the reaction free energy barrier separating the reactant state from the product state. In enzymatic reactions, such goal is achieved by the ability of enzymes to lower the reaction free energy barrier in comparison to equivalent uncatalysed chemical reactions. The enzyme-substrate complex (ES) from Equation 1.1, undergoes rearrangement to one or several transition states prior to the formation of the final product. These transition states possess a higher free energy than the enzyme-substrate complex and usually involve bond-breaking and bond-forming events [20]. The energy needed for bringing the free enzyme and the substrate to the highest transition state of the ES complex is called the activation energy. Enzymes accelerate the rate of chemical reactions by decreasing the activation energy and stabilizing transitions-states intermediates (Figure 1.7). One of the ways of achieving the decrease in activation energy is by providing catalytic residues that have catalytically active groups for a specific reaction mechanism. Single substrate enzyme kinetics was first investigated by Henri in 1902 and further generalized in 1913 by Michaelis and Menten. They proposed a mathematical model which describes single substrate enzyme kinetics by relating the rate of the reaction v to the concentration of the enzyme ($[E]$) and the concentration of the substrate ($[S]$) (Equation 1.2) assuming many approximations. In the equation, the rate constant (k_{cat}) represents the maximum number of substrate molecules which can be consumed per enzyme molecule per unit of time. Another parameter of the Michaelis-Menten equation is the Michaelis constant, also known as K_M. On the

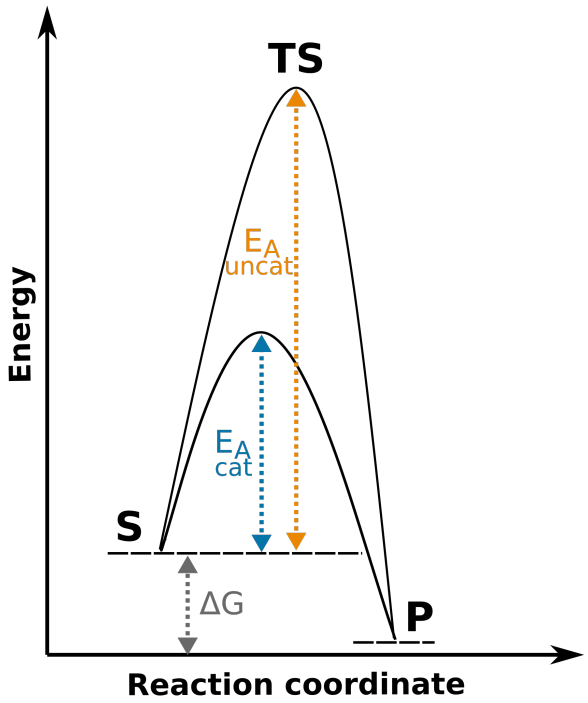

Figure 1.7: Free energy diagram - schematic comparison between catalyzed and uncatalyzed reactions. The standard free energy of reaction (ΔG) and activation energies for catalyzed reaction ($E_A$cat) and uncatalyzed reaction ($E_A$uncat).

assumption that the dissociation of the enzyme-substrate complex into the reaction products is the rate limiting step of the overall reaction scheme, K_M can be taken as a measure of the enzyme inverse affinity to the substrate. This equation, called the Michaelis-Menten equation, is one of the best known models of enzyme kinetics. The kinetic parameters K_M, k_{cat} and k_{cat}/K_M, which represents the specificity constant that provides a measure of the overall enzyme's efficiency, are the standard parameters that are commonly used for describing the properties of any enzymatic reaction. A more detailed presentation of enzymes and their catalytic power can be found in textbook references [21, 20].

$$v = \frac{k_{cat}[E][S]}{K_M + [S]} \tag{1.2}$$

Enzymes represent essential macromolecules that catalyse 99% of biochemical reactions that occur in biological systems [20]. These catalytic proteins are thus necessary in all living organisms.
Past years have been marked by an expansion of knowledge in the field of enzymology, focusing more particularly on enzyme properties and catalytic mechanisms. Enzymes also served as attractive models for fundamental studies that contributed to the understanding of proteins structure-function relationship [10, 12]. In this

regard, gaining deeper insights on catalytic mechanisms, but also on the enzyme's structure-function relationship had an important impact on the progress of modern biology.
The catalytic properties of enzymes made them very attractive for biotechnological developments and applications. "Exploiting" enzymes, isolated from natural sources and further mass-produced using recombinant DNA technologies through genetic engineering, has become of great interest for various industrial applications. Enzyme engineering consists in modifying selected properties of available enzymes to usually improve their activity and/or stability for a further large-scale use in industry. With the on-going development of enzyme engineering, the number of potential applications for enzyme catalysts in industry, in analytical techniques, and medicine keeps on growing. Many industrial processes have evolved by the use of enzymes. Enzyme engineering expanded the scope of applicability of many existing technologies and enabled the conception of enzymes with new features and with increased catalytic efficiency for large-scale biotransformations. Different types of industrial processes are now performed with enzymes: amylases are being used for starch processing, cellulases for cellulosic biomass conversion, pectinases and esterases for food industry etc. Recently PET depolymerase has been engineered for plastic degradation and recycling [22].

Being an essential catalyst in Nature, enzymes outperform traditional chemical methods for catalyzing complex stereospecific transformations. Enzymatic reactions also generate less by-products than pure chemistry-based reactions and enzymes are typically active in conditions closer to those of biological environments, which makes them environmental friendly. Therefore, motivated by environmental, technical and economical advancements, demands for improved biocatalysts have been numerous. Some studies have proposed that the optimal function of enzymes may be influenced by the conformational changes that their respective 3D structure undergo. As a protein structure and dynamics are often linked with its biological activity, getting insights into enzyme structure-function-dynamics relationship is fundamental for the conception of enzymes with new or improved properties.

Chapter 2

Molecular Modeling and Protein Design: Principles and Methods

Contents

2.1 Basic principles

Molecular Modeling regroups a wide variety of theoretical and computational techniques whose purpose is to mimic and/or simulate the behaviour of molecules. Models are abstract representations of reality, and in this case, molecular systems. Molecular modeling textbooks such as [23] introduce models by citing the *Oxford English Dictionary* definition which says that models are "a simplified or idealised description of a system or process, often in mathematical terms, devised to facilitate calculations and predictions".

Molecular models can be described at different levels of theory ranging from subatomic particles (protons, neutrons and electrons) to a more general atomistic level description. Consequently, two main types of molecular models exist: *ab initio* models described by quantum mechanics and classical models described by molecular mechanics and parametrized by an empirical force field. On one hand, the

quantum mechanics approach explicitly represents electrons, which makes this type of approach very precise. However, these methods are very time-consuming and not adapted for large systems such as macro-molecules. Molecular mechanics, on the other hand, is based on the principles of so-called Newton's classical physics. This classical model describes the nucleus and its electrons as a single entity, gives the representation of the system on the atomic level and allows the determination of the potential energy of the system. The potential energy of the system can be calculated using a set of parameters, a force field, which defines and models interactions between different atoms. These parameters are derived from experimental data as well as from quantum chemistry calculations.

2.1.1 Force field and energy function

A force field represents a set of parameters and equations that define the terms of the interaction energy and are used to model the potential energy of the system. The total potential energy of a system is determined as a sum of energies describing bonded and non-bonded interactions between atoms.

$$E_{total} = E_{bonded} + E_{non-bonded} \tag{2.1}$$

Within bonded-interactions, three different energy terms are considered: interactions between pairs of bonded atoms that involve bond-stretching, formation of bond angles between three consecutively bonded atoms and formation of dihedral angles created by four successively bonded atoms. In addition to these three terms, a term representing the improper dihedrals is also generally considered. This term describes a spatial constraint affecting a group of four atoms that do not sequentially follow each other and is generally used to enforce a relative planarity between these four atoms. The full equation of bonded-interaction energy is written below (Equation 2.2), where parameters l, θ and ϕ correspond to bond length, valence angle and value of the dihedral angle respectively. l_0, θ_0 and ϕ_0 refer to equilibrium values, specified in the force field and initially derived from QM calculations, while l_t, θ_t and ϕ_t are values calculated over the course of the simulation. In the dihedral term, n is a positive integer between 0 and 2π, and E_n is the value of the energy barrier of the torsion potential.

$$E_{bonded} = \underbrace{\sum \frac{1}{2} k_b (l_t - l_0)^2}_{\text{Bond length}} + \underbrace{\sum \frac{1}{2} k_a (\theta_t - \theta_0)^2}_{\text{Bond angle}} + \underbrace{\sum \frac{E_n}{2} [1 + cos(n\phi_t - \phi_0)]}_{\text{Dihedral angle}} + \underbrace{\sum E_{imp}}_{\text{Improper dihedral}} \tag{2.2}$$

Non-bonded interactions are determined as a sum of Van der Waals and electrostatic interactions that are modelled as Lennard-Jones and Coulomb potential.

Accordingly, the non-bonded interactions are calculated as:

$$E_{non-bonded} = \underbrace{4\varepsilon[(\frac{\sigma}{r})^{12} - (\frac{\sigma}{r})^{6}]}_{\text{Lennard-Jones potential}} + \underbrace{\frac{q_1 q_2}{4\pi\varepsilon_0 r}}_{\text{Coulomb potential}} \tag{2.3}$$

Parameters ε, σ, r, q_1 , and q_2 represent respectively the depth of the potential energy minima between two atoms, the distance at which the potential between two atoms is zero and the distance between the two atoms and the atoms charge.

A force field is a physics based energy function mostly used in molecular dynamics simulations. Commonly used force fields in protein science are AMBER, CHARMM, OPLS and GROMOS [24, 25, 26]. Along with physics based energy functions, there are also knowledge-based energy functions or so-called statistical energy functions. They are employed for other sorts of molecular modeling problems such as protein structure prediction or computational protein design. Parametrization of the knowledge-based terms present in this type of energy function relies on various statistical observations detected in the available experimental data. Many knowledge-based methods are currently available [27, 28, 29, 30, 31]. The large amount of data available in the PDB and the reduced computational cost required for knowledge-based methods make them more and more attractive than physics based methods. However, knowledge-based method can be biased by the static view of macro-molecules 3D structure as obtained from X-ray crystallography, and thus limited by their inability to take protein flexibility into account. To overcome this limitation, hybrid energy functions which combine statistical and physical terms have been developed. The ROSETTA [32] energy function is one example.

2.1.2 Solvation models

Macromolecules are functioning in a physiological environment which usually requires the modeling of proteins in water and ions at physiological concentrations. In this context, modeling the solvent represents an important aspect of molecular modeling studies. However, accurately modeling the solvent still remains an important challenge as it increases the complexity of the problem by adding new degrees of freedom for each water molecule in the system. Hence, two types of solvation models exist: implicit solvation [33] and explicit solvation model [34]. The explicit model is more realistic: it takes into account the effects of polarisation as the coordinates of the water molecules are explicitly defined. Explicit solvation models are frequently used in Molecular Dynamics simulations [35]. Nonetheless, the explicit representation of the solvent implies adding a significant number of solvent molecules and considering their contribution to the energy calculations. Contrarily, the implicit model simplifies this by omitting water molecules from the system and replacing them with an infinite continuum medium that has dielectric properties of water. It is represented by a specific energy term ($E_{solvent}$) that is added to the potential energy calculations. Some of the most prominent implicit models are the Generalized Born model [36] and the Poisson-Boltzmann model [37]. As a general rule, the more explicit and therefore accurate the description of the protein environment is, the more realistic molecular modeling of the system can be. However as the explicit models are very expensive computationally, implicit models remain an attractive alternative widely used in protein structure prediction as well as computational protein design problems.

2.2 Molecular Dynamics Simulations

Molecular dynamics simulations enable the study of protein dynamics which is described by a change of atomic coordinates as a function of time. Different configurations of the systems are generated by integrating over time the Newton's equations of motions, resulting in a trajectory that defines the variation of particle's positions and velocities over the simulation time. Velocities are applied on particles and forces are calculated as a negative gradient of the potential energy, as follows:

$$m_i \frac{\delta^2 r_i}{\delta t^2} = f_i = -\frac{\delta}{\delta r_i} E \tag{2.4}$$

where m_i represents the mass of the particle, r_i the position of the particle i, f_i the resulting force exerted on the particle i and E the potential energy associated with the particle displacement. The potential energy of the system is then calculated as a cumulative sum of non-bonded and bonded interaction energies as explained earlier (Equation 2.1). The initial configuration of the system must be defined. In order to initiate the movement of all atoms of the system at the beginning of the simulation, the attribution of the initial speed to each atom is necessary. These initial velocities are generally randomly assigned according to a probabilistic distribution (Maxwell-Boltzmann distribution) and are dependent on the simulated temperature. By solving Newton's equations of motion, the configuration of the system at time t (C_t) can be determined. From the C_t configuration, the configuration $C_{t+\delta t}$ of the system at time $t + \delta t$ can be computed at each step of the simulation. Thus, after a defined time-step interval δt, forces and velocities previously determined at a time t are being recalculated and then updated in order to permit the determination of new set of positions at time $t + \delta t$. Several algorithms exist and are used to numerically integrate the equations of motion. Most commonly used algorithms in MD simulations are the Verlet algorithm [38], the Leap-Frog algorithm [39] or the Beeman's algorithm [40]. This algorithmic framework is appropriate for molecular systems simulated to evolve in microcanonical ensemble, also called NVE ensemble where N represents the total number of particles in the system, V the system's volume, and E the total energy, and each of them being constant.

However, under the physiological conditions and in the context of laboratory experiments, systems are more subject to constant temperatures than constant energies. Thus, the microcanonical ensemble can be inappropriate for simulating systems that are subjected to constant pressure and/or temperature. For such situations, more appropriate ensembles exist such as the canonical (constant temperature and volume, NVT) and the isothermal-isobaric (constant temperature and pressure, NPT) ensembles. The NPT ensemble is the most commonly used ensemble in Molecular Dynamics simulations. In order to maintain the temperature and/or pressure constant, temperature and pressure coupling algorithms are added to the classical algorithms for numerical integration of equations of motion. Several temperature and pressure coupling methods have been developed and are widely used in MD simulations. The most commonly used algorithms for thermostats and

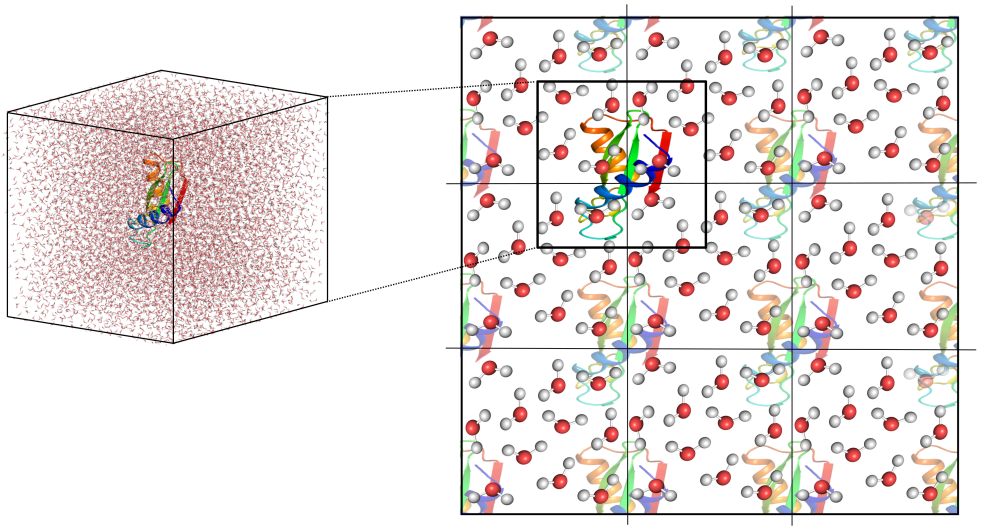

Figure 2.1: MD system representation: periodic boundary conditions allow a minimal representation of the system as it evolves in a virtual environment with cubic boxes (in this case) duplicated infinitely.

barostats are those of Brendensen [41], Nosé-Hoover [42, 43], Langevin [44] and Andersen[45].

In biology, MD simulations are generally employed for the study of the conformational dynamics of macromolecules in a solvated medium specifically modeled to mimic their actual physiological environment. Therefore, macromolecules are placed in a box of specific volume and with a finite number of particles. In order to avoid boundary artefacts, the simulation box must be large enough to encompass the volume of the system and satisfy the minimum image convention in the presence of Periodic Boundary Conditions (PBC). Applying PBC means that the primary box, also called a unit cell, is replicated to an infinite number of unit cells by translation in all the three Cartesian directions (Figure 2.1). This technique allows to overcome the finite size artifact which, in the absence of PBC, would be observed at the boundaries of the box while limiting the number of molecules in the simulation.

Over the last decades, MD simulations have become very famous and are nowadays widely used for studying molecular systems. By simulating dynamic behaviour of macromolecules these methods allow us to calculate, predict and thus better understand biological properties of the system that is being studied. Protein's ability to adapt to environmental variations or respond to the presence of other molecules is a crucial step for every biological process. This ability can be induced by protein motions ranging from local flexibility to large conformational rearrangements

which can play key roles on protein functions. With the help of MD simulations, a thorough study on a biological system can be performed at the atomic level to get a detailed structural description and quantified energy assessment of functionally relevant dynamic events occurring within the studied system. Quantitative information, generated at the microscopic level over the course of the simulation, can then be used to compute a macroscopic observable (pressure, energy etc.) using statistical mechanics. Thus, by defining and measuring different observables over the simulation time, different macroscopic or thermodynamic properties can be calculated. Assuming that the dynamics of a molecular system is described by a given observable, it is possible to estimate the average dynamical behavior of a given biological system by calculating the average of this observable over a sufficiently long time interval.

In the past twenty years MD simulations have proven their importance in structural biology studies. By providing detailed insights on individual particle motions over time, MD simulations allow to investigate relevant motions that have crucial roles for protein functions. In this regard, MD simulations can be utilized to address specific questions related to certain macromolecular properties and obtain information inaccessible from experimental data [46, 47]. Combining molecular dynamics studies with experiments, and in some way guiding and inspiring new experiments with preliminary MD results, may allow to gain a deeper understanding of biological systems structural dynamics, along with better comprehension of proteins structure-dynamics-function relationship.

2.3 Protein Structure Prediction

Since Anfinsen's experiment on ribonuclease A, the theory stating that the amino acid sequence determines the three-dimensional structure of proteins has been widely accepted [4]. However, the principle that explains how a given amino acid sequence of a protein can dictate its folding to a fully functional 3D structure remains unknown. It is since the 1970's that, "inspired by practical problems in biotechnology and medicine, researchers are attempting to figure out the rules that govern protein folding" [48]. The problem of predicting a protein's three dimensional structure became one of the most fascinating and greatest open questions in molecular biology to the point where it has been called "the Holy Grail" of molecular biology [49, 50]. Understanding how and why proteins fold in a specific way means understanding why some sequences fold into a specific α helix, β sheet, turn or loop, but also figuring out how these elements pack further together. Better understanding of the principles controlling the kinetics and thermodynamics of protein-folding would allow a much better comprehension of many diseases caused by the misfolding of essential proteins. In order to establish how a protein structure can only be determined only by its sequence, computational protein folding methods based on the thermodynamic hypothesis formulated by Anfinsen [4] are used. The main goal of these methods is to find, for a given amino acid sequence, a structure with the

lowest folding free-energy. There are two different approaches that are commonly used. The first approach consists in building a three-dimensional model directly from the amino acid sequence, without using any structural information from previously solved structures. This kind of approaches are called *ab initio* or *de novo* approaches. The second approach, the homology modeling method, is based on finding "similar" proteins in known protein structural databases and then constructing a 3D model by homology to these known structures. The main difficulty of the homology modeling methods relies on the determination of a suitable template, while the *ab initio* or *de novo* methods encounter difficulties with the considerable size of the search space.

2.3.1 *Ab initio* and *de novo* methods

Traditionally, *ab initio* prediction methods are based on a protocol in which different protein conformations are generated and further evaluated with the use of an energy function. The large search space that is required in order to explore different conformations makes this type of prediction method computationally very expensive. The vastness of the conformational search space has been illustrated by Levinthal's paradox expressing the theory of protein folding [51]. Cyrus Levinthal pointed out in 1968 that, due to a considerable number of degrees of freedom in a polypeptide chain, a protein has a theoretically astronomical number of possible conformations. Thus, the time needed to exhaustively search through every possible conformation would exceed the age of the known universe. Nonetheless, many proteins fold spontaneously within a millisecond or even a microsecond time scale. Therefore, a protein cannot fold by exhaustively sampling all the possible conformations and must have some sort of folding pathway [51]. *Ab initio* methods are based on physical principles and aim at predicting protein structures "from scratch" by applying stochastic methods in order to exhaustively search for possible solutions. The global optimisation of a suitable energy function will further enable to shorten the computational cost needed to find the lowest-energy conformation of a given protein sequence.

De novo structure prediction aims at exploring possible conformations using prior information obtained from the structure of small already known structural fragments. This type of approach is called fragment-based approach and has shown a great success [52, 53, 54, 55]. Taking small structural fragments from known protein structures in order to construct the final model enables a discretization of the search space. However, even with the reduction of its size, the search space remains important. With this respect, there is a crucial need for advanced sampling algorithms to accelerate the resulting predictions. The simulated annealing that is used in Rosetta macro-molecular modeling software is one example of available methods for such application [56]. Other algorithms, such as evolutionary algorithms and other population-based meta-heuristics are also used for this type of optimisation problems [57, 58, 59, 60, 61, 62].

2.3.2 Homology-based structure prediction methods

The foundation of the homology-based structure prediction methods relies on the observation that similar sequences adopt similar protein structures [63, 64, 65] and that protein structures/folds are more conserved than their sequences. These methods are generally conducted in five stages:

- identification of structural templates
- alignment of the query sequence to the template
- model building for the query sequence
- model evaluation
- model refinement

Identification of the structural template and alignment of template-query sequences represent the key steps in homology modeling. With the constant progress and growth of experimental data from experimentally determined protein structures, homology modeling methods have been improving and becoming more and more efficient. Different ways exist and have been used for improving the sensitivity of template identification and the quality of the template-query alignment. There are methods that consider structure information, called threading methods, whose goal is to align protein sequence with one or more structures in order to obtain the best sequence structure compatibility. Other methods are instead purely sequence-based and use multiple sequences from the same protein families. With the accumulation of new protein sequences and the development of Position Specific Iterative BLAST (PSI-BLAST) [66], profile-based homology studies have been enabled. Within PSI-BLAST, a Position Specific Scoring Matrix (PSSM) can also be generated. From the initial BLAST search and for a given multiple sequence alignment, this matrix can calculate the position-specific scores for each position in the alignment. The application of these techniques significantly increased the sensitivity of homology detection. Some of the most famous homology modeling software are I-Tasser [67], SWISS-MODEL [68], MODELLER [69], PHYRE2 [70].

Finally, it is worth mentioning that the long-standing race for solving one of biology's greatest challenges has been intensified in the past few years by the use of new artificial intelligence approaches, and more particularly deep learning [71]. Deep learning methods became very popular for their ability to process huge amount of information from the increasing mass of data available nowadays. In the past five years many studies showed that deep learning methods can improve the accuracy of protein structure predictions [72, 73, 74, 75, 76]. The use of deep learning for solving the protein structure prediction problem has been revolutionized in 2018 during the well-known Critical Assessment of Techniques for Protein Structure Prediction (CASP13) competition. Google's DeepMind work, a system called AlphaFold, made

an *unprecedented progress in the ability of computational methods to predict protein structures* and was ranked first during CASP13 competition [77]. Two years later at CASP14, AlphaFold2 results were so impressive that CASP organizers stated that the protein structure prediction problem is in some sense solved.

2.4 Computational Protein Design

2.4.1 Context and objective

Usually referred to as the *inverse folding problem*, Computational Protein Design exploits the sequence-structure-function relationship with the objective of identifying an amino acid sequence that folds into a known three-dimensional protein structure and ultimately performs a desired function. In order to identify proteins with the desired properties, an evaluation of different possible sequences is needed. For a protein of N residues and a choice between 20 amino acids per residue, the challenge of evaluating all possible sequences relies in the astronomical size of the search space (20^N).

Traditionally, protein engineering relies on directed evolution [78] (random mutagenesis or gene shuffling combined with high-throughput screening) and site-directed mutagenesis. Despite the power of these approaches and the advances they have enabled in the field of protein engineering, they still face a number of limitations. One important problem of these approaches is the limited diversity of protein sequences that can be generated and explored compared to the vastness of the sequence space. This problem is further compounded by the fact that, in a random mutant sequence library, the frequency of observing beneficial mutation is extremely low [79, 80]. Moreover, the exploration of protein diversity is also limited by the screening process. As high-throughput screening assays require extensive research and development, they are not always available and can be laborious and expensive to implement. Since it is only possible to test a very small fraction of a large number of possible protein sequences, the use of computational methods have become increasingly prominent in protein engineering strategies to explore *in silico* large sequence spaces and pre-filter the most relevant protein sequences for the targeted property/function prior to any experimental characterization. Such approaches aim at considerably narrowing down the number of mutants to consider for subsequently experimental testing while greatly increasing the chances of isolating a suitable mutant with the desired property. In this regard, Computational Protein Design (CPD) has become a powerful approach to fully rationalize and speed-up the conception of new tailored proteins.

2.4.2 Modeling Protein Design Problem

CPD seeks to identify sequences that adopt a desired tertiary structure which possesses sufficient stability and ultimately performs a desired function. Therefore, the CPD problem has been formulated as an optimisation problem which requires

an energy function that accurately reflects protein stability and a reliable search method to identify a sequence with a conformation of optimal stability (Global Minimum Energy Conformation or GMEC). Because of the intractable combination of the many degrees of freedom of a protein and the non-convex form of energy functions, this problem has been simplified by several assumptions: the energy is supposed to be described as a pairwise decomposable function, the protein backbone degrees of freedom are fixed to an idealized target backbone conformation and the side-chain of any given amino acid is assumed to adopt one conformation out of a finite set of possible conformations or rotamers. Despite these simplifications, the size of the search space remains exponentially large and the problem of searching for a sequence with a minimum energy conformation is known to be decision NP-complete [81]. For this reason, most CPD approaches rely on stochastic optimization algorithms such as Monte Carlo Simulated Annealing [82, 83] or Genetic algorithms [84], which provide only asymptotic convergence guarantees. However, recent progress in guaranteed discrete optimization techniques showed that such stochastic methods may durably fail to find or even get close to the GMEC when the problem becomes hard. Despite years of CPU-time, a tuned Simulated Annealing algorithm was unable to find the global energy optima that was identified and proved as optimal by Cost Function Networks (CFN) algorithms [85].

2.4.3 Successes and limitations

The chronological milestone road of CPD covers almost four decades [86]. It has been since 1985 and the very first CPD experiment on a calmodulin-binding peptide conducted by DeGrado [87] that CPD has been considered as a field in which computational knowledge and human expertise are fundamental. This yet evolving field consists of computationally designing proteins but also doing synthesis and experimental characterization of suggested designs. Even though the early experiments of CPD were mostly focusing on the methodological study that is at the core of this technique, the term "computational protein design" only entered the literature one decade later. Early attempts of CPD were marked by the use of the fundamental fact that the core of proteins is mainly composed of hydrophobic residues while the surface is rather populated by hydrophilic residues. With the objective of improving protein stability, CPD pioneers therefore focused on redesigning the core of existing proteins [88, 89, 90, 91, 92, 93]. A real breakthrough in the CPD field was presented in 2003 by Kuhlman and coworkers at the Baker lab with the successful design of a 93-residues long protein with a novel α/β topology, called TOP7. Kuhlman and coworkers achieved the very first systematic *de novo* CPD which folded with an atomic-level accuracy of 1,2 Å into the designed template.

The relatively short history of CPD was particularly marked by the advances made in the last decade. Development of new algorithms and energy functions allowed the field to expand its objectives and thus also (re)design proteins involved in complex interactions. Designing new drug delivery systems, enhancing catalytic activities of enzymes or binding activities of antibodies and self-assembling protein

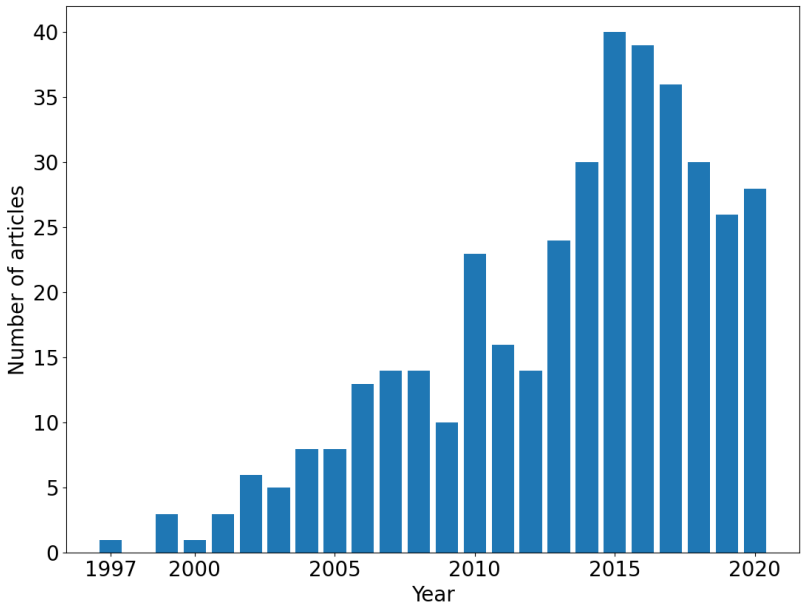

Figure 2.2: Number of articles with the term "Computational Protein Design" in the PubMed database since 1997 until today.

building blocks are just some of the successful examples of case studies in which CPD has been of great benefit for health, green chemistry and bio-nanotechnology applications [94, 95, 96, 97, 98, 99]. We have recently witnessed many exciting achievements in this field. The design of new transmembrane proteins that allow cells to take in certain chemicals, including charged ions and larger fluorescent molecules [100] is one of the examples. Also, Lajoie and coworkers addressed the problem of targeting only diseased cells by designing switches that bind to antigens on the cell surface and, through a conformational change, are activated only when there is a precise combination of antigens [101]. Just a while ago, CPD was also used to design small proteins that protect cells from severe acute respiratory syndrome coronavirus 2 (SARS-Cov-2), the virus that causes COVID-19 [102]. These success stories clearly highlight present achievements in the CPD but also show the great future potential that this field holds. Consequently, the significance of this field can be nicely illustrated by the increasing number of related publications in the past 30 years as shown in Figure 2.2.

However, despite these remarkable achievements, the success rate in CPD remains low. Numerous limitations need to be surpassed in order to allow CPD to achieve results with greater efficiency and reliability. Several factors may explain this: the CPD problem is ill-defined, suffers from a lack of expressiveness and lack of accuracy. The CPD problem is defined as an inverse problem: finding an amino acid sequence that folds in a given three-dimensional structure which performs a

given function. The stability of the sequence in the given fold and its suitability for the predetermined function are estimated by the energy function. Thus, we are looking for a sequence that minimizes the energy on the given structure without any certitude that the structure minimizes the energy of the sequence [103].

In other words, there could be another structure for which a given sequence would give a lower energy. The commonly accepted definition of CPD is thus inaccurate, even though proven to yield encouraging results in practice. Moreover, the target protein structure space is restricted to native-like conformations, with regular substructure patterns commonly observed in nature, which only represent a small fraction of the space defined by rigid body degrees of freedom, backbone torsion angles and amino acid side chains torsion angles. This inadequate definition of the CPD problem is a limiting factor that excessively restricts the design process.

Another limitation of CPD methods is their limited ability or inability to take protein flexibility into account. Protein flexibility is known to have a major impact on protein functions and properties. However, integrating protein flexibility into a CPD framework requires additional degrees of freedom to be taken into account and thus increases the already colossal size of the search space. In order to make the problem more manageable, and thus, to limit the exploration of the conformational space, the modeling of the problem can be simplified by considering the protein backbone "rigid" during the design process. This commonly used simplification illustrates the lack of expressiveness of the CPD problem.

Finally, it is also worth mentioning that the energy function which is used for the energy assessment of proteins is only an approximation. It typically includes statistical terms, which define different protein interactions, and are trained on naturally observed proteins. Such approximation of the energy function induces a lack of accuracy in CPD problems. Furthermore, the energy function commonly evaluates protein stability and ignores fitness objectives such as activity.

2.4.4 Stochastic and deterministic approaches

2.4.4.1 Stochastic approaches

Historically, the first computer simulation of a molecular system has been done using the Metropolis-Monte Carlo simulation method, one of the most famous stochastic approaches [104]. Monte Carlo methods are based on different computational algorithms that sample configurations of a system by making changes to the positions, orientations and conformations proportionally to an underlying distribution. The principle is to iteratively propose a modification of the system and then to decide if this modification is accepted or rejected according to a convergence criterion such the Metropolis criterion. For a given CPD problem and a given energy function to minimize, an MC algorithm randomly modifies the conformation of an amino acid residue in the sequence and calculates the new energy of the system. If the energy decreases, the modification is accepted. If the energy increases, the modification is accepted with Boltzmann probability, according to a temperature parameter.

Varying the temperature parameter allows crossing energy barriers and overcoming multiple local minima of the energy landscape. Simulated annealing utilizes this principle by heating the system and then cooling it down in order to gradually decrease the probability of accepting high energy conformations. Therefore, it is the most commonly used heuristic algorithm for the resolution of the CPD problem [82, 83], and is implemented in the well-known ROSETTA molecular modeling and design software [54]. Other types of stochastic algorithms are also used within other CPD frameworks such as genetic algorithms [105]. However, since stochastic methods only provide asymptotic convergence proofs, for a given CPD problem this type of method can give a different solution at each execution. Therefore, they do not guarantee that the solution returned corresponds to the optimal solution of the problem, the so-called Global Minimum Energy Conformation (GMEC). By exploring the search space locally, stochastic methods can get trapped in local minimum and give a solution that is far from the global minimum. Providing no finite-time guarantee on the quality of the solution is a disadvantage of such approaches. This is the reason why deterministic approaches have been developed in order to solve the combinatorial optimization problem of CPD. They provide guarantees of finding the GMEC or a sufficiently low-energy solution.

2.4.4.2 Deterministic approaches

Deterministic approaches do not involve any random process and perform a proof that the identified solution is the best or a sufficiently good solution for a given problem. Unlike stochastic methods which start from a random solution and explore the search space locally, deterministic methods explore a search tree. The nodes of the tree are sub-problems of the original problem defined by an increasingly restricted search space. A solution is found when the search space is reduced to a single assignment: a leaf of the tree is reached. Some of the most famous deterministic approaches are search tree algorithms such as Branch-and-Bound methods [106].

Search Trees and Branch-and-Bound

Search Tree algorithms are widely used for solving energy optimization problems. A search tree is represented by a hierarchical structure of linked nodes. Each node of the tree represents a sub-problem of the original problem. When a node has child nodes, it means that some variable has still several possible values. The domain of such a variable can be split into a collection of sub-domains and each child node has the domain of the corresponding variables restricted to a different subset. A node without child nodes is called a leaf: all its variables have only one possible value. A leaf defines a solution. The root of the tree corresponds to the initial problem, when all variables have their initial domain. A solution is given by the path from the root to one of the leaves. The aim of search tree algorithms is to find a solution by exploring the tree from the root node and examining child nodes. Several strategies exist for exploring a search tree. The Branch-and-Bound fam-

ily of algorithms represents a widely used algorithmic framework for finding exact solution to NP-hard optimization problems [107]. The branch-and-bound generic algorithm starts by initializing a global energy upper bound, which keeps track of the best solution found so far, from a heuristic solution (Algorithm 1). It maintains a list of so-called open nodes to explore. At each iteration, the algorithm branches on a node from the list. If the node is a leaf, it is evaluated. When a new best solution is found, a global upper bound on optimal energy is updated. If the current node is not a leaf, a local lower bound on the energy of all leaves below the current node is computed for each of its child nodes. If the lower bound of a child node is lower than the current upper bound, the node is added to the list of open nodes. Otherwise it is pruned, since it cannot lead to a better solution. The algorithm terminates when the list of candidate node is empty. At this point, every node in the tree has either been explored or pruned. Thus, the current best solution is the global optimum. The way in which the algorithm explores the tree is important and can follow different strategies. For example, the Best First Search strategy explores the most promising node first and the Depth-First Search strategy explores the deepest node first. When a node is selected, the choice of which variable sees its domain split is crucial in terms of efficiency. It is usually determined by sophisticated heuristics.

Algorithm 1 Branch and bound generic algorithm.

```
1: Inputs: P : problem
2: current_best = heuristicSolve(P)
3: upper_bound = eval(current_best)
4: L = candidateNodes(P)
5: while L ≠ ∅ do
6:    n = L.chooseCandidate()
7:    if n.isLeaf() then
8:       if eval(n.getSolution()) < upper_bound then
9:          current_best = n.getSolution()
10:         upper_bound = eval(current_best)
11:      end if
12:   else
13:      for all c = n.children() do
14:         if lowerBound(c) ≤ upper_bound then
15:            L.enqueue(c)
16:         end if
17:      end for
18:   end if
19: end while
20: Output: current_optimum
```

Deterministic methods in CPD

A well-known Protein design software that relies on a deterministic approach to solve CPD problem is *Osprey* [108]. In *Osprey*, a combination of Dead-End-Elimination (DEE) [109] and A* [110] (Best First Search Branch and Bound) is utilized in order to identify the GMEC. In the first step, DEE is used to eliminate rotamers that are energetically dominated by other rotamers and that are therefore not part of the optimal solution. This way, DEE reduces the search space of the problem. A* subsequently explores the remaining search tree in order to find the lowest energy solution. Even though this algorithm theoretically enables the identification of the GMEC, it is limited by its exponential time and space complexity leading to high memory consumption.

The exponential space complexity comes from the way the A* algorithm explores the search tree using the Best First Search method.
Other deterministic approaches employ Branch-and-Bound algorithms. Using the Depth First Search instead of the Best First Search, they can avoid memory issues. Besides the node exploring heuristics used, the efficiency of a Branch-and-Bound algorithm depends on its ability to quickly find a good solution (upper bound) and the ability to correctly estimate a floor value of the energy at each node of the search tree (lower bound). One of the mathematical frameworks that offers a nice and efficient way of calculating lower bounds is the graphical model called Cost Function Network (CFN). Furthermore, it has been shown that the CPD problem can be modeled as a CFN. This type of deterministic approach has been shown to solve the CPD problem more rapidly than DEE/A* approach [111, 112, 113].

Chapter 3

Computational Protein Design with Cost Function Networks

Contents

3.1 Cost Function Networks (CFN)

3.1.1 Graphical Models

Graphical Models are used to describe mathematical functions of many variables using decomposability [114]. Decomposability refers to the fact that some mathematical functions can be decomposed into a combination of small functions, involving few variables. Thus graphical models serve as a powerful tool for representing relationships between many variables, described as a graph. In this graph, nodes correspond to variables and edges represent dependencies between the variables. The graph then captures a description of the function decomposed as a combination of smaller functions, each depending only on a subset of the variables.

In the past years, graphical models became an influential tool in the field of Artificial Intelligence, Statistics and Statistical Physics for knowledge representation, learning and reasoning tasks. For example, in their deterministic variants, they are widely used for carrying out reasoning tasks such as planning, diagnosis and prediction, design etc. [115]

There are few classifications of graphical models but the main one separates them into probabilistic or deterministic graphical models. Some of the most famous probabilistic models are Bayesian networks and Markov Random Fields while Constraint Networks and Cost Function Networks are some of the most famous deterministic models. Probabilistic networks capture the joint probability distribution

of a set of random variables. This joint probability distribution can be decomposed into a product of factors each depending only on a subset of the variables.

Constraint Networks model constraints between variables using Boolean functions. Finding an assignment of variables which satisfies all constraints in a Constraint Network is known as the Constraint Satisfaction Problems (CSP) [116]. Cost Function Networks extend the notion of Constraint Networks by weighting the constraints with a numerical value. Optimizing the joint function defined by a CFN is the Weighted Constraint Satisfaction Problem (WCSP). This problem has been introduced in Artificial Intelligence for automated reasoning [117].

3.1.2 Definition of Cost function networks

Definition 1. *A CFN (X, W, k) is defined by:*

- *a set X of variables $x_i \in X$ indexed by $I = \{1, \ldots, n\}$, each variable x_i takes its values in a finite domain D_i of maximum cardinality d.*
- *a set of numerical cost functions $w_S \in W$ each involving a subset $\{x_i \in X \mid i \in S\}$ of all variables.*
- *The cost k is a finite or infinite upper bound on costs: a cost of k or above is considered as forbidden.*

The set $S \subset I$ of a cost function w_S is called the scope of the cost function. We denote by D^S the Cartesian product of the domains of all variables indexed in S: $D^S = \prod_{i \in S} D_i$.

The cost of an assignment t of all variables is defined as the sum $\sum_{w_S \in W} w_S(t[S])$ of all cost functions w_S, where $t[S]$ is the partial assignment of t with respect to the scope S of function w_S. If it is strictly less than k, it is said to be a solution. Notice that the upper bound k plays the role as an infinite cost: any assignment with cost k or above is considered as infeasible and is not a solution. A CFN model can be customized by adding constraints in the form of new cost functions that would yield a value greater than k if the constraint is not satisfied.

3.2 Modeling CPD as a Cost Function Network

CPD modeling as a cost function network is straightforward. In this section we describe the CFN representation of CPD problems. In this thesis, toulbar2, a CFN solver developed at MIAT, is used to solve CPD problems (`https://github.com/toulbar2/toulbar2`).

3.2.1 Modeling Single State CPD with CFN

As we mentioned earlier, the usual approach of CPD simplifies the CPD problem by fixing the protein backbone degrees of freedom to an idealized target backbone. This kind of approach is called Single State Protein Design (SSD).

We model the rigid discrete CPD problem using a CFN (X, W, k) with one variable x_i per position i in the design. At each position $1 \leq i \leq \ell$, where l is the number of residues, corresponds a set S_i of possible amino acids. For each amino acid $a \in S_i$, we are given a set $C_{i,a}$ of its allowed conformations. A pair $r = (a, c)$, where $a \in S_i$ and $c \in C_{i,a}$, is called a rotamer.

The domain of variable x_i is the set of rotamers $(a, c) \in S_i \times C_{i,a}$ available for design at position i and the set of functions W contains the terms of the pairwise decomposable energy functions: a constant term $E()$ for the rigid bodies, one-body terms $E(x_i)$ that capture internal side-chain energies and rotamer-backbone interactions at position i and two-bodies terms $E(x_i, x_j)$ which capture interactions between positions i and j. The objective is to find the combination of rotamers which minimizes the joint cost/energy of the backbone. This is the optimum solution of the WCSP [118, 119, 85].

3.2.2 Identification of guaranteed solution

Proving that a solution is optimal, means showing that no other solution can have a better energy. To do so, CFN algorithms use a Search Tree algorithm as described earlier and a mechanism that allows us to prune branches of the tree when we are sure that they cannot improve the current best solution.

In toulbar2, the default search tree strategy is Hybrid Best-First Search (HBFS). HBFS mixes the Depth-First Search and Best-First Search strategies: it maintains a list of open nodes, similarly to Best-First Search, but expands each selected open node in a Depth-First Search manner for a bounded number of nodes. Each unexplored node pending at the end of the Depth-First Search is added to the open node list.

Local consistency algorithms provide an efficient way of computing lower bounds, further used to prune the search tree. Local consistency are enforced using so-called *Equivalence Preserving Transformations* (EPTs). Different costs of a CFN can be manipulated, moved between different cost functions, in order to reveal properties which may improve the lower bound on the optimal cost. Thus, the new CFN has an increased lower bound and is equivalent to the original network. There are several levels of local consistencies which are defined by local consistency properties. These properties can be of variable strengths and thus provide more or less "tight" lower bounds. Two main types of strong local consistencies algorithms are used in toulbar2 : Existential Directional Arc Consistency (EDAC) [120] and Virtual Arc Consistency (VAC) [121].

3.3 Conclusion

The Single State Design method based on CFN framework has been implemented in toulbar2 and can use any decomposable energy function such as the energy functions available in ROSETTA molecular modeling and design software [54]. The recent

design of the hyper-stable self-assembling β-propeller "Ika" using the CFN technology [122] seems to indicate that guaranteed methods can also be useful in practice, combining efficiency with the assurance that optimization did not fail. However, considering a single rigid backbone as a target ignores backbone flexibility and certainly decreases the chances of designing a protein which folds and possesses desired properties. In order to take protein flexibility into account during CPD process, one should consider several backbone states simultaneously. This type of approach, called Multistate Design (MSD), defines challenging computational problems that are at the core of this thesis.

Part II

Modeling: MultiState Protein Design methods

Chapter 4

Positive Multistate Protein Design: modeling Protein Flexibility

Contents

4.1 Introduction

This chapter aims at combining the guarantees and efficiency of CFN algorithms with the idea of defining the target structure as an ensemble of backbone conformations instead of a single idealized structure. Indeed, the traditional single state protein design (SSD) contrasts with the increasing evidence that proteins do not remain fixed in a unique conformational state but rather sample conformational ensembles. Compared to the usual SSD approach, multistate design (MSD) has shown to provide enhanced design capacities [123] to stabilize an ensemble of backbones [124, 125], to design conformational switches [126, 101] or proteins with specific binding properties [127, 84, 128]. In 2017 Loffler and coworkers showed that Rosetta modular framework for multistate design offered a 15% higher performance than single-state design on a ligand-binding benchmark [129]. Multistate design

was also used for understanding thermal adaptation of enzymes by rational design of 100 adenylate kinases and prediction of their stability and adapted functions through multistate modeling [130]. The use of multistate also allowed the design of bispecific antibodies [131] or the design of switches that bind to antigens on the cell surface and, through a conformational change, are activated only when a precise combination of antigens is present [101]. In all these cases, MSD seeks to identify a sequence that optimizes a function of its optimal energies on the different considered states. This function, or "fitness", is itself non trivial to compute, as it requires the computation of optimal conformations of the sequence on several backbone states. Many SSD optimization algorithms have been extended to MSD, with more or less general fitness functions, including Monte Carlo with simulated annealing [126, 132, 133], genetic algorithms [134], the FASTER approach [135], cluster expansion [127], and dead-end-elimination [136], also in combination with A* [137]. These methods are often limited by the number of mutations they can explore across all states, usually going up to 30 mutations maximum. Recently, Sauer and coworkers proposed an MSD approach based on Monte Carlo Simulated Annealing which samples the sequence space on each state independently and then adds constraints so that the search on all states converges to a unique sequence [133].

The nature of the fitness function intimately depends on the design problem. The Boltzmann-weighted average of the energies in each state is ideal when the aim is to stabilize any of the backbone states. When instead, it is to design a sequence that must fit all conformational states, the fitness will typically be the average of the energies on all states. These two cases are identified as *positive* multistate design. The fitness improves when the energy of the sequence improves in any state. However, some design problems involve undesirable states for which this property is violated: the fitness worsen when the energy of the sequence in an undesirable state improves. This can occur for the design of protein-ligand binding or oligomeric association specificity. These design problems involve negative design against unwanted binding partners present in the medium. Specificity then arises from the preference for a given partner over the others. Thus, undesirable (negative) molecular states also have to be considered.

In this chapter, we observe that the type of the fitness function has a profound influence on the computational nature of the problem. The introduction of undesirable states makes the problem qualitatively more complex, shifting its complexity from NP-complete to the much harder $\mathsf{NP}^{\mathsf{NP}}$-complete category [138]. This result has several implications. Negative MSD being qualitatively harder than SSD, optimization methods may become unable to reach good quality solutions sooner than in the SSD case. It also shows that positive MSD is an interesting target: it is "just" NP-complete while capturing some backbone flexibility and dynamics [139]. Hence, we leverage the polynomial equivalence of NP-complete problems by introducing efficient reductions of two variants of positive multistate design to Cost Function Networks. The first variant uses a minimum energy fitness and the second one a (weighted) average energy fitness. Beyond saving programming efforts, this approach directly benefits from the advanced CFN processing machinery [140, 141].

On various positive MSD problems, we show that it is possible to identify an optimal MSD sequence with associated optimal conformations in reasonable time, on computationally extremely challenging design problems of a size far beyond what has been solved with existing state-of-the-art guaranteed multistate design methods [137], including recent CFN based methods with dedicated algorithms [142]. POMP^d is also natively able to exhaustively enumerate suboptimal sequences close to the MSD optimum, which is convenient for sequence library design. Contrarily to what has been previously described [124], we observe that the use of an ensemble of NMR structures as a positive ensemble of backbones provides strong improvements in terms of native sequence and sequence similarity recovery when an average energy criteria is used. We also show that this improvement is reduced but still present when a backrub generated ensemble derived from a single X-ray structure is used. These results show that Positive Multistate Design is essentially as hard to solve as Single State Design, both in theory and in practice. Given the significant improvement that the multistate approach brings, it is our feeling that positive MSD should be considered as a default design approach when specificity is not the main target.

4.2 Methods

4.2.1 Our definition of multistate design

In discrete rigid MSD, we are given a set of positive backbones that represent the target structure and a set of negative backbones that are undesirable. In either the positive or negative case, these states have also been called "sub-states" [142]. The final fitness of a sequence is then defined as the difference of the fitness on the positive and negative states. Various definitions of the fitness can be considered:

- If the set of states represents "possible backbones" that the sequence can (de)stabilize, with no prior knowledge on which one will be adopted in practice, the Boltzmann-weighted energy over all the considered states (defined as the sum of $e^{-\beta E}E$, where $\beta = \frac{1}{k_B T}$), is an attractive criteria. Because this gives an exponential advantage to the backbone with lowest energy, it has been approximated by the minimum energy [142]. This becomes equivalent to what is called Multistate Analysis (MSA) [143].

- If instead the set of states represents structures that must be jointly (de)stabilized, as in conformational switches design for example, it is important that the energy of every state contributes to the fitness: optimizing the average energy is more adequate.

More formally, we are given a set of positive and negative rigid backbone states $\mathbf{B} = \mathbf{B}^+ \cup \mathbf{B}^-$, all with the same number ℓ of residues. At each position $1 \leq i \leq \ell$, we have a set S_i of possible amino acids. For each $a \in S_i$ and each state $B_j \in \mathbf{B}$, we are given a set $C_{i,a}^j$ of allowed conformations for the amino acid a at position i

in state B_j. At position i, a pair $r = (a, c)$ where $a \in S_i$ and $c \in C^j_{i,a}$ is called a rotamer.

We also assume that the energy $E_b(\mathbf{a}, \mathbf{c})$ of a backbone B_b equipped with a given amino acid sequence $\mathbf{a} \in \prod_i S_i$ and conformations $\mathbf{c} \in \prod_i C^j_{i,a[i]}$ is described as a sum of terms that each involve at most two rotamers $r_i = (\mathbf{a}[i], \mathbf{c}[i])$ and $r_j = (\mathbf{a}[j], \mathbf{c}[j])$, for $1 \leq i, j \leq \ell$:

$$E_b(\mathbf{a}, \mathbf{c}) = \left[E_b() + \sum_{1 \leq i \leq \ell} E_b(r_i) + \sum_{1 \leq i < j \leq \ell} E_b(r_i, r_j) \right] \tag{4.1}$$

To capture the different criteria that have been used, such as minimum or (weighted) average energy, we imagine that a binary operator $\oplus$ is used to combine the energies of the backbones. Optimizing the minimum energy is obtained using $\oplus = \min$. This will be called min-MSD. Since the number of states is fixed, optimizing the average energy is obtained using $\oplus = +$. This will be called Σ-MSD.

More formally, the $\oplus$-MSD problem asks whether there exists a sequence $\mathbf{a} \in \prod_i S_i$ (the sequence design space) such that

$$\left(\bigoplus_{B_j \in \mathbf{B}^+} \min_{\mathbf{c} \in \prod_i C^j_{i,a[i]}} E_j(\mathbf{a}, \mathbf{c}) \right) - \left(\bigoplus_{B_j \in \mathbf{B}^-} \min_{\mathbf{c} \in \prod_i C^j_{i,a[i]}} E_j(\mathbf{a}, \mathbf{c}) \right) \leq k$$

When the set $\mathbf{B}^-$ is empty, we say that this is a positive $\oplus$-MSD problem. The problem is to identify a sequence $\mathbf{a} \in \prod_i S_i$ (the sequence design space) such that:

$$\left(\bigoplus_{B_j \in \mathbf{B}^+} \min_{\mathbf{c} \in \prod_i C^j_{i,a[i]}} E_j(\mathbf{a}, \mathbf{c}) \right) \leq k$$

.

Here, we consider three types of design approaches: SSD, min-MSD (equivalent to MultiState Analysis [143]) and Σ-MSD. These three approaches are described in Figure 4.1 showing how different backbones are used to score various sequences in each case.

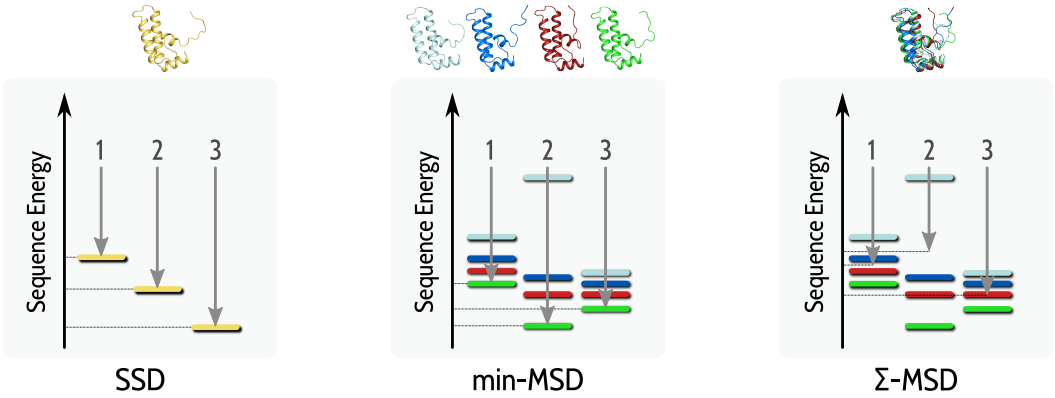

Figure 4.1: In SSD (left), a single state (yellow) is used to score and rank sequences (1,2 and 3) according to their energy, which defines the sequence fitness (grey arrow): the best sequence is sequence 3. In $\oplus$-MSD, an ensemble of (here) four backbone states (cyan, blue, red and green) is used to score and rank sequences. The fitness of each sequence (grey arrow) can be computed using the min (center) or the (weighted) sum of the sequence energies in each state (right). Depending on the used operator, the ranking may change and different sequences can be selected. In min-MSD sequence 2 is ranked first as it has the best energy on the green backbone. In Σ-MSD, it is ranked last because of its bad energy on the cyan backbone.

4.2.2 Computational complexity

Since Pierce's seminal paper [81], we know that the SSD problem is decision NP-complete: given an arbitrary rigid backbone, and arbitrary pairwise decomposable energy function and rotamer library, deciding whether there exists a sequence and associated side-chain conformations with energy lower than a given threshold k is NP-complete. This result proves that the SSD problem is among the hardest of all the problems in its class: any other problem in NP can be reduced to it efficiently (in a time that grows as a polynomial in the size of the problem).

Theorem 1. *Assuming energies are represented as finite objects and that addition, comparison and $\oplus$ can be computed in a time polynomial in the length of their arguments, the positive $\oplus$-MSD problem is NP-complete and the general $\oplus$-MSD problem is $\mathsf{NP}^{\mathsf{NP}}$-complete (or Σ_2^p-complete).*

This theorem was proved by Manon Ruffini, a colleague from MIAT laboratory, and the proof can be found in the Supplementary materials of our Pomp^{d}article [144].

4.2.3 Positive min-MSD as a CFN

In positive min-MSD, one seeks a sequence that best stabilizes one backbone among all backbones $B_i \in \mathbf{B}^+$ or equivalently that minimizes:

$$\min_{B_b \in \mathbf{B}^+} \min_{\mathbf{c} \in \prod_i C^j_{i,a[i]}} \left[E_b() + \sum_{1 \leq i \leq \ell} E_b(r_i) + \sum_{1 \leq i < j \leq \ell} E_b(r_i, r_j) \right]$$

where we use the decomposable form of the energy from Equation 4.1.

This problem can be tackled by solving the SSD problem on every backbone state $B_i \in \mathbf{B}^+$ and using the sequence of the backbone with minimum energy $E_{\min}$ as the solution. Given an energy gap of size $\Delta > 0$, a library of suboptimal sequences whose energy is less than $E_{min} + \Delta$ can be obtained by taking the union of the libraries obtained with energy threshold Δ on every backbone $B_i \in \mathbf{B}^+$. Because it allows to consider each state independently, this approach is often referred as just a "Multistate Analysis" (MSA) [123].

Instead of solving as many SSD problems as there are states, the problem can be modelled as a single Cost Function Network whose optimal solution will define both the optimal sequence and the state on which the optimum is reached. This model exploits the fact that CFNs solvers can deal with terms involving more than two variables.

For a set $\mathbf{B}^+$ of n states, we start from a model with the same variables as in the SSD case: one variable x_i per position i, with a domain equal to the set of available rotamers at this position. We also introduce a variable x_B with a domain $\{1, \ldots, b\}$ that represents an index in the set of positive states. The CFN for SSD of any of those backbones involves zero, one and two-bodies terms. We introduce the new variable x_b in the scope of each of these terms so that:

- all the constant terms $E_b()$ for state $B_b \in \mathbf{B}^+$ are transformed in a one-body function depending on the state index x_b and equal to the constant term for this state $E(x_b) = E_{x_b}$.
- for every position i, all the one-body terms $E_b(x_i)$ for states $B_b \in \mathbf{B}^+$ are transformed in two-bodies terms $E(x_b, x_i) = E_b(x_i)$.
- for every pair of positions (i, j), all the two-bodies terms $E_b(x_i, x_j)$ for all states $B_b \in \mathbf{B}^+$ are transformed in three-bodies terms $E(x_b, x_i, x_j) = E_b(x_i, x_j)$.

A solution of the resulting CFN defines a state through its index x_b and a sequence-conformation for every position in x_i. The cost of the solution is, by definition of the terms above, equal to the energy of this sequence-conformation on this backbone. An optimal solution minimizes the energy over all possible choices of states and sequence-conformations and is therefore a solution of the positive min-MSD problem.

This approach was tested but never found to outperform the simple approach where each backbone is solved independently. We therefore used this latter method. The reduction above has the advantage that it simplifies the construction of a sequence library: it suffices to enumerate all suboptimal sequences within Δ of the optimum of this Cost Function Network to directly build the joint library.

4.2.4 Positive Σ-MSD as a CFN

In positive Σ-MSD, one seeks a sequence that best simultaneously stabilizes all states $B_i \in \mathbf{B}^+$ or equivalently that minimizes:

$$\sum_{B_b \in \mathbf{B}^+} \min_{\mathbf{c} \in \prod_i C^j_{i,a[i]}} \left[E_b() + \sum_{1 \leq i \leq \ell} E_b(r_i) + \sum_{1 \leq i < j \leq \ell} E_b(r_i, r_j) \right]$$

This problem cannot be tackled by solving the SSD for every state $B_b \in \mathbf{B}^+$ and summing the energies because the optimal sequences for each SSD problem may differ. To avoid this issue, we exploit the capacity of CFNs to represent hard constraints using the cost "k". Contrarily to stochastic search algorithms (that could fail because of lack of ergodicity or require specific treatment to preserve it), CFN algorithms have the capacity to actively exploit these constraints to accelerate search by predicting inconsistent choices using local consistencies [140].

For each state $B_b \in \mathbf{B}^+$, we compute the SSD CFN defined in Chapter 3. We use a superscript for all variables in these CFNs to identify the state they correspond to: x_i^b is the variable representing position i in the SSD CFN of state B_b. We build a Σ-multistate CFN as follow:

- the set of variables of the multistate CFN is the union of all the sets of variables of each SSD CFN. For a positive Σ-MSD full redesign problem with n backbones of length ℓ, there will be $n\ell$ variables, each with the same domain as in the original SSD problems.

- the set of functions of the multistate CFN contains all the cost functions $E_b(), E_b(x_i^b), E_b(x_i^b, x_j^b)$ of every SSD CFN plus a set of two-bodies functions $SS(x_i^b, x_i^{b'})$ which, for every position i and every pair of state B_b and $B_{b'} \in \mathbf{B}^+$, enforce that the rotamers used in the states B_b and $B_{b'}$ for position i should represent the same amino acid. $SS(x_i^b, x_i^{b'})$ is equal to zero if x_i^b and $x_i^{b'}$ represent the same amino acid and is equal to the upper bound k in the formal definition of cost function networks given in Chapter 3 otherwise.

A solution of the multistate CFN contains a solution defining a sequence and conformation for every state $B_b \in \mathbf{B}^+$. By definition, the cost of this solution is the sum of all energy terms over all states. Additionally, the $SS(x_i^b, x_i^{b'})$ functions impose that the same sequence is used in all states: an optimal solution defines a sequence that minimizes the sum of energies. It therefore solves the positive Σ-MSD problem.

This generates a CFN with $n\ell$ variables, $n\frac{\ell(\ell+1)}{2}$ energy terms and $\ell\frac{n(n-1)}{2}$ additional $SS(x_i^b, x_i^{b'})$ constraints. Since the $SS(x_i^b, x_i^{b'})$ constraints define an equivalence relation, transitivity implies that it is sufficient to only enforce this constraint for every pair $b, b+1$ of successive states. This requires $\ell(n-1)$ constraints instead of $\ell\frac{n(n-1)}{2}$.

This reduction is used in the rest of the chapter to solve positive Σ-MSD: the MSD problem is transformed in a CFN and the CFN solved. The use of a single

CFN also allows to easily generate suboptimal sequences using the dedicated SCP-branching strategy [145].

4.2.5 Benchmark Preparation

Two datasets have been prepared. The first one contains 15 NMR structures and the second one 15 X-ray structures (see Figure 4.2) that have been extracted from the Protein Data Bank (PDB) [8] and filtered with following criteria:

- monomeric proteins, no missing or nonstandard residues, no ligand
- maximum sequence length of 100 amino acid residues
- NMR resolved structures must contain at least 20 conformations
- X-ray structures must be resolved below 2 Å

The set of backbones in the NMR ensemble was submitted to RMSD-based hierarchical clustering using the Durandal software [146] in order to select the four most diverse conformations.

The X-ray ensembles have been generated by RosettaBackrub [147] which uses the BackrubEnsemble method for flexible protein backbone modeling in Rosetta [148, 149]. One hundred conformations were generated for each structure. This step was followed by the same RMSD-based hierarchical clustering as for the NMR ensembles in order to select the four most diverse among given conformations. Clustering distance thresholds were set to reach the desired number of clusters (see Table 4.1). The structures were relaxed using RosettaFastRelax with harmonic constraints, resulting in output structures which are typically within 2 Å RMSD of the initial structure. Pairwise energy matrices were computed with Dunbrack2010 rotamer library [150] and `beta_nov16` scoring function [32], using PyRosetta 171 [151]. These problems define huge search spaces (Table 4.2) with sizes that can exceed 10^{900} or 10^{540} if the effect of the SS constraints is taken into account.

We also used the 4 multistate problems provided with the multistate iCFN solver at `https://shen-lab.github.io/software/iCFN`. All 4 problems include eleven states of 3QDJ, a complex between TCR DMF5 and human Class I MHC HLA-A2 with a bound MART-1(27-35) nonameric peptide, produced by an MD simulation [142]. Each problem has a single residue to redesign (from 20 possible amino acids, with 7 protonation states for Asp, Glu and His), all close residues are considered as flexible. Because of a dense rotamer library ($4,731$ rotamers), these problems define large search spaces (Table 4.3).

System name	PDB ID	Number of conformations	Number of residues
Saccharomyces cerevisiae J-domain	5vso	20	75
Human SNF5/INI1 domain	5l7b	20	75
Trypanosoma brucei Pex14 N-terminal domain	5mmc	20	70
Immunoglobulin binding domain of streptococcal protein G	1gb1	60	56
Cytotoxin-I from the venom of cobra N. oxiana	5t8a	20	61
E2 lipoyl domain from Thermoplasma acidophilum	2l5t	33	77
Spider toxin U4-hexatoxin-Hi1a	2n6r	20	76
Phl PII from timothy grass pollen	1bmw	38	96
Antibacterial factor-2	5ix5	20	68
Peptide toxin SsTx from Scolopendra subspinipes mutilans	5x0s	20	53
Rhabdopeptide NRPS Docking Domain Kj12A-NDD	6ews	20	63
Platelet integrin-binding C4 domain of von Willebrand factor	6fwn	20	85
Human ubiquitin at 298K	6qf8	20	79
Ubiquitin (Q41N variant)	6jlt	20	76
Sushi 1 domain of GABAbR1a	6hkc	20	75

System name	PDB ID	R(Å)	Number of residues
Hydrophobic protein from Soybean	1hyp	1.8	80
Alpha-amylase inhibitor hoe-467A	1hoe	2	74
E.coli Cold-shock protein A	1mjc	2	69
B1 immunoglobulin-binding domain of streptococcal protein G	1pga	2.07	56
PAS Factor from Vibrio vulnificus	2b8i	1.8	77
Apo-GolB	4y2k	1.7	65
Toxin isolated from the Malayan Krait	1f94	0.97	63
Allergen phl p2	1who	1.9	96
Alpha-spectrin src homology 3 domain	1tud	1.77	62
Headpiece Domain of Chicken Villin	1yu5	1.4	67
Ribosomal protein L30 from thermus thermophilus	1bxy	1.9	60
C-TERMINAL DOMAIN OF THE RIBOSOMAL PROTEIN L7/L12	1ctf	1.7	74
C-Myb DNA-Binding Domain	1guu	1.6	52
Domain 3 of human alpha polyC binding protein	1wvn	2.1	82
Type III Antifreeze Protein RD1 from an Antarctic Eel Pout	1ucs	0.62	64

Figure 4.2: Description of protein systems: For each instance: system name, reference PDB id, crystallographic resolution or number of conformations for NMR structures, number of amino acid residues(N), SCOP stuctural classification(Class).

Table 4.1: User-defined clustering distance thresholds (d) for each protein structure.

NMR structures		X-ray structures	
PBD ID	d (Å)	PBD ID	d (Å)
5vso	2.0	1hyp	0.5
5l7b	0.4	1hoe	0.4
5mmc	2.0	1mjc	0.6
1gb1	0.3	1pga	0.4
5t8a	0.2	2b8i	0.4
2l5t	1.0	4y2k	0.4
2n6r	0.3	1f94	0.4
1bmw	1.2	1who	0.4
5ix5	0.6	1tud	0.5
5x0s	1.5	1yu5	0.15
6ews	0.5	1bxy	0.5
6fwn	0.8	1ctf	0.3
6qf8	1.2	1guu	0.3
6jlt	0.5	1wvn	0.5
6hkc	1.0	1ucs	0.3

4.2.6 Solving SSD, min-MSD and Σ-MSD with Pomp^d

Thanks to the three reductions of SSD, positive min-MSD and Σ-MSD to CFNs, it is now possible to solve these problems using a CFN solver such as toulbar2.

For our benchmarking NMR and X-ray instances, we downloaded toulbar2 from its repository, using its 'cpd' branch. All instances were solved using the "`-dee: -O=-3 -B=1 -A -cpd`" taken in a recent paper [85]. Compared to the default behavior, this command line deactivates Dead End Elimination and activates the exploitation of the interaction structure (treewidth) and the strong 'Virtual Arc Consistency' bounds [140]. Computations were done on an Intel(R) Xeon(R) CPU E5-2630 at 2.30GHz with 24GB of RAM. The overall workflow is described in Figure 4.3.

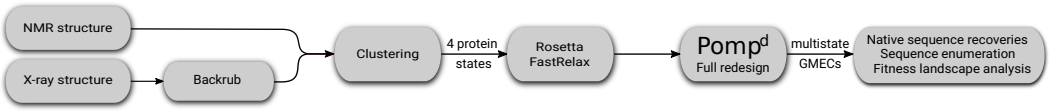

Figure 4.3: Overall workflow for Xray and NMR structures.

4.2.7 Solving positive min -MSD with iCFN

Recently, a guaranteed CFN-based algorithm for both positive and negative min-MSD was introduced as the iCFN method [142]. The authors did not use a re-

Table 4.2: Multistate design problems: for each problem we give the average search space of four SSD problems, search space for the min-MSD problem, defined as the sum of all SSD search space sizes, the raw Σ-MSD search space size, defined by the product of the size of all variable domains and the search space size reduced by the SS constraints that impose that all states use the same sequence.

PBD ID	average $\overline{SSD}$ search space	min-MSD search space	Σ-MSD search space	Σ-MSD reduced search space
		NMR structures		
5vso	$1.3\ 10^{181}$	$5.4\ 10^{181}$	$8.5\ 10^{723}$	$1.6\ 10^{431}$
5l7b	$2.6\ 10^{170}$	$1.0\ 10^{171}$	$6.4\ 10^{680}$	$3.1\ 10^{411}$
5mmc	$5.6\ 10^{158}$	$2.3\ 10^{159}$	$4.1\ 10^{634}$	$8.3\ 10^{380}$
1gb1	$2.5\ 10^{137}$	$1.0\ 10^{138}$	$5.9\ 10^{547}$	$1.6\ 10^{329}$
5t8a	$9.4\ 10^{133}$	$3.8\ 10^{134}$	$5.9\ 10^{535}$	$4.8\ 10^{297}$
2l5t	$2.2\ 10^{185}$	$9.0\ 10^{185}$	$7.2\ 10^{738}$	$2.1\ 10^{438}$
2n6r	$3.4\ 10^{168}$	$1.3\ 10^{169}$	$4.0\ 10^{673}$	$9.3\ 10^{376}$
1bmw	$5.2\ 10^{229}$	$2.1\ 10^{230}$	$1.1\ 10^{914}$	$1.4\ 10^{547}$
5ix5	$4.4\ 10^{148}$	$1.7\ 10^{149}$	$2.0\ 10^{593}$	$7.8\ 10^{327}$
5x0s	$1.2\ 10^{119}$	$4.9\ 10^{119}$	$1.4\ 10^{470}$	$2.0\ 10^{263}$
6ews	$8.1\ 10^{155}$	$3.2\ 10^{156}$	$4.3\ 10^{622}$	$5.4\ 10^{376}$
6fwn	$2.8\ 10^{188}$	$1.1\ 10^{189}$	$2.4\ 10^{751}$	$4.3\ 10^{419}$
6qf8	$2.3\ 10^{188}$	$9.1\ 10^{188}$	$4.0\ 10^{750}$	$9.3\ 10^{453}$
6jlt	$6.9\ 10^{188}$	$2.8\ 10^{189}$	$1.5\ 10^{755}$	$3.5\ 10^{458}$
6hkc	$4.5\ 10^{174}$	$1.8\ 10^{175}$	$6.6\ 10^{694}$	$1.2\ 10^{402}$
		Xray structures		
1hyp	$2.1\ 10^{166}$	$8.2\ 10^{166}$	$3.2\ 10^{664}$	$4.8\ 10^{375}$
1hoe	$2.2\ 10^{171}$	$8.9\ 10^{171}$	$5.8\ 10^{684}$	$8.6\ 10^{395}$
1mjc	$4.3\ 10^{165}$	$1.7\ 10^{166}$	$10.0\ 10^{661}$	$4.9\ 10^{392}$
1pga	$4.7\ 10^{137}$	$1.9\ 10^{138}$	$1.5\ 10^{550}$	$4.0\ 10^{331}$
2b8i	$7.7\ 10^{189}$	$3.1\ 10^{190}$	$5.1\ 10^{758}$	$1.5\ 10^{458}$
4y2k	$2.5\ 10^{161}$	$9.8\ 10^{161}$	$1.7\ 10^{644}$	$3.4\ 10^{390}$
1f94	$2.9\ 10^{134}$	$1.2\ 10^{135}$	$1.4\ 10^{537}$	$1.8\ 10^{291}$
1who	$2.6\ 10^{227}$	$1.0\ 10^{228}$	$6.1\ 10^{907}$	$7.8\ 10^{540}$
1tud	$3.1\ 10^{146}$	$1.3\ 10^{147}$	$5.0\ 10^{585}$	$3.3\ 10^{351}$
1yu5	$9.4\ 10^{165}$	$3.8\ 10^{166}$	$2.9\ 10^{663}$	$9.0\ 10^{401}$
1bxy	$5.8\ 10^{147}$	$2.3\ 10^{148}$	$7.6\ 10^{590}$	$5.0\ 10^{356}$
1ctf	$5.2\ 10^{164}$	$2.1\ 10^{165}$	$1.9\ 10^{658}$	$9.38\ 10^{388}$
1guu	$1.2\ 10^{123}$	$4.7\ 10^{123}$	$1.4\ 10^{492}$	$1.2\ 10^{293}$
1wvn	$8.0\ 10^{179}$	$3.2\ 10^{180}$	$4.5\ 10^{718}$	$6.8\ 10^{429}$
1ucs	$5.9\ 10^{153}$	$2.4\ 10^{154}$	$7.9\ 10^{614}$	$1.3\ 10^{365}$

duction of the problem to CFN but proposed and implemented a new algorithm that exploits some of the underlying machinery of CFN algorithms (arc consisten-

Table 4.3: iCFN multistate design problems: for each problem we give the position of the redesigned residue, the number of flexible residues around the redesigned residue and the search space for the min-MSD problem, defined as the sum of all SSD search space sizes, the raw Σ-MSD search space size, defined by the product of the size of all variable' domains and the actual search space size, reduced by the SS constraints that impose that all states use the same sequence.

redesigned position	# of flexible residues	min-MSD search size	Σ-MSD search size	Σ-MSD reduced search size
26	18	$7.6\ 10^{30}$	$1.6\ 10^{323}$	$7.7\ 10^{308}$
28	18	$3.1\ 10^{34}$	$6.3\ 10^{362}$	$3.1\ 10^{348}$
98	19	$4.9\ 10^{31}$	$7.7\ 10^{334}$	$3.7\ 10^{320}$
100	29	$1.4\ 10^{42}$	$5.2\ 10^{447}$	$2.5\ 10^{433}$

cies [140]). The authors showed that their method outperforms the guaranteed COMETS software [137]. We therefore decided to compare POMP^d against iCFN only.

The iCFN website (`https://shen-lab.github.io/software/iCFN/`) gives access to both the software in binary format and to multistate design energy matrices. We wrote a first python script to translate iCFN-formatted problems into the `cfn.gz` CFN format that can be directly read by the CFN solver toulbar2. iCFN uses double resolution floating point energies and the `cfn.gz` format relies on a fixed point representation of energies. We used a "6 digits after the decimal point" representation. We wrote a second python/PyRosetta script to generate energy matrices in iCFN-format directly from PyRosetta. These scripts make it possible to either apply POMP^d to the positive min-MSD instances available on the iCFN website or to apply the iCFN algorithm on our benchmark set (for the min-MSD problem only as iCFN is not able to tackle Σ-MSD).

The iCFN command line used on the positive min-MSD problems was `iCFN -just_pos -ecutDEE=2 -ecutDEE_across=2 -ecutDEE_seq=10 -ecut_stability=5 -max_conf_seq=1 -max_dis_seq=9999` ⟨`files`⟩ which asks for one solution of the min-MSD problem, with no limitation on the number of mutations in the produced design sequence. Except for the effect of the various pruning thresholds used by iCFN that reduce computing time, this precisely matches the min-MSD problem we solve using CFN reductions.

The iCFN multistate designs use a specific rotamer library that includes 2 extra protonated states for glutamate (Glu) and aspartate (Asp) as well as 3 protonated states for histidine (His). Because the 'cpd' branch of toulbar2 relies on the one letter code of amino acids, it is currently unable to process the corresponding energy matrices. We therefore used the 'master' branch of toulbar2 to solve these problems. The command line used in this case is simply `-m -hbfs:` which deactivates the default Hybrid Best First Search algorithm [152] for a simple Depth-First Search and activates the median cost variable ordering heuristic [113]. All computations

were done on a laptop equipped with 16GB of RAM and a Intel(R) Core(TM) i7-7600U CPU at 2.80GHz.

4.3 Results and Discussion

4.3.1 Comparing SSD, min-MSD and Σ-MSD

Protein design problems can be modeled as SSD, min-MSD or Σ-MSD. SSD has the advantage of simplicity: there is only one backbone to design. MSD approaches have the advantage of accounting for protein backbone flexibility, with additional modeling and computing costs. As we already mentioned, min-MSD seems more suitable for situation of uncertainty: it is not known which, among all the available backbones, is the suitable one. Instead Σ-MSD seems more suitable when there is an explicit requirement that all states should be stabilized.

We assessed POMP^d on our benchmark backbone conformational state ensembles, either extracted from NMR structures or generated by backrub motions from X-ray structures. Notice that our benchmark dataset represents a selection of full protein design problems for structures of size varying between 53 and 96 residues.

Σ-MSD outperforms SSD and min-MSD in terms of sequence recovery In order to compare the accuracy of these methods, we used the native sequence recovery (*nsr*) and native sequence similarity recovery (*nssr*), which have been used extensively to evaluate protein design methods [153, 154, 129]. Native sequence recovery is defined as the fraction of positions where the native and designed sequences are identical. Native sequence similarity is defined as the fraction of positions where the native and designed sequences have a positive similarity score in BLOSUM62 protein similarity matrix. For SSD, *nsr* and *nssr* have been computed as the average of the recovery for the four SSD conformations. The results of these comparisons are shown in Table 4.4. Σ-MSD achieves on average a *nsr* of 64.7% and 66.4% and a *nssr* of 74.4% and 73.9% for respectively back-rubbed X-ray and NMR structure datasets. For every protein design in the X-ray structure dataset and for 13 out of the 15 protein designs in the NMR structure dataset, Σ-MSD provides the best native sequence recovery (p-value when comparing to respectively average SSD and min-MSD over all proteins of 2.5 10^{-6} and 1.3 10^{-5}, Wilcoxon signed rank test). For NMR structures, Σ-MSD performs 15.6% better on average than SSD and 8% better for X-ray structures. min-MSD achieves native sequence recovery rates which can almost not be differentiated from those obtained by SSD (p-value of 0.6 on the 30 proteins, Wilcoxon signed rank test). While min-MSD and SSD achieve a better sequence recovery rate on the X-ray dataset than on the NMR structures ($7-9$% better on average), Σ-MSD is less sensitive to the dataset type (1% better on average on X-ray dataset).

We expected Σ-MSD to perform better on NMR, given that the NMR ensemble corresponds to likely states of the observed proteins and min-MSD to be more adapted to the back-rubbed X-ray structures that just define a set of possible states.

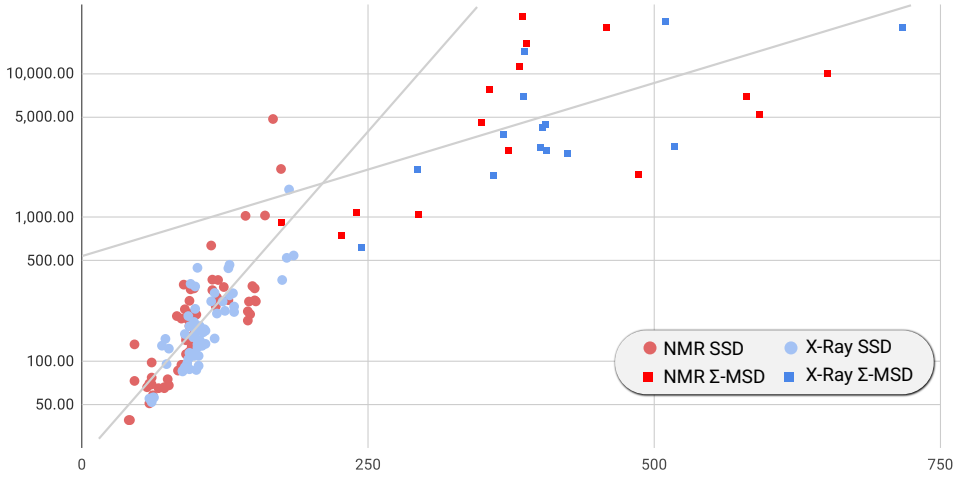

Figure 4.4: CPU time in seconds (Y logscale axis) vs. problem size (MB) for SSD and Σ-MSD problems (X axis). Each point represents one instance, NMR structures are in red, X-ray in blue. SSD problems are represented as circles, Σ-MSD problems as squares.

Instead, Σ-MSD dominates even in the back-rubbed case. min-MSD is worse than SSD on NMR ensembles but at least improves over SSD on back-rubbed X-ray structures. It is possible that a set of 4 states is too small for min-MSD to have a chance to find a suitable backbone while the more consensual approach of Σ-MSD is able to extract local information from every backbone.

Analyzing the efficiency of Pomp^d on SSD and Σ-MSD: Because SSD and ⊕-MSD are NP-complete, we expect an exponential cpu-time growth as the sizes of the problems solved increase. We plotted the cpu-time taken by Pomp^d to solve the SSD and Σ-MSD problems against the problem size represented as the size (in bytes) of the compressed file that contains the description of the problem solved in `wcsp` format (see Figure 4.4). Empirically, we observe that for each class of problem (SSD and Σ-MSD), an exponential function fits the CPU-time reasonably well and that the Σ-MSD problems tend to be simpler to solve than the SSD problems, given their larger size. In the end, the relatively slow increase in CPU time as the size grows shows that full redesign problems using an SSD or a positive min-MSD and Σ-MSD approach can be solved on a standard computer for proteins of size less than 100 amino acids in reasonable time, with guarantee on the fitness of the produced sequence.

Comparing the computational efficiency of iCFN and Pomp^d We compared Pomp^d to the recent iCFN solver [142]. iCFN can solve min-MSD problems but not Σ-MSD problems. In our first comparison, we converted the 4 positive multistate problems available on the iCFN web site (see Section 4.2.7) to a format that we could process. We tackled the min-MSD problem on these four instances

Table 4.4: Native sequence recoveries and similarity recoveries for SSD, min-MSD and Σ-MSD on both NMR structure (left) and X-ray structure (right) datasets. The protein sequences have length that vary from 53 to 96.

NMR structures				X-ray structures			
PBD ID	$\overline{SSD}$	min-MSD	Σ-MSD	PBD ID	$\overline{SSD}$	min-MSD	Σ-MSD
	Native sequence recoveries				Native sequence recoveries		
5vso	48.0%	44.0%	62.7%	1hyp	61.8%	67.5%	68.9%
5l7b	52.9%	53.6%	59.5%	1hoe	60.1%	59.4%	77.0%
5mmc	54.2%	60.0%	58.5%	1mjc	56.5%	53.6%	59.4%
1gb1	48.7%	46.4%	60.7%	1pga	52.7%	58.9%	73.2%
5t8a	39.7%	39.3%	37.7%	2b8i	48.4%	42.8%	57.1%
2l5t	58.4%	49.3%	83.1%	4y2k	62.7%	67.7%	70.8%
2n6r	53.9%	52.6%	67.1%	1f94	64.3%	65.1%	71.4%
1bmw	53.4%	56.4%	79.8%	1who	61.4%	62.7%	68.1%
5ix5	52.6%	48.5%	63.2%	1tud	45.8%	45.0%	48.3%
5x0s	42.9%	50.9%	58.5%	1yu5	64.1%	65.7%	68.6%
6ews	46.8%	46.1%	69.8%	1bxy	53.3%	50.0%	60.0%
6fwn	55.8%	54.1%	72.9%	1ctf	57.9%	50.7%	60.9%
6qf8	54.3%	51.3%	68.4%	1guu	53.4%	66.6%	66.6%
6jlt	54.9%	57.8%	65.7%	1wvn	41.9%	44.5%	50.0%
6hkc	45.0%	44.0%	66.6%	1ucs	66.1%	70.3%	70.3%
Average	50.8%	50.3%	66.4%	Average	56.7%	58.0%	64.7%
Native sequence similarities				Native sequence similarities			
5vso	58.6%	54.6%	70.6%	1hyp	71.3%	75.7%	81.1%
5l7b	69.6%	65.2%	75.4%	1hoe	67.2%	68.9%	82.4%
5mmc	59.6%	63.1%	61.5%	1mjc	67.5%	66.7%	69.6%
1gb1	62.9%	60.7%	67.8%	1pga	62.9%	69.6%	80.3%
5t8a	52.5%	52.5%	49.2%	2b8i	63.3%	58.4%	71.4%
2l5t	71.1%	62.3%	90.9%	4y2k	66.1%	69.2%	75.4%
2n6r	64.5%	59.2%	73.7%	1f94	74.2%	73.0%	80.9%
1bmw	64.9%	64.9%	84.1%	1who	72.1%	73.4%	80.9%
5ix5	62.8%	58.8%	72.1%	1tud	55.0%	56.7%	55.0%
5x0s	51.9%	60.4%	75.5%	1yu5	72.4%	73.1%	74.6%
6ews	65.8%	73.0%	82.5%	1bxy	64.5%	58.3%	73.3%
6fwn	63.2%	61.1%	78.8%	1ctf	65.6%	59.4%	68.1%
6qf8	64.4%	64.4%	77.6%	1guu	69.6%	72.5%	82.4%
6jlt	62.5%	61.8%	71.1%	1wvn	57.1%	63.5%	62.2%
6hkc	59.3%	58.6%	78.6%	1ucs	73.0%	76.6%	78.1%
Average	62.2%	61.4%	73.9%	Average	66.8%	67.7%	74.4%

with 11 states with iCFN and POMP^d. Since sequence recovery was shown to be better on Σ-MSD, we also tried to solve Σ-MSD problem with POMP^d only. This is also the criteria that COMETS [137] uses.

The results are presented in Table 4.5. We observe that POMP^d is much faster than iCFN, by a non constant factor that increases with problem size. Furthermore, the Σ-MSD variant can also always be solved in reasonable time by POMP^d despite the vast search spaces (See Table 4.3). A possible explanation for this surprising capacity to explore vast spaces of size larger than 10^{440} is that the several backbones in each problem are sufficiently similar to define correlated regions of low energies that enable both quick identification of optimal sequences and fast optimality proof. To check if this intuition is true, we computed, for each protein in the benchmark set, the difference ΔE between the average energy of the SSD optimal sequences ($\overline{SSD}$) and the optimal average energy provided by Σ-MSD (see Table 4.6). With an average of respectively 19.2 $kcal.mol^{-1}$ and 10.5 $kcal.mol^{-1}$ for respectively NMR and back-rubbed X-ray structures, these exact differences show that the Σ-MSD sequences have higher energies than the SSD sequences: there is a non negligible frustration generated by trying to fit all backbones together. This frustration is also more important for NMR structures than X-ray structures (p-value = 0.03, Wilcoxon rank sum test) indicating that the back-rubbed structures are more compatible with each other energy-wise than the NMR structures.

Table 4.5: Comparison of the CPU-times (in seconds) for iCFN and POMP^d for solving min-MSD and for POMP^d to solve the corresponding Σ-MSD.

redesigned position	iCFN min-MSD	POMP^d min-MSD	speedup	POMP^d Σ-MSD
26	445.4	25.7	17.3	55.4
28	594.9	32.7	18.1	99.9
98	640.3	22.7	28.2	89.6
100	719.8	29.5	24.4	105.1

We also converted our benchmarking problems to a format suitable for iCFN min-MSD algorithm. After 65 hours of computing, none of the full-redesign min-MSD problems could be solved by iCFN. This was even the case for the smallest protein of our dataset (PDB id: 1pga) which is solved by POMP^d in less than 20 minutes. We therefore prepared several design problems with increasingly smaller search spaces by decreasing the number of mutable amino acid residues, leaving non-mutable residues as flexible. With a number of mutable residues reduced to 5, iCFN was still unable to provide a solution after 24 hours. It's only after fixing all non-mutable residues in a rigid position that iCFN could finally produce a solution in 247 seconds. POMP^d solves this problem in 14.59 seconds.

Table 4.6: Difference in energy for each protein in the benchmark between the average of all SSD optimal sequences and the energy of the optimal Σ-MSD sequence (kcal/mol).

NMR PDB	Σ-MSD-$\overline{SSD}$	X-ray PDB	Σ-MSD-$\overline{SSD}$
5vso	16.0	1hyp	12.7
5l7b	10.4	1hoe	14.8
5mmc	14.3	1mjc	8.9
1gb1	11.6	1pga	12.8
5t8a	5.6	2b8i	13.3
2l5t	25.4	4y2k	5.5
2n6r	11.7	1f94	9.2
1bmw	44.9	1who	17.5
5ix5	21.7	1tud	4.2
5x0s	30.3	1yu5	6.3
6ews	14.8	1bxy	8.6
6fwn	31.0	1ctf	10.1
6qf8	16.9	1guu	8.0
6jlt	18.7	1wvn	20.4
6hkc	27.7	1ucs	9.9
Mean	20.1	Mean	10.8

4.3.2 Sequence enumeration for min-MSD and Σ-MSD

In addition to the optimal sequence, Pomp^d can provide an exhaustive list of sub-optimal sequences within a given energy threshold of the MSD optimum. In order to characterize the energy landscape of the min and Σ-MSD approaches, we enumerated all sequences within a 1 $kcal.mol^{-1}$ of the optimum for the largest protein of our dataset (96 amino acid residues) whose structure has been solved by both NMR (1bmw) and X-ray crystallography (1who). As expected, Σ-MSD enumerations are computationally more costly than min-MSD enumerations (Table 4.7).

Table 4.7: Number of enumerated sequences and CPU-time taken for the enumeration for 1who and 1bmw

	min-MSD		Σ -MSD	
	# of seq.	CPU-time	# of seq.	CPU-time
1bmw	131,616	2' 50"	94,522	43'30"
1who	56,790	2'32"	143,457	67'16"

Different important features of the fitness landscape of SSD problems have already been studied in [155]. We used some of these features to analyze the landscapes of min-MSD and Σ -MSD. The distribution of the Hamming distances to

the optimal sequence (number of substitutions compared to the optimum) shows a similar uni-modal distribution for both methods (Figure 4.5).

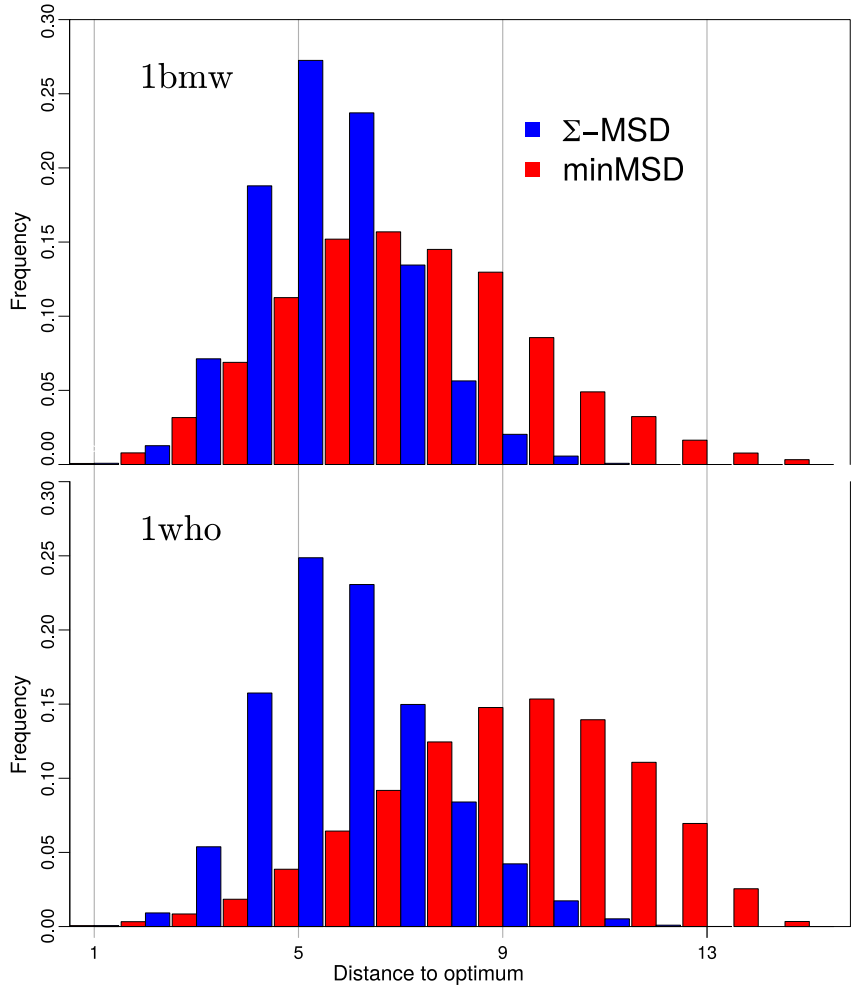

Figure 4.5: Distribution of Hamming distances to GMEC for 1bmw (top) and 1who (bottom). min-MSD is shown in red and Σ-MSD in blue.

However, Σ-MSD shows a narrower distribution, with more solutions close to the optimum (mode at distance 5 of the optimum instead of 7 and 10 respectively for 1bmw and 1who in min-MSD). These results are consistent with the *nsr* and *nnsr* computed for all enumerated sequences (average values shown in Table 4.8).

We also computed the local optima network defined by the enumerated sequences and a neighborhood at a Hamming distance of 1 (See Figure 4.6). For both proteins, the networks for the Σ-MSD landscapes are much more densely connected than the min-MSD networks. In min-MSD, the basin of the global optimum is often disconnected from most of the other basins. Instead, the Σ-MSD landscapes show far less wider basins which can be reached by all or a large fraction of the other basins. This may explain why, despite the frustration generated by the requirement

Table 4.8: Average *nsr* and *nssr* over all enumerated sequences.

	nsr(%)		*nssr*(%)	
PBD ID	min-MSD	Σ-MSD	min-MSD	Σ-MSD
1who	62.9%	67.9%	74.7%	80.3%
1bmw	55.6%	79.5%	63.2%	84.3%

of fitting several backbones, Σ-MSD are easier to solve given their size: they more clearly identify the globally optimal sequence.

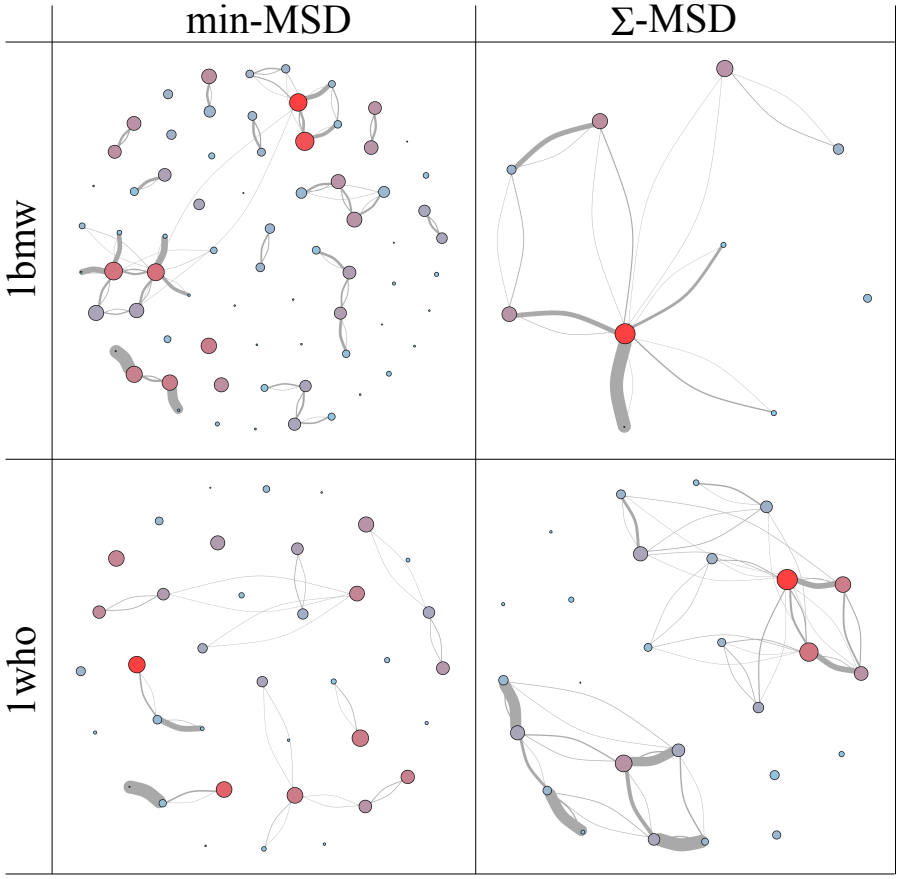

Figure 4.6: 2D views of the local optima networks of 1bmw and 1who for min-MSD and Σ-MSD. The size of a node is proportional to the log size of the attraction basin of the local minima. The energy of the local minima is represented as a color gradient from blue (high energy) to red (low energy). Edge thickness is proportional to the probability of escaping a basin to another basin assuming that the probability to go from a solution to any of its neighbors is uniform.

From a biological perspective, the Σ-MSD fitness landscapes seem more relevant. Considering that evolution occurs by random mutations, one can interpret these

networks as an abstract representation of the possible mutational paths that can be explored by evolution. The densely connected Σ-MSD local optima networks allow random mutations to easily escape local minima. By capturing the natural flexibility of proteins in a more realistic manner, Σ-MSD leads to more natural fitness landscapes.

4.4 Conclusions

We have shown that multistate protein design problems can be intrinsically much harder than the usual NP-complete Single State Protein Design problem [81]. This additional complexity can be precisely pinned down to the introduction of negative states. While negative states are crucial when the design target is to generate specificity, when the aim is just to stabilize an ensemble of backbones, or to design conformational switches, positive states suffice. The positive MSD problem is therefore a soft spot of multistate protein design, offering the ability to capture some of the flexibility of protein backbones while remaining "only" NP-complete, just as SSD.

To exploit this situation, we designed efficient reductions of the optimization problem defined by positive MSD problems to the generic discrete optimization framework of Cost Function Networks [140], a framework introduced in Artificial Intelligence that has already shown its efficiency on SSD problems [119, 85]. On a mixture of NMR and back-rubbed X-ray structures, POMP^d shows that the average energy criteria is clearly superior to the MSA approach in terms of native sequence recovery. In terms of efficiency, it also outperforms a very recent guaranteed multistate algorithm such as iCFN [142], which is also restricted to the simple min-MSD (or MSA) problem. In our knowledge, this is the first time that it is possible to access guaranteed optimal average energy full multistate redesigns of proteins of size close to 100 amino acids, defining search space of size larger than 10^{500}. Because it just relies on a reduction to CFN, this approach also inherits all the capabilities of CFN solvers such as toulbar2, including the ability to exhaustively enumerate sequences within a threshold of the optimum to directly produce a sequence library.

Chapter 5

Customizing Pomp^d with constraints

Contents

In this chapter we present additional features that were implemented during this thesis. These additional features represent different functionalities that were introduced in the previously developed algorithm to customize it for specific applications.

5.1 Hpatch

With the objective to perform a specific biological function, proteins must adopt a stable folded conformational state and be soluble and functional in water for most of them. Protein solubility is an important physicochemical property which is related to other functional properties and which can be influenced by a number of environmental and internal factors. While the environmental factors may include pH, ionic strength, temperature, and the presence of various solvent additives, the internal factors that mostly influence protein solubility are defined by the physicochemical properties of the amino acids present at the protein surface [156]. It has been shown that protein solubility is indeed determined by the amount of exposed hydrophobic surface area in the protein folded state [157, 158]. In 2003, it has been demonstrated that the rate of aggregation of proteins and peptides increases as the amount of exposed hydrophobic surface area increases [159]. Therefore, with protein solubility being strongly influenced by exposed hydrophobic surface area, computational protein design tools must consider protein's surface hydrophobicity when designing new sequences. There are two key components that are required for any computational protein design program: an energy function that accurately reflects protein stability and a reliable search method that identifies a sequence with a conformation of optimal stability. Jaramillo and co-workers showed that in protein design procedures, the free-energy function tend to select native-like protein

sequence fragments for the design of the proteins core as opposed to their surface [160]. Since then, the Rosetta molecular modeling software has implemented a new scoring term, called hpatch. This term penalizes the formation of hydrophobic patches on the surface of designed proteins [161]. Pomp^d can only use decomposable functions, and the addition of this scoring term within Pomp^d was not possible due to its non-pairwise-decomposability. To overcome this limitation, we have exploited the fact that a CFN model can be customized by adding constraints in the form of new cost functions. In this respect, we have implemented in Pomp^d an additional feature also called hpatch. The hpatch option allows Pomp^d to control and prevent creation of hydrophobic patches at the protein surface. If the option is used, Pomp^d will disallow neighboring hydrophobic residues at the protein surface. The algorithm 2 describes the hpatch procedure implemented in Pomp^d. Firstly, all residues of the protein are mutated to leucine. Exposed surface residues are then predicted with the pyrosetta software package. For each residue, the list of all its exposed neighbors is computed. Finally, neighboring hydrophobic pairs identified at the protein surface are forbidden with CFN constraints. The solvent exposure of residues is quantified by its relative Solvent Accessible Surface Area (SASA). The relative SASA of a residue is computed by normalizing its corresponding SASA by its reference Gly-X-Gly tripeptide value [162]. If the relative SASA value is greater than 0.5, the residue is considered exposed. In the wild type structure, it may happen that residues with bulky side chains cover some other residues and prevent them from being identified as exposed. For this reason the whole protein sequence is mutated to leucine prior to relative SASA calculation. Two residues are considered neighbors if there exist a binary cost function involving them which has at least one non-zero value. In this work we consider A, V, I, L, M, F, P and W as hydrophobic residues.

5.2 Weight attribution

In Pomp^d, by default, all states equally contribute to the total energy. By working on different applications and evaluations, we noticed that one may want to control the contribution of each state to the total energy. Weight attribution can be very useful in the case of enzyme design, for example. When applying multistate design to enzymes in free state and in complex with a given ligand, additional weight can be added on enzyme/complex conformational state in order to ensure that the enzymatic activity (binding) is preserved during the CPD procedure. Accordingly, we have added a feature which allows this by weighting the contribution of each state in the multi-state design problem. Concretely, the costs in all cost functions applied on a given state s are multiplied by a weight w_s.

Algorithm 2 Hpatch in POMP^d.

```
1: Inputs: C : protein conformation, cfn : cost function network
2: exp_residues = ∅
3: exp_neighbors = ∅
4: C = mutate_all_residues(C,"Leu")
5: for res in residues(C) do
6:    if is_exposed(res) then
7:       exp_residues = exp_residues ⋃ res
8:    end if
9: end for
10: for res in residues(C) do
11:    current_res_exp_neighbors = ∅
12:    for all neighbor = neighbors(res) do
13:       if neighbor ∈ exp_residues then
14:          current_res_exp_neighbors = current_res_exp_neighbors ⋃ neighbor
15:       end if
16:    end for
17:    exp_neighbors = exp_neighbors ⋃ current_res_exp_neighbors
18: end for
19: for res in residues(C) do
20:    if res ∈ exp_residues then
21:       for all neighbor ∈ exp_neighbors[res] do
22:          add_constraints(cfn,res,neighbor)
23:       end for
24:    end if
25: end for
```

5.3 Diversity constraints

In many applications of constraint programming, it is often impossible to capture all the relevant information in one numerical criterion. In this case, it is useful to produce a set of high-quality, yet diverse, solutions. In CPD, as in many other real problems, the actual potential energy of the protein, that the algorithm aims at optimizing, can only be approximated. This makes the protein design process unreliable, as a typical workflow includes the expensive production and experimental testing of a library containing several proteins. Ideally, this library should be a set of diverse and low energy solutions, with the hope that a sufficient sequence diversity will improve the likelihood that a functional protein is found. The sequence diversity can be quantified by calculating the Hamming distance between selected sequences. As an alternative, a "bio-chemical" diversity can also be estimated with the use of existing protein dissimilarity matrices. Because of their important applications, protein sequences can also be subject to patents. In such a case, a newly designed sequence must absolutely satisfy a certain Hamming distance constraint with respect to existing patented sequences. In this respect, I took part in a research project whose objective was to consider the general problem of producing a diverse set of high-quality solutions of a given Weighted Constraint Satisfaction Problem, with guarantees both on solution quality and diversity. This new feature, implemented in Pomp^d by Manon Ruffini, makes the program able to generate large sets of diverse and high quality (low energy) optimized sequences for a given CPD problem. By evaluating *in silico* the efficiency of this method on real protein design problems, we have observed that sufficiently large diversity requirements do improve the quality of sequence libraries when native proteins are fully redesigned [163].

Part III

Applications: optimized enzymes and new nanobodies

Chapter 6

Thermal stability and activity of GH-11 xylanases

Contents

6.1 Context

6.1.1 GH11 Xylanases

6.1.1.1 General and Biochemical Properties

Xylanases are enzymes degrading polysaccharides that are mainly composed of xylans. Xylans represent a group of hemicelluloses that is one of the most abundant biopolymers on Earth. Commonly known as xylanases, endo-1,4-β-xylanases catalyse the hydrolysis of the β-1,4 glycosidic linkage of the xylane backbone in heteroxylans (constituting the lignocellulosic plant cell wall) and produce mainly xylobiose and, to a lesser extent, short xylo-oligosaccharides (XOS) [**dumon2012progress**].

Xylanases are widely used in industrial processes. The first industrial applications of xylanases were in pulp and paper industry, food industry and animal feed

[164]. However, with the need of renewable and sustainable sources of fuels and chemicals that could help reduce pollution and reduce global warming linked to industrial activities, the importance of xylanases in bio-refinery processes has been rapidly increasing [165]. Xylanases are classified within the Carbohydrate Active Enzymes database (CAZy - common classification system of glycoside hydrolases organized in different families according to sequence similarities) [166] in the Glycoside Hydrolase (GH) families 5, 8, 10, 11 and 43. Xylanases produced by bacteria and fungi mostly belong to GH10 and GH11 families and are the ones that have been widely studied. In this thesis, we are interested in xylanases from the GH11 family. The main feature that differentiates the GH11 family of xylanases from others is that the GH11 family gathers all xylanases capable of exclusively hydrolysing endo β-1,4 bonds. This family of enzymes is also characterized by a catalytic mechanism which leads to the retention of the configuration of the anomeric carbon at the cleavage point.

6.1.1.2 Overall structure

GH11 xylanases are defined by a low molecular weight, generally ranging between 20 and 30 kDa. The first three dimensional X-ray crystallographic structures were available in 1993 and allowed the first detailed studies on these enzymes [167, 168]. To date, there are 133 PDB entries in the Protein Data bank that correspond to GH11 xylanases of bacterial and fungal origins. The three dimensional structure of xylanases has been compared with the shape of a partially closed right hand and different elements, such as fingers, thumb and palm, have been named accordingly. The fold has a β-jelly roll architecture, which is highly conserved in all GH11 xylanases [165] (Figure 6.1). It is composed of 2 anti-parallel β-sheets (β-sheets A and β-sheets B) which form the fingers of the hand and a unique α-helix packed under the β-sheet B, which forms the palm together with a part of the twisted β-sheet B. As the secondary structure elements dominate within the overall structure of the enzyme, loops that connect these elements are quite short. There are however two important exceptions called the “thumb” and the “cord” that are 10 to 12 residues long.

6.1.1.3 Active site and catalytic dyad mechanism

The active site of GH11 xylanases is a deep cleft where substrate recognition and binding occur thanks to the presence of aromatic residues that are tightly packed together in order to form a hydrophobic cleft and be able to fit the substrate. A catalytic dyad is located in the middle of the cleft. The active site is composed of at least four xylose-binding subsites, each of them accommodating one xylose moiety from suitable xylan substrates. The enzymatic activity is mostly established by the organization of the active site into subsites. A subsite defines the region of the active site that is able to accommodate a single unit of the substrate. The subsites are assigned with a positive or negative number depending on whether they bind the

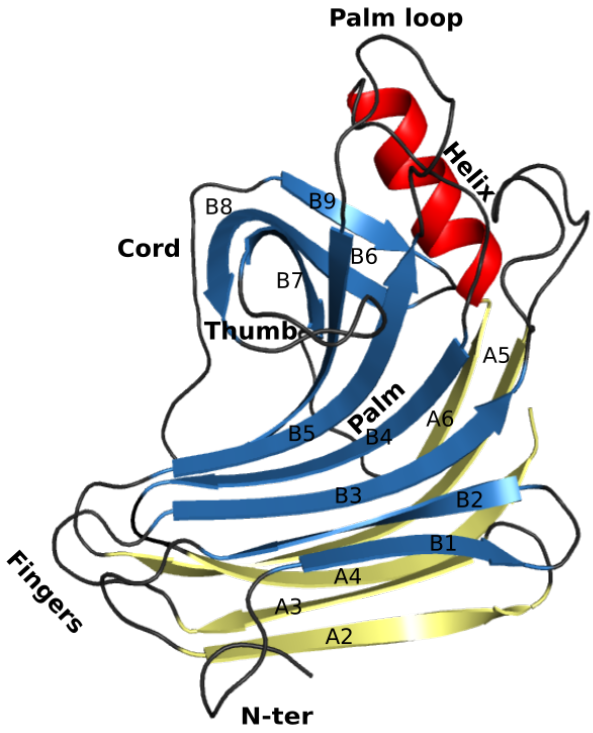

Figure 6.1: Typical three dimensional structure of a GH11 xylanase showing the jellyroll fold with a visual representation of right-hand analogy regions (fingers, palm, thumb, cord, helix). β-sheets A are shown in yellow while β-sheets B are shown in blue and a unique α-helix in red. Crystal structure of *Np*Xyn11A (PDB ID 2C1F) is taken as example.

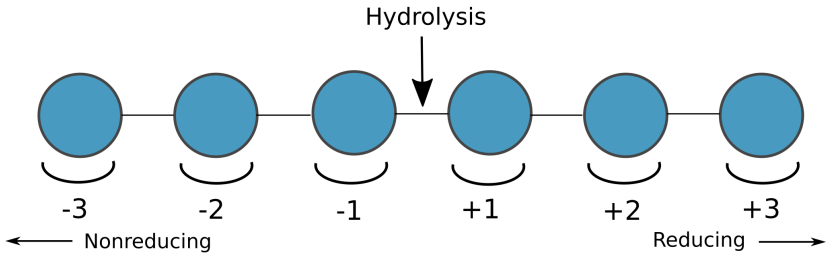

Figure 6.2: Schematic representation of substrate binding subsites in glycosidases. Circles represent xylose moieties linked to each other by β-1-4 bonds.

non-reducing (glycone) or reducing (aglycone) moiety of the substrate [169]. The hydrolysis of glycosidic bonds occurs between the -1 and +1 subsites (Figure 6.2). GH11 xylanases are usually limited to 3 subsites where amino acids such as tyrosine and tryptophan establish π-hydrogen interactions with the pyranose ring of xylose moities. Polar amino acids form hydrogen bonds with the hydroxyl groups of xylose moities [170].

X-ray crystallographic studies are a good way of investigating enzyme-substrate interactions and have allowed a better understanding of substrate binding in the active site of the GH11 family of enzymes. By studying crystallographic structures of GH11 xylanases in complex with xylobiose and xylotriose it has been suggested that these enzymes can have up to six carbohydrate-binding subsites [171]. Binding of oligosaccharide substrates at subsites -3 through +1 was also shown in another crystallographic study using inactive variants [172]. Another study, using six xylose subunits, provided unambiguous structural evidence that the active site of one particular GH11 xylanase, has six possible sugar-binding subsites from -3 to +3 [173]. Within the active site, the catalysis proceeds with retention of stereochemistry at the anomeric carbon of the nonreducing (glycone) moiety of the product. Hydrolyis involves two catalytic carboxylic amino acids (usually two glutamic acids) which act as acid/base and nucleophile residues. The reaction occurs via a two-step mechanism which involves the formation of a covalent glycosyl-enzyme intermediate. This intermediate, formed during the first step of the mechanism, displays an inverted anomeric configuration which is further inverted one more time during the second step of the mechanims to lead to the final configuration identical to the ground state (Figure 6.3). One glutamic acid acts as a general acid/base catalyst and has a pKa value that is necessarily high (≈ 7) while the other glutamic acid plays the role of a catalytic nucleophile and has a low pKA value (<5). Therefore, in order for catalysis to take place, one glutamic acid must be protonated and the other negatively charged.

6.1.1.4 Highly conserved regions

To date, there are more than a thousand GH11 xylanase sequences available in the CAZy database. Some studies have analysed mature sequences and carried

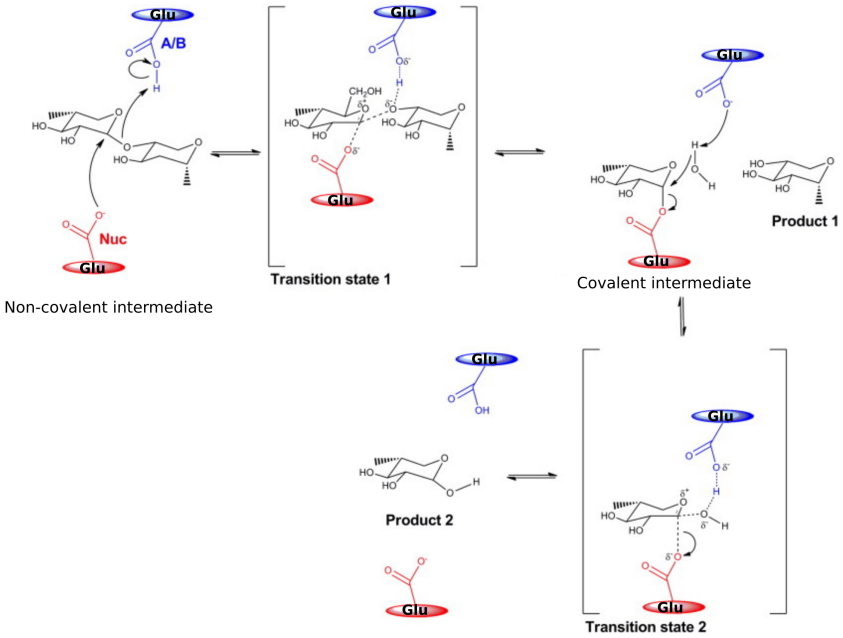

Figure 6.3: Catalytic mechanism of GH11 xylanses. Catalytic residues are represented in red and blue. Nuc is the nucleophile catalytic residue and A/B is the acid/base catalytic residue. Adapted from [165].

out a comprehensive analysis in order to highlight the sequence homology observed within the GH11 xylanase family [174, 165]. The conclusions made in these studies, conducted on a total of 82 and 164 sequences respectively, underline the fact that the active site of GH11 family of xylanases is highly conserved and thus explains their restricted substrate specificity. Important conserved residues are also found in the thumb and cord regions. With the increasing number of sequences that have become available over the years following these respective studies, we have decided to update the sequence homology analysis of this family of enzymes. To do so, we selected more than 1000 sequences from the CAZy database and ran a blastp search on each of them [175]. Using an E value threshold of 10, a final non-redundant set of 510 hits was collected. These 510 sequences have then been subjected to a multiple sequence alignment with MAFFT [176]. Figure 6.4 shows the sequence entropy of the selected 510 sequences, calculated with Sequester [177] and mapped on the 3D structure of the GH11 xylanase from *Neocallimastix patriciarum*. Conserved residues are shown in red and less conserved residues in blue. The outcome of this analysis confirms the results obtained previously on a much smaller number of sequences [174, 165]. Many of the conserved residues are found in buried, solvent inaccessible regions, while highly conserved residues are also found in the thumb loop, palm loop and the cord region. This suggests that these regions might play an important role in xylanases function.

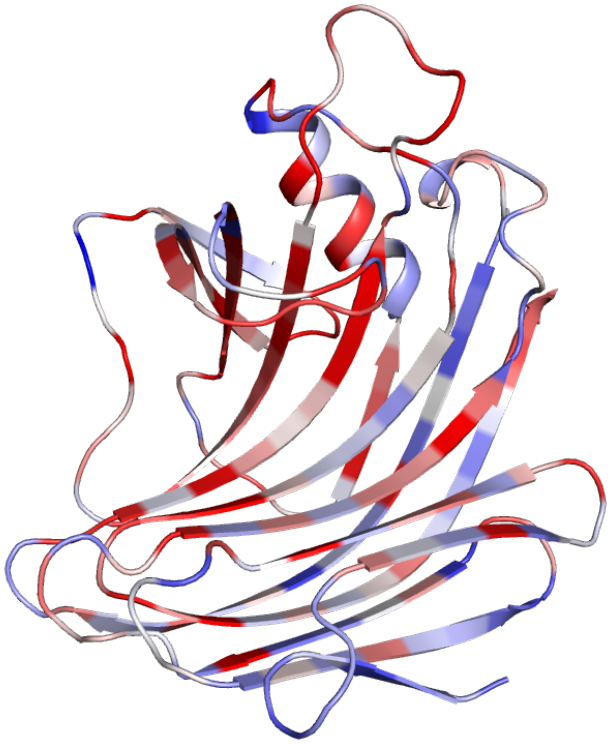

Figure 6.4: Cartoon representation of the GH11 xylanase from *Neocallimastix patriciarum* indicating (in red) highly conserved regions and (in blue) less conserved regions.

6.1.1.5 Dynamics

The importance of protein flexibility and its role in protein function has already been discussed in Chapter 1. In the specific case of enzymes, movements representing different conformational changes have been demonstrated. They indicate that enzymes are highly dynamical macromolecules whose structures are in close relationship with their dynamics and catalysis [178]. Many enzymes undergo large conformational changes which can play an essential role in promoting substrate binding and catalysis. Therefore, investigating enzymes motions is essential for having more comprehensive knowledge of their structure-dynamics-activity relationship. In this regard, Molecular Dynamics simulations are often used in order to explore possible enzyme conformations and the respective transitions from one conformation to another. Combining different multi-scale modeling methods has proven to be a valuable strategy for analyzing important molecular motions and deciphering the molecular basis underlying particular biochemical properties. Several Molecular Dynamics simulations have been performed on GH11 xylanases, and different investigations have been undertaken. One study showed that the increase in temperature induces a significant change in the dynamics of the thumb region [179]. This study revealed that both the thumb loop and the palm regions separate at higher temperatures, going from a close conformation at lower temperature to an open one, thus facilitating the substrate access to the active-site pocket. Similar works showed that such conformational change was not only dependent on the temperature, but also on the presence of the substrate [180, 181]. Molecular Dynamics simulations were also used to analyse differences in thermotolerance between twelve members of GH11 xylanases, including thermophilic and mesophilic ones [182]. Intramolecular hydrogen bonds and salt bridges were analysed and revealed to be an important factor, responsible for different thermostabilites between two structurally similar GH11 xylanses [183]. Simulations of the free-enzyme, non-covalently bound and covalently bound xylobiose intermediate showed that covalently bound substrate induces a change in the structural conformation of the receptor and demonstrated a high flexibility of the thumb region in the non-covalent complex compared to the covalent complex [184]. When investigating the structural basis of catalysis and other biochemical properties in enzymes, it is often required to characterize functional molecular motions and understand how they contribute to enzymes functions. The Molecular Dynamics simulations presented in studies previously mentioned try to answer to this major question. Nonetheless, even though some useful information about conformational changes and side chain movements can be seen over the nanosecond range, the timescale of these simulations (maximum 45ns) is quite far from the time needed for biological events to occur.

6.1.1.6 Thermostability of GH11 xylanases

Thermostable enzymes are very important as they can be used at high temperatures and are therefore suitable for different industrial applications. Thermostable enzymes are usually created using rational design by site-directed mutagenesis [185]

or directed evolution by random mutagenesis on mesophilic enzymes [186]. Directed evolution does not require any prior knowledge of an enzyme 3D structure to be conducted unlike rational design, which is based on the introduction of site-directed mutations at specific locations in an enzyme amino acid sequence, while taking into consideration the impact a given mutation might have on the structure of an enzyme. Thermostabilization of enzymes at the experimental level includes numerous demanding steps, such as evaluation of residual activity after heat treatment of the mutants, determination of the melting temperature analysis (T_m) or optimal temperature (T_{opt}). The GH11 family of xylanases contains mesophilic, thermophilic but also hyperthermophilic enzymes. The optimal temperature of xylanases generally ranges from 35°C to 85°C [165]. Among them, the thermophilic GH11 xylanases have an optimal temperature ranging from 62°C to 85°C. Some of these thermophilic xylanases are natural thermophilic or hyperthermophilic enzymes from organisms such as *Thermopolyspora flexuosa* and *Chaetomium thermophilium* that have T_{opt} of 80°C [187]. Many engineering studies have been conducted on GH11 mesophilic xylanases in order to turn them into thermophilic ones. Different factors responsible for the thermostability of these enzymes have been exploited but they seem to be quite unique to a given enzyme. However, some general features have been identified as important for thermostability in all thermostable enzymes such as the presence of more hydrogen bonds, disulfide bridges or salt bridges. Important regions have also been identified (N-ter, C-ter and α-helix) and are considered as "hot spots" where unfolding preferentially occurs [182]. Despite the fact that not all features conferring thermostability are fully understood, many studies focused on engineering the thermostability of these enzymes. Interactions or structural motifs specific to thermophilic xylanases have been transferred to mesophilic enzymes [188, 189, 190]. Hot spot regions have been stabilized, and disulfide bridges have been introduced [191, 192, 193]. Some studies also focused on rational design of glycoside hydrolases based on structural analysis, by linking the N- and C-terminal ends or by optimizing β-turn structures to promote hydrophobic interactions [194]. In this respect, long molecular dynamics simulations also remain a promising strategy to unravel the molecular determinants governing the thermostability of xylanases. Indeed, analysis of MD trajectories can assist in the identification of the regions in a given enzyme 3D structure that are less stable than others and that can be engineered in order to improve enzyme thermal stability.

6.2 Motivations

The three dimensional structure of an enzyme is intrinsically linked to its function. By analyzing the structure of an enzyme, we can gain insights on its role on the enzyme's function. However, enzymes, as all proteins, possess a dynamic nature. Their thermodynamic and kinetic properties define the conformational states they are likely to adopt and the energy necessary to switch between these respective conformational states. Thus, analyzing enzymes structural dynamics has already

proven to be of great importance for understanding the interplay between their structural features and their specific properties. In the context of this project, understanding the structural basis for thermostability or enzymatic activity of GH11 xylanases is of paramount interest for their biological and biotechnological applications. This chapter is devoted to a computational study which aims at investigating the structure-dynamics-activity relationship of GH11 xylanases using long MD simulations. We focus on two different enzymes of the GH11 family of xylanases. The first one is a thermostable mutant of environmentally isolated GH11 xylanase, *Ev*Xyn11^{TS} and is known to be hyperthermostable [195]. The other one is known to be a particularly active GH11 xylanase from *Neocallimastix patriciarum* (*Np*Xyn11A) [171] but mesophilic. This study focuses on a comparative analysis of both structural and dynamics properties that differentiate these two enzymes. More specifically, the main objective is to identify a set of unique characteristics that could explain their respectively high thermostability and activity. To fulfill this objective, 1μs MD simulations of the free-enzymes and the enzymes in complex with xylohexaose were performed at 310K and 340K. MD simulations at 310K allow us to perform a comparative analysis at a temperature where both enzymes are known to be active and stable. MD simulations at 340K allow us to enhance conformational sampling, observe the impact that higher temperature can have on the mesophilic *Np*Xyn11A and analyze the physicochemical properties of thermostable *Ev*Xyn11-^{TS}. Shorter MD simulations were also done at very high temperature (500K) with the objective of comparing the thermal resistance of these two enzymes and eventually observe details of the initial unfolding process.

From different MD trajectories, we have conducted a thorough comparative analysis of the structural properties of these enzymes based on the comparison of different structural and geometrical features which include intramolecular interactions stabilizing their respective structures. An analysis focusing on the structural differences of their active sites was also carried out.

6.3 Materials and Methods

6.3.1 Molecular modeling and molecular dynamics procedures

MD simulations of the ligand-free enzymes and the enzyme-xylohexaose complexes were carried out at three temperatures: 310K, 340K and 500K. These simulations were performed using the AMBER 18 suite using pmemd.CUDA on GPU [196, 197, 198, 199]. The AMBER package was preferred over other MD simulation packages as it includes the state-of-the-art GLYCAM06 force field [200] for an optimal description of carbohydrate molecules and as it also supports the mixed scaling of 1-4 non-bonded electrostatic and van der Waals terms which is required for a correct treatment of 1-4 non-bonded interactions in systems mixing proteins and carbohydrates.

The high-resolution structures of the particularly active GH11 xylanase from *Neocallimastix patriciarum*, *Np*Xyn11A (PDB code: 2C1F) [171], and the ther-

mostable mutant of the environmentally isolated GH11 xylanase, *Ev*Xyn11^{TS} (PDB code: 2VUL) [195], were used as starting models for MD simulations. Xylohexaose (X6) was manually docked in the binding cleft of *Np*Xyn11A and *Ev*Xyn11^{TS}, using as template the X-ray structure of the E177Q catalytic acid/base mutant of the xylanase from *Trichoderma reesei* co-crystallized with xylohexaose (X6) (PDB code: 4HK8)[173].

In both enzyme-xylohexaose complexes, the catalytic acid/base residue, (that is GLU 201 for *Np*Xyn11A and GLU 181 for *Ev*Xyn11^{TS}) was protonated. MD simulations were performed with the AMBER ff14SB [26] and the GLYCAM_06j-1 force fields [200], respectively used for describing the proteins and the xylohexaose substrate. To neutralize the net charge of the simulated systems, an appropriate number of counter-ions was included. Explicit solvation was performed with TIP3P water molecules, using an octahedral box [201] with a minimum distance of 10 Å between the solute and the simulation box edges.

Preparation of simulations consisted of four energy minimization steps (using both steepest descent and conjugate gradient methods), a gradual heating of the respective systems to a target temperature (310K, 340K or 500K depending on the simulation) under constant volume over a period of 100 ps followed by an equilibration of 100 ps under constant pressure (1 bar) and temperature. Harmonic potential restraints of 25kcal/mol/Å^2 were first applied on the solute atoms and then subsequently gradually removed along the equilibration procedure. The simulations productions were carried out in the NPT ensemble for 1 μs at constant temperature of 310K or 340K and for 100 ns at constant temperature of 500K. The temperature and the pressure were controlled by using the Berendsen algorithms [41]. Long-range electrostatic interactions were handled by using the Particle-Mesh Ewald method [202]. A 9 Å cut-off was used for non-bonded interactions. The integration time-step was 2 fs and the SHAKE algorithm was used to constrain the lengths of all covalent bonds involving hydrogen atoms to their equilibrium values [203]. Atomic coordinates for each simulation were saved every 10 ps.

6.3.2 Molecular dynamics trajectory analysis

The CPPTRAJ module implemented in AMBER [204] was utilized for processing all MD trajectories, calculating different structural and geometrical properties as well as for performing dynamic cross correlations and principal component analyses. The "saltbr" plugin within VMD [205] was used to quantify the number of salt bridges formed over the course of the simulation.

The first three residues at the N-ter region as well as the last three residues of the C-ter region were excluded from these analyses, which only considered the remaining 213 amino acid residues for *Np*Xyn11A and 187 residues for *Ev*Xyn11^{TS}. Hence, in this chapter the protein residues have been renumbered accordingly.

Structural and Geometrical Properties

A set of geometrical and topological properties of each enzyme's active site have

been calculated with the CASTp 3.0 web-server [206]. More precisely, the residues composing the respective active site have been identified and the negative volume (the space encompassed by the atoms that form the active site) as well as the surface area of the active site have been calculated. The PLIP web-server was used to investigate the enzyme-xylohexahose interactions [207].

The root mean square deviation (RMSD) of backbone atoms, relative to the starting structure, was calculated for each enzyme (free form or in complex with X6) along each MD simulation. Per-residue B-factors averaged over the entire trajectories were derived from the respective root mean square fluctuations (RMSF) calculated on all backbone atoms. RMSF calculations provide a crude estimation of the average atomic positional fluctuations over the course of a given MD simulation trajectory. Prior to RMSF calculations, the MD snapshots were RMS-fitted onto the average structure to remove all degrees of translations and rotations. The mass-weighted fluctuations of the backbone atoms (C, $C\alpha$, and N) and B-factors for each residue were calculated as follows:

$$RMSF = \sqrt{\frac{1}{nsteps} \sum_{i=1}^{nsteps} \|r_i(t) - \langle r_i \rangle\|^2} \tag{6.1}$$

and

$$B - factor = RMSF^2 (\frac{8}{3})\pi^2 \tag{6.2}$$

where r_i is the position of atom i at time t and $\langle r_i \rangle$ the average position of the atom. For comparative purposes and as the two studied enzymes differ in their amino acid sequence length, the calculated B-factor values have been aligned between the two enzymes by aligning their respective sequences and introducing gaps in regions corresponding to insertions/deletions.

The number of hydrogen bonds (HBs) formed in each MD snapshot between two molecular entities was calculated using the following geometric criteria: the distance from a donor heavy atom D and an acceptor heavy atom A is less than 3 Å and the valence angle between A, a donor hydrogen atom H and D (A-H-D) is greater than 135°. Dynamic and static HBs were determined. As the two enzymes do not have the same amino acid sequence length, the number of static and dynamic hydrogen bonds formed was normalized by the number of amino acid residues in each enzyme. Thus, the results are given for each enzyme as the average number of static or dynamic hydrogen bonds per residue. Static HBs represent the per-residue average number of hydrogen bonds observed during the MD trajectory and weighted by their probability of occurrence. In other words, it is the expectation of hydrogen bonds per residue. Dynamic HBs correspond to the per-residue number of hydrogen bonds observed in at least one snapshot of the MD trajectory. Enzyme intramolecular HBs and enzyme-solvent HBs were determined from 1000 regularly spaced snapshots taken along the 1 μs MD trajectory of each enzyme, carried out at T310K and T340K. Enzyme-substrate HBs were calculated from 1000 snapshots of

the first 100 ns of the respective MD trajectories performed at T310 K and T340K. The number of salt bridges formed over the course of the MD simulations was calculated assuming that a salt bridge can only be formed if the distance measured between the oxygen atoms of acidic residues and the nitrogen atoms of basic residues does not exceed 4 Å.

Dynamic Cross correlation

Dynamic cross-correlation method has been widely used in MD simulation analysis [208] to quantify the correlation coefficients of motions between atoms in molecular structures [209]. In this study, dynamic cross-correlation matrices were calculated using the Cα atomic coordinates to quantify the correlated motions in the studied enzyme's backbone and identify potential domain motions over the course of the respective MD trajectories. Cross correlation elements for Cα atoms of two residues i and j are given by the following equation:

$$C_{ij} = \frac{\langle r_i \cdot r_j \rangle - \langle r_i \rangle \langle r_j \rangle}{\sqrt{[(\langle r_i^2 \rangle - \langle r_i \rangle^2)(\langle r_j^2 \rangle - \langle r_j \rangle^2)]}} \tag{6.3}$$

Highly correlated motions are denoted by $C_{ij} = 1$ while $C_{ij} = -1$ denotes highly anti-correlated motions.

Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction mathematical method commonly used in Molecular Modeling for describing and classifying molecular motions in macromolecules. It is also called Essential Dynamics method [210]. In general data analysis, the objective of a PCA is to apply a lossy compression to an initial collection of points (here, atomic coordinates), and store the points in a way that is observable in few dimensions by loosing as little precision as possible. From a MD simulation trajectory, PCA is applied to reduce the number of dimensions needed to describe protein's dynamics, and extract the largest amplitude protein motions, also called collective motions [211]. A covariance matrix is first constructed from the atomic coordinates of a selected set of atoms over the course of a given MD trajectory. The diagonalization of the covariance matrix results in a complete set of eigenvectors or principal components (directions of the atomic motions in the conformational space) with corresponding eigenvalues (amplitude of the respective atomic motions).

In this study, PCA was performed on all MD trajectories simulated at T310K and T340K. Only Cα atoms were considered for the analysis. To remove global proteins rotations and translations, the snapshots of each trajectory were aligned to their calculated average coordinates.

For each MD simulation, the Kullback-Leibler divergence (KLD) between the principal component histograms corresponding to the first and second half of each

MD simulation was calculated using CPPTRAJ to assess the system convergence [212].

The time dependent KLD is calculated as follows:

$$KLD_{(t)} = \sum_i (P_{(t,i)} \ln \frac{P_{(t,i)}}{Q_{(t,i)}} \tag{6.4}$$

where $P_{(t,i)}$ and $Q_{(t,i)}$ represent different probability distributions of atomic coordinates, i represents a histogram bin index (here 400 bins were used) and t represents the time at which the histogram is being constructed.

Free-energy landscape

Once the collective motions are identified with PCA, the Free-Energy Landscape (FEL) of a protein can be derived from a probability density function. MD simulation serves as a sampling method that allows the exploration of conformations near the native state structure [213, 214, 215]. The FEL was here constructed along the first two Principal Components (PCs) using the following equation:

$$G_a = -kT \ln (\frac{P(q_a)}{P_{max}(q)}) \tag{6.5}$$

where k is the Boltzmann constant, T is the temperature of the simulation, $P(q_a)$ is the probability of a state a, and $P_{max}(q)$ is the probability of the most probable state. Considering two PCs, i and j, the free-energy landscapes were obtained from the joint probability distributions $P(i, j)$ of the system.

6.4 Results and Discussion

6.4.1 Structural and biochemical properties

*Np*Xyn11A is 26 residues longer than *Ev*Xyn11^{TS}. Their respective 3D structures are very similar and as in all GH11 enzyme members, it can be compared with the shape of a partially closed right hand. Different elements such as fingers, thumb and palm, have been named accordingly. The residues that compose these regions have been identified in each enzyme and their corresponding index are given in Table 6.1 and shown in Figure 6.5. A feature worth mentioning is the presence of a disulfide bridge at the N-terminal region of *Ev*Xyn11^{TS}, absent in *Np*Xyn11A. This disulfide bridge is also present in the wild type thermostable *Ev*Xyn11 and is thus not the unique responsible of *Ev*Xyn11^{TS} hyperthermostability.

As it can be observed in Figure 6.5,the 3D structures are dominated by β-sheets and one α-helix. Loops that connect these secondary structure elements may play important roles on the stability and function of these enzymes by regulating their structural dynamics. It is widely accepted by different studies which focused on the comparison of the three-dimensional structures of thermophilic and mesophilic enzymes that thermostable enzymes tend to be have more compact structure with shorter loops and a more densely packed hydrophobic core [216]. When comparing

Regions	*Np*Xyn11A	*Ev*Xyn11^{TS}
Fingers	1-59;72-92;197-213	1-49;59-74;175-187
Thumb	134-164	113-143
Palm; Palm loop	93-97,110-118,192-196 ; 98-109	75-79,85-91,168-174 ; 80-84
Helix; Helix loop	173-183 ; 184-191	151-162 ; 163-167
Cord	119-126	92-103
Loop B3-A5	60-72	50-58

Table 6.1: Definitions of structural regions of the *Np*Xyn11A and *Ev*Xyn11^{TS} enzymes by residue number

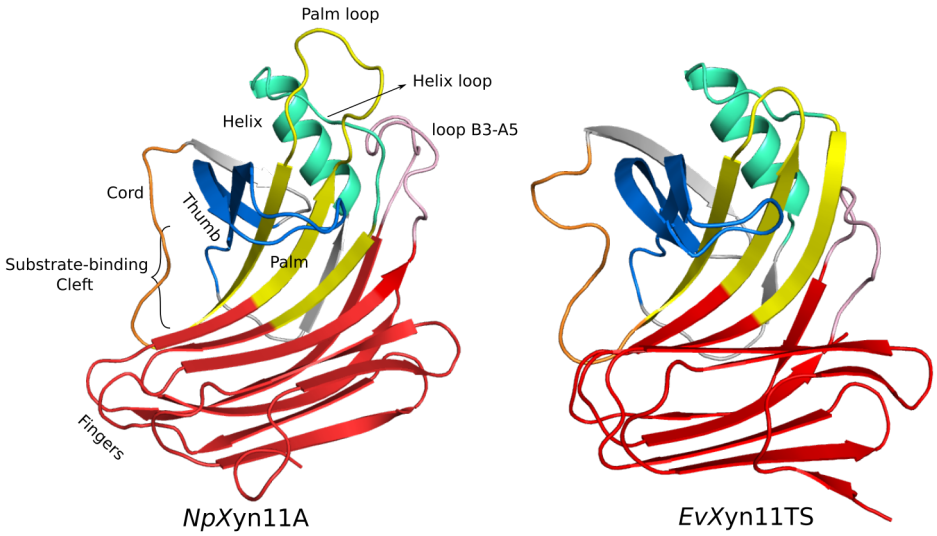

Figure 6.5: 3D structures of *Np*Xyn11A (left) and *Ev*Xyn11^{TS} (right). Visual representation of the right-hand analogy regions are indicated: the fingers in red, palm region in yellow, the thumb region in blue, the cord in orange, the helical region in cyan green and the loop B3-A5 in pink. The substrate binding cleft is also shown.

the 3D structures of *Np*Xyn11A and *Ev*Xyn11^{TS} (Figure 6.5), we can see that loops present major structural differences between the two enzymes, being generally longer in *Np*Xyn11A. More specifically, the length of the loops in the thumb, palm and helix regions vary the most. In *Np*Xyn11A, the thumb loop and the helix loop are 9 residues long, whereas they are 6 residues long in *Ev*Xyn11^{TS}. The loop of the palm region varies the most with 12 residues in *Np*Xyn11A and only 5 in *Ev*-Xyn11^{TS}. This palm loop is particularly long in *Np*Xyn11A compared to any other GH11 xylanases. The loop between the β-sheet B3 and A5 (loop colored in pink in the figure 6.5) is also much longer in *Np*Xyn11A, with the length of 13 residues versus 9 in *Ev*Xyn11^{TS}.

Specific activity on wheat arabinoxylan (WAX) and melting temperature (T_m) of *Np*Xyn11A and *Ev*Xyn11^{TS} were previously measured in the TBI laboratory.

*Np*Xyn11A has an average specific activity of 3916.94 IU.mg^{-1} (standard deviation of 43.22) while that of *Ev*Xyn11^{TS} is 1012.0 IU.mg^{-1} (standard deviation of 12.39). T_m values are 55.7°C and 101.1°C respectively.

6.4.2 System stability and convergence

PCA analysis was based on the first two principal components which explain around 30% of the total protein motions. For each frame, the projection of the transformed coordinates along all eigenvectors (PCs) was calculated, and each eigenvector's contribution was derived from its respective eigenvalue.

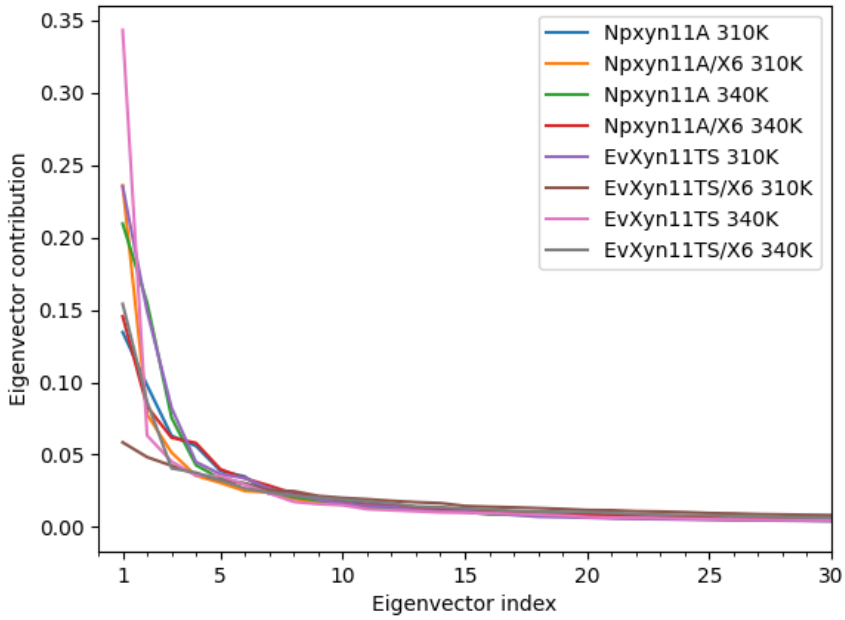

Figure 6.6: Eigenvector contribution as a function of the eigenvector index. Only the first 30 eigenvectors are shown.

Figure 6.6 shows that the contribution of PCs quickly decays to 0. Time dependent KL divergence analysis was performed on the first two PCs (Figure 6.7). In the free-enzyme form of *Np*Xyn11A at 340K, the simulation probably converges but the shape of the first PC curve does not allow us to conclude with certainty. In all other simulations, we observe a flat curve after a few hundred nanoseconds which indicates convergence.

MD simulations were further used to compare and investigate mesophilic *Np*Xyn11A and thermophilic *Ev*Xyn11^{TS}. The stability of the studied systems at different temperatures was determined by monitoring the backbone root mean square deviation (RMSD) as a function of time. This was firstly done for simulations performed at very high temperature (500K) for 100 ns in order to compare the re-

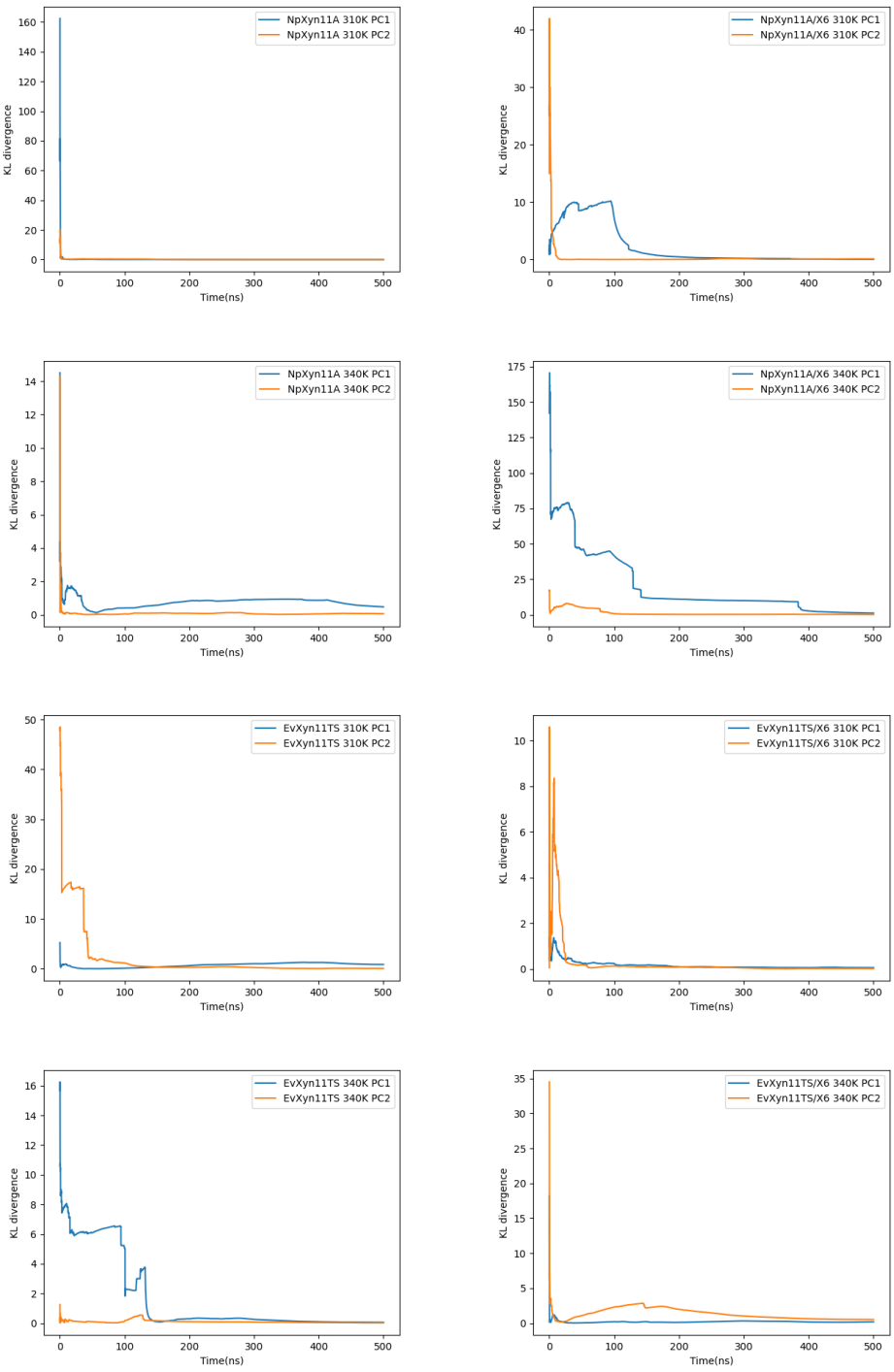

Figure 6.7: KL divergence on the first two PCs as a function of time.

sistance of these two enzymes with respect to thermal denaturation and eventually observe their unfolding (see Figure 6.8). In the corresponding RMSD time series,

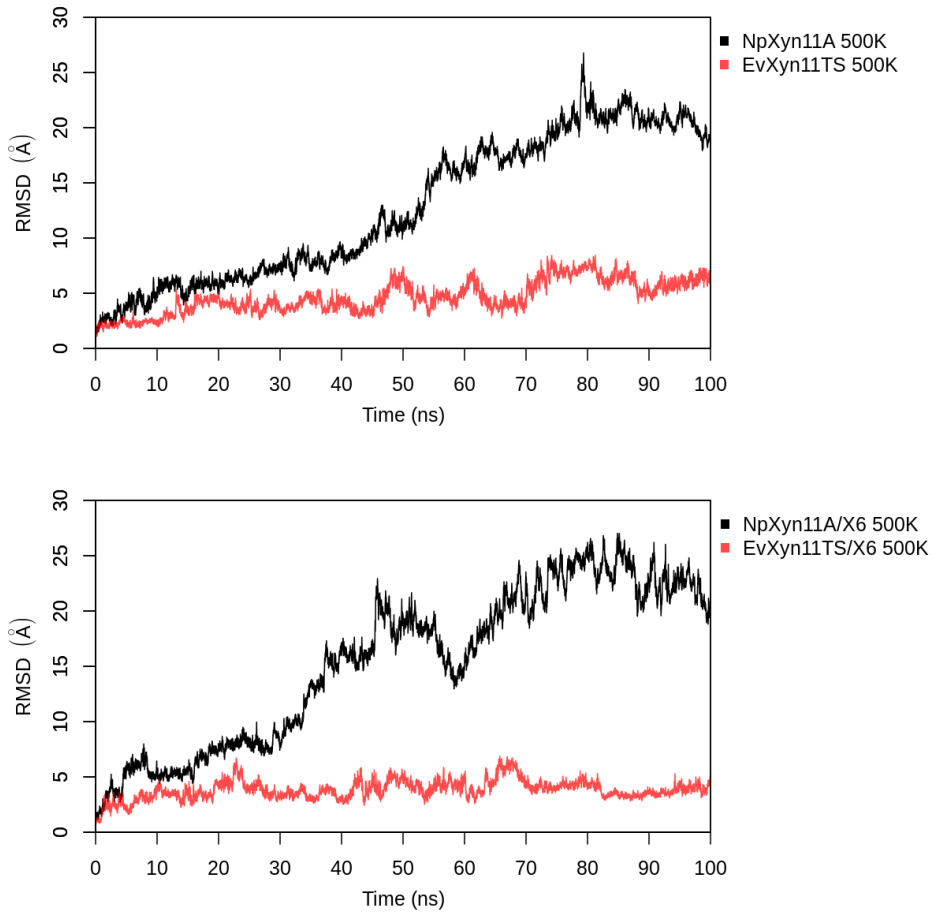

Figure 6.8: Backbone Root Mean Square Deviation of systems in free-enzyme forms and enzyme-substrate complexes at 500K.

we can clearly see a drastic increase in the RMSD of *Np*Xyn11A from the beginning of the simulation in the free-enzyme and the enzyme-substrate complex. In opposition to *Np*Xyn11A, *Ev*Xyn11^{TS} presents a relatively stable RMSD at such high temperature, especially in the case of the enzyme-substrate complex. Figure 6.9 displays the secondary structure propensities over 100 ns of simulation time. It confirms the greater thermostability of *Ev*Xyn11^{TS} at high temperature and shows the unfolding of *Np*Xyn11A, which is mainly observed in the the N-terminal, α-helix and thumb regions. This analysis reveals a striking difference in structural stability between these two enzymes with respect to increasing temperature, in accordance with previous experimental results. It also confirms the hyperthermostability of *Ev*Xyn11^{TS} in comparison to the mesophilic *Np*Xyn11A.

RMSD analyses were further done on the free-enzymes and enzymes-substrate complexes at 310K and 340K over 1 μ of simulation time (see Figure 6.10). In general, we observe fluctuations in RMSD values during the first 200 ns, which are

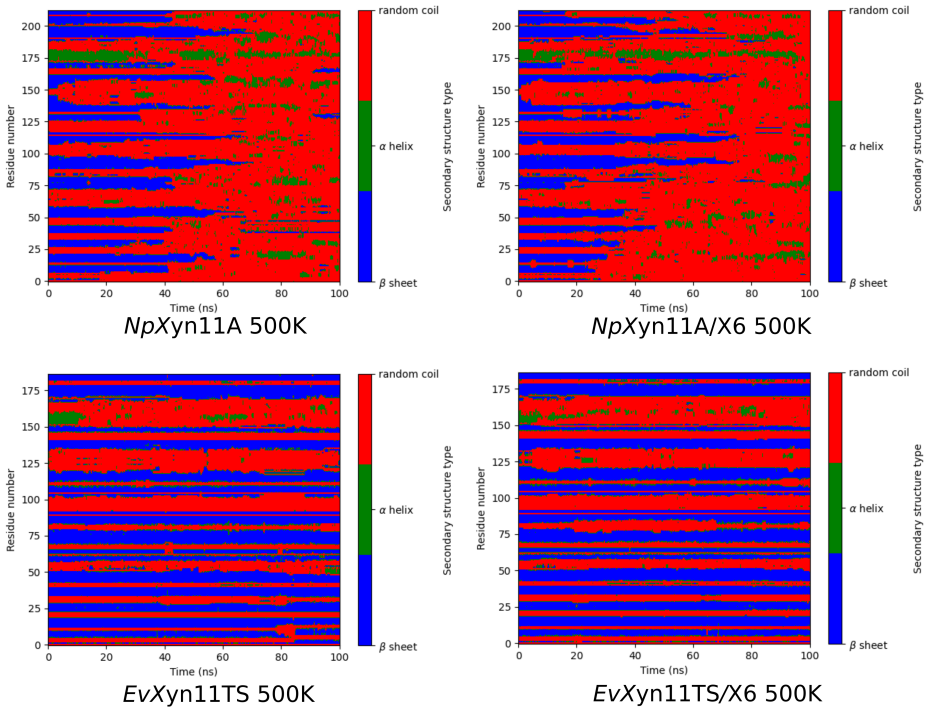

Figure 6.9: Secondary structure propensities during 100 ns of simulation time for *Np*Xyn11A (up) and *Ev*Xyn11^{TS} (down) in free-enzyme forms (left) and enzyme-substrate complexes (right) forms at 500K.

further followed by a stabilisation marked by a characteristic plateau reaching an equilibrium value between 0.5 and 1.5 Å depending on the systems. However, in the case of *Np*Xyn11A at 340K, significant fluctuations in RMSD values can still be observed between 600 ns and 900 ns in both free-enzyme and complex forms. As the maximum RMSD for all systems does not exceed 2.0 Å and remains relatively stable over time, one can conclude that the respective systems are all stable with respect to the chosen MD parameters.

The RMSD fluctuations observed during the first 200 ns are visible in the KL divergence calculations between first 500 ns and last 500 ns on the first PC only. This suggests that these fluctuations are due to motions along a single direction.

As shown in the RMSD time series, both considered mesophilic or hyperthermophilic xylanases did not show any signs of denaturation over the course of the simulation (1μs) at 310 and 340 K. Some structural changes were observed but the structures did not show any significant unfolding at these temperatures. Figure 6.11 shows key regions RMSD and highlights differences in conformational rearrangements between the two enzymes. The largest variations along time occur in the cord region of the free-enzyme form of *Np*Xyn11A, while in the enzyme-substrate complex the helical region induces the most variations. In the case of *Ev*Xyn11^{TS}, RMSD values of the thumb region are the highest and contribute the most to the

flexibility of this enzyme in its free from. In complex, this region becomes stable but RMSD values stay high in the cord region and induce the largest variations.

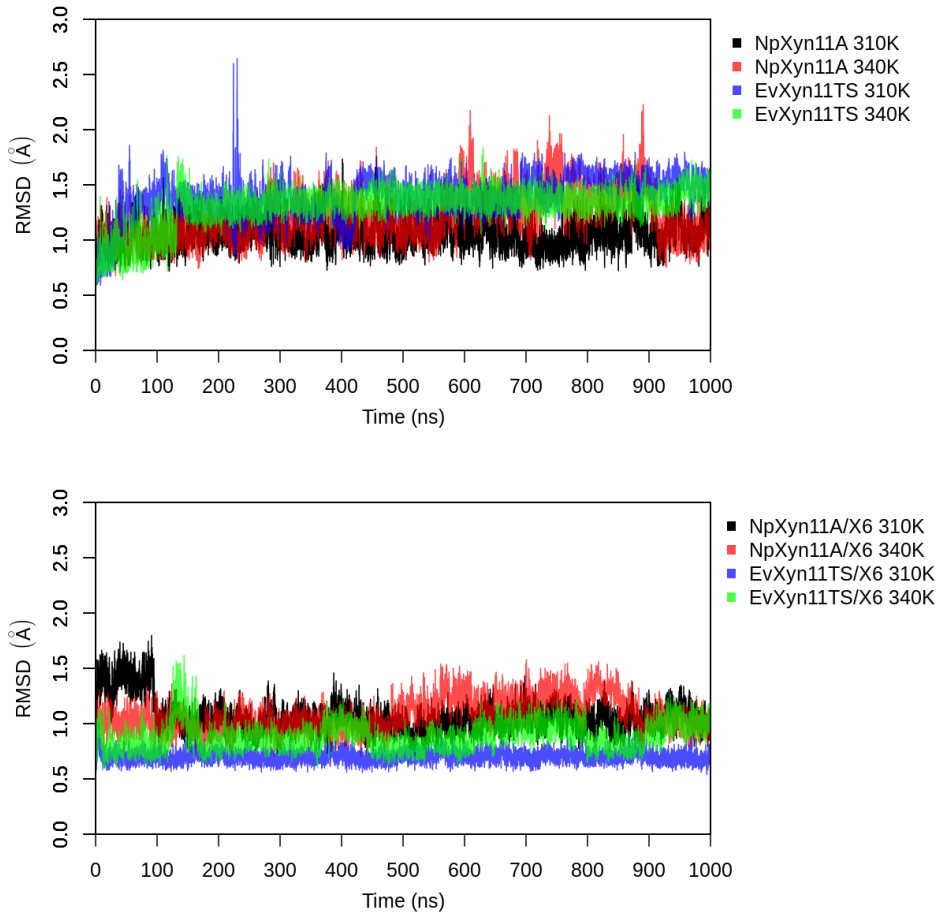

Figure 6.10: Backbone Root Mean Square Deviation (in Å) of systems in the free-enzyme and enzymes-substrate complex forms at different temperatures.

6.4.3 Flexibility analysis

In order to compare the backbone flexibility of these two enzymes, per-residue B-factors were calculated and monitored over the course of the simulations from the Root Mean square fluctuations (RMSF) on all backbone atoms.

Figure 6.12 shows the backbone B-factor values as a function of the residue index for the free-enzyme and the enzyme-substrate complexes. To facilitate the analysis, both *Np*Xyn11A and *Ev*Xyn11^{TS} structures have been aligned with respect to their corresponding per-residue B-factor values.

The same backbone B-factor patterns can be observed in both forms of *Np*-Xyn11A at 310K and 340K, although fluctuations of greater amplitude are noticed

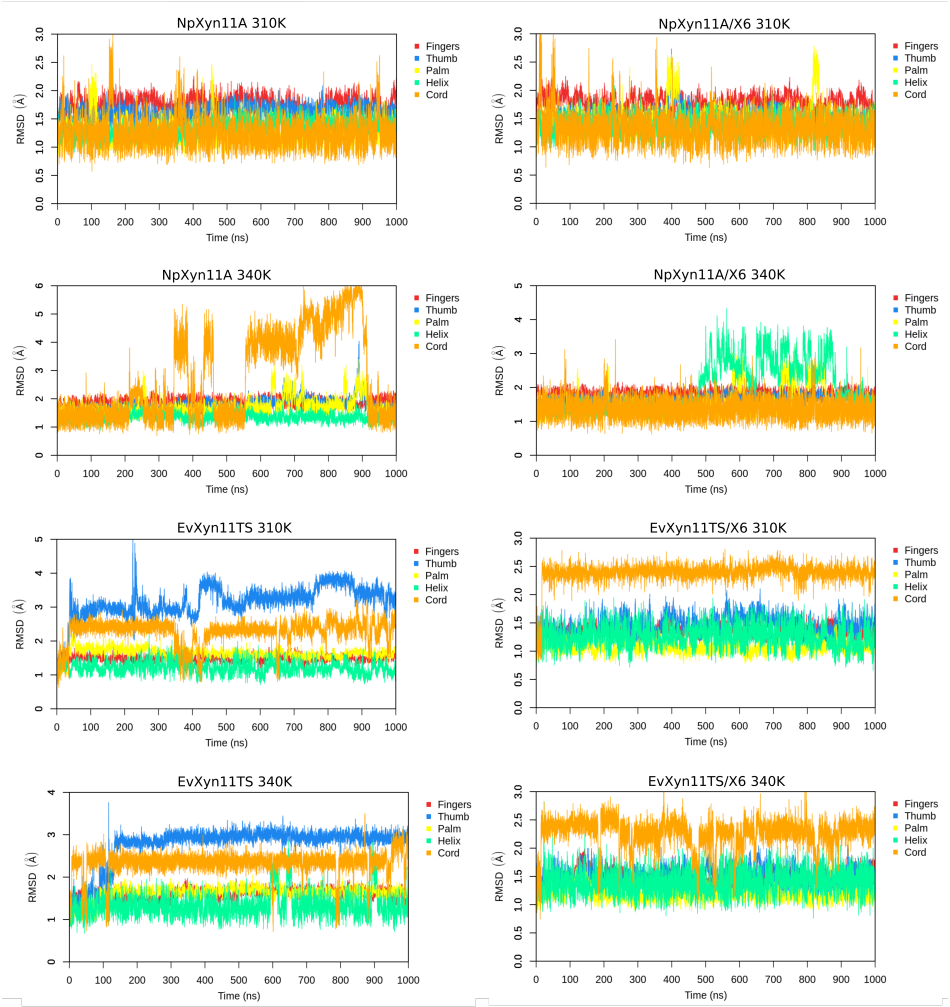

Figure 6.11: Backbone Root Mean Square Deviation (in Å) of each key region in the free-enzyme and enzyme-substrate complex forms at different temperatures.

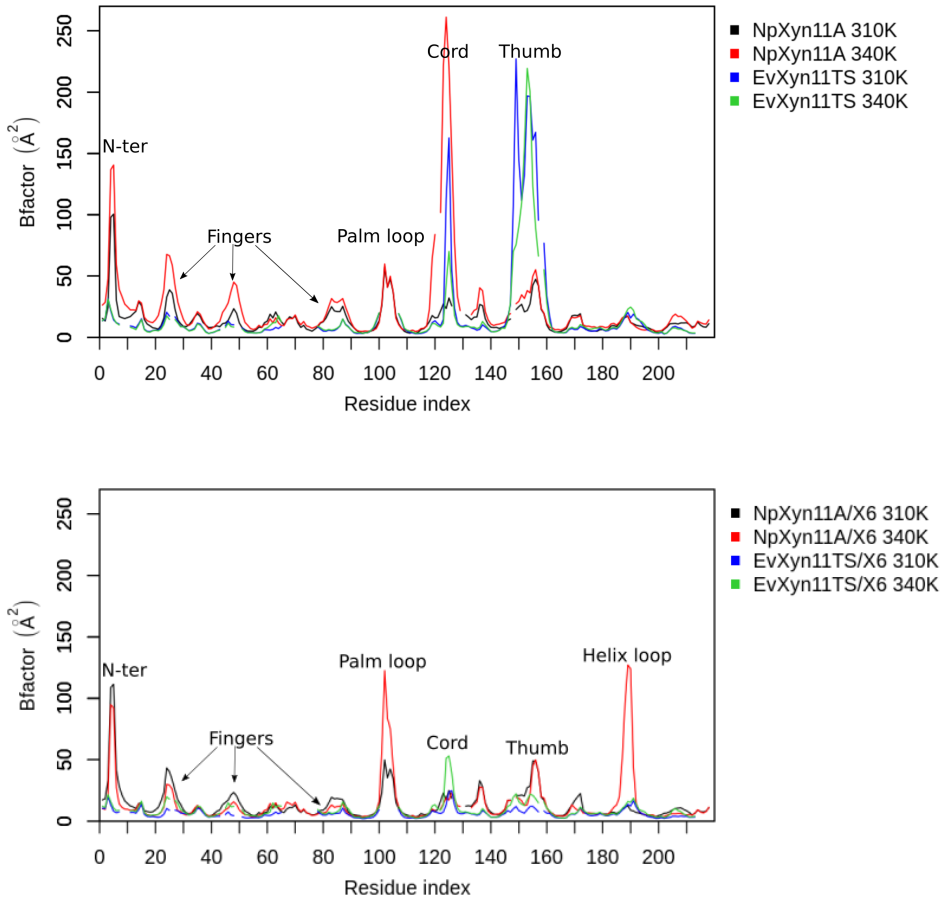

Figure 6.12: Per-residue average backbone B-factor profiles calculated from MD trajectories of *Np*Xyn11A and *Ev*Xyn11^{TS} in their free-enzyme and enzyme-substrate complex forms at two different temperatures (310K and 340K). Gaps in curves correspond to the gaps introduced in the alignment.

at higher temperature. Overall, *Ev*Xyn11^{TS} exhibits lower backbone B-factor values with smaller fluctuations compared to *Np*Xyn11A. The B-factor profiles of *Ev*Xyn11^{TS} in the free-enzyme and the enzyme-substrate complex are very similar at 310K and 340K. In its free form, *Np*Xyn11A has an average B-factor value of 14.7 Å^2 at 310K and 23.7 Å^2 at 340K, while the free form of *Ev*Xyn11^{TS} has an average B-factor value of 18.2 and 15.5 Å^2 respectively. In the respective enzyme-substrate complexes, both enzyme's backbones tend to be less flexible. Average B-factor values are 13.3 Å^2 for *Np*Xyn11A at 310K, 14.2 Å^2 at 340K and only 6.2 and 8.9 Å^2 for *Ev*Xyn11^{TS} at 310K and 340K respectively. One can observe a total of 7 major B-factor peaks in the B-factor profile of the free-enzyme form of *Np*Xyn11A. They are located in the N-ter, the fingers, the palm loop, the cord and the thumb regions. We can observe that the cord region exhibits higher B-factor values at

higher temperature, judging by the significantly higher B-factor peak observed in this region at 340K in comparison to 310K. When observing the B-factor profiles of the *Np*Xyn11A enzyme-substrate complex, we can see that the backbone B-factor values are generally smaller than in the free-enzyme. The thumb region and the N-ter region present approximately the same flexibility as in the free-enzyme form at both studied temperatures. The cord is less flexible in enzyme-substrate complex, even at 340K. However, a new B-factor peak can be noticed in the region spanning from the residue 180 to 200 in the enzyme-substrate complexes. This increase in B-factor corresponds to the loop located between the α-helix and the β-sheet B4 (here referred to as the helix loop). The palm loop B-factor is much higher than in the free-enzyme forms, which suggests that this region is also more flexible in the enzyme-substrate complexes. When comparing *Np*Xyn11A with its hyper-thermostable counterpart *Ev*Xyn11$^{\mathrm{TS}}$, we can clearly see that *Ev*Xyn11$^{\mathrm{TS}}$ is more stable and presents a much lower number of flexible regions. The N-ter region as well as the fingers, palm loop and α-helix loop regions do not present any apparent backbone flexibility in *Ev*Xyn11$^{\mathrm{TS}}$. The very low B-factor of the N-ter region can explained by the presence of a disulfide bridge restraining the backbone dynamics in this region. It is well known that disulfide bridges play an important role on the stability of all xylanases of the GH11 family. Therefore, the presence of this disulfide bridge at the N-ter region of *Ev*Xyn11$^{\mathrm{TS}}$ plays a crucial role on the stabilization of this particular region but may also play a role on the general stability of this enzyme.

When looking at the B-factor profiles of *Ev*Xyn11$^{\mathrm{TS}}$ in its free form, a high peak can be noticed in the thumb region. Even though the thumb region seems to be very flexible in the free-enzyme form of this hyper-thermostable mutant, this peak becomes almost insignificant in complex form. This suggests that the presence of the ligand stabilizes this region. In both, the free-enzyme and enzyme-substrate complex forms of *Ev*Xyn11$^{\mathrm{TS}}$, the relatively high B-factor of the cord region indicates that the binding of the ligand does not completely reduce its overall flexibility.

The flexibility analysis of the mesophilic *Np*Xyn11A and the hyperthermophilic *Ev*Xyn11$^{\mathrm{TS}}$ revealed that *Ev*Xyn11$^{\mathrm{TS}}$ is globally less flexible than *Np*Xyn11A, thus more stable with respect to an increase of temperature. High B-factor values only apply to the thumb region and only in the free-enzyme form. The greater stability of this region in the enzyme-substrate complex can be explained by the presence of the ligand and its important interactions with the thumb.

*Np*Xyn11A has a greater number of flexible regions than *Ev*Xyn11$^{\mathrm{TS}}$ and is thus globally less stable. In order to confirm the previous results and get more insights on the conformational dynamics of these flexible regions identified in both enzymes, a comparison of the dynamic cross correlation of the backbone of their respective 3D structures, in the presence and in the absence of the substrate, has been performed.

6.4.4 Dynamic cross correlation

Dynamically cross correlated motions have been analyzed for both enzymes at 310K and 340K in their free-enzyme and enzyme-substrate complex forms (see Figure 6.13 and Figure 6.14). Both enzymes exhibit very similar global dynamics, with very similar regions showing highly correlated motions. The fingers regions tend to be dynamically correlated with the N-ter region (zone a in the Figure 6.13A and B). A correlation of the cord region backbone dynamics with the finger region can be observed in both enzymes (zone b). The backbone dynamics of the thumb region also seems highly correlated with the one of the palm loop region in *Np*Xyn11A (Fig.6.13A zone c). This less pronounced correlation in *Ev*Xyn11^{TS} (Fig.6.13B zone c), is probably due to the short length of the palm loop in this enzyme. A correlation involving of the β-sheet region 151-161 of the thumb with the cord region and its prolongation can also be noticed (zone d). Finally, other correlations involving to the helix region and its surroundings can be observed (zone e1 and e). Compared to the free-enzyme forms, the enzyme-substrate complex forms present a higher proportion of positively correlated motions. An important correlation is detected between the region corresponding to the loop between β-sheets B3 and A5 with the palm loop (zone f) and with the α-helix region (zone e1). This correlation is observed in the free-enzyme and in the enzyme-substrate complex forms of *Np*Xyn11A but is much more pronounced in latter form (Fig.6.13A). In *Ev*Xyn11^{TS}, the loop B3-A5 is correlated with the helix region (Fig.6.13B zone e1) but the correlation with the palm loop is almost nonexistent (Fig.6.13B zone f). As mentioned previously, this could be caused by the shorter length, thus minor mobility of the palm loop in *Ev*Xyn11^{TS}.

These results suggest that the palm loop may play an important role on the higher flexibility of *Np*Xyn11A. To validate this hypothesis, this region could be engineered in further studies focusing on improving the thermal stability of GH11 xylanases.

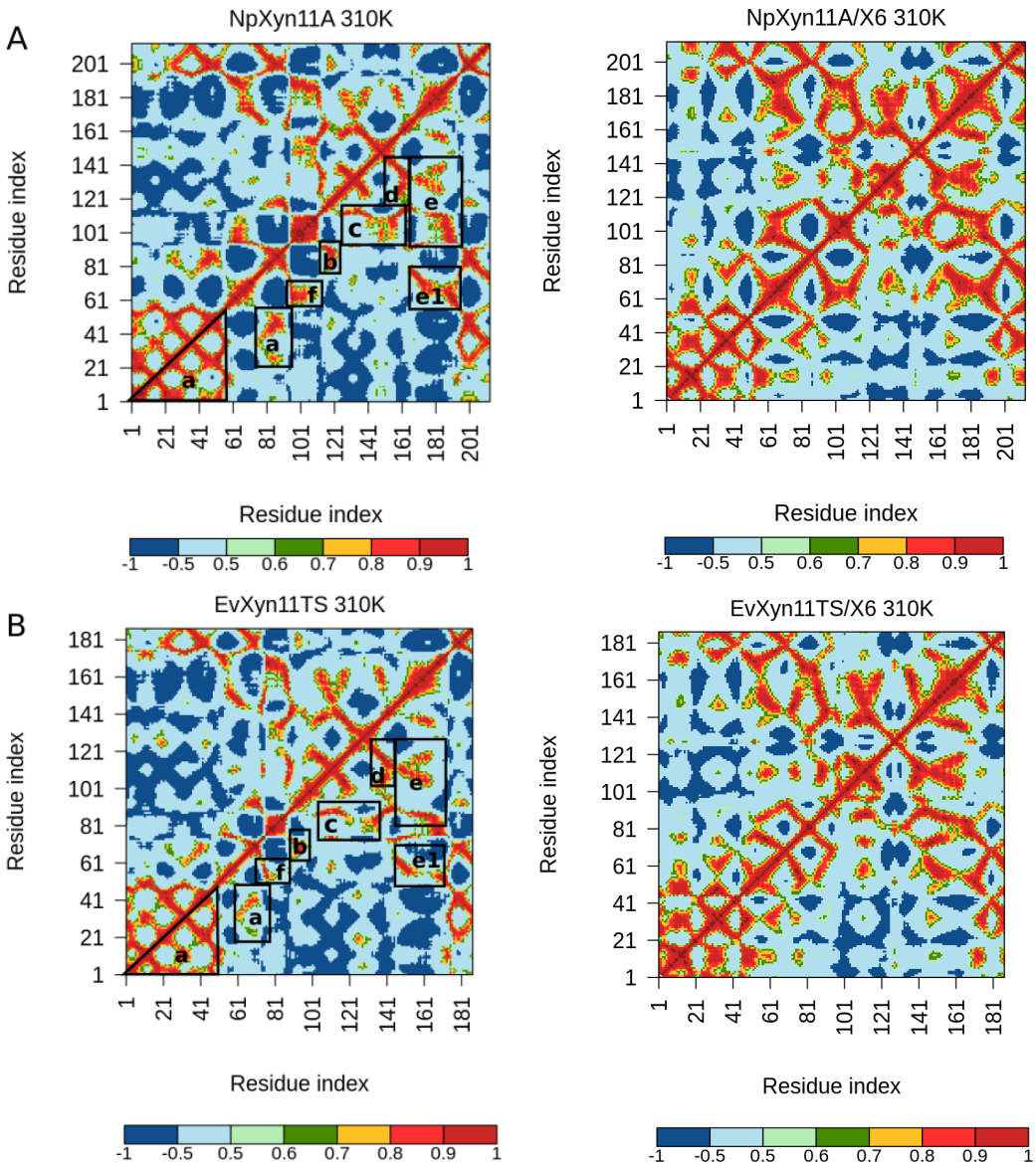

Figure 6.13: Dynamic cross-correlation analysis for *Np*Xyn11A and *Ev*Xyn11^{TS} in their free-enzyme and enzyme-substrate complex forms at 310K.

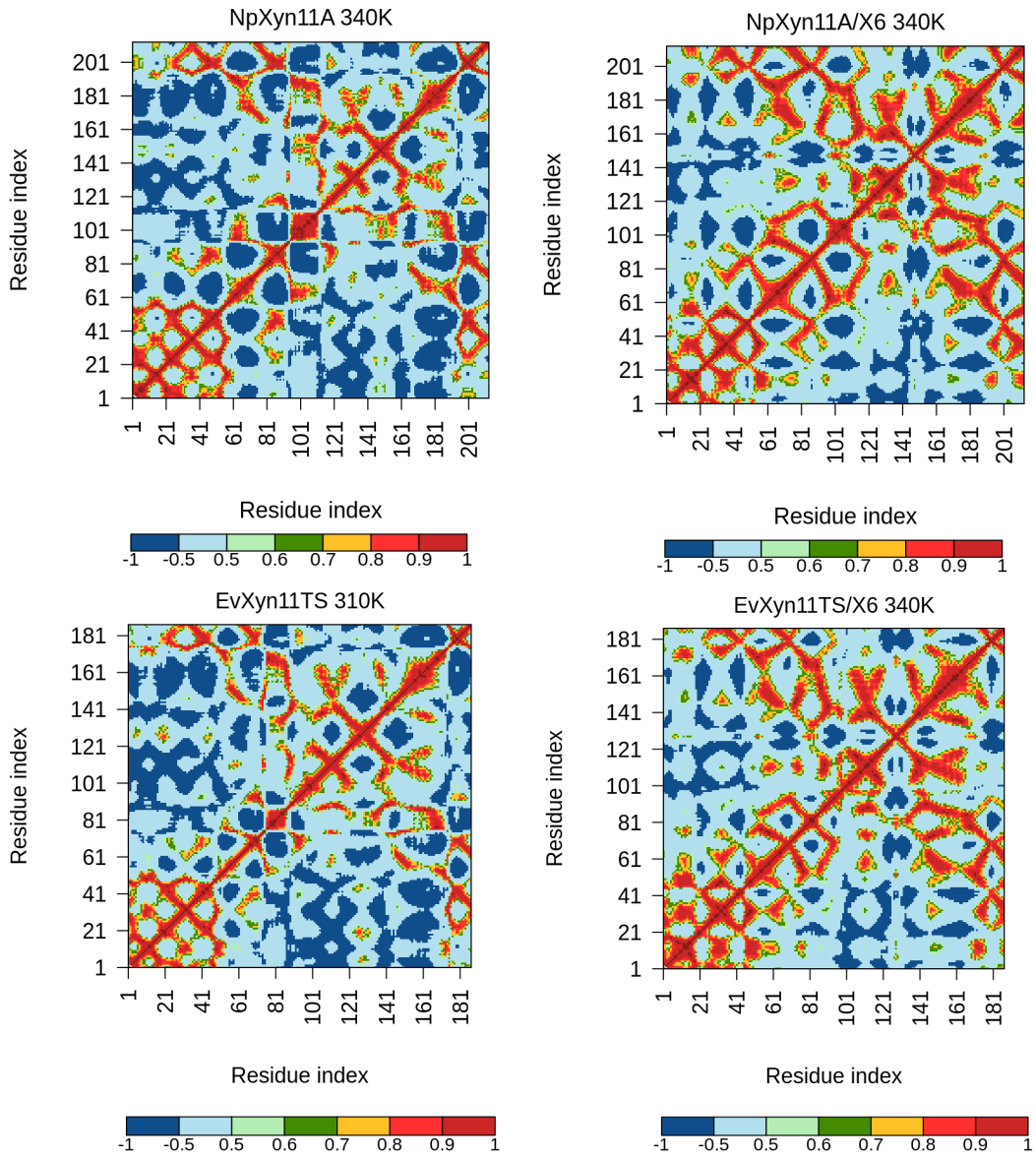

Figure 6.14: Dynamic cross-correlation analysis for *Np*Xyn11A and *Ev*Xyn11^{TS} in their free-enzyme and enzyme-substrate complex forms at 340K.

6.4.5 Free energy landscapes

The FELs of both enzymes have been constructed based on the projections of the first (PC1) and second (PC2) eigenvectors. Figures 6.15 and 6.16 show the FELs of *Np*Xyn11A and *Ev*Xyn11^{TS} in their free-enzyme and enzyme-substrate complex forms at 310K and 340K respectively. At 310K, only one free energy basin can be observed for *Np*Xyn11A in its free enzyme form, indicating the presence of one major ensemble of conformational substates. Two basins can however be observed for the

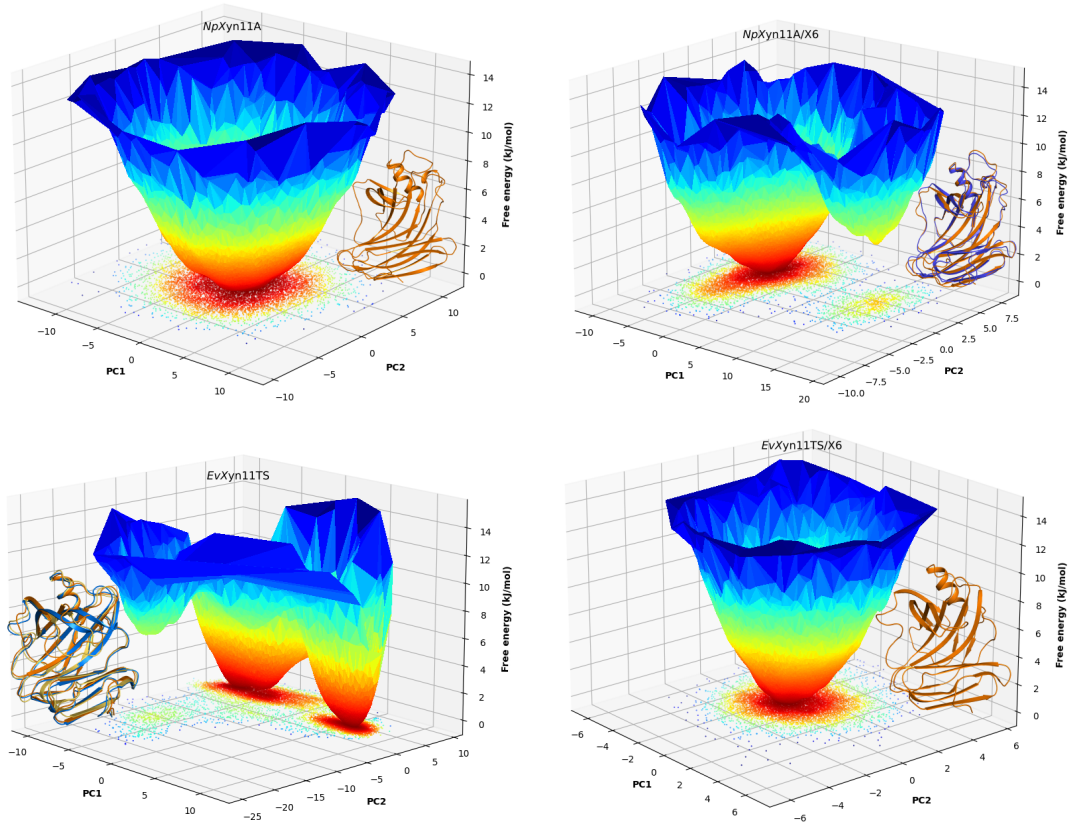

Figure 6.15: Free energy landscapes of *Np*Xyn11A and *Ev*Xyn11^{TS} at 310K in their free-enzyme and enzyme-substrate complex forms as a function of the first (PC1) and second (PC2) eigenvectors. The colorbar represents the free energy values in kJ/mol. The 3D structure corresponding to the global free-energy minimum is displayed in orange cartoon, while the blue one refers to the alternative conformation corresponding to the free-energy minimum of the second basin for the enzyme-substrate complexes.

enzyme-substrate form although the smallest basin presents a much higher energy minima than the other one, thus corresponding to much less stable conformational substates. At higher temperature (340K), two distinguished basins are observed for both forms of the enzymes. This put in evidence the increased flexibility of this enzyme at higher temperature, resulting from the enhanced conformational sampling. *Ev*Xyn11^{TS} presents more conformational sampling in its free-enzyme form, in accordance with the previous B-factor and per region RMSD analyses from which important movements of the thumb region in the free-enzyme form were characterized. In complex with X6, *Ev*Xyn11^{TS} presents only one main free-energy basin at 310K, while having a greater number of conformational substates at 340K.

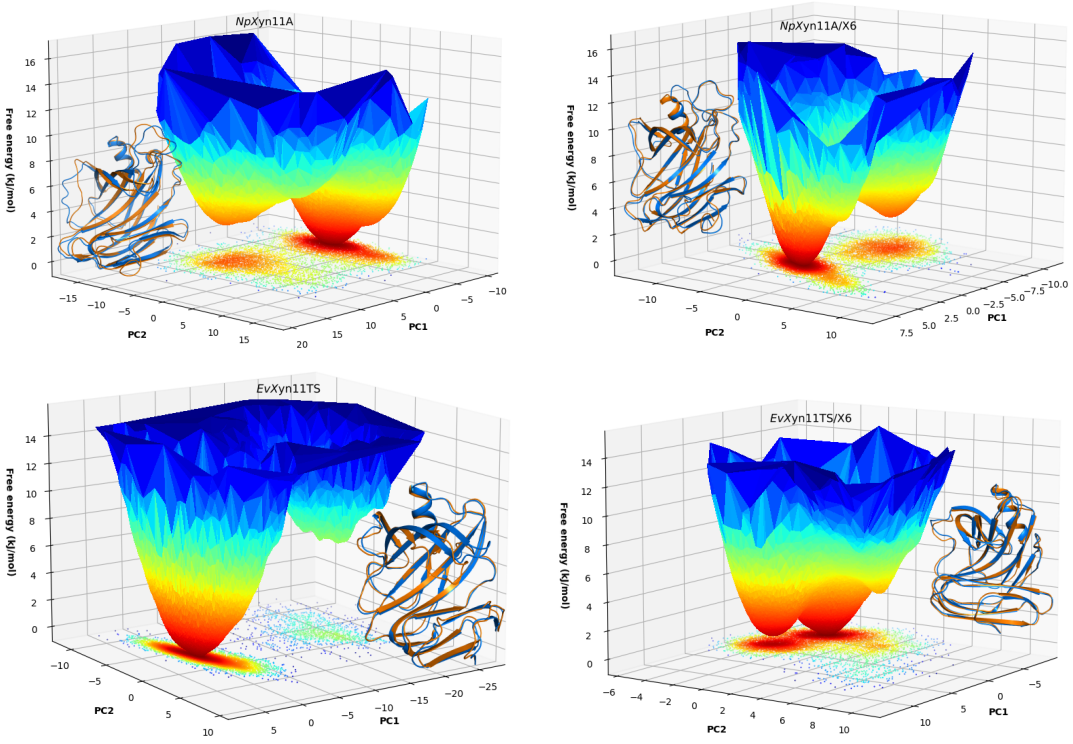

Figure 6.16: Free energy landscapes of *Np*Xyn11A and *Ev*Xyn11^{TS} at 340K in their free-enzyme and enzyme-substrate complex forms as a function of the first (PC1) and second (PC2) eigenvectors. The colorbar represents the free energy values in kJ/mol. The 3D structure corresponding to the global free-energy minimum is displayed in orange cartoon, while the blue one refers to the alternative conformation corresponding to the free-energy minimum of the second basin for the enzyme-substrate complexes.

6.4.6 Salt bridges, Hydrogen bonding and SASA

Ionic interactions have been identified as one of the main factors contributing to thermostability within the GH11 family of xylanases [217, 183]. Here we have monitored different salt bridges which are formed within the protein structures over 1 μs of simulation time. The identified salt bridges are shown on the three dimensional structure of each enzyme in Figure 6.17. Surprisingly, a total of 8 salt bridges stabilizes the structure of *Np*Xyn11A against 4 in *Ev*Xyn11^{TS} despite its greater thermostability. However, salt bridges are present in 33 to 98% of the total simulation time for *Np*Xyn11A, against 94 to 99% of the total simulation time for *Ev*Xyn11^{TS} (Table 6.2). Interestingly, the most stable salt bridge is formed by the residue pair Asp142-Lys156 in the thumb region of *Np*Xyn11A. The average frequency of occurrence of this salt bridge is around 80%, which may explain the relatively moderate flexibility of this region in *Np*Xyn11A at both studied temperatures. Another salt bridge that may play a role in the stability of *Np*Xyn11A is formed by the residue pair Asp123-Lys137 located between the cord and the thumb region. Analysis of B-factors in Section 6.4.3 revealed a very high backbone flexibility of the cord region in the free-enzyme form of *Np*Xyn11A at 340K. The high backbone flexibility in this region may be explained by a lower ability for the residue pair Asp123-Lys137 to form a salt bridge at 340K due to the increased mobility of their respective side chains at this temperature (6.2)

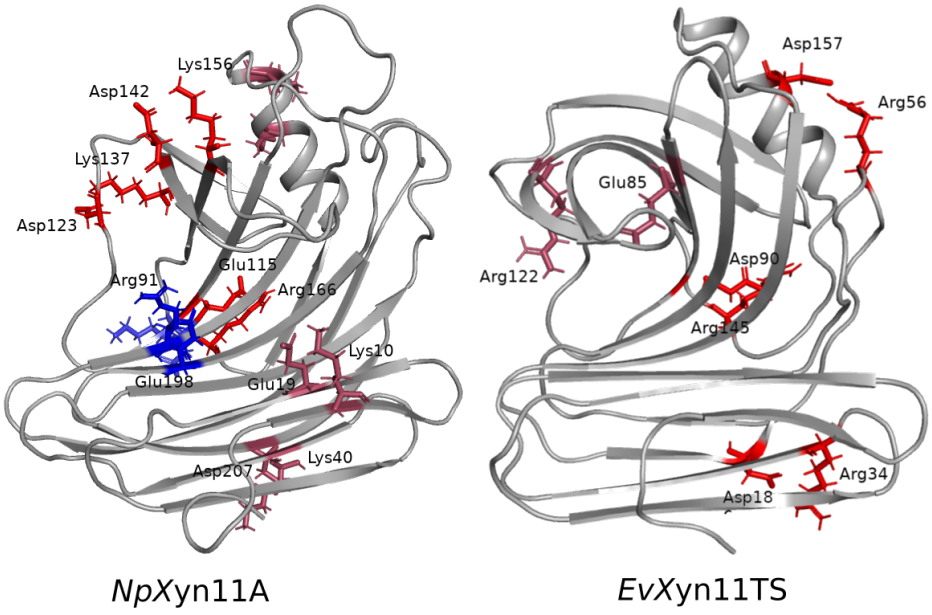

Figure 6.17: Location of salt bridges in *Np*Xyn11A and *Ev*Xyn11^{TS}, colored by their frequency of occurrence (red > 80%, pink between 50% and 80% and blue < 50%).

In *Ev*Xyn11^{TS}, the residue pair Asp157-Arg56 forms a salt bridge located be-

tween the α-helix and the loop B3-A5 (loop in pink in Figure 6.5). This salt bridge is formed around 98% of the total simulation time. Its location is probably very important for the global stability of the enzyme as the α-helix is considered to be a "hot spot" where unfolding preferentially occurs. Overall, *Ev*Xyn11^{TS} possesses more salt bridges with higher frequencies of occurrence which may explain the greater thermostability of this enzyme in comparison with *Np*Xyn11A.

	Free enzyme		Complex/X6 enzyme	
	310	340	310	340
*Np*Xyn11A				
Asp116-Lys165	38.9	-	33.8	35.8
Asp-123-Lys137	92.3	68.5	90.2	91.9
Asp142-Lys156	82.6	81.8	78.5	77.4
Asp207-Lys40	54.6	54.71	55.3	60.4
Glu115-Arg166	98.7	96.9	98.8	96.9
Glu179-Lys182	68.9	72.0	78.2	72.8
Glu19-Lys10	54.9	45.6	48.4	33.1
Glu198-Arg91	38.1	46.50	-	-
*Ev*Xyn11^{TS}				
Asp157-Arg56	95.7	97.7	99.6	97.7
Asp186-Arg34	99.7	99.20	99.8	99.6
Asp90-Arg145	99.9	98.9	99.9	99.5
Glu85-Arg122	94.6	96.1	-	-

Table 6.2: Occurrence fraction in percentage of salt bridges identified in MD simulations.

The number of intra-molecular hydrogen bonds as well as the number of enzyme-solvent hydrogen bonds in *Np*Xyn11A and *Ev*Xyn11^{TS} have been calculated over the course of their respective MD trajectories. Table 6.3 shows the number of calculated intramolecular static, dynamic and enzyme-solvent hydrogen bonds per residue in the free-enzyme and enzyme-substrate complex forms at 310K and 340K. *Ev*Xyn11^{TS} has more static hydrogen bonds, which contribute to a higher number of stabilizing interactions. *Np*Xyn11A possesses a higher number of dynamic HBs which reflect the dynamic formation of competitive HB interactions. As opposed to static HBs, the transient existence of a greater number of dynamic HBs in *Np*Xyn11A may contribute to explain its greater flexibility. Furthermore, the number of enzyme-solvent hydrogen bonds is also higher for *Np*Xyn11A. This results may also suggest that the enzyme has more interactions with the solvent and thus less static interactions within the protein what makes its structure more dynamic. Another property that may explain the difference in stability between these two enzymes is the Solvent accessible surface area (SASA). Table 6.4 shows average SASA values for each enzymes. When averaged over the course of their respective MD trajectories, *Np*Xyn11A has a higher average SASA than *Ev*Xyn11^{TS}. This reveals once again that *Np*Xyn11A is less tightly packed than *Ev*Xyn11^{TS} and which establishes a greater number of static HBs.

	Static HB	Dynamic HB	HBs with solvent
*Np*Xyn11A 310K	0.57	17.01	1473.03
*Np*Xyn11A 340K	0.56	20.87	1356.64
*Ev*Xyn11^{TS} 310K	0.62	16.53	1228.94
*Ev*Xyn11^{TS} 340K	0.60	18.17	1125.52
*Np*Xyn11A/X6 310K	0.56	16.54	1427.75
*Np*Xyn11A/X6 340K	0.56	19.75	1327.13
*Ev*Xyn11^{TS}/X6 310K	0.62	14.62	1115.26
*Ev*Xyn11^{TS}/X6 340K	0.61	17.37	1068.98

Table 6.3: Number of hydrogen bonding intercations in *Np*Xyn11A and*Ev*Xyn11^{TS}. Intra-molecular static, dynamic and the number of enzyme-solvent hydrogen bonds are given.

	SASA(Å^2)
*Np*Xyn11A 310K	8714.9
*Np*Xyn11A 340K	8840.3
*Ev*Xyn11^{TS} 310K	7224.6
*Ev*Xyn11^{TS} 340K	7368.1
*Np*Xyn11A/X6 310K	8549.5
*Np*Xyn11A/X6 340K	8430.2
*Ev*Xyn11^{TS}/X6 310K	6664.35
*Ev*Xyn11^{TS}/X6 340K	6900.1

Table 6.4: Values of SASA of each system in their free-enzyme and enzyme-substrate complex forms at 310K and 340K.

6.4.7 Analysis of enzyme/substrate interactions

One of the main features of the compact globular structure of these enzymes is the presence of a long cleft located in the center of the enzyme which contains the active site (also shown in Figure 6.5). The active site of each enzyme has been analyzed in terms of residues composition, negative volume and area of the pocket. The figure 6.18 shows the negative volume of each enzyme's active site as well as the residues that compose it. In Table 6.5, we summarize the number of residues that compose the active site of each enzyme as well as the values of negative volume and area of each pocket. The volume of the active site of *Np*Xyn11A is almost six times bigger than the volume of the active site of *Ev*Xyn11^{TS}. It encompasses 41 residues with a pocket area of 521.62 Å^2 while *Ev*Xyn11^{TS} possesses only 23 residues, and a pocket area of 173.27 Å^2. The size of the pocket where substrate binding occurs, shows that the three-dimensional structure of *Ev*Xyn11^{TS} is more compact. However, the size of the cleft may have an important role on the unusually high activity displayed by the *Neocallimastix* enzyme. Given its size, the substrate binding cleft of *Np*Xyn11A is more extended and may better accommodate xylose residues in each of its subsites.

	Nb residues	Volume (SA)	Area (SA)
*Np*Xyn11A	41	461.24	521.68
*Ev*Xyn11^{TS}	23	77.12	173.27

Table 6.5: Geometrical and topological properties of each enzyme's active site.

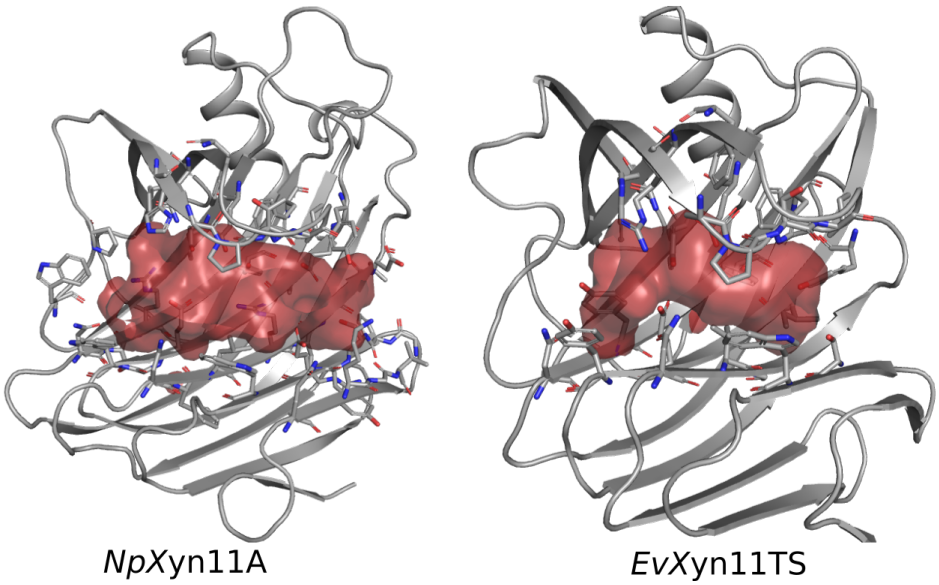

Figure 6.18: Volume of enzymes active site pockets in *Np*Xyn11A and *Ev*Xyn11^{TS}.

Different enzyme/substrate non-covalent interactions have been evaluated and monitored using the PLIP webserver. To consider the most catalytically favor-

able conformation of xylohexaose, we have chosen to use the 3D structure of *Np*-Xyn11A/X6 and *Ev*Xyn11^{TS}/X6 generated after the equilibration phase. Different predicted interactions are presented in Figure 6.19 where we display the initial equilibrated configuration of the respective enzymes in complex with xylohexaose. There are in total 17 different hydrogen bonding interactions with xylohexaose in *Np*Xyn11A, versus 13 in *Ev*Xyn11^{TS}. The list of residues involved in hydrogen bonding interactions with xylohexaose in each enzyme is given in Table 6.6. One salt bridge is observed in *Ev*Xyn11^{TS} but there are no other types of non-covalent enzyme/substrate interactions.

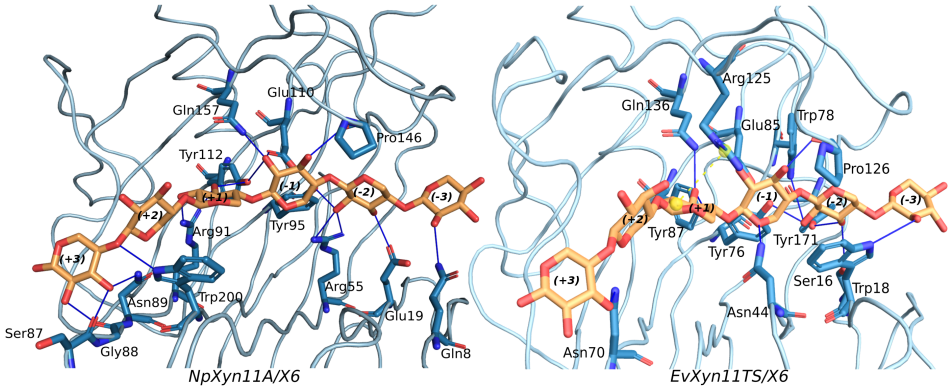

Figure 6.19: Non-covalent enzyme/substrate contacts found with the Plip web-server on the equilibrated structure structures of *Np*Xyn11A and *Ev*Xyn11^{TS}. Hydrogen bonds are shown in blue lines and salt bridges in yellow lines.

*Np*Xyn11A/X6	*Ev*Xyn11^{TS}/X6
Gln8 (SC)	Ser16 (SC)
Glu19 (SC)	Trp18 (SC)
Arg55 (SC)	Asn44 (SC)
Ser87 (BB)	Asn70 (SC)
Gly88 (BB)	Tyr76 (SC)
Asn89 (SC)	Trp78 (SC)
Arg91 (SC)	Glu85 (SC)
Tyr95 (SC)	Tyr87 (SC)
Glu110 (SC)	Arg122 (SC)
Tyr112 (SC)	Pro126 (BB)
Pro146 (SC)	Gln136 (SC)
Gln157 (SC)	Tyr171 (SC)
Trp200 (SC)	

Table 6.6: List of residues involved in hydrogen bonding with X6 in *Np*Xyn11A and *Ev*Xyn11^{TS} equilibrated structures. The side chain (SC) or Backbone(BB) atoms that contribute to the hydrogen bonding are given in parenthesis.

The intermolecular hydrogen bonds formed between the enzymes and the xylo-

hexaose substrate were also monitored during the first 100ns of the respective MD trajectories at 310K. A residue is counted as involved in a hydrogen bonding interaction with the ligand if the interaction between the residue and the ligand is present in at least one frame of the MD trajectory. By counting the number of frames in which an interaction is formed between a given residue pair, we can calculate its frequency of occurrence. Over the course of the MD trajectories, we do not observe the formation of any other intermolecular hydrogen bonds than the ones already observed in the initial equilibrated configuration. The tables 6.7 and 6.8 show the percentage of occurrence (greater than 10%) of the intermolecular hydrogen bonds established between *Np*Xyn11A and xylohexaose and *Ev*Xyn11^{TS} and xylohexaose respectively. Figure 6.20 displays the residues involved in hydrogen bonding in *Np*Xyn11A and*Ev*Xyn11^{TS} over the course of their respective simulations. These residues are colored by their percentage of occurrence.

HB_{inter}	*Np*Xyn11A/X6 310K
Gln8-X6	22
Glu19-X6	43
Ser87-X6	58
Asn51-X6	65
Asn89-X6	29
Tyr95-X6	21
Glu110-X6	77
Pro146-X6	79
Trp200-X6	10

Table 6.7: The percentage of occurrence of the inter-molecular hydrogen bonds between *Np*Xyn11A and xylohexaose at 310K.

HB_{inter}	*Ev*Xyn11^{TS}/X6 310K
Asn44-X6	51
Asn70-X6	38
Tyr72-X6	24
Tyr76-X6	67
Glu85-X6	95
Asp101-X6	28
Arg122-X6	29
Pro126-X6	77
Gln136-X6	30
Tyr171-X6	80

Table 6.8: The percentage of occurrence of the inter-molecular hydrogen bonds between *Ev*Xyn11^{TS} and xylohexaose at 310K.

Some interactions, observed in the initial configuration, are present in less than 10% of the 100 ns of both *Np*Xyn11A and*Ev*Xyn11^{TS}. In *Np*Xyn11A, this is the

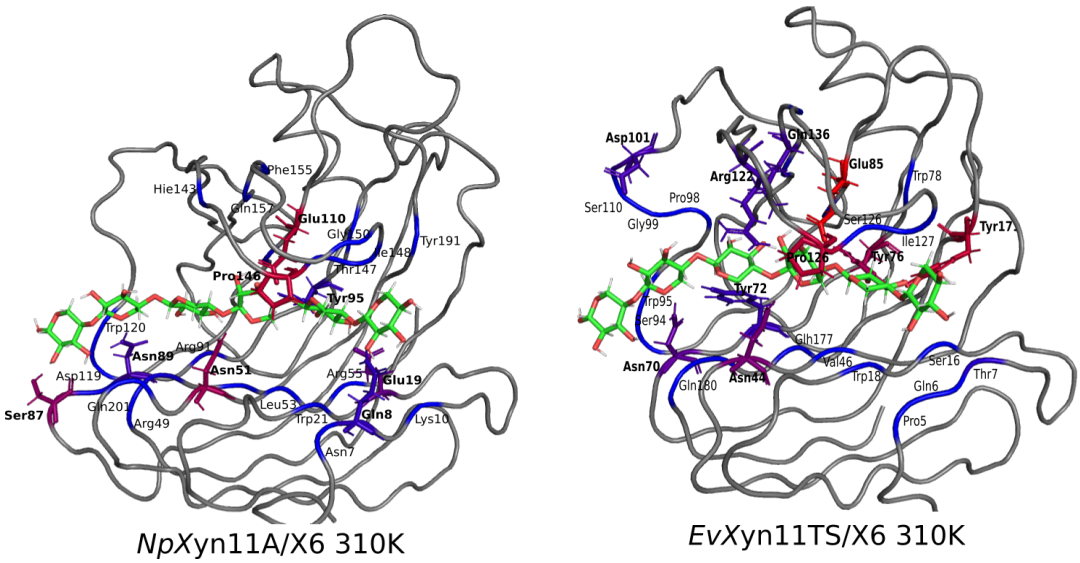

Figure 6.20: Residues involved in hydrogen bonding interactions with xylohexaose in *Np*Xyn11A and *Ev*Xyn11^{TS} over the course of their respective MD simulations. Residues are colored by their frequency of occurrence. Residues colored in blue represent the residues involved in the less frequent interactions and residues colored in red represent the residues involved in the most frequent interactions.

case for Arg55 which initially interacts with the xylose subunit in the subsite -2, Arg91 which interacts with the xylose subunit in the subsite +1, Tyr112 which interacts with both xylose subunits in both subsites +1 and -1 and Gln157 which interacts with the xylose subunit in the subsite -1. In *Ev*Xyn11^{TS} this is the case for both Ser16 and Trp78 which initially interact with the xylose subunit in the subsite -2, Trp18 which initially interacts with the xylose subunit in the susbiste -3 and Tyr87 with the xylose subunit in the subsite +1. In comparison with *Np*Xyn11A, this result shows that the interactions with the -2 and -3 (glycone) subsites are less conserved in *Ev*Xyn11^{TS}. As HB interactions involving the glycone subsites residues are crucial for the substrate binding and catalysis, the loss of interactions identified in *Ev*Xyn11^{TS} might explain the lower catalytic activity of this enzyme in comparison with *Np*Xyn11A. Another interesting observation is the presence of an hydrogen bonding interaction with the amino acid Asp101 in *Ev*Xyn11^{TS}. This interaction is observed in MD analysis in 28% of time. Asp101 is located in the cord region of the enzyme (Figure 6.20). In the initial structure, this amino acids is located at 7Å distance from the ligand. This observation confirms that a conformational adjustment of this flexible region is required to allow this amino acid to interact with the substrate in the *Ev*Xyn11^{TS}/X6 complex. It is probable that the binding of the ligand triggers this conformational change and stabilizes the cord region which has been shown to be highly flexible in the free-enzyme form.

The presence of a tryptophan residue at position 24 in *Np*Xyn11A and at position 18 in *Ev*Xyn11^{TS} seems important as it promotes a stacking interaction of

the indole ring with the xylose moiety. Not surprisingly, this residue is highly conserved in all GH11 xylanases and therefore seems crucial for the catalytic activity of these enzymes [218]. A highly conserved proline in the GH11 family of xylanase can also be found at position 149 in *Np*Xyn11A and 126 in *Ev*Xyn11^{TS}. This proline is located in the thumb loop and is involved in forming a conserved pattern of HB interactions in more than 70% of both enzymes MD trajectories (Table 6.7 and Table 6.8). Arg122 in *Ev*Xyn11^{TS}, which is also located in the thumb region close to this proline in *Ev*Xyn11^{TS}, makes a polar contact with the +1 subsite (Figure 6.19). A histidine (His143) residue is found at this position in *Np*Xyn11A. Another study has already suggested that this residue should be involved in forming polar contacts with surrounding residues [171]. Along with Glu110 (*Np*Xyn11A) and Glu85 (*Ev*Xyn11^{TS}), Tyr95 (*Np*Xyn11A) and Tyr79 (*Ev*Xyn11^{TS}) are also known as catalytically important residues as they form intermolecular hydrogen bonds with xylohexaose. These interactions are also frequent in the first 100 ns of both enzyme-substrate complexes MD trajectories. Tyr95 interacts less frequently with xylohexaose in *Np*Xyn11A. This may be explained by the bigger *Np*Xyn11A active site pocket, especially in the glycone region, thus enabling the substrate to interact with more residues over the course of the simulation.

Compared with *Ev*Xyn11^{TS}, these results suggest that *Np*Xyn11A establishes more hydrogen bonding interactions with the substrate, possibly explaining its unusually high catalytic activity. The analyses of the unusually active *Np*Xyn11A and hyperthermostable *Ev*Xyn11^{TS} in complex with xylohexaose revealed new details on the substrate binding interactions in these two enzymes. The glycone-binding subsites of xylanases are not well known due to the lack of structural and biochemical data. Here, by providing detailed analysis of these two enzymes and their specific interactions with xylohexaose, we were able to get new insights on the binding mode of xylohexaose which could be transposed to other similar xylanases in the GH11 family.

6.5 Conclusion

In this study we investigated some dynamic properties of two different xylanases from the GH11 family: the particularly active GH11 xylanase from *Neocallimastix patriciarum*, *Np*Xyn11A [171], and the thermostable mutant of environmentally isolated GH11 xylanase, *Ev*Xyn11^{TS}. We performed MD simulations of the respective free-enzymes and enzymes-xylohexaose complexes. Diverse techniques for analyzing these MD simulations were used to explore the differences in dynamics influencing the activity and stability of these two enzymes. Analysis of backbone flexibility combined with monitoring of some specific structural and geometrical properties reveled that *Ev*Xyn11^{TS} is more tightly packed and that its thermal stability is enhanced by a higher number of intramolecular interactions. Some structural differences, such as shorter loops or the presence of a disulfide bridge at the N-ter region may also explain the increased stability of *Ev*Xyn11^{TS}. Analysis at a very

high temperature (500K) showed that during the first 100 ns, *Ev*Xyn11^{TS} does not undergo unfolding, while the unfolding occurs after a few ns in *Np*Xyn11A. This confirmed that *Ev*Xyn11^{TS} has a higher capacity for resisting heat denaturation.

Analyses on enzymes flexibility and conformational rearrangements revealed that *Ev*Xyn11^{TS} is globally more stable, but has greater conformational variability due to the conformational sampling observed in the thumb region. In comparison with the free enzyme, the lower flexibility of the thumb region in the complex form can be explained by the presence of the substrate which stabilizes the thumb region. *Np*Xyn11A has more flexible regions. N-terminal region seems to be very flexible contrary to *Ev*Xyn11^{TS} which possesses a disulfide bridge and a salt bridge helping the stabilization of this part of the enzyme. Surprisingly, the thumb region of *Np*-Xyn11A is moderately flexible in both free enzyme and complex forms, probably due to the presence of an important salt bridge. This suggests that *Np*Xyn11A, in both bound and unbound forms, possesses a thumb conformation which is competent for catalysis. Thus, the conformation of the thumb region is more stable in *Np*-Xyn11A and might allow better catalytic efficiency. However, other regions were found to have an important impact on the general flexibility of this mesophilic enzyme. The cord region presents very high flexibility in free enzyme form but seems to be stabilized in complex with xylohexaose. A quite high flexibility of the palm loop and the helix loop is observed in *Np*Xyn11A even when it is in its complex form. A cross-correlation analysis showed that the global dynamics of both enzymes is very similar. It confirmed the flexibility of previously identified regions, and showed that these movements are correlated in *Np*Xyn11A. Some important correlations involving the palm loop or the B3-A5 loop are absent in *Ev*Xyn11-^{TS}. They represent, together with other identified regions, potential stabilization hotspots.

In light of these analyses the thumb region and the larger catalytic site pocket of *Np*Xyn11A seem to play a major role on the activity of this enzyme. Its lower thermal stability may be caused by higher flexibility of certain regions located further from the active site. Regions such as the N-ter, β-turns located in the fingers region, the palm loop, the helix and the B3-A5 loop seem to be less stable than in hyperthermophilic *Ev*Xyn11^{TS} and thus represent interesting targets for engineering studies.

Chapter 7

Improving thermal stability of a GH11 xylanase

Contents

7.1 Motivations

The previous Chapter devoted to GH11 xylanases highlighted the importance of this enzyme and its applications in industrial processes. As we mentioned, xylanases, as most enzymes, need to suit specific conditions to be integrated in different industrial processes. We have studied two GH11 xylanases with MD techniques and we have identified potentially destabilizing regions in the mesophilic *Np*Xyn11A. In this Chapter, we try to combine the use of the multistate CPD method developed during this thesis and the knowledge gained by MD simulations and analysis. Our objective is to improve the thermal stability of particularly active *Np*Xyn11A and render it suitable for industrial applications. By using its 3D structure, conformational states previously generated with MD, and Pomp^d, we want to define a specific CPD approach and deliver sequences that possess desired properties: improved thermal stability and preserved enzymatic activity. We apply our multistate CPD method for thermostabilization of the xylanase *Np*Xyn11A. For all real applications of CPD methods and procedures, experimental validation is a mandatory step. Therefore, computationally generated mutant sequences have been experimentally tested and

validated by our colleagues from TBI: Manon Darribere, Thomas Enjalbert, Sophie Bozonnet, Cédric Montanier and Claire Dumon. Here, we present different computational but also experimental methods that have been used in this study and discuss obtained results.

7.2 Context

In the past years diverse techniques have been applied in order to enhance the thermostability of enzymes [219]. Different techniques of protein engineering have been explored, mostly based on rational design and site-directed mutagenesis. This type of enzyme engineering is guided by the knowledge and information on protein 3D structure, sequence, and catalytic mechanism. The advantage of this kind of methods, compared to traditional directed evolution, resides in the fact that they focus on several specific mutation sites. High-throughput screening is not needed and chances of obtaining active variants are higher. Rigidifying flexible sites (RFS) is another engineering strategy that specifically targets flexible regions in enzymes and tries to find mutations in these regions that would improve their thermostability [220, 221]. The success rate of this strategy remains low as it is not easy to determine the best mutation candidates due to the impact they can have on enzyme's activity. Therefore, the structure-dynamics-activity relationship is of great importance as a trade-off has to be found between the rigidity, essential to improve stability, and flexibility, essential to keep the enzymes active.

In the specific case of GH11 family of xylanases, many engineering studies have been done and mostly by site-directed mutagenesis. As we mentioned in the previous Chapter 6, GH11 xylanases were mainly engineered by transforming characteristics of thermophilic enzymes into mesophilic ones, by introducing disulfide bonds, or by introducing point mutations in regions considered as "hot spots" [222]. Different studies showed the important contribution of the N-ter region to the general stability of enzymes by substituting the whole N-ter of mesophilic xylanases with the N-ter of thermophilic ones, or by mutating residues in order to mimic the corresponding region from thermophilic xylanase. These experiences, as well as other rational design approaches, induced an increased thermostability in GH11 xylanases [223, 189, 224].

However, site-directed mutagenesis methods remain time-consuming and are limited by the diversity of protein sequences that can be generated and explored compared to the vastness of the sequence space. Alternatively, CPD methods possess great potential for this type of challenging problems. The CPD method and different options, developed during this thesis represent a promising approach to fully rationalize and speed-up the conception of optimized enzymes, and more precisely in this case, GH11 xylanases. Recently, an engineering study improved thermal stability of a GH11 xylanase via computational library design [225]. Bu and co-workers identified potentially stabilizing mutations by energy calculations with three different programs (FoldX [226], Rosetta_ddg [227] and ABACUS [228]). The design

protocol proposed by the authors requires an additional step of filtering chemically unreasonable mutations by visual inspection and MD simulations. Experimental verification is then done in two steps, first by testing mono-mutations and than by recombining them. Here we propose a multistate CPD protocol for improving thermal stability and specific activity of the GH11 xylanase from *Neocallimastix patriciarum*. With POMP^d [144], enzyme flexibly was taken into account and enzymatic activity was preserved by taking conformational states of the enzyme in complex with its substrate during the CPD procedure. With our approach, 20 mutant variants were generated and directly assessed for their impact on specific activity and thermostability. From these, 4 variants were found to possess better specific activity and better thermostability than the wild-type *Np*Xyn11A. These 4 xylanase variants possess improved properties and may represent more suitable candidates for industrial applications.

7.3 Material and methods

7.3.1 Computational methods using Pomp^d

The high-resolution structure of *Np*Xyn11A (PDB code: 2C1F) [171] was used to construct our starting models for the design procedure. Two different models have been used: free enzyme and enzyme/substrate model. Enzyme/substrate model has been constructed with the xylohexaose substrate as described in the previous Chapter (Chapter 6). Free enzyme and enzyme/xylohexaose complex were then equilibrated and minimized with the AMBER ff14SB force-field [26] for the free enzyme and GLYCAM_06j-1 force field [200] for the xylohexaose substrate. In order to use a multistate design (MSD) approach, conformational states were generated with two different procedures. The first one consisted in generating conformational states by Molecular Dynamics simulations. For each system (free enzyme and enzyme/substrate complex), we have generated one hundred conformations from the first 100ns of MD simulations previously done on *Np*Xyn11A and *Np*Xyn11A/X6 (described in Chapter 6). Four of the most diverse conformational states were kept (using RMSD-based hierarchical clustering [146]). The second procedure uses the Rosetta Backrub [147] software for flexible protein backbone modeling [148, 149]. We have generated one hundred conformations for each structure and as with the MD procedure, four of the most diverse conformational states were kept. The design strategy was mostly based on the analysis and information obtained from the previous MD study. This study allowed us to reveal key regions of this enzyme, in terms of stability, flexibility, interactions with the substrate etc. Thus, diverse mutations were allowed or disallowed in each key region. Designable residues were combined in 20 different ways, providing 20 scenarios which were further given to POMP^d for computational protein design. Computational protein design was done by taking into account multiple conformational bound and/or unbound states simultaneously but also by using the additional features (described in the Chapter 5). The hpatch option was used in some of the scenarios to prevent formation

of hydrophobic patches at the enzyme's surface and weight attribution was used in different ways on enzyme/complex conformational states in order to ensure that the enzymatic activity (binding) was preserved during the CPD procedure. A general workflow describing our CPD procedure is shown in Figure 7.1. Finally, 20 sequences were predicted and sent for experimental testing.

7.3.2 Materials, strains, media, and growth conditions

Unless otherwise stated, all chemicals were of analytical grade and purchased from Sigma-Aldrich (St. Louis, MO, UA). The genes encoding for *Np*Xyn11A11A mutants were synthesized by GeneCust (Boynes, France) and sub-cloned in pET22 expression plasmids. The expression strains Escherichia coli BL21 (DE3) and Top10 were prepared using a commercial kit from Zymo Research (Irvine, U.S.A.). Wheat arabinoxylan (WAX) was purchased from Megazyme(Bray, Ireland). Plasmid extraction was performed using QIAprep Spin Miniprep kit (Qiagen, Germany).

7.3.3 Expression and purification of enzymes

A streak of E. coli BL21 (DE3) colonies harbouring an appropriate plasmid were inoculated into 5 mL LB in the presence of appropriate antibiotics (kanamycin or ampicillin at 50 μg/mL final) and grown with aeration (180 rpm) at 37 °C for 16 h. The culture was then used as the inoculum for a 250 mL baffled flask containing 50 mL of Terrific Broth supplemented with the appropriate antibiotics at an optical density (OD600) of 0.1 nm and incubated at 37°C, 120 rpm. When the optical density reached a value between 0.4 and 0.6, expression of the enzymes of interest was induced by the addition of 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG, Sigma-Aldrich). Cultures were stopped by centrifugation (15 min, 750 g at 18°C, Eppendorf Centrifuge 5804R). The cell pellet was resuspended in 1 mL of 50 mM sodium phosphate, 300 mM NaCl at pH 7.5, supplemented with proteases inhibitor mixture (PIM x100). The cell lysis was achieved by supersonic vibration (40 s, 6.5 M/s, FastPrep-24TM 5G, MP Biomedicals). The supernatant was retrieved by centrifugation at 20,000xg for 10 min and then stored at 4°C. The enzymes of interest were purified by Immobilized Metal Affinity Chromatography (TALON® Metal Affinity Resin, Clontech, Clonetech). Aliquot of 0.5 mL of resin was equilibrated in a column (20 ml, Clonetech) with the equilibration buffer (50 mM sodium phosphate, 300 mM NaCl, pH 7.5). The resin was washed with 10 column volumes using the equilibration buffer. Protein of interest was then eluted with 2 column volumes of the equilibration buffer containing 200 mM imidazole. Enzyme conformity and purity were assessed using SDS-PAGE (Any kD, Mini-PROTEAN TGX Stain-Free. Protein Gels, Bio-rad, Hercules, CA, USA). Purified enzymes were extensively dialysed against 50 mM sodium phosphate, pH7.5 (Pur-A-Lyzer Midi Dialysis Kit, Sigma-Aldrich).

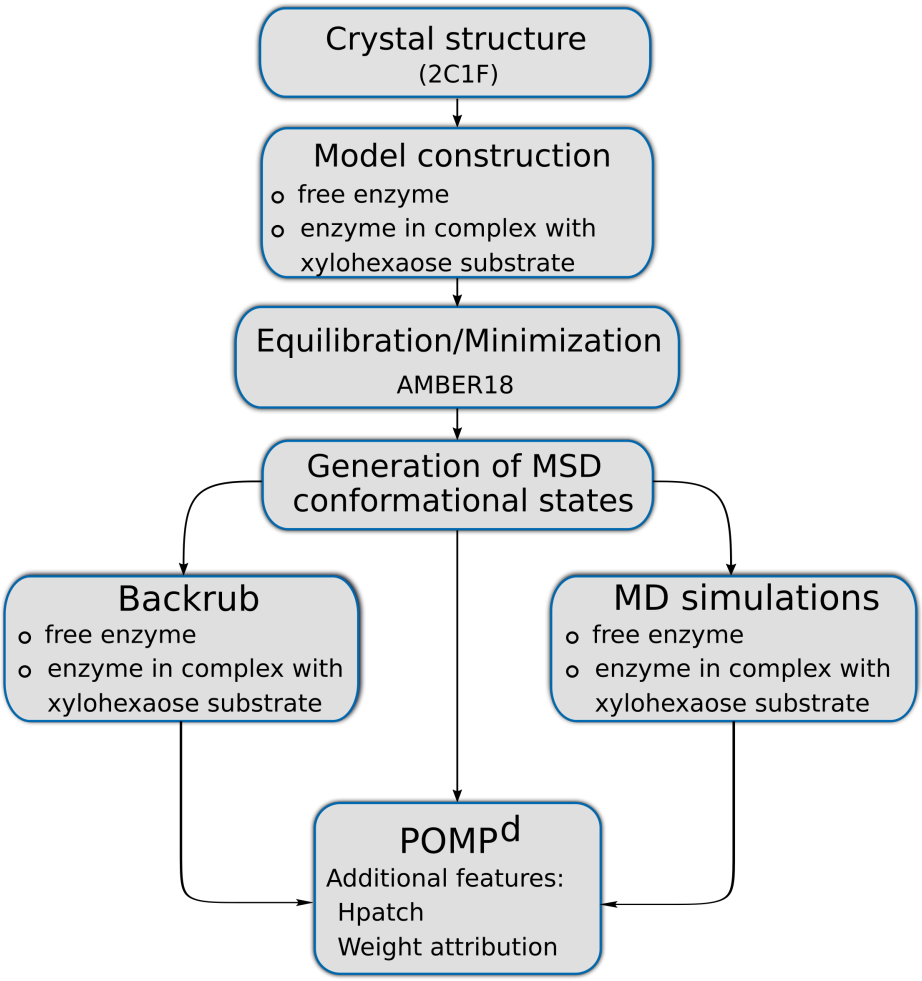

Figure 7.1: General workflow of the CPD procedure.

7.3.4 Activity assays on arabinoxylane

The activity of xylanase mutants was determined by measuring the release of reducing sugar from WAX with 3,5-dinitrosalicylic acid (DNS). The reaction mixture (1450 μL) was pre-incubated for 10 minutes at 37 °C and the reaction was initiated by the addition of 50 μL of enzyme (5 nM final), under orbital agitation (1600 rpm, Thermomixer, Eppendorf). Aliquots of 100 μL were regularly retrieved (t= 1, 2, 4, 6, 8, 10, 12, 14 min) and instantly mixed with 100 μL of DNS solution. At the end of kinetics, all the samples were heated at 95 °C for 10 min, cool downed and centrifuged (1500xg for 1 min) before 1 mL of milliQ water was added. Aliquots of 300 μL of each sample were transferred in a 96-well microplate and absorbance at 540 nm was read using a microplate reader (Infinite, M200 pro, Tecan, Männedorf, Switzerland). A xylose dyeset (0, 0.1, 0.2, 0.4, 0.5, 1, 1.5 and 2 g/L) was systematically carried out in parallel as a standard curve. All reactions were performed in triplicate.

7.3.5 Thermostability assay

To measure the thermostability of xylanase mutants, enzyme solutions were incubated at 60 °C for 50 min. At intervals of 10 min, samples of 70 μL were collected and stored at 4 °C. Samples were then used to measure residual activity on WAX using the DNS method described in the previous section.

7.3.6 Determination of melting temperature

CFX96 Real-Time PCR Detection System (Bio-Rad) was used with Fluorescence Resonance Energy Transfer (FRET) mode (excitation wavelengths: 450 - 490 nm and detection wavelengths: 560-580 nm). A TCEP (tris(2-carboxyethyl) phosphine) free PCR plate (in triplicate) was prepared by mixing 10 μM of protein, 50 mM phosphate buffer pH 7.5 and 2 μL of SYPRO Orange (2.5x). A triplicate with TCEP was also prepared (0.2 nM TCEP). The plate was filmed and centrifuged for 1 min at 4500xg. In the presence of TCEP, the plate was incubated for 1 hour at 4 °C. After centrifugation, SYPRO Orange was added.

7.4 Results and Discussion

*Np*Xyn11A is used in this study as a suitable candidate for different biotechnological applications achievable with GH11 xylanases. As described in the previous Chapter, *Np*Xyn11A possesses high specific activity and melting temperature which is 55.7°C. Here, our objective was to preserve particularly high specific activity of this enzyme while improving its thermal stability. To do so, we have used CPD approach with multistate design procedure. Modeling multiple conformational states offers several benefits. Taking multiple conformational states into account during a CPD procedure allows considering protein flexibility or modeling big conformational changes. In the case of enzymes, multistate design allowed us to take into account

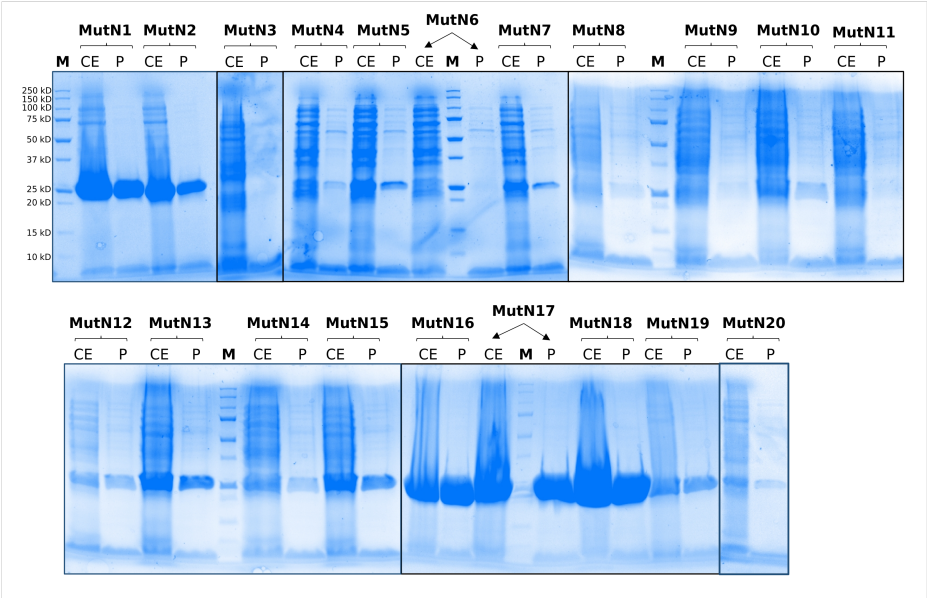

Figure 7.2: SDS-PAGE expression analysis of 20 mutants generated with CPD methods. CE: cell extract; P: purified protein; M: weight marker.

different systems simultaneously: the free enzyme and the enzyme in complex with the substrate. In our protocol, we have generated different conformational states of *Np*Xyn11A in its free enzyme and enzyme/substrate form and thus provided data for multistate CPD. We have analysed the typical β-jelly roll fold of *Np*Xyn11A and defined different important regions (Figure 6.1) in which design hotspots were determined.

20 Mutants generated with Pomp[d]

With our computational protocol, 20 sequences were generated containing between 7 and 12 mutations. Therefore, 20 *Np*Xyn11A mutants were submitted to experimental verification of improved thermal stability and preserved activity. The specific activity on the Wheat ArabinoXylan (WAX) substrate was measured as well as the melting temperature (T_m) which represents the temperature at which 50% of the protein is denatured/unfolded. In order to enable an easy comparison with the template sequence, all experiments were simultaneously done on the *Np*Xyn11A template. All of the mutants have been successfully expressed and purified. The purified mutants were electrophoretically homogenous (SDS-PAGE) with a molecular weight of 25kDa, which corresponds to the molecular weight of *Np*Xyn11A (Figure 7.2).

Specific activities have been examined for each mutant. Two mutants possess a specific activity equal to the one of the template sequence while three mutants have higher specific activity than the template. The other 15 mutants possess a lower specific activity than that of the *Np*Xyn11A template sequence (Figure 7.3A).

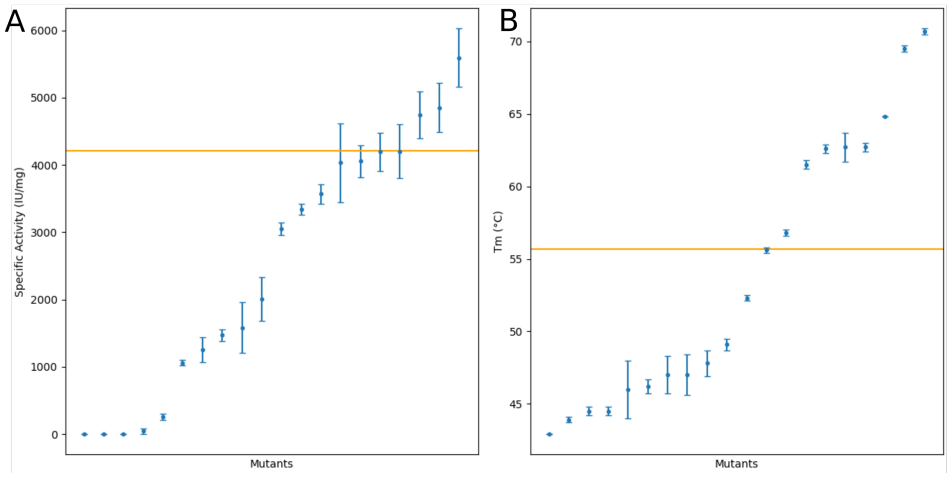

Figure 7.3: Specific activity (A) and T_m (B)of 20 mutants generated with POMP^d. Orange line represents the value of the template *Np*Xyn11A enzyme.

Results on the melting temperature (T_m) show that the T_m varies between 42.9 °C and 70.1 °C. T_m of *Np*Xyn11A template is 55.7 °C. As shown in Figure 7.3B, 8 mutants have a T_m value that is superior to 55.7 °C and thus possess significantly improved thermal stability. However, in four of these mutants the specific activity was reduced. This suggests that the mutations introduced in these sequences allowed a strong gain in terms of thermal stability but simultaneously introduced an important loss of specific activity. This is particularly the case for the mutant number 19 which has a T_m of 70.1°C and specific activity of 257 IU/mg (loss of 93%).

4 Mutants have improved thermal stability and catalytic activity

When both properties, specific activity and melting temperature, are taken into account, there are 4 very interesting mutants. As it is shown in Figure 7.4, 4 mutants possess higher T_m but also higher or equally good specific activity. Mutant number 17 exhibited 13.8 °C higher T_m compared to the template. This indicates that the mutations predicted with POMP^d in different regions of the enzyme promoted overall stability but also activity of this enzyme whose average specific activity was improved by 13%. Other three mutants also represent very interesting variants (Table 7.1) even though the improvement in T_m is slightly lower (varies between 5.8 and 9.7°C).

Thermal tolerance of these four mutants was determined by measuring their residual activity after 10 minutes of incubation at 60 °C. As shown in the Figure 7.5, resistance to temperature of these four mutants is confirmed. Compared to the template, all four mutants showed a clear improvement. Mutants 2, 16 and 18 have residual activity that is around 80%, compared to 40% of the template. Once again mutant number 17 shows the best results with residual activity of 100%. Number

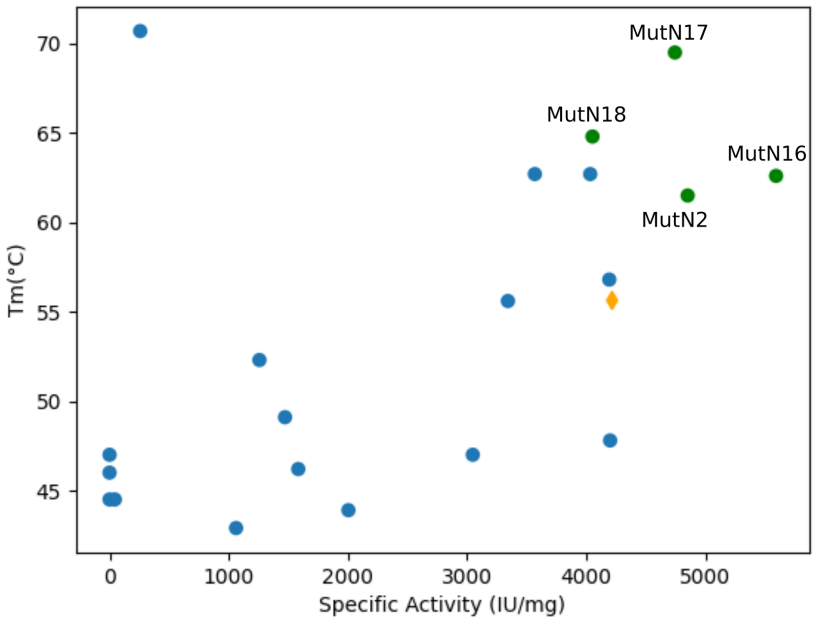

Figure 7.4: Analysis of the best mutants in terms of T_m and specific activity. The best variants (MutN2, 16, 17 and 18) are colored in green and the template *Np*-Xyn11A enzyme is colored in orange.

	Specific activity (IU/mg)	Residual activity (%)	T_m (°C)	Nb of mutations
Template	4209 ±266	40±15	55.7±0.2	-
MutN2	4853±362	81±3	61.5±0.3	8
MutN16	5595±433	86±5	62.6±0.3	10
MutN17	4746±347	100±3	69.5±0.2	9
MutN18	4054±242	87±5	64.8	8

Table 7.1: Specific activity (average of triplicate), Residual activity (average of duplicate) and Melting Temperature (T_m) of the four most interesting mutants

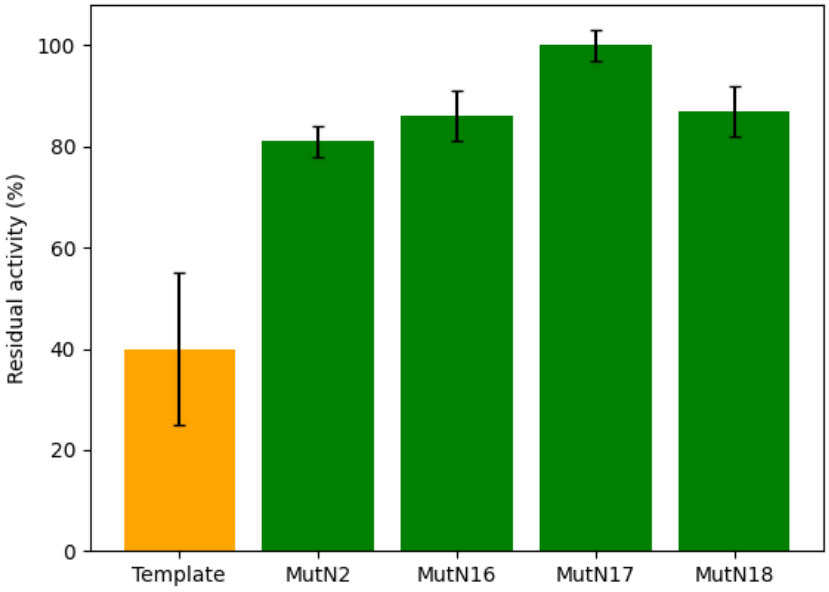

Figure 7.5: Residual activity of the 4 best variants (in green) and template *Np*-Xyn11A enzyme (in orange).

of mutations of each of these mutants compared to the template sequence is also shown in Table 7.1 which summarizes different properties of the 4 mutants that are considered as hits. A list of specific mutations for each mutant is given in Table 7.2. Their location is shown in Figure 7.6.

Mutant number 17 represents the most interesting variant with improved thermal stability by 14 °C, improved specific activity by 13% and 100% of residual activity. Predicted stabilizing mutations of this mutant have been analysed more in details in order to understand molecular basis for this improvement. On the basis of predicted sequence and 3D structure of *Np*Xyn11A template, we have constructed a 3D model of this mutant. Thus, different mutations were structurally analysed. Mutant 17 possesses in total 9 mutations. These mutations are mostly located in the N-ter, fingers, palm loop and helix regions. In the previous section, our MD analysis revealed that these regions are thermally sensitive. N-ter, fingers and the palm loop regions showed to be quite unstable while the region around the α-helix was the first to go through unfolding at 500K MD simulation. This mutant possesses two mutations that are located in the N-ter region of the enzyme. The first (N16H) allows the introduction of a salt bridge instead of the polar interaction between the asparagine 16 with aspartate at position 17 (Figure 7.8B). The replacement of asparagine by a positively charged amino acid, a histidine (N16H), enables the formation of a salt bridge between this histidine in position 16 and the negatively charged aspartate at position 17. This molecular interaction is stronger

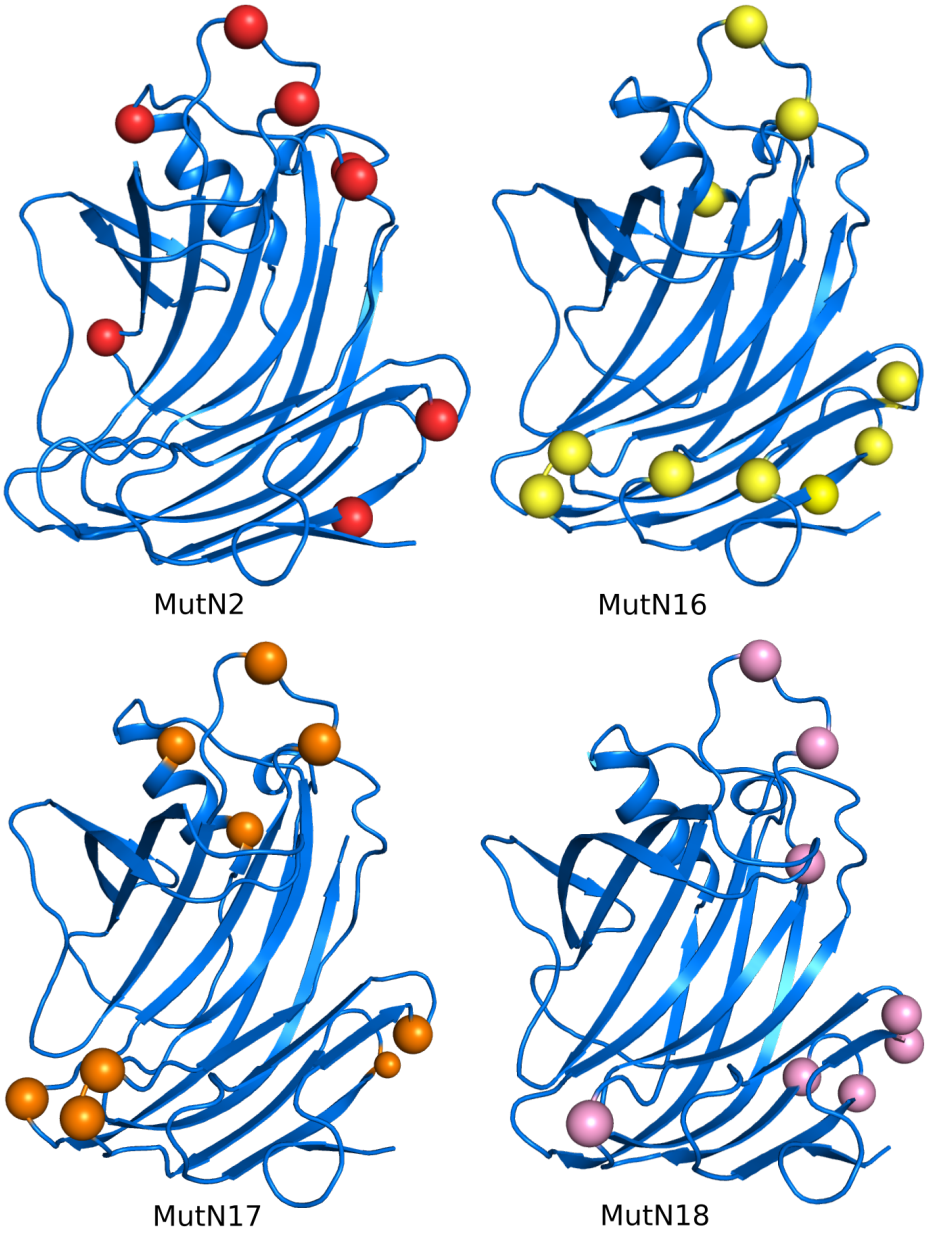

Figure 7.6: Location of stabilizing mutations for each variant are shown in the crystal structure of *Np*Xyn11A.

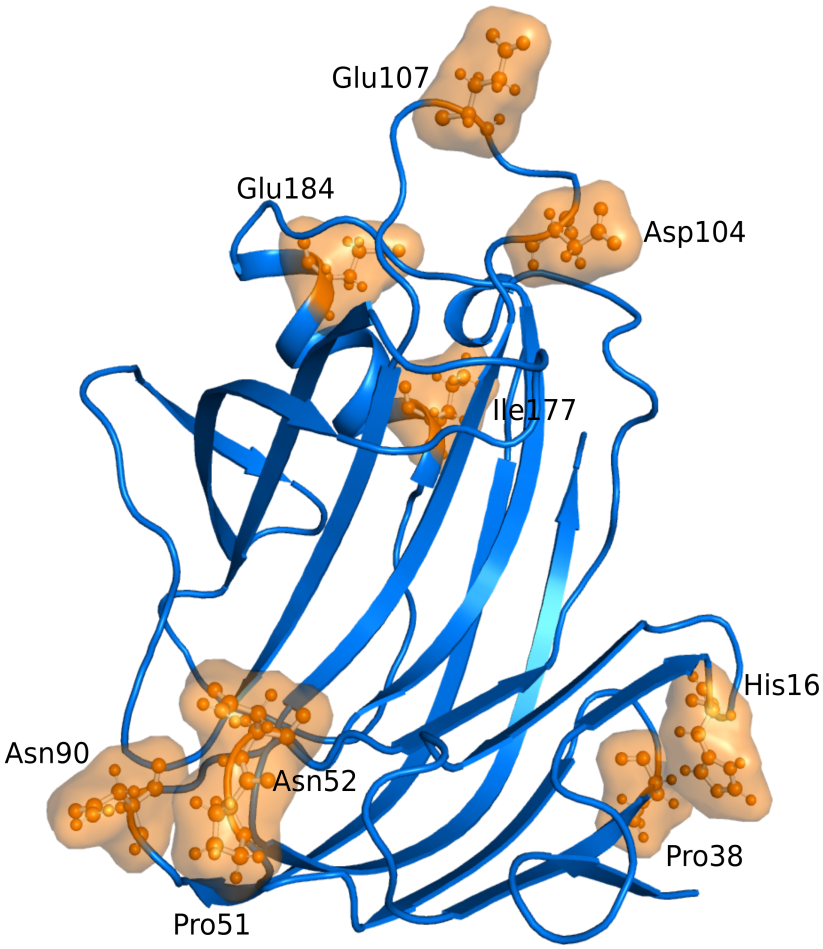

Figure 7.7: Model structure of the mutant 17, stabilizing mutations are shown in orange.

	Mutations
MutN1	N16H; F19Y; D62E; R104D; N107E; S177I; A184E; Y194H
MutN2	N16H; T35I; S65A; D73E; R104K; N107D; Q167E; Q186E
MutN3	N10H; N16F; T35I; D78R; T82I; R104D; N107E; Q166M; Q167D; S177I
MutN4	N16H; F19Y; S38E; S65A; A80S; R104D; N107E; G189P; L213K
MutN5	N16H; T35V; A49P; N51E; R52N; A89L; S90N; R104D; N107E; S177I
MutN6	N16H; N31R; K68P; D71A; R104D; N107E; K130L; Q167E; Q186L; V216I
MutN7	V15F; N16H; F19Y; S20F; S38E; D62Q; R104K; N107E
MutN8	V15T; N16H; F19Y; S20Y; N51P; N107E; A184E; V216I
MutN9	N16H; F19Y; T35I; N51P; D73E; R104D; N107P; D178E
MutN10	V15T; N16H; N51P; R52N; D62E; R104D; N107E; Y194H
MutN11	Q9K; N10H; N16F; T35I; S38P; S65A; Q66R; K68I; R104D; N107E; Q167D; S177I
MutN12	T35I; S38P; T41A; S65A; R104D; N107E; Q186E; L213V
MutN13	N16H; F19Y; S38P; T41A; D73E; R104D; N107E; A184E; Y194H
MutN14	Q9K; N27H; T35V; S38P; N51D; R52N; S177I; G189P
MutN15	S38P; T41A; N51P; R104D; N107E; Q167E; Y194H
MutN16	N10F; V15T; N27H; T35I; S38P; N51P; R52N; R104K; N107E; D178E
MutN17	N16H; S38P; N51P; R52N; S90N; R104D; N107E; S177I; A184E
MutN18	V15T; N16H; S38P; T41A; N51P; R104D; N107E; Y194H
MutN19	V15Y; N16H; F19Y; S20F; T35V; Y83F; R104K; N107D; Q167E; S177I; Q186M; V216I
MutN20	V15F; N16H; S20Y; S38E; D62E; A80S; R104K; N107E; S177I; D178E; Q186E

Table 7.2: List of introduced mutations in all of the 20 variants.

and probably stabilizes the N-ter domain of *Np*Xyn11A. The second mutation in this region corresponds to the substitution of a serine at position 38 by a proline (S38P). The region where the proline is introduced corresponds to a loop fragment which is connecting two anti-parallel β-strands and which is allowing the change of direction of the polypeptide chain. Loops as well as terminal tails are known to be the least rigid fragments composing secondary structure of proteins. Prolines significantly reduce the flexibility of the polypeptide chain by restricting rotation around the N-Cα bond to a relatively small region of conformational space [229, 230, 231]. Therefore, introduction of proline residue in this region may provide more rigidity. Another proline mutation (N51P) is also introduced in the fingers region. Once again, this mutation is found in a loop fragment connecting two anti-parallel β-strands. At position 184, alanine is mutated to a negatively charged glutamic acid. This mutation can certainly allow the introduction of another salt bridge (Glu184 with Lys68) located between the α-helix and the B3-A5 loop (Figure 7.8A). We have seen in the previous study that the region around the α-helix represents a hotspot where folding preferentially occurs. Also, in our comparison of *Np*Xyn11A with hyper-thermostable *Ev*Xyn11^{TS}, we have observed a presence of a salt bridge in this exact region. This salt bridge in *Ev*Xyn11^{TS} (Asp157-Arg56) was present 97.7% of time during 1μs MD simulation which suggested that this salt bridge plays an important role on the stability of this enzyme (Figure 6.17 and Table 6.2). Hence, the introduction by POMP^{d} of a salt bridge in this critical domain certainly

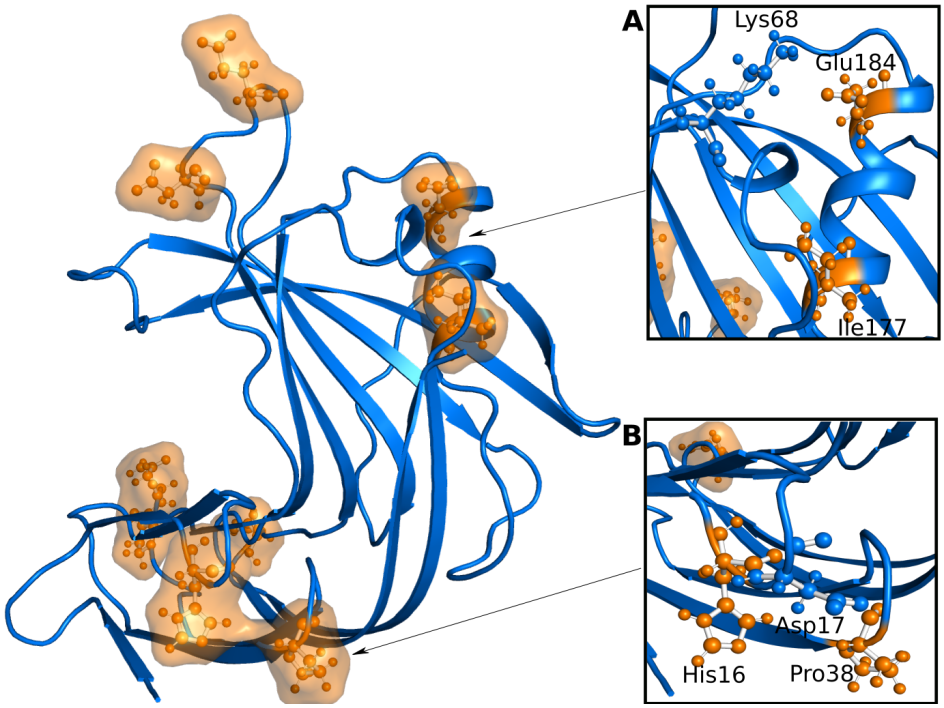

Figure 7.8: Location of two salt bridges introduced in mutant 17. **A** Salt bridge between the α-helix and B3-A5 loop. Interaction between mutated glutamic acid (in orange) and lysine (in blue) **B** Salt bridge in the N-ter domain. Interaction between mutated histidine (in orange) and asparate (in blue).

helps improve the thermal stability of *Np*Xyn11A. Furthermore, mutations R104D and N107E are found in the palm loop region. This region has previously (Chapter 6) been identified as particularly long in *Np*Xyn11A compared to other xylanases and also as one of the most flexible regions in *Np*Xyn11A. Therefore, introducing these mutations in the palm loop must lead to a reorganization of the polar and/or ionic interaction network, favorable for the stabilization of this very flexible region. Finally, at position 177 there is a substitution of a polar residue, serine, by apolar isoleucine (S177I). This mutation is probably important for the improvement of hydrophobic packing in the protein core and introduction of hydrogen-bonding interactions (Ile177-Lys181;Ile177-Ile75).

7.5 Conclusion

In this chapter, we exploit the knowledge obtained by a detailed MD analysis in order to define computational design strategies for *Np*Xyn11A. This analysis allowed targeting regions whose redesign may have a positive impact on thermostabilization of the enzyme. Protein flexibility, which is important for the enzyme's function, was taken into account in our MSD procedure. Our strategy aimed at proposing stabilizing mutations while trying to keep a certain flexibility. 20 sequences were gen-

erated and directly given for experimental verification. Experimental tests showed that out of 20 sequences, 8 possess better thermal stability. However, in some of these sequences catalytic activity was not preserved. When comparing both wanted properties, catalytic activity and thermal stability, 4 sequences were found to have improved properties. These four sequences were submitted to additional experiments of residual activity which showed that all of the 4 mutant variants possess better residual activity than the wild-type enzyme. One mutant was particularly interesting with 14 °C improved T_m, improved specific and residual activity. This mutant is considered as the most interesting variant and could be used as efficient biocatalyst in harsh conditions of industrial and biotechnological processes.

Chapter 8

Computational Design of a nanobody scaffold

Contents

8.1 Context

8.1.1 Antibodies

Antibodies are highly specialized protein molecules essential for the immune system. These proteins, also called immunoglobulins, are secreted by B-cells or expressed on the surface of their membrane. Antibodies identify and help neutralize foreign pathogens such as viruses and bacteria. The pathogen that is being targeted by an antibody is called the antigen. Antibodies are produced by the immune system in response to the presence of an antigen and every single antibody typically recognizes a specific foreign antigen. In order to provoke an efficient immune response and also avoid targeting self-proteins, antibodies must possess high affinity but also high specificity to their antigens. They have a particular quaternary structure. An antibody is a "Y"-shaped protein (Figure 8.1), formed by the association of two identical heavy chains and two identical light chains. These chains contain different domains called the Variable (V) or Constant (C) domains. Each heavy chain is composed of one variable domain (V_H) and several constant domains (C_H1,

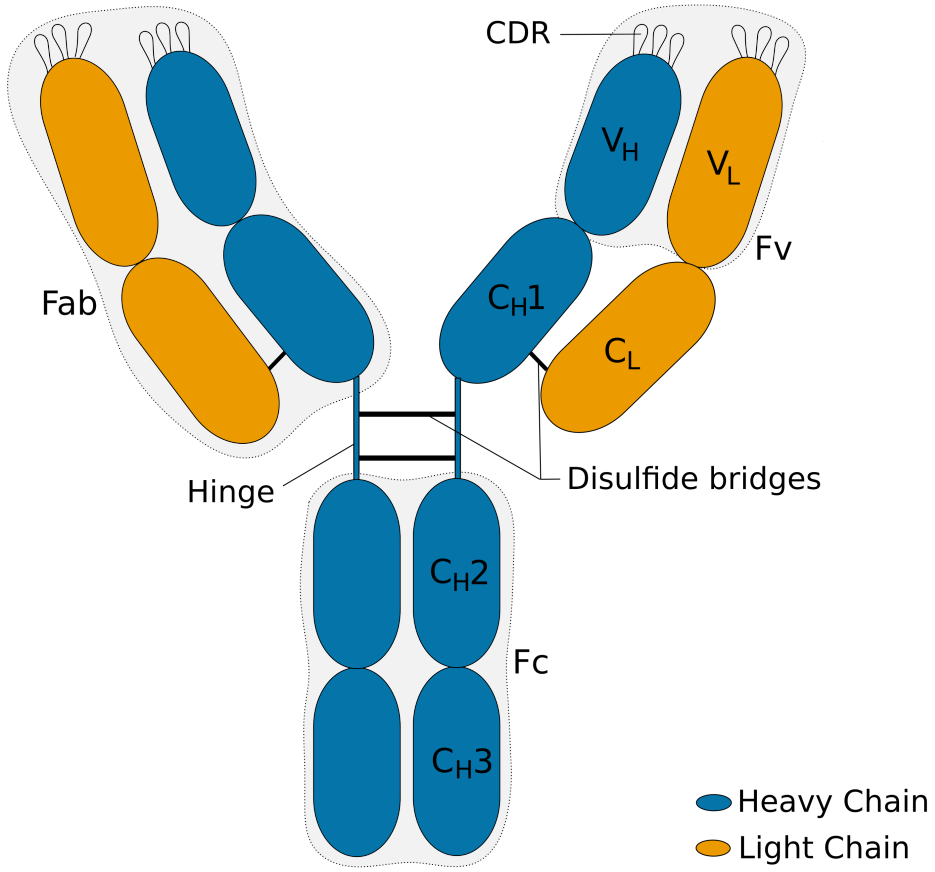

Figure 8.1: Schematic representation of a conventional antibody structure. Heavy and light chains are shown in blue and orange respectively. Fc region corresponds to the crystallizable fragment, Fab region to the fragment antigen binding domains, Fv to the variable fragments. CDR: Complementarity-Determining regions.

C_H2, C_H3). Each light chain contains one variable domain (V_L) and one constant domain (C_L). The tail region of the antibody, or the base of the "Y" is called the crystallizable fragment (Fc). This region binds to a variety of receptor molecules and is responsible for the activation of the immune system. There are also two fragment antigen binding domains (Fabs), at each side of the "Y", directly linked to the Fc region by a hinge region. Variable fragments (Fv), located at the tips of the Fab region, are composed of a pair of variable domains (V_H and V_L). These variable domains are the ones that directly interact with the antigen. Each variable domain contains three hypervariable loops (H1, H2, H3 for V_H and L1, L2, L3 for V_L). They are called Complementarity-Determining regions (CDRs) and are known for playing a crucial role in the antigen binding.

Because of their high specificity for the target antigen, and possibilities to bind

a wide range of potential antigens, antibodies have been revolutionizing the medical sector in the past decades. Therapeutic antibodies, that are nowadays being developed almost exponentially by the pharmaceutical industries, represent as a matter of fact years of research and development. The production of antibodies for diagnostic or therapeutic purposes has been revolutionized in 1975 by Georges Kohler and Cesar Milstein [232] by a method called hybridoma. This method, awarded in 1984 by a Nobel Prize, allows *in vitro* production of a large number of identical antibodies, called monoclonal antibodies. However, the limiting point of this approach remains the necessity to immunize the animal at the beginning of the process in order to provoke an immune response and retrieve the antibody producing B-cells. Also, in the late 80's, a multitude of mouse monoclonal antibodies were developed but reported disappointing therapeutic results especially by inducing the production of anti-mouse immunoglobulin antibodies (HAMA from Human Anti-Mouse Anti-bodies) in patients treated for cancer [233]. Remarkable progresses in genetic engineering and molecular biology enabled the cloning of genes that encode the heavy and the light chains of antibodies. In order to avoid undesirable immune reactions against injected mouse antibodies, combining DNA from mice with DNA from genes encoding human antibodies allowed the creation of antibodies that are closer to human antibodies. Today, thanks to this technological progresses, fully humanized antibodies are mostly developed. This type of antibodies are called recombinant antibodies because they are cloned in eukaryotic or prokaryotic expression vectors. Different forms of recombinant antibodies exist, but they usually represent antibody fragments that consist of one Fab domain or heavy and light chain of the variable region, also called single domain antibodies. Their reduced size can improve their bioavailability and facilitate their production by bacteria or yeasts. Indeed, it has been shown that the Fab domain possesses an increased capacity to penetrate dense tissues such as solid tumors, and the single-chain Fv domain seems to be even more effective [234]. Thus, some efforts have been done in order to further reduce the size of fragments into a monomeric single domain entity such as V_H or V_L domains only [235]. However, some properties such as solubility and affinity seem to lack in this type of antibody fragments. Recombinant antibodies also allowed the development of new type of antibodies that are directly expressed in living cells as intracellular antibodies, called intrabodies. In addition to classical antibodies found in mammalian species, llamas, other camelidae (i.e *Camelus dromedarius, Camelus bactrianus, Lama glama, Lama guanoco, Lama alpaca* and *Lama vicugna*) and sharks produce a considerable fraction of heavy-chain antibodies (HCAbs). This unusual type of antibodies that completely lack the light chain, is composed of three instead of four globular domains. Their antigen-binding site is formed only by a single domain called VHH (Variable domain of camelid heavy chain antibody) in camelidae and VNAR (Variable domain of the shark new antigen receptor) in sharks. Since it has been demonstrated that the VHH domain alone, cloned and expressed in bacteria, is a monomeric single domain antigen binding entity [236], many companies and research groups focused on their therapeutic applications [237].

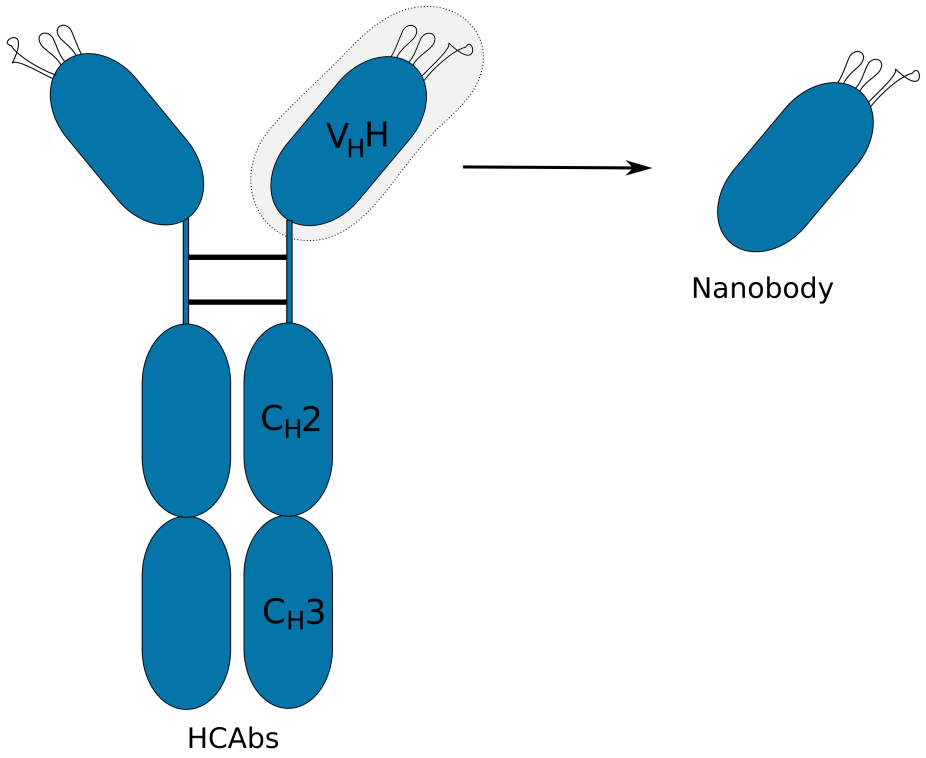

Figure 8.2: Schematic representation of a heavy-chain antibody (HCAbs) and Nanobody

8.1.2 Nanobodies

The scientific breakthrough that actually stands behind the basis of the "Nanobody technology" dates back to late 1980's. As in many scientific discoveries, serendipity played a crucial role in this one as well. At the Free University of Brussels, during a practical course, a group of students working on the extraction of antibodies from dromedary serum, discovered a new type of antibodies, they were smaller and did not correspond to anything that was known then. In 1993 it has been confirmed by Hamers-Casterman and his colleagues that camels, llamas and dromedaries contain a special type of antibody which does not contain a light chain [238]. As we just mentioned, this type of antibody is called heavy-chain antibody (HCAbs) and possesses a heavy chain that has lower molecular weight than the conventional antibody. As shown in the Figure 8.2 , the heavy chain of HCAbs does not contain three constant domains, but only two (C_H1, C_H2). Antigen-binding domain corresponds to the VHH region. When it was discovered that VHH can function as a single entity, because of its particularly small size (nanometer range), it has been named Nanobody.

Figure 8.3: Schematic representation of a nanobody sequence organisation composed of framework regions (FR1-4) and three CDR loops (CDR1-3). Mutations in the framework 2 region (stars) correspond to the residues that are substituted by hydrophilic residues in VHH compared to conventional antibody V_H region where these residues are hydrophobic. Orange lines represent disulfide bonds. There is the conserved disulfide bond between framework 1 region (Cys22) and framework 3 region (Cys97) and an additional interloop disulfide bond between CDR1 and CDR3 that is present in many dromedary VHHs

8.1.2.1 Structural and biochemical properties

The first crystal structure, in the early 2000's, enabled a detailed structural analysis of a nanobody [239, 240, 241]. It revealed that its structure is very similar to the one of the standard V_H domain. Just like V domain in the conventional antibodies, the VHH domain contains three CDRs which are connected by four framework regions (Figure8.3). They display a typical IgV fold with nine β strands and contain a conserved disulfide bond between the framework 1 and 3 which stabilizes the structure.

There are several very important features that differentiate structurally similar V_H and VHH domains [242]. The architecture of CDR loops is more diverse than in standard V_H domains. In fact, the loop CDR3 seems to play a crucial role for the antigen-binding as it is much longer in nanobodies than in standard V_H. In standard antibodies, six loops of V_H-V_L contribute more or less equally to antigen-binding, while in nanobodies, it has been shown that CDR3 loop dominates in the antigen-binding [243]. The antigen-binding paratope of the VHH domain has usually a convex form and the binding typically occurs in protein clefts or at the domain-domain interfaces. The CDR3 loop, which possesses greater flexibility due to its longer size, can sometimes be stabilized by an additional disulfide bridge formed between CDR1 and CDR3 (Figure 8.3). The CDR3 loop can also fold over the framework 2 region and thus form a flatter paratope. This enables nanobodies to bind an antigen in many different ways while having only three CDR loops. The affinity with which nanobodies bind their targets is very similar to the affinity detected in the binding of conventional antibody [244].

Also, even though the sequence homology with the human V domain is particularly notable, there are some important difference within the framework 2 region of VHH [245, 244] sequence. In fact, in conventional antibodies, framework 2 region is mainly composed of hydrophobic residues (such as Val37, Gly44, Leu45 and Trp47). Conserved hydrophobic residues at these positions exist because they facilitates the pairing with V_L domain and thus form a hydrophobic interface with V_L. These hydrophobic residues are substituted by hydrophilic residues in VHH (Val37→Phe or

Tyr, Gly44→Glu or Gln, Leu45→Arg and Trp47→Gly, Phe or Leu), which makes the former V_L interface more hydrophilic. It has been shown that these substitutions contribute to VHH's high solubility with low aggregation propensity [246, 247]. The presence of hydrophilic amino acids in the framework 2 region is also detected in VNAR shark domain. VNAR and VHH sequences are quite different, however, the presence of polar and charged residues in the framework 2 (the V_L side of the domain) highlights structural and functional convergent evolution in this region [248, 249].

Another important characteristic of nanobodies is their high stability. It has been shown that nanobodies display high T_m values (60-80°C) and can also retain their functionality after exposure to elevated temperatures (up to 90 °C) [247, 250, 251].

8.1.2.2 Generation of synthetic nanobodies

Although immunization techniques generally provide high affinity antibodies because of the benefit of the affinity maturation in the immune system of the host, these conventional cloning techniques are quite confined. They depend on animal experimentation (immunization phase for each antigen of interest), and are limited by natural immunogenicity or toxicity of antigens. Synthetic nanobody libraries confront these limitations by using totally *in vitro* techniques. They offer greater diversity and so access to larger repertories. Recently, a synthetic VHH library has been developed [252]. This library called "NaLi-H1: Nanobody Library Humanized 1" uses a humanized scaffold that has been selected for its stability and ease of expression. Even though stabilized nanobodies [253] and libraries [254] have been described before, this library is the first synthetic library that produces at high frequency functional intrabodies, while still containing two canonical cysteine residues.

8.1.2.3 Applications with nanobodies

The unique properties of nanobodies such as high stability, affinity, small size and ease of modification have enabled diverse applications. These applications range from fundamental research to diagnostics and therapeutics. When they are expressed as intrabodies, with the ability to be stable in the reducing cytoplasmic environment, they can serve as tools to trace and visualize antigens [249]. Thus, with nanobodies it is possible to target protein-protein interactions, disrupt signaling pathways, or directly observe and follow protein dynamics [255]. Nanobodies can also be utilized as tools to crystallize proteins. Used as crystallization chaperones, they can serve to investigate different protein conformational states [256]. Their small size also allows them to be used in super-resolution microscopy. By using GFP with nanobodies, it was possible to analyze dynamics of microtubules, living neurons and yeast cells [257]. Nanobodies have also been used as probes in biosensor applications or as *in vivo* imaging agents in imaging techniques such

as radionuclide-based, optical and ultrasound [258]. Therapeutically, nanobodies have been used as antagonistic drugs, or as targeting moieties of drug delivery systems [259]. Many nanobodies are under clinical trials for a very wide range of human diseases such as inflammation, infectious diseases, cancer therapy for brain tumor, breast tumor and lung diseases. A more detailed presentation of diverse nanobodies applications can be found in recent scientific review articles [260, 244, 261].

8.2 Motivations

Computational Design methods have been previously applied on antibody design. Methods such as OptCDR [262], OptMAVEn [263], Abdesign[264] and RosettaAntibody design [265] can be categorized as *ab initio* methods that aim at designing new paratopes to improve antibody stability and affinity [266]. Some of the successful antibody designs have been already mentioned in the Computational Design section of this manuscript. Here, we do not aim at redesigning paratope regions but at creating new optimized scaffold that could be universal for many CDR loops. Recently a novel nanobody scaffold has been designed, based on conserved framework sequences and starting from a sequence dataset of llama VHHs. This scaffold has been validated by grafting the CDRs from two known nanobodies and seems exploitable as universal scaffold for specific VHH bacterial expression and for the construction of a large ($> 10^{12}$ individual members) ribosome display DNA library [267]. However, this nanobody scaffold was not obtained using an automated and generalizable computational approach.

In this chapter we aim at using our computational design tools to design new universal nanobody scaffold that could potentially allow the development of new synthetic library of nanobodies. Our starting point was the humanized nanobody scaffold created by our colleagues from the Cancer Research Center of Toulouse (CRCT) in 2016. This scaffold, optimized for intracellular stability was used for the development of a highly diverse library which provides high affinity binders without animal immunization [252]. This "NaLi-H1" library was screened against various targets, and highly specific antibodies were selected against EGFP, mCherry, b-tubulin, b-actin, heterochromatin protein HP1a, GTP-bound RHO, p53 and HER2. The authors showed that this nanobody scaffold is usable as fluorescent intrabody to track antigens in cells. Overall, in this study our colleagues reported for the first time a large and diverse synthetic single domain antibody library enabling fully *in vitro* selection of highly functional antibodies and intrabodies [252]. A very important notion about intrabodies remains their dependence on the stability of antibody fragments in the reducing environment of the cytosol which does not allow a formation of the disulfide bond. Despite the presence of two canonical cysteine residues, their library produced functional intrabodies at high frequency. This scaffold was further used in another study for engineering an analytical tool to selectively degrade the GTP-bound form of endogenous RHOB. A phage display library

with nanobodies that bind to RHOB-GTP was enriched, and cell-based assay for screening the protein degradation of antigen/intrabody complexes was developed. Our colleagues from CRCT identified several intrabodies that recognize RHO-GTP proteins, and characterized one nanobody that showed greater selectivity to RHOB-GTP [268]. The crystal structure of this complex has been obtained and is available in the PDB. However, this nanobody scaffold along with the commercial use of the library is under a patent application (filled under ref: WO/2015/063331). Our objective was to create new cysteine-less nanobodies for ease of intracellular expression. This scaffold should be more or as stable as the old one and the sequence should be beyond the patent framework. Therefore, in this chapter we present the methods that have been employed in order to create this new cysteine-less nanobody scaffold. We show different techniques of *in silico* evaluation that have been done in order to chose between the most promising designs, that were further experimentally tested by our colleagues at CRCT: Claudine Tardy, Coralie Morand, Patrick Chinestra and Aurélien Olichon.

8.3 Materials and Methods

8.3.1 Computational Design

Preparation of the initial template

The crystal structure of human RHOB-GTP in complex with nanobody B6 was used for the preparation of our initial nanobody template (PDB code: 6SGE). The 3D structure of RHOB-GTP (chain A) was removed and only the 3D structure of the nanobody (chain B) was kept for further studies. The initial nanobody template contains 126 residues.

Generation of conformational ensemble

In order to generate conformational states of the initial nanobody template structure, we used Rosetta Backrub protocol as described in Chapter 4, and generated 100 protein models. Protein models were then clustered using Durandal software [146]. A clustering radius of 0.3 Å was used to obtain the cluster centers of the four biggest clusters.

Design strategies

Different multi-state and single-state design strategies were used for the redesign of the nanobody scaffold. Here we present the 6 multi-state strategies that led to experimentally characterized artificial nanobody scaffolds. It is well-known that the presence of hydrophobic patches on the surface of a protein usually leads to poor expressability. We confirmed this for nanobodies in a quick pilot experiment where we quickly redesigned the target nanobody using a standard single state design approach and Rosetta energy. Design was poorly expressed and poorly purified with signs of aggregation, while the WT sequence was purified with high yield and more than 95% purity. Therefore, we concentrated on defining strategies principally with

multistate design approach but also by explicitly trying to prevent formation of hydrophobic patches at the protein surface. CDR loops are never designed because their composition of amino acids is essential for specific recognition. The following list provides an explanation and preparation details on each of the 6 strategies. Unless stated otherwise, each design strategy used an ensemble of four conformational states generated by Rosetta Backrub. Cysteine residues are always mutated as the objective of the design is to create a cysteine-less nanobody scaffold. In each strategy, residues that are not allowed to mutate are considered as flexible (this includes CDR loops as well). This means that their side chains can adopt any rotamer conformation available for the natural amino acid types in the rotamer library. Mutable residues are allowed to mutate to any of the 20 natural amino acids.

1. **Design by forbidding mutations of conserved residues and VH destabilization hotspots**
   In this strategy a total of 49 residues were allowed to mutate. 40 residues identified as "conserved" in an alignment of all crystallized VHH domains were forbidden to mutate. Another 5 residues identified as VH destabilizing hotspots [269] were also forbidden to mutate. Figure 8.4 represents this first design strategy on the nanobody template structure with mutable residues shown in blue, forbidden residues from the scaffold shown in red and CDR loops in gray.

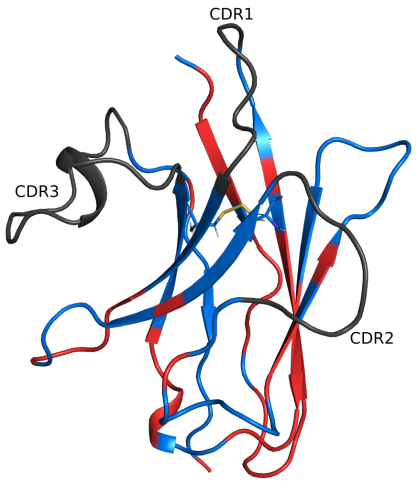

Figure 8.4: First design strategy: forbidding mutations of conserved residues and VH destabilization hotspots. Illustration of the design strategy on the template 3D structure of the nanobody. Residues that are allowed to mutate to any of the 20 amino acids are shown in blue, residues from the scaffold that are forbidden to mutate are shown in red and CDR loops are shown in gray. The disulfide bridge (also allowed to mutate) is represented in sticks.

2. **Design by forbidding mutations of VH destabilization hotspots**

In this strategy, 90 residues are allowed to mutate. Here, in comparison the the previous strategy, only VH destabilization hotspots are forbidden to mutate. Figure 8.5 represents this strategy.

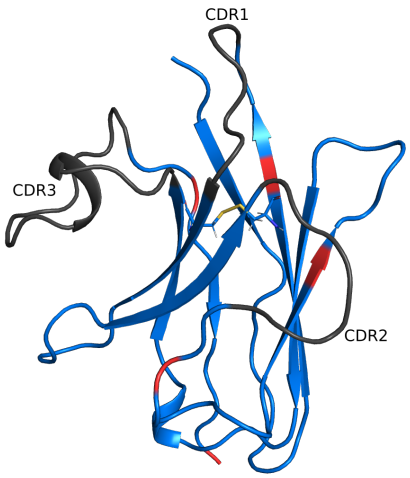

Figure 8.5: Second design strategy: forbidding mutations of VH destabilization hotspots. Illustration of the design strategy on the template 3D structure of the nanobody. Residues that are allowed to mutate to any of the 20 amino acids are shown in blue, residues from the scaffold that are forbidden to mutate are shown in red, and CDR loops are shown in gray. The disulfide bridge (also allowed to mutate) is represented in sticks.

3. **Design with diverse CDR loops ensembles**
   For this strategy, our collaborators from CRCT provided 6 sets of new CDR loops sequences. These loops are already known to be functional on our template nanobody structure as they derive from the synthetic VHH library "NaLi-H1" [252]. However, 3D structures with these diverse CDR loops do not exist. Therefore, we generated structure models of these CDR loops using I-TASSER[67]. We have first separated CDR loops in 3 groups according to their sequence length. For each group, we have generated one model using one set of CDR loops. 3 models in total were generated with I-TASSER web-server, one for each group, and other sequences were mapped on the model generated for their group. A total of 6 new models with diverse CDR loops was obtained (Figure 8.6). Short MD simulations of 20 ns at 310K were then performed on each of the 6 models with Amber ff14SB force-field [26]. For each model, the conformation from the last frame of the MD simulation was taken. It was then prepared with FastRelax and Rosetta `beta_nov16` scoring function [32] for design procedure with PROMP^d. The objective of this design strategy was to model and take into account multiple and diverse CDR loops in order to increase the chances of designing a new universal nanobody scaffold. Thus, in this strategy, instead of taking multiple conformational states of the

same nanobody structure/sequence, we perform multistate design by taking multiple nanobody structures with diverse CDR loops. This implied some changes in POMP^d. Initially in POMP^d, amino acid type constraints were added for each variable in the CFN, in order to ensure that each variable has the same amino acid type in each conformational state. An upgrade of this model was made by separating variables in two categories: mutable variables and flexible variables. Hence, constraints need to be added only for mutable variables. This modification allowed us to take into account sequences which possess different length of flexible variables. Here, the 6 previously described structures were given to POMP^d as 6 conformational states. All residues in the scaffold (95 residues) were allowed to mutate.

Figure 8.6: Third design strategy: design with diverse CDR loop ensembles. We show 6 generated models, each having a unique set of CDR loops (shown in different colors). All residues of the scaffold are allowed to mutate to any of all 20 natural amino acids (blue part of the structure).

4. **Design by allowing mutations at all positions (except CDR loops)** In this strategy we also allow all mutations. In total, 95 residues are allowed to mutate.

5. **Design by not allowing hydrophobic mutations observed in 4.** This strategy is guided by the solution/sequence obtained for the previous strategy (strategy 4). In order to avoid hydrophobic patches at the protein surface, we have visually inspected the surface of the nanobody and the impact of mutations introduced using the strategy number 4 (Figure 8.7). 14 hydrophilic surface residues from the WT sequence are mutated into hydrophobic residues using strategy number 4 (the location of these residues is represented by spheres in Figure 8.7). In this new strategy, we disallow these

14 residues to mutate and impose the native residue from the WT sequence instead. Therefore in this strategy, 81 residues are allowed to mutate and 14 selected residues are kept flexible.

Figure 8.7: Fifth design strategy: not allowing hydrophobic mutations observed in the design strategy 4. Representation of the nanobody structure and surface of the variant designed with strategy number 4. Hydrophobic surface of this design is shown in red as well as 14 residues that were mutated into hydrophobic amino acids. For these 14 residues (locations represented in red spheres), the type of the amino acid found in the WT sequence is now imposed. The rest of the scaffold is allowed to mutate to any of the 20 natural amino acids.

6. **Design with new Hpatch option**
   In this design strategy, we activated the hpatch constraint that was described in Chapter 5 to prevent the formation of hydrophobic patches in an automated manner and we allow all (95) residues to mutate.

Computational Nanobody Design
For each design strategy a multistate design procedure was performed with POMP^d. Each input (conformational state) was submitted to an additional relaxation step using RosettaFastRelax with harmonic constraints on backbone atoms.Pairwise energy matrices were computed with Dunbrack2010 rotamer library [150] and `beta_nov16` scoring function [32], using PyRosetta [151].

8.3.2 *In silico* evaluation of designed nanobodies

Each of the designed sequences was mapped on the initial nanobody template in order to obtain a model of designed variants. Each of them was further evaluated

in silico with two molecular modeling methods: Molecular Dynamics Simulation and Forward Folding.

Molecular Dynamics Simulations
MD simulations were performed on the designed sequences as well as on the WT nanobody with the following MD protocol: MD simulations were performed using AMBER ff14SB force-field [26]. To obtain a neutral charge of the simulated systems, a number of counter-ions were included. Each protein together with the counter-ions was solvated with TIP3P water molecules, using an octahedral box [201] with a minimum distance of 10 Å between the solute and the simulation box edges. The system was energy-minimized with a restraint potential of 25kcal/mol/Å^2 on the solute. This minimization consisted of 500 steepest descent steps, followed by 500 steps of conjugate gradient. The entire system was then gradually heated from 100K to 310K during 100ps with the same harmonic positional restraints of 25kcal/mol/Å^2 on the solute atoms. Energy-minimization and equilibration of the system has been done during 100 ps in the NVT ensemble and the positional restraints have been gradually removed and followed by production MD run of 20 ns. During the production run, MD simulations were carried out at constant temperature (310K) and pressure (1bar) using Berendsen algorithm [41]. Each production run has a time-step of 2 fs, periodic boundary conditions, a 9 Å cut-off for nonbonded interactions, and the Particle-Mesh Ewald (PME) method for treating long range electrostatic interactions [202]. SHAKE algorithm [203] was used to constrain hydrogens.

***Ab initio* forward folding**
Forward folding experiments were performed on each of the designed sequences with EdaRose software [270]. EdaRose is an *ab initio* fragment based protein structure prediction software. Forward folding techniques in general, aim at assessing the quality of a protein design by predicting whether it will fold in the target structure or not. The advantage of *ab initio* structure prediction methods is that they predict protein structures exclusively based on their amino acid sequences. However, their drawback remains their difficulty to deal with the astronomical size of the conformational search space. This is the reason why *ab initio* structure prediction methods are more efficient on small proteins (up to 150 residues). In this study, the nanobody is 126 residues long, which makes its *in silico* evaluation with *ab initio* forward folding method adequate and possible. 60 000 protein models were predicted for each design using EdaRose software with default parameters, and RMSD to the template structure was computed for the 1000 top scoring models.

8.3.3 Experimental validation

Plasmids
All nanobodies sequences were gene synthesized (Twist Bioscience) and cloned using NcoI and NotI into bacterial expression vector pAOT7-hs2dAb-6His-Myc-6His

for monovalent expression, or into the pFUSE-rIgG-Fc (Invivogen) for bivalent expression with a rabbit IgG Fc.

Protein expression and purification

2SHA-RHO protein purification: 2SHA-RHOA or 2SHA-CDC42 were expressed in BL21(DE3) E.coli cells from a pET vector as previously described 26. Transformed bacteria cells were used to grow 3mL LB-carbenicillin (100 μg/ml) cultures overnight at 37°C prior to inoculation in baffled flasks containing 1 L of the same media. Cells were allowed to grow at 37°C until OD600 reached 0.5-0.7. Cells were then induced with IPTG at a final concentration of 100 μM and grown for an additional 20 hours at 25°C. Cells were harvested by centrifugation at 4000g for 20 min. The pellets were resuspended in lysis buffer (50 mM TrisHCl pH 8, 150 mM NaCl, 5 mM MgCl2, 0.1% triton, 1mM DTT, 1X lysozyme and DNase I, protease inhibitors) and lyzed by sonication on ice prior to centrifugation (30 min, 15000g, 4°C). Strep-Tactin® SuperFlow Plus (IBA) matrix was equilibrated in buffer A (50 mM Tri-sHCl pH 8.0, 150 mM NaCl, 5 mM MgCl2) and was incubated with supernatant for 2 hours at 4°C. Then supernatant and matrix were loaded on a simple column in order to maximise capture of 2SHA fused proteins. Matrix was washed by 15 mL of washing buffer (50 mM TrisHCl pH 8.0, 300 mM Nacl, 5 mM MgCl2, 0.1% tween20). RHO proteins were then eluted in buffer A containing 10 mM Biotin (Sigma). Dialysis was proceeded overnight against buffer A containing 15% glycerol.

For nanobody production, cytosolic expression of Hs2dAb-6His-myc-6His was performed in BL21(DE3) E.coli cells from the pAOT7 vector [252]. Transformed bacteria cells were used to grow 3 mL TB-kanamycin (35 μg/mL) cultures overnight at 37°C prior to dilution of the pre-culture in baffled flasks containing 1 L of the same media. Cells were allowed to grow at 37°C until OD600 reached 0.5 to 0.7. Cells were then induced with IPTG at a final concentration of 100 μM and grown for an additional 16 h at 20°C. Cells were harvested by centrifugation at 4000g for 20 min. The pellets were re-suspended in lysis buffer (50 mM Na2HPO4 pH 8.0, 300 mM NaCl, 1X lysozyme and DNase I, protease inhibitors) and lysed by sonication on ice prior to centrifugation (30 min, 15000g, 4°C). The protein extract was incubated for 2 hours in the presence of complete His-Tag purification beads (ROCHE®, Basel, Switzerland) previously equilibrated with an equilibration buffer (50 mM Na2HPO4 pH 8.0, 300 mM NaCl, 10 mM imidazole). The beads were washed with 30 mL of washing buffer (50 mM Na2HPO4 pH 8.0, 300 mM NaCl, 10 mM imidazole). Hs2dAb were then eluted with elution buffer (50 mM Na2HPO4 pH 7.0, 500 mM NaCl, 300 mM imidazole) and dialyzed against PBS containing 20% glycerol for 16 hours at 4°C, and purity was assessed by SDS-PAGE fol-lowed by InstantBlueTM (Expedeon, Cambridgeshire, UK) Coomassie staining.

Bivalent hs2dAb were produced as fusion proteins with the Fc domain of Rabbit IgG2. hs2dAb were sub-cloned in pFuse-RIgG-Fc2 plasmid (NcoI/NotI) inframe between the interleukin-2 (IL2) secretion signal and the Fc domain [271]. 4 days

after transient transfection in HEK293T cells seeded in 12 well plates, supernatants were recovered and used directly in the ELISA assay.

Immunofluorescence
HeLa S3 cells expressing Histone H2B-GFP were grown on coverslip for 24 hours then fixed in 3% paraformaldehyde and permeabilized with PBS (plus 0.05% saponin and 0.2% BSA). hs2dAbs were co-incubated with 9E10 anti-Myc tag monoclonal for 90 min on cells. Cells were then washed quickly twice and incubated with secondary antibodies for 30 min (Invitrogen - Thermofisher).

ELISA assays
Wells of strepTactin coated plates (IBA®, 2-4101-001) were coated with 100 nM of recombinant proteins 2S-HA fused RHOAQ61L, RHOT19N or CDC42Q61L (200 μl in TBS by well) during 2 hours at room temperature (RT) and then blocked with 5% milk in TBS-Tween 0.05% (blocking buffer) for 1 hour at RT. Several dilutions of hs2dAb 6His-Myc-6His in blocking buffer were applied to the ELISA plates in duplicates for 1 hour at RT. Next, we added 1 μg/ml anti-myc HRP antibody (QED Biosciences®, #18824P) in blocking buffer for 1 hour at RT. Alternatively, hs2dAb Rabbit-Fc 6 fusion secreted in HEK293T supernatant were diluted 1/1 in blocking buffer and further detected using Goat anti-Rabbit HRP-conjugated secondary antibody (Sigma). Plates were washed three times with washing buffer (TBS containing 0.05% (v/v) Tween 20) after each step. The reaction was revealed by the addition of 100 μl chromogenic substrate (Thermoscientific®, 1-step ultraTMB, #34028) for 1 min. The reaction was stopped with 50 μl H 2 SO 4 1N and absorbance at 450 nm was measured using FLUOstar OPTIMA microplate reader. All steps are performed under agitation (400 rpm).

8.4 Results and Discussion

8.4.1 *In silico* analysis of selected designs

In silico evaluation of our designs consisted of MD simulations and forward folding experiments. After the short MD simulations, RMSD values of backbone atoms relative to the starting structure were calculated for each design. Also, per-residue B-factors along the MD trajectory were calculated from RMSF values, on all backbone atoms. For each design, these results were compared with the WT nanobody on which the same procedure and calculations were performed. Furthermore, forward folding experiments were carried out in order to assess the chances that the design sequences possess to fold into the target structure. Out of 60 000 generated structures, 1000 lowest energy structures were taken. Predicted structures plots show their energy as a function of their RMSD. The lowest the energy and the smallest the RMSD, the better the chances for the designed sequence to fold into the target structure. All these results are presented in Figure 8.8, where a graph

compares RMSD values with the WT nanobody, another compares B-factor values with the WT nanobody, and a last one shows forward folding results for each selected design.

We can see that MD profiles are generally very similar to the WT MD profile, which seems to be very stable. The RMSD profiles suggest that designs 2, 4, 5 and 6 are a bit less stable than the WT. However, all of them are stabilized after few nanoseconds. In general 4 peaks are observed in the B-factor analysis. 3 of them correspond to the 3 CDR loops (shown on the b-factor graph of design 1 in Figure 8.8). The fourth peak, observed between the CDR2 and CDR3 loops correspond to a flexible loop that is in the vicinity of CDR loops. We can see that this loop is generally more flexible in the designed sequences. This is particularly the case for design 2, but similar observations can be found for designs 4 and 5. In some designs, CDR loops also tend to be more flexible, such as CDR1 in design 2 or CDR2 in design 4. On the contrary, designs 1, 3 and 5 seem to be more or as stable as the WT nanobody.

Forward folding figures show a cloud of points representing the RMSD to native of structural models as a function of energy. This cloud of points can be interpreted in order to evaluate the folding propensity of sequences. We consider an evaluation as successful if models at a distance of around 5Å (or less) from the native structure are present among the lowest energy models. 5Å may seem a lot, but it is important to note that any structural knowledge of homologous sequences has been excluded from the prediction process. Therefore, designs 3, 4 and 6 possess satisfying forward folding profiles with promising results. We can see that some of their lowest energy structures have RMSD values around 5Å which means that these sequences are likely to fold. On the contrary, designs 1, 2 and 5 do not present very good profiles. Their lowest energy models are around 8 Å distance from the native structure.

8.4.2 Sequence screening and experimental characterisation

8.4.2.1 Sequence screening

The goal of this study was to design a new universal nanobody scaffold that could potentially allow the development of new synthetic library of nanobodies. Thus, the main objective was to create a cysteine-less nanobody scaffold which would, in comparison with the WT nanobody, have preserved or improved biochemical properties such as stability and solubility. Experimental validation in this study was under different time and budgetary constraints and was possible for only 6 computationally designed sequences. Diverse nanobody sequences have been generated with CPD techniques. Each of them was generated with different design strategies, by using SSD, MSD or options such as Hpatch. 6 sequences were chosen based on their sequence profiles, MD analysis and forward folding results (Figure 8.9) and were further experimentally tested by our colleagues from CRCT.

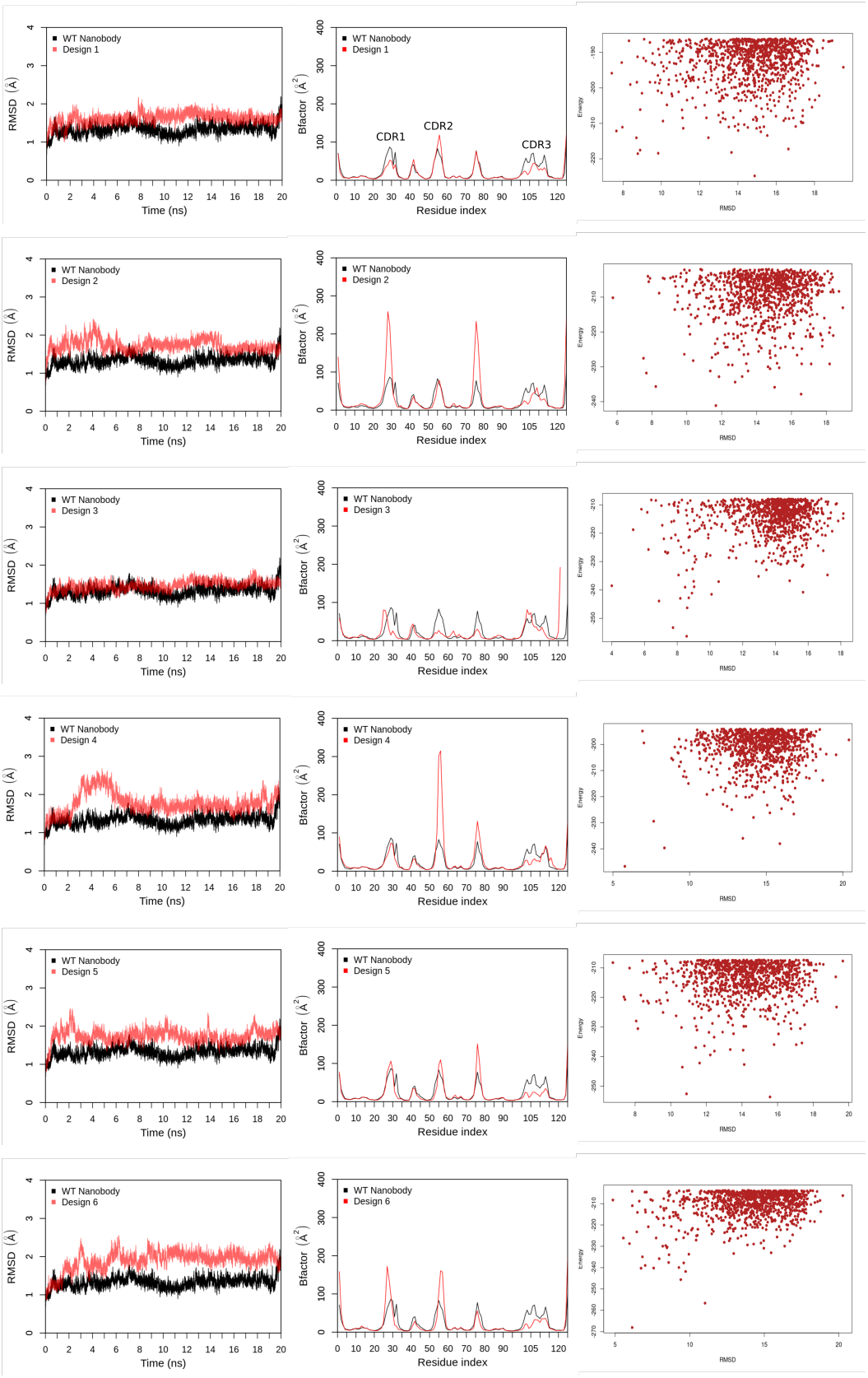

Figure 8.8: *In silico* evaluation of designed sequences with MD simulations and Forward Folding experiments. Each mutant is evaluated in terms of backbone RMSD profiles (left), per-residue average B-factor profiles (center) and forward folding profiles (right). For the forward folding evaluation, 60 000 protein models were predicted for each design. RMSD to the template structure was computed for the 1000 top scoring models.

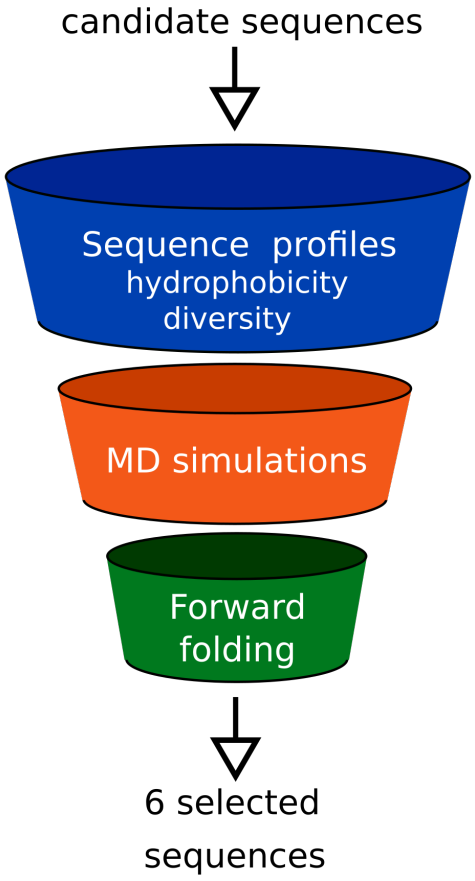

Figure 8.9: Filtering procedure for choosing 6 sequences.

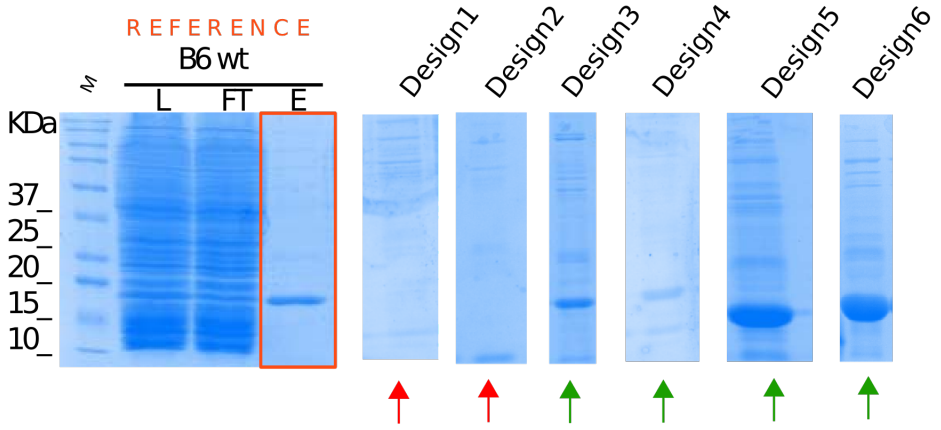

Figure 8.10: Protein purification analysis of 6 selected sequences on SDS-PAGE electrophoresis stained with Coomassie Blue. L = Cell Lysate; FT = Flow Through; E = Elution. Behaviour of the WT nanobody is used as reference (orange rectangle).

8.4.2.2 Experimental characterization

Protein purification analysis on SDS-PAGE showed that 4 out of 6 designed proteins were successfully purified and expressed (Figure 8.10). Designs 1 and 2 could not be expressed and purified. This is in accordance with forward folding results which showed quite bad profiles with lowest energy models around 8 Å away from the native structure. MD results showed that design 1 was more or as stable as the WT nanobody, suggesting that MD simulations were possibly too short to evaluate our designs and that longer simulations should be considered. Designs 3, 4 and 6 were correctly predicted by our forward folding experiments. Finally, design 5 forward folding profile was quite bad, with the lowest energy models being at 10Å distance from the native structure. However, this sequence is the one with the best experimental results. This false negative underlines the fact that performing forward folding with *ab initio* protein structure prediction methods remains a challenging task. It would be interesting to see how recent deep learning based methods perform for this task.

Elisa tests and Immunofluorescence assays haven't been done yet on designs 3 and 6. These experiments are ongoing and will be soon published with other results presented in this manuscript.

To test whether the designed scaffolds were functional recombinant proteins, they were gene synthesized with B6 CDR loops, cloned into a cytoplasmic expression vector under the control of T7 promotor, and expressed them in E.coli Bl21de3 strain. Following NiNTA purification in batch, the resulting nanobodies were tested in an ELISA assay for the detection of one of the GTPase that the wild type B6 hs2dAb binds with a KD of 80nM, RHOA Q63L constitutively active mutant [268].

As positive control we included the wild type B6 hs2dAb (Figure 8.11A) To check the conformational selectivity and the specificity of target recognition, we also assayed the binding to the inactive state of RHA GTPase using the T19N mutant or to the related active GTPase CDC42 (Figure 8.11B). As shown in Figure 8.11, the design 5 was able to give a dose response effect on active RHOA as the wild type B6, with similar selectivity and specificity. No signal was observed with any of the other monovalent binders tested, indicating that most of the scaffold mutant lost the binding capacities of the B6 or lost affinity.

To test the latter hypothesis, a simple way to increase binding capacities of nanobodies in immunoassays consist in increasing their avidity by expressing them as bivalent IgG like antibodies. Thus we subcloned and expressed all the constructs into a mammalian expression vector of Rabbit IgG [271]. Hs2dAb-RFc recombinant protein were secreted in cell culture supernatant and directly tested in a similar ELISA assay for the detection of active RHOA, inactive RHOA or CDC42. Again, only the design number 5 gave a signal similar to the wild type B6 and none of the other construct were able to give a signal (Figure 8.12). This result demonstrated that mutations in the scaffold most often lead to total loss of binding if the parameters do not take into account hydrophobic patches on the surface of the nanobody.

As the design 5 scaffold appeared, so far, the only one to keep the binding properties of the wild type B6 hs2dAB, we wondered whether this was only due to a preferential display of the CDR loops in the right orientation or if this scaffold could withstand several combination of CDR loops. Therefore, we grafted by gene synthesis the CDR loops sequence of other previously characterised hs2dAb (RH12, Tub2, HGX44 in [252], RHB15 unpublished under CISBIO patent) or some published llama VHH targeting GFP, LaminA/C [255] or HistoneH2A/H2B [272].

Design 5 scaffold grafted with CDR loops of various hs2dAb targeting RHO GTPase with different selectivities is shown in Figure 8.13. RH12 wild type was reported to bind RHOA, RHOB, RHOC with subnanomolar affinities, while RHB15 had a preferential binding to RHOB but poorly bind RHOA or RHOC. Design 5 -RH12 was able to bind active mutant of all three RHO GTPase, indicating that this design can display other set of CDR loops than the one of the B6 clone.

For the other grafted loop, target antigens were intracellular proteins such as ectopically expressed GFP, or endogenous Histone, Lamin or Tubulin. Thus, we tested their capacity to give a signal in immunofluorescence on fixed cells. In a preliminary experiment, the design 5 grafted with the CDR loops of the TUB2 hs2dAb [252] kept the potential binding to microtubule as the wild type binder reported (Figure 8.14). However, no staining was observed with none of the other grafted design 5 that we tested, although some lama VHH targeting chromatin (S12) or lamin (lam) were reported to stain their antigen in fixed cells. It is not trivial to conclude because the binding mode of these nanobodies are not known and the paratope can involve residues from their original scaffold or can be constrained by their scaffold. The display of the loops may also be disturbed due to adjacent residues. The fact that the RH12 or Tub2 loops, as well as the B6 ones could be

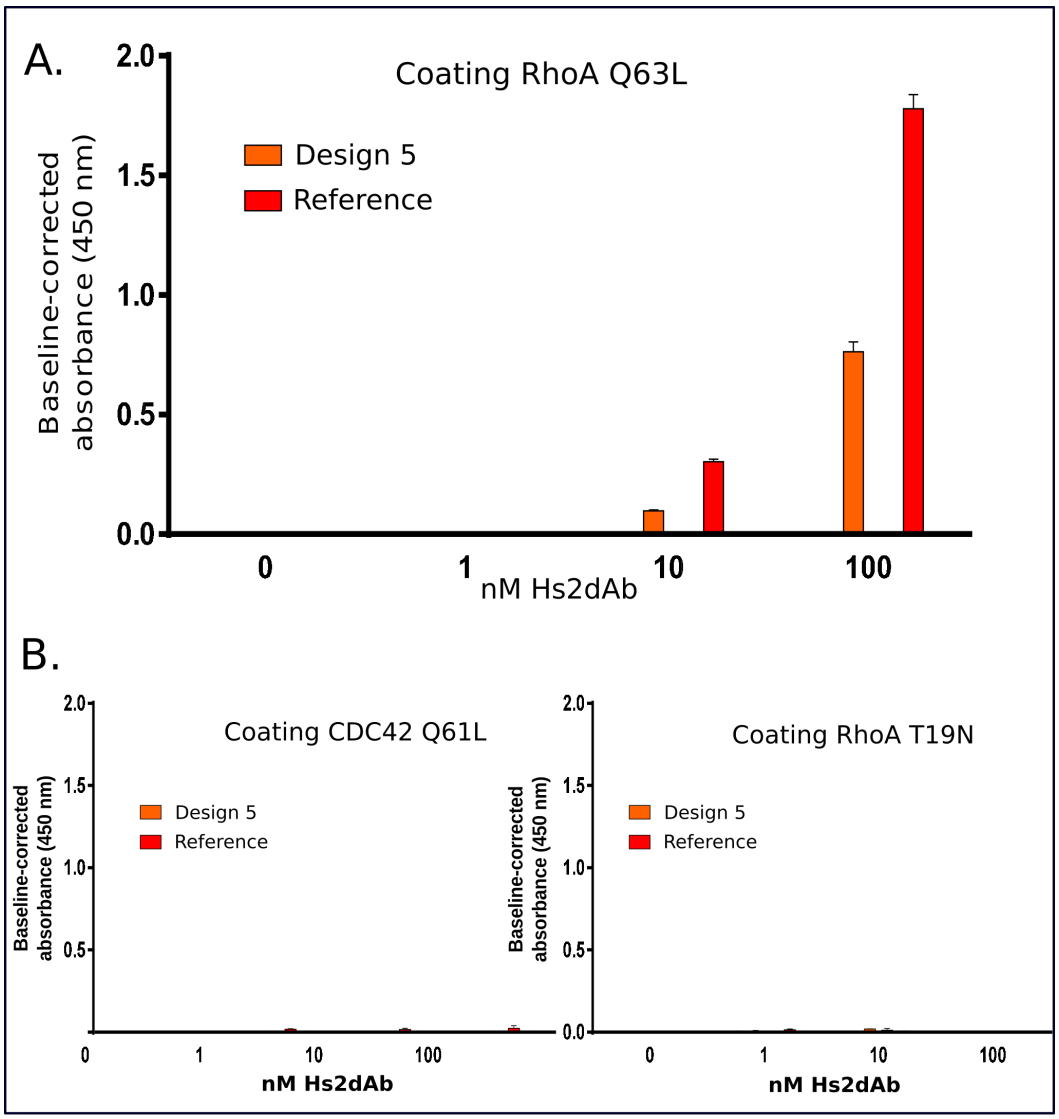

Figure 8.11: (A) Design 5 - selective detection of active conformation of RHOA recombinant protein. Streptactin plates were coated at saturation with 2S HA RHOA Q63L active GTP-bound mutant. (B) As a control, the inactive state mutant 2S HA RHOA T19N or the related GTPase active mutant 2S HA CDC42 Q61L were used. Absorbance at 405 nm reflects myc signal after hs2dAb-6his-Myc-6His dose-effect (0, 1nM, 10 nM or 100 nM)

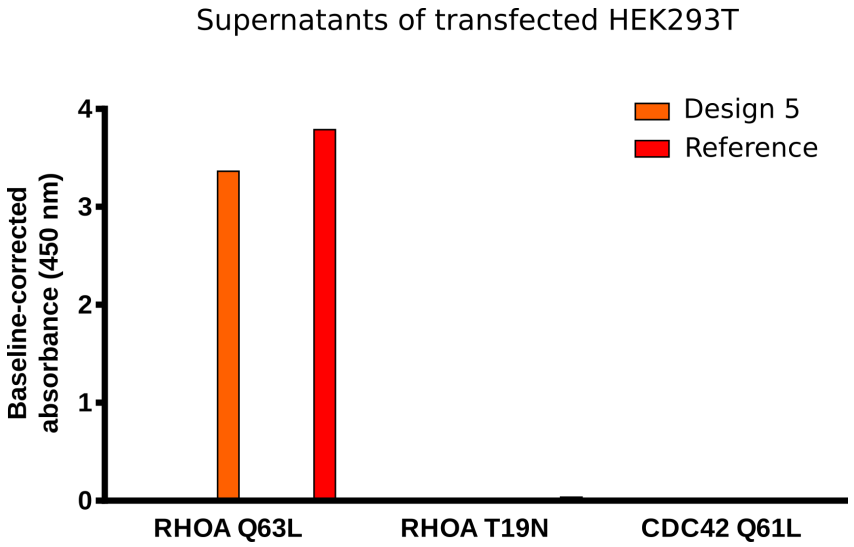

Figure 8.12: Design5 selective detection of active conformation of RHOA recombinant protein. Streptactin plates were coated at saturation with either 2S HA RHOA Q63L active GTP-bound mutant, the inactive state mutant 2S HA RHOA T19N or the related GTPase active mutant 2S HA CDC42 Q61L. Absorbance at 405 nm reflects rabbit Fc-hs2dAb fusion detection from HEK293 cells supernatant incubation.

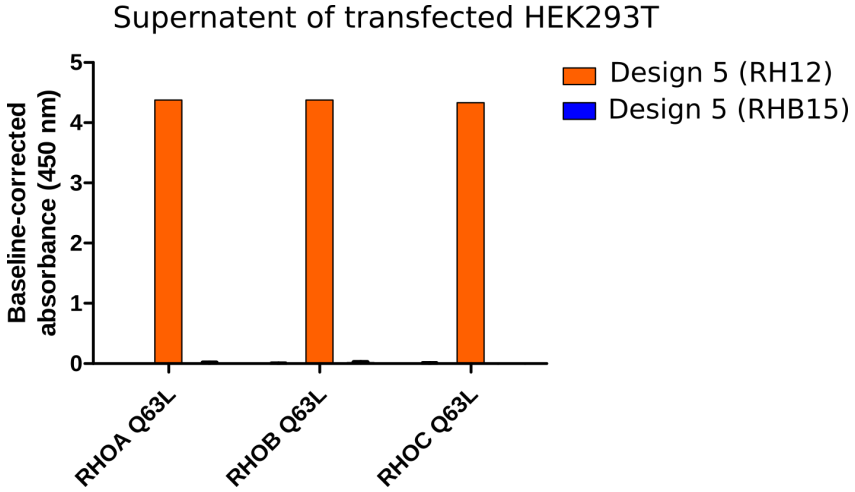

Figure 8.13: ELISA using design5 scaffold grafted with CDR loops form RH12 or RHB15 hs2dAb. Streptactin plates were coated at saturation with either 2S HA RHOA, RHOB, RHOC Q63L active GTP-bound mutant. Absorbance at 405 nm reflects rabbit Fc-hs2dAb fusion detection from HEK293 cells supernatant incubation.

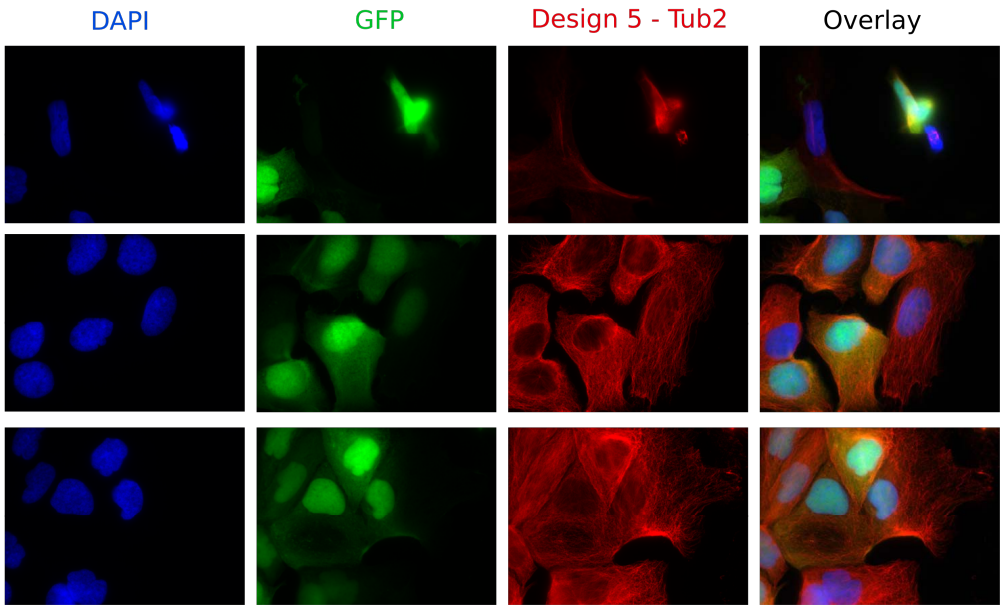

Figure 8.14: Immunofluorescence stainng with design 5 scaffold grafted with Tub2 CDR loops. HeLa S3 H2B-GFP cells were seeded for 24hours, fixed using paraformaldehyde, permeabilized using saponin and stained with non-purified myc-tagged MSD10-Tub2-6His-Myc-6His and revealed with anti-Myc-Tag (9E10) and secondary Rabbit-anti-mouse conjugated to Alexa564 (red), and stained with DAPI to stain nuclear DNA (blue). Design 5 -Tub2 stained microtule cytoskeleton in mitotic cell (upper panel) or interphase cells lower panels.

efficiently grafted reveals that the design 5 is a suitable scaffold for proper display of CDR like aminoacid sequences.

8.5 Conclusion

In this chapter we focused on the design of a new universal nanobody scaffold that could potentially be further used for a development of new synthetic library of nanobodies. This work was based on an already known humanized nanobody scaffold that has been under a patent application. Therefore, this study involved many different design constraints. The new nanobody scaffold had to be beyond the old patented framework, it had to be more or as stable compared to the original scaffold, while being cysteine-less. It is important to mention that the disulfide bridge present in nanobodies represent a trademark, present in 99.4% of all aligned nanobody sequences. In this work, we showed that it is possible to computationally design new stable cysteine-less nanobody scaffolds. Different experiments done by our colleagues showed that some of the scaffolds designed with Pomp^d are highly expressed, and that one of them possesses suitable affinity with different CDR loops.

This study is still ongoing as another scaffold designed with hpatch (described in Chapter 5) shows promising results.

Conclusions and perspectives

This thesis addressed the problem of computational protein design at different levels. First, a new approach that takes protein flexibility into account during the CPD procedure has been developed. In this approach, we define an average energy criteria that needs to be satisfied by multiple conformational states simultaneously. By providing multiple inputs for a given CPD problem, and not only one static structure, this method improves the quality of CPD predictions. On a benchmark composed of NMR and X-ray back-rubbed structures, we showed the superiority in terms of native sequence recovery and efficiency of this method compared to previous SSD, but also to state-of-the-art MSA approaches. For the first time, we showed that it is possible to access guaranteed optimal average energy solutions on full multistate design problems of proteins of size up to 100 amino acids, but also to exhaustively enumerate sequences for a given energy threshold. We named this method POMP^d, for Positive Multistate Protein Design.

Experimental validation is a necessary step for CPD techniques. Furthermore, perceptive feedback from the experimental evaluations helps improving computational modeling. Multiple interactions with our experimental collaborators inspired the development of new functionalities in our software. Besides the possibility of taking into account several conformational states, POMP^d can prohibit hydrophobic patches at the protein surface, give a greater weight to certain states and generate a set of diverse, good quality solutions.

In this context, POMP^d was applied on two different projects in white biotechnology and health domains. In the first one, the objective was to use our CPD method to engineer a new GH11 xylanase enzyme, with improved thermal stability. In the conception of this new enzyme and computational design procedure, molecular dynamics simulations played a major role. Molecular dynamics simulations at the atomic scale have allowed us to study and understand in greater depth the molecular and structural basis of these systems. Thanks to our simulations, we have been able to identify regions that are critical for the stability of the systems we have studied, and thus to define appropriate design strategies. These design strategies were applied and experimental evaluations showed that 4 enzymes mutants possess improved thermal stability and catalytic activity.

Finally, in the last part, POMP^d was applied to design a synthetic humanized nanobody scaffold. This project included many design constraints. One of the most important constraints was the objective to redesign a stable cysteine-less nanobody scaffold that could be expressed as intrabody and be stable in the reducing cytoplasmic environment. The disulfide bridge usually found in the scaffold contributes to its general stability. Redesigning this scaffold without it represented a clear challenge. Results on this new scaffold showed that the new nanobodies we designed are highly expressed and possess suitable affinity with different CDR loops.

Perspectives

The work presented in this thesis revealed multiple directions for future research. First of all, our methods offer many possibilities that have yet to be explored. Experimentally validated artificial proteins presented in this work demonstrate the great potential of our CPD approach and we have only barely scratched the surface of what could be done in terms of biotechnological or biomedical applications.

Molecular Dynamics simulations were exploited for identification of redesignable regions. Various measures such as B-factors, RMSD and cross-correlations could be used as input in an automated enzyme design protocol. Such a protocol would identify designable regions targeting thermostability without impacting the dynamicity observed on very active enzymes nor the catalytic site itself.

The ability of our software to exhaustively enumerate sequences within a threshold of the optimum, represents an important feature that could be exploited in many ways. We could directly produce a batch of sub-optimal sequences and thus augment the chances of success. Knowing that proteins are highly evolvable macromolecules, sequence enumeration could also be used in order to anticipate escape mutations of pathogens for example [273, 274].

The Computational Protein Design field is today mature enough so that experimental synthesis of completely artificial protein sequences is possible. Meanwhile, the success rate of CPD relies on the nature of the application, and many exciting challenges lie ahead of us. Enzyme design, for example, still represents a great challenge. During this thesis we successfully designed new optimized GH11 xylanases. We achieved good results by designing enzymes regions that are quite far from the active site and by an in-depth analysis of the dynamics of this enzyme, which guided us through the design procedure. Being able to explicitly take into account the catalytic activity into the design process would represent a landmark towards the *de novo* design of highly active catalysts with new functions.

Positive Multistate design allows modeling local flexibility, large conformational changes or molecular systems in free and complex forms. However, many applications require the ability to promote some conformations or molecular systems while discouraging some others (negative design). For example, it is the case when modeling ligand binding specificity or oligomeric association specificity. Negative design represents a challenging task and is highly in demand for designing new therapeutics and biosensors. Our Multistate approach could be extended in order to address negative design problems.

Deep learning methods have recently become very popular for learning from large datasets, doing both feature extraction and prediction. As we already mentioned in this manuscript, the use of deep learning recently revolutionized protein structure prediction. Over the past years, the development of new algorithms, sophisticated architectures such as graphical neural networks and high-performance computing tools have improved the performance of deep artificial neural networks as a learning technique. These methods are capable of learning complex characteristics from data, with a sufficient mass of data as a prerequisite. By freeing ourselves

from the energy function and learning directly from structure and sequence data, deep learning methods represent an interesting alternative to address the problem of computational protein design. Newly learned energy potentials could replace or expand the energy function used in our computational protein design framework.

Résumé long en français

Les protéines sont des composants fondamentaux de la vie. Elles sont indispensables à la structure et au fonctionnement des cellules vivantes et des virus et sont responsables de nombreux processus essentiels dans tous les organismes vivants. Elles peuvent transporter de l'énergie, transmettre des signaux, fournir une structure aux cellules ou favoriser des réactions chimiques particulières. Au cours des milliards d'années d'évolution, les protéines ont évolué pour remplir mieux et plus rapidement certaines fonctions ou pour réaliser de nouvelles fonctions afin de répondre aux besoins biologiques dans des conditions diverses et changeantes.

La plupart des protéines ont une structure tridimensionnelle particulière qui est directement liée à leur fonction spécifique. La structure et la fonction d'une protéine proviennent d'un ensemble d'éléments constitutifs qui composent une séquence de protéines, appelés résidus d'acides aminés. Pour une longueur de séquence donnée, l'espace de séquence de la protéine décrit un ensemble de combinaisons possibles de résidus d'acides aminés à chaque position séquentielle. Par exemple, pour une protéine de 100 résidus, l'espace de séquences contient 20^{100} séquences. Les protéines naturelles couvrent une très petite partie de cet espace. Une grande partie des séquences est inexplorée par la nature et de nombreuses protéines fonctionnelles restent certainement à découvrir. Ces dernières années, l'intérêt pour les protéines ayant des propriétés nouvelles ou améliorées s'est accru dans de nombreux domaines. Cependant, la synthèse de toutes les séquences possibles reste inimaginable. L'approche par évolution dirigée, couronnée par le prix Nobel de Frances Arnold en 2018, a des capacités limitées d'exploration des séquences malgré son succès. Par conséquent, le besoin de méthodes computationelles précises est crucial afin de rationaliser et d'accélérer la conception de nouvelles protéines.

La dernière décennie a été marquée par des avancées scientifiques majeures qui ont permis une compréhension plus approfondie des protéines à différents niveaux. De nombreuses données biochimiques et cinétiques ont permis de mieux comprendre les propriétés structurelles et fonctionnelles des protéines, ce qui a conduit à une extension du paradigme structure-fonction pour inclure la dynamique structurelle des protéines. La cristallographie aux rayons X, la spectroscopie par résonance magnétique nucléaire et la microscopie électronique cryogénique ont permis de mettre en évidence un très grand nombre de structures protéiques. Les méthodes computationnelles ont complètement révolutionné le domaine de la prédiction de structures des protéines [1]. De plus, les simulations de dynamique moléculaire sur les protéines, récompensées par le prix Nobel de Martin Karplus et Michael Levitt en 2013, ont permis d'étudier les protéines au niveau atomique. Toutes ces avancées ont largement contribué à affiner notre compréhension de la relation séquence-structure-fonction des protéines. La quantité de structures protéiques disponibles et notre compréhension de leurs fonctions rendent aujourd'hui possible le design computationnel de protéines (CPD).

En raison de la taille prohibitive de l'espace de recherche des séquences et de la combinaison de nombreux degrés de liberté d'une protéine, les approches CPD les plus courantes modélisent les protéines comme un seul squelette rigide, et ignorent généralement la flexibilité des protéines. Cette approche traditionnelle "Single State Design" (SSD) contraste avec la vision aujourd'hui admise des protéines comme étant des entités flexibles et dynamiques. En outre, les mouvements des protéines à grande échelle, allant de la flexibilité locale à de grands réarrangements conformationnels, sont connus pour jouer des rôles clés sur les propriétés et les fonctions des protéines. L'objectif de cette thèse est, premièrement, de développer une nouvelle méthode qui allège les limitations du SSD en considérant plusieurs états conformationnels simultanément, et deuxièmement, de démontrer l'intérêt d'appliquer cette méthode sur des exemples pertinents de conception de protéines pour des applications en santé et en biotechnologie blanche. Le manuscrit est structuré en 3 parties et 8 chapitres.

La première partie introduit les concepts permettant de comprendre le travail présenté dans cette thèse. Le chapitre 1 présente les protéines avec des notions plus générales sur leur fonction, structure et flexibilité. Le chapitre 2 fournit quelques détails sur les principes de base des techniques de modélisation moléculaire. Le design computationnel de protéines est ensuite présenté, ainsi que différentes approches de l'état de l'art. Le chapitre 3 présente les méthodes de design computationnel de protéines basées sur l'optimisation de réseaux de fonctions de coûts (CFN).

La deuxième partie de cette thèse décrit le développement de nouvelles méthodologies de design computationnel. Le chapitre 4 décrit une approche de design multi-états (MSD) qui permet de prendre en compte simultanément plusieurs états conformationnels des protéines. Au cours de cette thèse, de nombreuses interactions avec nos collaborateurs expérimentaux ont permis d'améliorer notre méthode par l'introduction de nouvelles fonctionnalités qui sont présentées dans le chapitre 5.

La troisième et dernière partie de cette thèse présente deux études de cas validant expérimentallement les prédictions de la méthodologie de design computationnel. La première étude de cas se concentre sur la conception de GH11 Xylanases, une enzyme largement utilisée dans les processus de bioraffinage industriel. Le chapitre 6 décrit une étude de dynamique moléculaire menée dans le but de mieux comprendre la relation structure-dynamique-activité de cette classe d'enzymes et d'identifier les déterminants moléculaires régissant leur stabilité thermique et leur activité. Dans le chapitre 7, les caractéristiques révélées par cette dernière étude sont exploitées pour concevoir de nouvelles xylanases présentant une thermostabilité et une activité catalytique améliorées. Enfin, le chapitre 8 présente la deuxième étude de cas où les stratégies de design ont été utilisées pour concevoir un échafaudage de nanocorps humanisés synthétiques. Les résultats ont montré que ce nouveau nanocorps est hautement exprimé et possède une affinité appropriée avec différentes boucles CDR.

À la fin de ce manuscrit, une conclusion générale fournit un résumé des différentes études réalisées au cours de ce doctorat et donne quelques perspectives et orientations de recherche futures.

Cette thèse aborde le problème de design computationnel de protéines à différents niveaux. Tout d'abord, une nouvelle approche qui prend en compte la flexibilité des protéines pendant la procédure de CPD a été développée. Dans cette approche, nous définissons un critère d'énergie moyenne qui doit être satisfait par plusieurs états conformationnels simultanément. En fournissant plusieurs entrées pour un problème de CPD donné, et non pas seulement une structure statique, cette méthode améliore la qualité de prédicition des aproches CPD. Sur un benchmark composé de structures RMN et X-ray, nous avons montré l'efficacité de cette méthode par rapport aux méthodes SSD, mais aussi aux approches de l'état de l'art. Pour la première fois, nous avons montré qu'il est possible d'accéder à des solutions garanties d'énergie moyenne optimale sur des problèmes de design multi-états complets de protéines de taille allant jusqu'à 100 acides aminés, mais aussi d'énumérer exhaustivement les séquences pour un seuil d'énergie donné. Nous avons appelé cette méthode Pomp^d, pour Positive Multistate Protein Design.

La validation expérimentale est une étape nécessaire pour les techniques de CPD. Les multiples interactions avec nos collaborateurs expérimentaux ont inspiré le développement de nouvelles fonctionnalités dans notre logiciel. Outre la possibilité de prendre en compte plusieurs états conformationnels, Pomp^d peut interdire les patchs hydrophobes à la surface des protéines, donner un poids plus important à certains états et générer un ensemble de solutions diverses et de bonne qualité.

Dans ce contexte, Pomp^d a été appliqué sur deux projets différents dans les domaines de la biotechnologie blanche et de la santé. Dans le premier, l'objectif était d'utiliser notre méthode CPD pour concevoir une nouvelle enzyme xylanase GH11, avec une stabilité thermique améliorée. Dans la conception de cette nouvelle enzyme, les simulations de dynamique moléculaire ont joué un rôle majeur. Des simulations de dynamique moléculaire à l'échelle atomique nous ont permis d'étudier et de comprendre plus en profondeur les bases moléculaires et structurales de ces systèmes. Grâce à nos simulations, nous avons pu identifier les régions qui sont critiques pour la stabilité des systèmes que nous avons étudiés, et ainsi définir des stratégies de design appropriées. Ces stratégies de design ont été appliquées et des évaluations expérimentales ont montré que 4 mutants d'enzymes possèdent une stabilité thermique et une activité catalytique améliorées.

Enfin, dans la dernière partie, Pomp^d a été appliqué pour concevoir un échafaudage de nanocorps humanisés synthétiques. Ce projet comportait de nombreuses contraintes de conception. L'une des contraintes les plus importantes était l'objectif de concevoir un échafaudage stable de nanocorps sans cystéine qui pourrait être exprimé sous forme d'intra-corps et être stable dans l'environnement cytoplasmique réducteur. Le pont disulfure que l'on trouve habituellement dans l'échafaudage contribue à sa stabilité générale. La conception de cet échafaudage sans ce pont a représenté un défi évident. Les résultats obtenus sur ce nouvel échafaudage ont montré que les nouveaux nanocorps que nous avons conçus sont hautement exprimés et possèdent une affinité appropriée avec différentes boucles CDR.

Ce travail révèle de multiples directions de recherche. Tout d'abord, nos méthodes offrent de nombreuses possibilités qui doivent encore être explorées. Les pro-

téines artificielles validées expérimentalement présentées dans ce travail démontrent le grand potentiel de notre approche de CPD et nous n'avons fait qu'effleurer la surface de ce qui pourrait être fait en termes d'applications biotechnologiques ou biomédicales.

Les simulations de dynamique moléculaire ont été exploitées pour l'identification de régions pouvant être redesigner. Diverses mesures, telles que les facteurs B, le RMSD et les corrélations croisées, pourraient etre utilisées dans un protocole automatisé de conception d'enzymes. Un tel protocole permettrait d'identifier automatiquement des espaces de design qui cibleraient la thermostabilité sans trop nuire à la dynamicité observée sur les enzymes très actives ni au site catalytique lui-même.

La capacité de notre logiciel à énumérer de manière exhaustive des séquences à l'intérieur d'un seuil optimal représente une caractéristique importante qui pourrait être exploitée de nombreuses manières. Nous pourrions produire directement un lot de séquences sous-optimales et augmenter ainsi les chances de succès. Sachant que les protéines sont des macromolécules hautement évolutives, les enumérations de séquences pourraient également être utilisées, par exemple, afin d'anticiper les mutations d'échappement des agents pathogènes [273, 274].

Le design computationnel de protéines est un domaine qui est aujourd'hui suffisamment mature pour permettre la synthèse expérimentale de séquences protéiques entièrement artificielles. Toutefois, le taux de réussite du design computationnel de protéines dépend de la nature de l'application, et de nombreux défis passionnants nous attendent. La conception d'enzymes, par exemple, représente toujours un grand défi. Au cours de cette thèse, nous avons réussi à concevoir de nouvelles xylanases GH11 optimisées. Nous avons obtenu de bons résultats en redesignant des régions d'enzymes assez éloignées du site actif et après une analyse approfondie de la dynamique de cette enzyme. Pouvoir prendre en compte explicitement l'activité catalytique dans le processus de conception représenterait un jalon vers la conception *de novo* de catalyseurs hautement actifs dotés de nouvelles fonctions.

Le design multi-états positif permet de modéliser la flexibilité locale, les grands changements conformationnels ou les systèmes moléculaires sous des formes libres et complexes. Toutefois, de nombreuses applications nécessitent la capacité de promouvoir certaines conformations ou systèmes moléculaires tout en décourageant certaines autres (design négatif). C'est le cas, par exemple, de la modélisation de la spécificité de liaison d'un ligand ou de la spécificité d'une association oligomérique. Le design négatif est une tâche difficile qui trouve néanmoins de nombreuses applications comme la conception de nouvelles thérapies et de nouveaux biocapteurs. Notre approche multi-états pourrait être étendue afin de résoudre les problèmes de design négatif.

Les méthodes d'apprentissage profond sont de nos jours très populaires pour apprendre à partir de grands ensembles de données. L'utilisation de l'apprentissage profond a récemment révolutionné le domaine de la prédiction de structure de protéine. Au cours des dernières années, le développement de nouveaux algorithmes et d'architectures sophistiquées telles que les réseaux de neurones relationnels ainsi

que les progrès matériels ont permis d'améliorer les performances des réseaux de neurones artificiels profonds. En s'affranchissant de la fonction d'énergie et en apprenant directement à partir des données de structures et de séquences, les méthodes d'apprentissage profond peuvent s'avérer très intéressantes pour répondre au problème de design computationnel de protéines. Les potentiels énergétiques nouvellement appris pourraient remplacer ou étendre la fonction d'énergie utilisée dans notre cadre de design computationnel de protéines.

Bibliography

[1] John Jumper et al. “High Accuracy Protein Structure Prediction Using Deep Learning”. In: *abstractbookCASP14* (2020) (cit. on pp. 3, 149).

[2] Harvey Lodish et al. *Molecular cell biology*. Macmillan, 2008 (cit. on p. 7).

[3] Jenny Gu and Philip E Bourne. *Structural bioinformatics*. Vol. 44. John Wiley & Sons, 2009 (cit. on p. 7).

[4] Christian B Anfinsen. “Principles that govern the folding of protein chains”. In: *Science* 181.4096 (1973), pp. 223–230 (cit. on pp. 8, 26).

[5] William Ramsay Taylor. “The classification of amino acid conservation”. In: *Journal of theoretical Biology* 119.2 (1986), pp. 205–218 (cit. on p. 10).

[6] GN t Ramachandran and V Sasisekharan. “Conformation of polypeptides and proteins”. In: *Advances in protein chemistry*. Vol. 23. Elsevier, 1968, pp. 283–437 (cit. on p. 11).

[7] Charles Tanford. “The hydrophobic effect and the organization of living matter”. In: *Science* 200.4345 (1978), pp. 1012–1018 (cit. on p. 15).

[8] Helen M Berman et al. “The protein data bank”. In: *Nucleic acids research* 28.1 (2000), pp. 235–242 (cit. on pp. 15, 50).

[9] H Hartmann et al. “Conformational substates in a protein: structure and dynamics of metmyoglobin at 80 K”. In: *Proceedings of the National Academy of Sciences* 79.16 (1982), pp. 4967–4971 (cit. on p. 16).

[10] Katherine Henzler-Wildman and Dorothee Kern. “Dynamic personalities of proteins”. In: *Nature* 450.7172 (2007), pp. 964–972 (cit. on pp. 16, 18).

[11] Gordon G Hammes, Stephen J Benkovic, and Sharon Hammes-Schiffer. “Flexibility, diversity, and cooperativity: pillars of enzyme catalysis”. In: *Biochemistry* 50.48 (2011), pp. 10422–10430 (cit. on p. 16).

[12] Ahmet Bakan and Ivet Bahar. “The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding”. In: *Proceedings of the National Academy of Sciences* 106.34 (2009), pp. 14349–14354 (cit. on pp. 16, 18).

[13] Katherine A Henzler-Wildman et al. “A hierarchy of timescales in protein dynamics is linked to enzyme catalysis”. In: *Nature* 450.7171 (2007), pp. 913–916 (cit. on p. 16).

[14] Kaare Teilum, Johan G Olsen, and Birthe B Kragelund. “Functional aspects of protein flexibility”. In: *Cellular and Molecular Life Sciences* 66.14 (2009), p. 2231 (cit. on p. 16).

[15] Rieko Ishima and Dennis A Torchia. “Protein dynamics from NMR”. In: *Nature structural biology* 7.9 (2000), pp. 740–743 (cit. on p. 16).

[16] Donald J Jacobs et al. “Protein flexibility predictions using graph theory”. In: *Proteins: Structure, Function, and Bioinformatics* 44.2 (2001), pp. 150–165 (cit. on p. 16).

[17] Elisa Cilia et al. “The DynaMine webserver: predicting protein dynamics from sequence”. In: *Nucleic acids research* 42.W1 (2014), W264–W270 (cit. on p. 16).

[18] Tarun J Narwani et al. “In silico prediction of protein flexibility with local structure approach”. In: *Biochimie* 165 (2019), pp. 150–155 (cit. on p. 16).

[19] Dominik Schwarz et al. “Co-evolutionary Distance Prediction for Flexibility Prediction”. In: *bioRxiv* (2020) (cit. on p. 16).

[20] Narayan S Punekar. *Enzymes*. Springer, 2018 (cit. on pp. 17, 18).

[21] Trevor Palmer and Philip L Bonner. *Enzymes: biochemistry, biotechnology, clinical chemistry*. Elsevier, 2007 (cit. on p. 18).

[22] V Tournier et al. “An engineered PET depolymerase to break down and recycle plastic bottles”. In: *Nature* 580.7802 (2020), pp. 216–219 (cit. on p. 19).

[23] Andrew R Leach and Andrew R Leach. *Molecular modelling: principles and applications*. Pearson education, 2001 (cit. on p. 21).

[24] Jay W Ponder and David A Case. “Force fields for protein simulations”. In: *Advances in protein chemistry*. Vol. 66. Elsevier, 2003, pp. 27–85 (cit. on p. 23).

[25] Kenno Vanommeslaeghe et al. “CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields”. In: *Journal of computational chemistry* 31.4 (2010), pp. 671–690 (cit. on p. 23).

[26] James A Maier et al. “ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB”. In: *Journal of chemical theory and computation* 11.8 (2015), pp. 3696–3713 (cit. on pp. 23, 78, 107, 130, 133).

[27] Yuedong Yang and Yaoqi Zhou. “Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions”. In: *Protein science* 17.7 (2008), pp. 1212–1219 (cit. on p. 23).

[28] Ionel A Rata, Yaohang Li, and Eric Jakobsson. “Backbone statistical potential from local sequence-structure interactions in protein loops”. In: *The Journal of Physical Chemistry B* 114.5 (2010), pp. 1859–1869 (cit. on p. 23).

[29] Guang Qiang Dong et al. “Optimized atomic statistical potentials: assessment of protein interfaces and loops”. In: *Bioinformatics* 29.24 (2013), pp. 3158–3166 (cit. on p. 23).

[30] Mikhail Karasikov, Guillaume Pagès, and Sergei Grudinin. “Smooth orientation-dependent scoring function for coarse-grained protein quality assessment”. In: *Bioinformatics* 35.16 (2019), pp. 2801–2808 (cit. on p. 23).

[31] José Ramón López-Blanco and Pablo Chacón. “KORP: knowledge-based 6D potential for fast protein and loop modeling”. In: *Bioinformatics* 35.17 (2019), pp. 3013–3019 (cit. on p. 23).

[32] Rebecca F Alford et al. “The Rosetta all-atom energy function for macromolecular modeling and design”. In: *Journal of chemical theory and computation* 13.6 (2017), pp. 3031–3048 (cit. on pp. 23, 50, 130, 132).

[33] Marcin Hoffmann and Jacek Rychlewski. “Effects of solvation for (R, R) tartaric-acid amides”. In: *New Trends in Quantum Systems in Chemistry and Physics*. Springer, 2000, pp. 189–210 (cit. on p. 23).

[34] Ronald M Levy and Emilio Gallicchio. “Computer simulations with explicit solvent: recent progress in the thermodynamic decomposition of free energies and in modeling electrostatic effects”. In: *Annual review of physical chemistry* 49.1 (1998), pp. 531–567 (cit. on p. 23).

[35] CL Brooks III. “III; Karplus, M.; Pettitt, BM. Proteins: A theoretical perspective of dynamics, structure, and thermodynamics”. In: *Advances in Chemical Physics* 71 (1988) (cit. on p. 23).

[36] Di Qiu et al. “The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii”. In: *The Journal of Physical Chemistry A* 101.16 (1997), pp. 3005–3014 (cit. on p. 23).

[37] Wonpil Im, Dmitrii Beglov, and Benoit Roux. “Continuum solvation model: computation of electrostatic forces from numerical solutions to the Poisson-Boltzmann equation”. In: *Computer physics communications* 111.1-3 (1998), pp. 59–75 (cit. on p. 23).

[38] Loup Verlet. “Computer" experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules”. In: *Physical review* 159.1 (1967), p. 98 (cit. on p. 24).

[39] Roger W Hockney. “The potential calculation and some applications”. In: *Methods Comput. Phys.* 9 (1970), p. 136 (cit. on p. 24).

[40] David Beeman. “Some multistep methods for use in molecular dynamics calculations”. In: *Journal of computational physics* 20.2 (1976), pp. 130–139 (cit. on p. 24).

[41] Herman JC Berendsen et al. “Molecular dynamics with coupling to an external bath”. In: *The Journal of chemical physics* 81.8 (1984), pp. 3684–3690 (cit. on pp. 25, 78, 133).

[42] Shuichi Nosé. “A unified formulation of the constant temperature molecular dynamics methods”. In: *The Journal of chemical physics* 81.1 (1984), pp. 511–519 (cit. on p. 25).

[43] William G Hoover. "Canonical dynamics: Equilibrium phase-space distributions". In: *Physical review A* 31.3 (1985), p. 1695 (cit. on p. 25).

[44] Scott E Feller et al. "Constant pressure molecular dynamics simulation: the Langevin piston method". In: *The Journal of chemical physics* 103.11 (1995), pp. 4613–4621 (cit. on p. 25).

[45] Hans C Andersen. "Molecular dynamics simulations at constant pressure and/or temperature". In: *The Journal of chemical physics* 72.4 (1980), pp. 2384–2393 (cit. on p. 25).

[46] Martin Karplus and J Andrew McCammon. "Molecular dynamics simulations of biomolecules". In: *Nature structural biology* 9.9 (2002), pp. 646–652 (cit. on p. 26).

[47] Martin Karplus and John Kuriyan. "Molecular dynamics and protein function". In: *Proceedings of the National Academy of Sciences* 102.19 (2005), pp. 6679–6685 (cit. on p. 26).

[48] Gina Kolata. "Trying to crack the second half of the genetic code". In: *Science* 233 (1986), pp. 1037–1040 (cit. on p. 26).

[49] Herman JC Berendsen. "A glimpse of the holy grail?" In: *Science* 282.5389 (1998), pp. 642–643 (cit. on p. 26).

[50] Gregory A Petsko. "The grail problem". In: *Genome biology* 1.1 (2000), comment002–1 (cit. on p. 26).

[51] Cyrus Levinthal. "Are there pathways for protein folding?" In: *Journal de chimie physique* 65 (1968), pp. 44–45 (cit. on p. 27).

[52] Carol A Rohl et al. "Protein structure prediction using Rosetta". In: *Methods in enzymology*. Vol. 383. Elsevier, 2004, pp. 66–93 (cit. on p. 27).

[53] Philip Bradley et al. "Free modeling with Rosetta in CASP6". In: *Proteins: Structure, Function, and Bioinformatics* 61.S7 (2005), pp. 128–134 (cit. on p. 27).

[54] Andrew Leaver-Fay et al. "ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules". In: *Methods in enzymology*. Vol. 487. Elsevier, 2011, pp. 545–574 (cit. on pp. 27, 33, 39).

[55] Philip Bradley, Kira MS Misura, and David Baker. "Toward high-resolution de novo structure prediction for small proteins". In: *Science* 309.5742 (2005), pp. 1868–1871 (cit. on p. 27).

[56] Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. "Optimization by simulated annealing". In: *science* 220.4598 (1983), pp. 671–680 (cit. on p. 27).

[57] Yoshitake Sakae et al. "Protein structure predictions by parallel simulated annealing molecular dynamics using genetic crossover". In: *Journal of Computational Chemistry* 32.7 (2011), pp. 1353–1360. ISSN: 1096-987X. DOI: 10.1002/jcc.21716. URL: http://dx.doi.org/10.1002/jcc.21716 (cit. on p. 27).

[58] Sameh Saleh, Brian Olson, and Amarda Shehu. "A population-based evolutionary search approach to the multiple minima problem in de novo protein structure prediction". In: *BMC Structural Biology* 13.1 (2013), S4. ISSN: 1472-6807. DOI: 10.1186/1472-6807-13-S1-S4 (cit. on p. 27).

[59] Brian Olson and Amarda Shehu. "Multi-Objective Stochastic Search for Sampling Local Minima in the Protein Energy Surface". In: *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*. BCB'13. Washington DC, USA: ACM, 2013, 430:430–430:439. ISBN: 978-1-4503-2434-2. DOI: 10.1145/2506583.2506590. URL: http://doi.acm.org/10.1145/2506583.2506590 (cit. on p. 27).

[60] Rudy Clausen and Amarda Shehu. "A Multiscale Hybrid Evolutionary Algorithm to Obtain Sample-based Representations of Multi-basin Protein Energy Landscapes". In: *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. BCB '14. Newport Beach, California: ACM, 2014, pp. 269–278. ISBN: 978-1-4503-2894-4. DOI: 10.1145/2649387.2649390 (cit. on p. 27).

[61] Shaun M Kandathil, Simon C Lovell, and Julia Handl. "Toward a detailed understanding of search trajectories in fragment assembly approaches to protein structure prediction". In: *Proteins: Structure, Function and Bioinformatics* 84.4 (Apr. 2016), pp. 411–26. ISSN: 0887-3585. DOI: 10.1002/prot.24987 (cit. on p. 27).

[62] Mario Garza-Fabre et al. "Generating, Maintaining and Exploiting Diversity in a Memetic Algorithm for Protein Structure Prediction". In: *Evolutionary Computation* (2016). ISSN: 1063-6560. DOI: 10.1162/EVCO_a_00176 (cit. on p. 27).

[63] Wynne J Browne et al. "A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme". In: *Journal of molecular biology* 42.1 (1969), pp. 65–86 (cit. on p. 28).

[64] Cyrus Chothia and Arthur M Lesk. "The relation between the divergence of sequence and structure in proteins." In: *The EMBO journal* 5.4 (1986), pp. 823–826 (cit. on p. 28).

[65] Chris Sander and Reinhard Schneider. "Database of homology-derived protein structures and the structural meaning of sequence alignment". In: *Proteins: Structure, Function, and Bioinformatics* 9.1 (1991), pp. 56–68 (cit. on p. 28).

[66] Mark Johnson et al. "NCBI BLAST: a better web interface". In: *Nucleic acids research* 36.suppl_2 (2008), W5–W9 (cit. on p. 28).

[67] Jianyi Yang and Yang Zhang. "I-TASSER server: new development for protein structure and function predictions". In: *Nucleic acids research* 43.W1 (2015), W174–W181 (cit. on pp. 28, 130).

[68] Andrew Waterhouse et al. “SWISS-MODEL: homology modelling of protein structures and complexes”. In: *Nucleic acids research* 46.W1 (2018), W296–W303 (cit. on p. 28).

[69] Benjamin Webb and Andrej Sali. “Comparative protein structure modeling using MODELLER”. In: *Current protocols in bioinformatics* 54.1 (2016), pp. 5–6 (cit. on p. 28).

[70] Lawrence A Kelley et al. “The Phyre2 web portal for protein modeling, prediction and analysis”. In: *Nature protocols* 10.6 (2015), p. 845 (cit. on p. 28).

[71] Surbhi Dhingra et al. “A glance into the evolution of template-free protein structure prediction methodologies”. In: *Biochimie* (2020) (cit. on p. 28).

[72] James O’Connell et al. “SPIN2: Predicting sequence profiles from protein structures using deep neural networks”. In: *Proteins: Structure, Function, and Bioinformatics* 86.6 (2018), pp. 629–633 (cit. on p. 28).

[73] Sheng Wang et al. “Accurate de novo prediction of protein contact map by ultra-deep learning model”. In: *PLoS computational biology* 13.1 (2017), e1005324 (cit. on p. 28).

[74] Evangelia I Zacharaki. “Prediction of protein function using a deep convolutional neural network ensemble”. In: *PeerJ Computer Science* 3 (2017), e124 (cit. on p. 28).

[75] Jinbo Xu. “Distance-based protein folding powered by deep learning”. In: *Proceedings of the National Academy of Sciences* 116.34 (2019), pp. 16856–16865 (cit. on p. 28).

[76] Jianyi Yang et al. “Improved protein structure prediction using predicted interresidue orientations”. In: *Proceedings of the National Academy of Sciences* (2020) (cit. on p. 28).

[77] Andrew W Senior et al. “Improved protein structure prediction using potentials from deep learning”. In: *Nature* (2020), pp. 1–5 (cit. on p. 29).

[78] Jeffrey C Moore and Frances H Arnold. “Directed evolution of a para-nitrobenzyl esterase for aqueous-organic solvents”. In: *Nature biotechnology* 14.4 (1996), pp. 458–467 (cit. on p. 29).

[79] Misha Soskine and Dan S Tawfik. “Mutational effects and the evolution of new protein functions”. In: 11.8 (2010). Exported from https://app.dimensions.ai on 2018/11/12, pp. 572–582. DOI: 10.1038/nrg2808. URL: https://app.dimensions.ai/details/publication/pub.1044757133 (cit. on p. 29).

[80] S. Bershtein and D. S. Tawfik. “Ohno’s model revisited: measuring the frequency of potentially adaptive mutations under various mutational drifts”. In: *Mol Biol Evol.* 25.11 (2008), 2311–8. Epub 2008 Aug 6.+ (cit. on p. 29).

[81] Niles A Pierce and Erik Winfree. “Protein design is NP-hard”. In: *Protein engineering* 15.10 (2002), pp. 779–782 (cit. on pp. 30, 47, 62).

[82] Brian Kuhlman and David Baker. "Native protein sequences are close to optimal for their structures". In: *Proceedings of the National Academy of Sciences* 97.19 (2000), pp. 10383–10388 (cit. on pp. 30, 33).

[83] Christopher A Voigt, D Benjamin Gordon, and Stephen L Mayo. "Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design". In: *Journal of molecular biology* 299.3 (2000), pp. 789–803 (cit. on pp. 30, 33).

[84] James J Havranek and Pehr B Harbury. "Automated design of specificity in molecular recognition". In: *Nature structural biology* 10.1 (2003), pp. 45–52 (cit. on pp. 30, 43).

[85] David Simoncini et al. "Guaranteed discrete energy optimization on large protein design problems". In: *Journal of chemical theory and computation* 11.12 (2015), pp. 5980–5989 (cit. on pp. 30, 39, 52, 62).

[86] Ilan Samish. "Achievements and challenges in computational protein design". In: *Computational Protein Design*. Springer, 2017, pp. 21–94 (cit. on p. 30).

[87] William F DeGrado et al. "The design, synthesis, and characterization of tight-binding inhibitors of calmodulin". In: *Journal of cellular biochemistry* 29.2 (1985), pp. 83–93 (cit. on p. 30).

[88] James H Hurley, Walter A Baase, and Brian W Matthews. "Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme". In: *Journal of molecular biology* 224.4 (1992), pp. 1143–1159 (cit. on p. 30).

[89] Pehr B Harbury, Bruce Tidor, and Peter S Kim. "Repacking protein cores with backbone freedom: structure prediction for coiled coils". In: *Proceedings of the National Academy of Sciences* 92.18 (1995), pp. 8408–8412 (cit. on p. 30).

[90] John R Desjarlais and Tracy M Handel. "De novo design of the hydrophobic cores of proteins". In: *Protein Science* 4.10 (1995), pp. 2006–2018 (cit. on p. 30).

[91] Stephen F Betz and William F DeGrado. "Controlling topology and native-like behavior of de novo-designed peptides: design and characterization of antiparallel four-stranded coiled coils". In: *Biochemistry* 35.21 (1996), pp. 6955–6962 (cit. on p. 30).

[92] Bassil I Dahiyat and Stephen L Mayo. "Protein design automation". In: *Protein Science* 5.5 (1996), pp. 895–903 (cit. on p. 30).

[93] Bassil I Dahiyat and Stephen L Mayo. "De novo protein design: fully automated sequence selection". In: *Science* 278.5335 (1997), pp. 82–87 (cit. on p. 30).

[94] Louis A Clark et al. "Affinity enhancement of an in vivo matured therapeutic antibody using structure-based computational design". In: *Protein science* 15.5 (2006), pp. 949–960 (cit. on p. 31).

[95] Daniela Röthlisberger et al. “Kemp elimination catalysts by computational enzyme design”. In: *Nature* 453.7192 (2008), pp. 190–195 (cit. on p. 31).

[96] Lin Jiang et al. “De novo computational design of retro-aldol enzymes”. In: *science* 319.5868 (2008), pp. 1387–1391 (cit. on p. 31).

[97] Andrew H. Ng et al. “Modular and tunable biological feedback control using a de novo protein switch”. In: *Nature* (July 24, 2019). DOI: 10.1038/s41586-019-1425-7 (cit. on p. 31).

[98] Gottfried Palm et al. “Structure of the plastic-degrading Ideonella sakaiensis MHETase bound to a substrate”. In: *Nature Communications* 10 (Apr. 2019). DOI: 10.1038/s41467-019-09326-3 (cit. on p. 31).

[99] Hiroki Noguchi et al. “Computational design of symmetrical eight-bladed β-propeller proteins”. In: *IUCrJ* 6.1 (2019) (cit. on p. 31).

[100] Chunfu Xu et al. “Computational design of transmembrane pores”. In: *Nature* 585.7823 (2020), pp. 129–134 (cit. on p. 31).

[101] Marc J Lajoie et al. “Designed protein logic to target cells with precise combinations of surface antigens”. In: *Science* 369.6511 (2020), pp. 1637–1643 (cit. on pp. 31, 43, 44).

[102] Longxing Cao et al. “De novo design of picomolar SARS-CoV-2 miniprotein inhibitors”. In: *Science* 370.6515 (2020), pp. 426–431 (cit. on p. 31).

[103] Christoffer Norn et al. “Protein sequence design by explicit energy landscape optimization”. In: *bioRxiv* (2020) (cit. on p. 32).

[104] Nicholas Metropolis et al. “Equation of state calculations by fast computing machines”. In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092 (cit. on p. 32).

[105] Kaushik Raha et al. “Prediction of amino acid sequence from structure”. In: *Protein Science* 9.6 (2000), pp. 1106–1119 (cit. on p. 33).

[106] Eugene L Lawler and David E Wood. “Branch-and-bound methods: A survey”. In: *Operations research* 14.4 (1966), pp. 699–719 (cit. on p. 33).

[107] David R Morrison et al. “Branch-and-bound algorithms: A survey of recent advances in searching, branching, and pruning”. In: *Discrete Optimization* 19 (2016), pp. 79–102 (cit. on p. 34).

[108] Pablo Gainza et al. “OSPREY: protein design with ensembles, flexibility, and provable algorithms”. In: *Methods in enzymology*. Vol. 523. Elsevier, 2013, pp. 87–107 (cit. on p. 35).

[109] Johan Desmet et al. “The dead-end elimination theorem and its use in protein side-chain positioning”. In: *Nature* 356.6369 (1992), pp. 539–542 (cit. on p. 35).

[110] Andrew R Leach and Andrew P Lemon. "Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm". In: *Proteins: Structure, Function, and Bioinformatics* 33.2 (1998), pp. 227–239 (cit. on p. 35).

[111] David Allouche et al. "Computational protein design as a cost function network optimization problem". In: *International Conference on Principles and Practice of Constraint Programming*. Springer. 2012, pp. 840–849 (cit. on p. 35).

[112] Seydou Traoré et al. "A new framework for computational protein design through cost function network optimization". In: *Bioinformatics* 29.17 (2013), pp. 2129–2136 (cit. on p. 35).

[113] David Allouche et al. "Computational protein design as an optimization problem". In: *Artificial Intelligence* 212 (2014), pp. 59–79 (cit. on pp. 35, 54).

[114] Martin Cooper, Simon de Givry, and Thomas Schiex. "Graphical Models: Queries, Complexity, Algorithms". In: *Leibniz International Proceedings in Informatics* 154 (2020), pp. 4–1 (cit. on p. 37).

[115] Rina Dechter. "Reasoning with probabilistic and deterministic graphical models: Exact algorithms". In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 7.3 (2013), pp. 1–191 (cit. on p. 37).

[116] T. Schiex, H. Fargier, and G. Verfaillie. "Valued Constraint Satisfaction Problems: hard and easy problems". In: *Proc. of the 14^{th} IJCAI*. Montréal, Canada, Aug. 1995, pp. 631–637 (cit. on p. 38).

[117] Francesca Rossi, Peter Van Beek, and Toby Walsh. *Handbook of constraint programming*. Elsevier, 2006 (cit. on p. 38).

[118] David Allouche et al. "Computational protein design as an optimization problem". In: *Artif. Intell.* 212 (2014), pp. 59–79 (cit. on p. 39).

[119] Seydou Traoré et al. "A New Framework for Computational Protein Design through Cost Function Network Optimization". In: *Bioinformatics* 29.17 (2013), pp. 2129–2136 (cit. on pp. 39, 62).

[120] Simon De Givry et al. "Existential arc consistency: Getting closer to full arc consistency in weighted CSPs". In: *IJCAI*. Vol. 5. 2005, pp. 84–89 (cit. on p. 39).

[121] Martin C Cooper et al. "Virtual Arc Consistency for Weighted CSP." In: *AAAI*. 2008, pp. 253–258 (cit. on p. 39).

[122] Hiroki Noguchi et al. "Computational design of symmetrical eight-bladed β-propeller proteins". In: *IUCrJ* 6.1 (2019) (cit. on p. 40).

[123] James A Davey and Roberto A Chica. "Multistate approaches in computational protein design". In: *Protein Science* 21.9 (2012), pp. 1241–1252 (cit. on pp. 43, 48).

[124] Benjamin D Allen, Alex Nisthal, and Stephen L Mayo. "Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles". In: *Proceedings of the National Academy of Sciences* (2010) (cit. on pp. 43, 45).

[125] James A Davey et al. "Prediction of stable globular proteins using negative design with non-native backbone ensembles". In: *Structure* 23.11 (2015), pp. 2011–2021 (cit. on p. 43).

[126] Xavier I Ambroggio and Brian Kuhlman. "Computational design of a single amino acid sequence that can switch between two distinct protein folds". In: *Journal of the American Chemical Society* 128.4 (2006), pp. 1154–1161 (cit. on pp. 43, 44).

[127] Christopher Negron and Amy E Keating. "Multistate protein design using CLEVER and CLASSY". In: *Methods in enzymology*. Vol. 523. Elsevier, 2013, pp. 171–190 (cit. on pp. 43, 44).

[128] Alexander M Sevy et al. "Multistate design of influenza antibodies improves affinity and breadth against seasonal viruses". In: *Proceedings of the National Academy of Sciences* 116.5 (2019), pp. 1597–1602 (cit. on p. 43).

[129] Patrick Löffler et al. "Rosetta: MSF: a modular framework for multi-state computational protein design". In: *PLoS computational biology* 13.6 (2017), e1005600 (cit. on pp. 43, 55).

[130] Stanley C Howell et al. "Understanding thermal adaptation of enzymes through the multistate rational design and stability prediction of 100 adenylate kinases". In: *Structure* 22.2 (2014), pp. 218–229 (cit. on p. 44).

[131] Andrew Leaver-Fay et al. "Computationally designed bispecific antibodies using negative state repertoires". In: *Structure* 24.4 (2016), pp. 641–651 (cit. on p. 44).

[132] Alexander M Sevy et al. "Design of protein multi-specificity using an independent sequence search reduces the barrier to low energy sequences". In: *PLoS computational biology* 11.7 (2015), e1004300 (cit. on p. 44).

[133] Marion F Sauer et al. "Multi-state design of flexible proteins predicts sequences optimal for conformational change". In: *PLoS computational biology* 16.2 (2020), e1007339 (cit. on p. 44).

[134] Navin Pokala and Tracy M Handel. "Energy functions for protein design: adjustment with protein–protein complex affinities, models for the unfolded state, and negative design of solubility and specificity". In: *Journal of molecular biology* 347.1 (2005), pp. 203–227 (cit. on p. 44).

[135] Benjamin D Allen and Stephen L Mayo. "An efficient algorithm for multistate protein design based on FASTER". In: *Journal of computational chemistry* 31.5 (2010), pp. 904–916 (cit. on p. 44).

[136] Chen Yanover, Menachem Fromer, and Julia M Shifman. "Dead-end elimination for multistate protein design". In: *Journal of computational chemistry* 28.13 (2007), pp. 2122–2129 (cit. on p. 44).

[137] Mark A Hallen and Bruce R Donald. "Comets (Constrained Optimization of Multistate Energies by Tree Search): A provable and efficient protein design algorithm to optimize binding affinity and specificity with respect to sequence". In: *Journal of Computational Biology* 23.5 (2016), pp. 311–321 (cit. on pp. 44, 45, 54, 58).

[138] Larry J Stockmeyer. "The polynomial-time hierarchy". In: *Theoretical Computer Science* 3.1 (1976), pp. 1–22 (cit. on p. 44).

[139] James A Davey et al. "Rational design of proteins that exchange on functional timescales". In: *Nature chemical biology* 13.12 (2017), p. 1280 (cit. on p. 44).

[140] Martin C Cooper et al. "Soft arc consistency revisited". In: *Artificial Intelligence* 174.7-8 (2010), pp. 449–478 (cit. on pp. 44, 49, 52, 54, 62).

[141] Barry Hurley et al. "Multi-language evaluation of exact solvers in graphical model discrete optimization". In: *Constraints* 21.3 (2016), pp. 413–434 (cit. on p. 44).

[142] Mostafa Karimi and Yang Shen. "iCFN: an efficient exact algorithm for multistate protein design". In: *Bioinformatics* 34.17 (2018), pp. i811–i820 (cit. on pp. 45, 50, 52, 56, 62).

[143] James A Davey and Roberto A Chica. "Multistate computational protein design with backbone ensembles". In: *Computational Protein Design*. Springer, 2017, pp. 161–179 (cit. on pp. 45, 46).

[144] Jelena Vucinic et al. "Positive multistate protein design". In: *Bioinformatics* 36.1 (2020), pp. 122–130 (cit. on pp. 47, 107).

[145] Seydou Traoré et al. "Fast search algorithms for computational protein design". In: *Journal of computational chemistry* 37.12 (2016), pp. 1048–1058 (cit. on p. 50).

[146] Francois Berenger et al. "Durandal: fast exact clustering of protein decoys". In: *Journal of computational chemistry* 33.4 (2012), pp. 471–474 (cit. on pp. 50, 107, 128).

[147] Ian W Davis et al. "The backrub motion: how protein backbone shrugs when a sidechain dances". In: *Structure* 14.2 (2006), pp. 265–274 (cit. on pp. 50, 107).

[148] Gregory D Friedland et al. "A correspondence between solution-state dynamics of an individual protein and the sequence and conformational diversity of its family". In: *PLoS computational biology* 5.5 (2009), e1000393 (cit. on pp. 50, 107).

[149] Elisabeth L Humphris and Tanja Kortemme. "Prediction of protein-protein interface sequence diversity using flexible backbone computational protein design". In: *Structure* 16.12 (2008), pp. 1777–1788 (cit. on pp. 50, 107).

[150] Maxim V Shapovalov and Roland L Dunbrack. "A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions". In: *Structure* 19.6 (2011), pp. 844–858 (cit. on pp. 50, 132).

[151] Sidhartha Chaudhury, Sergey Lyskov, and Jeffrey J Gray. "PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta". In: *Bioinformatics* 26.5 (2010), pp. 689–691 (cit. on pp. 50, 132).

[152] David Allouche et al. "Anytime hybrid best-first search with tree decomposition for weighted CSP". In: *International Conference on Principles and Practice of Constraint Programming*. Springer. 2015, pp. 12–29 (cit. on p. 54).

[153] James J Havranek, Carlos M Duarte, and David Baker. "A simple physical model for the prediction and design of protein–DNA interactions". In: *Journal of molecular biology* 344.1 (2004), pp. 59–70 (cit. on p. 55).

[154] Elisabeth L Humphris and Tanja Kortemme. "Design of multi-specificity in protein interfaces". In: *PLoS computational biology* 3.8 (2007), e164 (cit. on p. 55).

[155] David Simoncini et al. "Fitness landscape analysis around the optimum in computational protein design". In: *Proceedings of the Genetic and Evolutionary Computation Conference*. ACM. 2018, pp. 355–362 (cit. on p. 59).

[156] Ryan M Kramer et al. "Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility". In: *Biophysical journal* 102.8 (2012), pp. 1907–1915 (cit. on p. 63).

[157] Susan Jones and Janet M Thornton. "Principles of protein-protein interactions". In: *Proceedings of the National Academy of Sciences* 93.1 (1996), pp. 13–20 (cit. on p. 63).

[158] Joël Janin, Susan Miller, and Cyrus Chothia. "Surface, subunit interfaces and interior of oligomeric proteins". In: *Journal of molecular biology* 204.1 (1988), pp. 155–164 (cit. on p. 63).

[159] Fabrizio Chiti et al. "Rationalization of the effects of mutations on peptide andprotein aggregation rates". In: *Nature* 424.6950 (2003), pp. 805–808 (cit. on p. 63).

[160] Alfonso Jaramillo et al. "Folding free energy function selects native-like protein sequences in the core but not on the surface". In: *Proceedings of the National Academy of Sciences* 99.21 (2002), pp. 13554–13559 (cit. on p. 64).

[161] Ron Jacak, Andrew Leaver-Fay, and Brian Kuhlman. "Computational protein design with explicit consideration of surface hydrophobic patches". In: *Proteins: Structure, Function, and Bioinformatics* 80.3 (2012), pp. 825–838 (cit. on p. 64).

[162] Susan Miller et al. “Interior and surface of monomeric proteins”. In: *Journal of Molecular Biology* 196.3 (1987), pp. 641–656. ISSN: 0022-2836. DOI: https://doi.org/10.1016/0022-2836(87)90038-6. URL: http://www.sciencedirect.com/science/article/pii/0022283687900386 (cit. on p. 64).

[163] Manon Ruffini et al. “Guaranteed Diversity & Quality for the Weighted CSP”. In: *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE. 2019, pp. 18–25 (cit. on p. 66).

[164] Tony Collins, Charles Gerday, and Georges Feller. “Xylanases, xylanase families and extremophilic xylanases”. In: *FEMS microbiology reviews* 29.1 (2005), pp. 3–23 (cit. on p. 70).

[165] Gabriel Paës, Jean-Guy Berrin, and Johnny Beaugrand. “GH11 xylanases: structure/function/properties relationships and applications”. In: *Biotechnology advances* 30.3 (2012), pp. 564–592 (cit. on pp. 70, 73, 76).

[166] Vincent Lombard et al. “The carbohydrate-active enzymes database (CAZy) in 2013”. In: *Nucleic acids research* 42.D1 (2014), pp. D490–D495 (cit. on p. 70).

[167] RL Campbell et al. “A comparison of the structures of the 20 kDa xylanases from Trichoderma harzianum and Bacillus circulans”. In: *Foundation for Biotechnical and Industrial Fermentation Research Publication* 8 (1993), pp. 63–72 (cit. on p. 70).

[168] Warren W Wakarchuk et al. “Mutational and crystallographic analyses of the active site residues of the Bacillus circulans xylanase”. In: *Protein Science* 3.3 (1994), pp. 467–475 (cit. on p. 70).

[169] Gideon J Davies, Keith S Wilson, and Bernard Henrissat. “Nomenclature for sugar-binding subsites in glycosyl hydrolases”. In: *Biochemical Journal* 321.2 (1997), pp. 557–559 (cit. on p. 72).

[170] Bharat Madan and Sun-Gu Lee. “Sequence and Structural Features of Subsite Residues in GH10 and GH11 Xylanases”. In: *Biotechnology and Bioprocess Engineering* 23.3 (2018), pp. 311–318 (cit. on p. 72).

[171] Maria Vardakou et al. “Understanding the structural basis for substrate and inhibitor recognition in eukaryotic GH11 xylanases”. In: *Journal of molecular biology* 375.5 (2008), pp. 1293–1305 (cit. on pp. 72, 77, 103, 107).

[172] Elien Vandermarliere et al. “Crystallographic analysis shows substrate binding at the- 3 to+ 1 active-site subsites and at the surface of glycoside hydrolase family 11 endo-1, 4-β-xylanases”. In: *Biochemical journal* 410.1 (2008), pp. 71–79 (cit. on p. 72).

[173] Qun Wan et al. “X-ray crystallographic studies of family 11 xylanase Michaelis and product complexes: implications for the catalytic mechanism”. In: *Acta Crystallographica Section D: Biological Crystallography* 70.1 (2014), pp. 11–23 (cit. on pp. 72, 78).

[174] Amalia Sapag et al. “The endoxylanases from family 11: computer analysis of protein sequences reveals important structural and phylogenetic relationships”. In: *Journal of Biotechnology* 95.2 (2002), pp. 109–131 (cit. on p. 73).

[175] Stephen F Altschul et al. “Basic local alignment search tool”. In: *Journal of molecular biology* 215.3 (1990), pp. 403–410 (cit. on p. 73).

[176] Kazutaka Katoh et al. “MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform”. In: *Nucleic acids research* 30.14 (2002), pp. 3059–3066 (cit. on p. 73).

[177] Christopher Topham, Jeremy Esque, and Isabelle André. “SEQUESTER: A Software Tool for the Analysis of Protein Sequence-Structure-Function Relations”. In: *TouCAM*. Nov. 2017 (cit. on p. 73).

[178] Gabriel Paës et al. “Thumb-loops up for catalysis: a structure/function investigation of a functional loop movement in a GH11 xylanase”. In: *Computational and Structural Biotechnology Journal* 1.2 (2012), e201207001 (cit. on p. 75).

[179] Mário T Murakami et al. “Correlation of temperature induced conformation change with optimum catalytic activity in the recombinant G/11 xylanase A from Bacillus subtilis strain 168 (1A1)”. In: *Febs Letters* 579.28 (2005), pp. 6505–6510 (cit. on p. 75).

[180] Davi Serradella Vieira, Léo Degrève, and Richard John Ward. “Characterization of temperature dependent and substrate-binding cleft movements in Bacillus circulans family 11 xylanase: a molecular dynamics investigation”. In: *Biochimica et Biophysica Acta (BBA)-General Subjects* 1790.10 (2009), pp. 1301–1306 (cit. on p. 75).

[181] Davi Serradella Vieira and Richard John Ward. “Conformation analysis of a surface loop that controls active site access in the GH11 xylanase A from Bacillus subtilis”. In: *Journal of molecular modeling* 18.4 (2012), pp. 1473–1479 (cit. on p. 75).

[182] Mikko Purmonen et al. “Molecular dynamics studies on the thermostability of family 11 xylanases”. In: *Protein Engineering, Design & Selection* 20.11 (2007), pp. 551–559 (cit. on pp. 75, 76).

[183] Davi Serradella Vieira and Leo Degreve. “An insight into the thermostability of a pair of xylanases: the role of hydrogen bonds”. In: *Molecular Physics* 107.1 (2009), pp. 59–69 (cit. on pp. 75, 96).

[184] Ndumiso N Mhlongo et al. “Dynamics of the thumb-finger regions in a GH11 xylanase Bacillus circulans: Comparison between the Michaelis and covalent intermediate”. In: *RSC advances* 5.100 (2015), pp. 82381–82394 (cit. on p. 75).

[185] Kentaro Miyazaki and Frances H Arnold. "Exploring nonnatural evolutionary pathways by saturation mutagenesis: rapid improvement of protein function". In: *Journal of molecular evolution* 49.6 (1999), pp. 716–720 (cit. on p. 75).

[186] Keqin Chen and Frances H Arnold. "Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide". In: *Proceedings of the National Academy of Sciences* 90.12 (1993), pp. 5618–5622 (cit. on p. 76).

[187] Nina Hakulinen et al. "Three-dimensional structures of thermophilic β-1, 4-xylanases from Chaetomium thermophilum and Nonomuraea flexuosa: Comparison of twelve xylanases in relation to their thermal stability". In: *European journal of biochemistry* 270.7 (2003), pp. 1399–1412 (cit. on p. 76).

[188] Jacques Georis et al. "Purification and properties of three endo-β-1, 4-xylanases produced by Streptomyces sp. strain S38 which differ in their ability to enhance the bleaching of kraft pulps". In: *Enzyme and microbial technology* 26.2-4 (2000), pp. 178–186 (cit. on p. 76).

[189] Jian-Yi Sun et al. "Improvement of the thermostability and catalytic activity of a mesophilic family 11 xylanase by N-terminus replacement". In: *Protein expression and purification* 42.1 (2005), pp. 122–130 (cit. on pp. 76, 106).

[190] Shan Zhang et al. "Five mutations in N-terminus confer thermostability on mesophilic xylanase". In: *Biochemical and biophysical research communications* 395.2 (2010), pp. 200–206 (cit. on p. 76).

[191] Ossi Turunen et al. "A combination of weakly stabilizing mutations with a disulfide bridge in the α-helix region of Trichoderma reesei endo-1, 4-β-xylanase II increases the thermal stability through synergism". In: *Journal of biotechnology* 88.1 (2001), pp. 37–46 (cit. on p. 76).

[192] Fred Fenel et al. "A de novo designed N-terminal disulphide bridge stabilizes the Trichoderma reesei endo-1, 4-β-xylanase II". In: *Journal of biotechnology* 108.2 (2004), pp. 137–143 (cit. on p. 76).

[193] Gabriel Paës and Michael J O'Donohue. "Engineering increased thermostability in the thermostable GH-11 xylanase from Thermobacillus xylanilyticus". In: *Journal of biotechnology* 125.3 (2006), pp. 338–350 (cit. on p. 76).

[194] Masahiro Watanabe, Tomohiko Matsuzawa, and Katsuro Yaoi. "Rational protein design for thermostabilization of glycoside hydrolases based on structural analysis". In: *Applied microbiology and biotechnology* 102.20 (2018), pp. 8677–8684 (cit. on p. 76).

[195] Claire Dumon et al. "Engineering hyperthermostability into a GH11 xylanase is mediated by subtle changes to protein structure". In: *Journal of Biological Chemistry* 283.33 (2008), pp. 22557–22564 (cit. on pp. 77, 78).

[196] DA Case et al. "AMBER 2018; 2018". In: *University of California, San Francisco* () (cit. on p. 77).

[197] Romelia Salomon-Ferrer et al. “Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald”. In: *Journal of chemical theory and computation* 9.9 (2013), pp. 3878–3888 (cit. on p. 77).

[198] Andreas W Götz et al. “Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized born”. In: *Journal of chemical theory and computation* 8.5 (2012), pp. 1542–1555 (cit. on p. 77).

[199] Scott Le Grand, Andreas W Götz, and Ross C Walker. “SPFP: Speed without compromise—A mixed precision model for GPU accelerated molecular dynamics simulations”. In: *Computer Physics Communications* 184.2 (2013), pp. 374–380 (cit. on p. 77).

[200] Karl N Kirschner et al. “GLYCAM06: a generalizable biomolecular force field. Carbohydrates”. In: *Journal of computational chemistry* 29.4 (2008), pp. 622–655 (cit. on pp. 77, 78, 107).

[201] William L Jorgensen et al. “Comparison of simple potential functions for simulating liquid water”. In: *The Journal of chemical physics* 79.2 (1983), pp. 926–935 (cit. on pp. 78, 133).

[202] Tom Darden, Darrin York, and Lee Pedersen. “Particle mesh Ewald: An N . log (N) method for Ewald sums in large systems”. In: *The Journal of chemical physics* 98.12 (1993), pp. 10089–10092 (cit. on pp. 78, 133).

[203] WF Van Gunsteren and HJC Berendsen. “Algorithms for macromolecular dynamics and constraint dynamics”. In: *Molecular Physics* 34.5 (1977), pp. 1311–1327 (cit. on pp. 78, 133).

[204] Daniel R Roe and Thomas E Cheatham III. “PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data”. In: *Journal of chemical theory and computation* 9.7 (2013), pp. 3084–3095 (cit. on p. 78).

[205] William Humphrey, Andrew Dalke, and Klaus Schulten. “VMD – Visual Molecular Dynamics”. In: *Journal of Molecular Graphics* 14 (1996), pp. 33–38 (cit. on p. 78).

[206] Wei Tian et al. “CASTp 3.0: computed atlas of surface topography of proteins”. In: *Nucleic acids research* 46.W1 (2018), W363–W367 (cit. on p. 79).

[207] Sebastian Salentin et al. “PLIP: fully automated protein–ligand interaction profiler”. In: *Nucleic acids research* 43.W1 (2015), W443–W447 (cit. on p. 79).

[208] PH Hünenberger, AE Mark, and WF Van Gunsteren. “Fluctuation and cross-correlation analysis of protein motions observed in nanosecond molecular dynamics simulations”. In: *Journal of molecular biology* 252.4 (1995), pp. 492–503 (cit. on p. 80).

[209] Kota Kasahara, Ikuo Fukuda, and Haruki Nakamura. "A novel approach of dynamic cross correlation analysis on molecular dynamics simulations and its application to Ets1 dimer–DNA complex". In: *PloS one* 9.11 (2014), e112419 (cit. on p. 80).

[210] Andrea Amadei, Antonius BM Linssen, and Herman JC Berendsen. "Essential dynamics of proteins". In: *Proteins: Structure, Function, and Bioinformatics* 17.4 (1993), pp. 412–425 (cit. on p. 80).

[211] Charles C David and Donald J Jacobs. "Principal component analysis: a method for determining the essential dynamics of proteins". In: *Protein dynamics*. Springer, 2014, pp. 193–226 (cit. on p. 80).

[212] Daniel R Roe, Christina Bergonzo, and Thomas E Cheatham III. "Evaluation of enhanced sampling provided by accelerated molecular dynamics with Hamiltonian replica exchange methods". In: *The Journal of Physical Chemistry B* 118.13 (2014), pp. 3543–3552 (cit. on p. 81).

[213] Ivano Tavernelli, Simona Cotesta, and Ernesto E Di Iorio. "Protein dynamics, thermal stability, and free-energy landscapes: a molecular dynamics investigation". In: *Biophysical journal* 85.4 (2003), pp. 2641–2649 (cit. on p. 81).

[214] Martin Gruebele. "Downhill protein folding: evolution meets physics". In: *Comptes rendus biologies* 328.8 (2005), pp. 701–712 (cit. on p. 81).

[215] Elena Papaleo et al. "Free-energy landscape, principal component analysis, and structural clustering to identify representative conformations from molecular dynamics simulations: the myoglobin case". In: *Journal of molecular graphics and modelling* 27.8 (2009), pp. 889–899 (cit. on p. 81).

[216] Abbas Razvi and J Martin Scholtz. "Lessons in stability from thermophilic proteins". In: *Protein Science* 15.7 (2006), pp. 1569–1578 (cit. on p. 81).

[217] Karl Gruber et al. "Thermophilic xylanase from Thermomyces lanuginosus: high-resolution X-ray structure and modeling studies". In: *Biochemistry* 37.39 (1998), pp. 13475–13485 (cit. on p. 96).

[218] Tariq A Tahir et al. "Specific Characterization of Substrate and Inhibitor Binding Sites of a Glycosyl Hydrolase Family 11 Xylanase fromAspergillus niger". In: *Journal of Biological Chemistry* 277.46 (2002), pp. 44035–44043 (cit. on p. 103).

[219] Andreas S Bommarius and Marietou F Paye. "Stabilizing biocatalysts". In: *Chemical Society Reviews* 42.15 (2013), pp. 6534–6565 (cit. on p. 106).

[220] Haoran Yu and He Huang. "Engineering proteins for thermostability through rigidifying flexible sites". In: *Biotechnology advances* 32.2 (2014), pp. 308–315 (cit. on p. 106).

[221] Haoran Yu et al. "Two strategies to engineer flexible loops for improved enzyme thermostability". In: *Scientific reports* 7 (2017), p. 41212 (cit. on p. 106).

[222] Vishal Kumar, Julia Marin-Navarro, and Pratyoosh Shukla. "Thermostable microbial xylanases for pulp and paper industries: trends, applications and further perspectives". In: *World Journal of Microbiology and Biotechnology* 32.2 (2016), p. 34 (cit. on p. 106).

[223] Hajime SHIBUYA, Satoshi KANEKO, and Kiyoshi HAYASHI. "Enhancement of the thermostability and hydrolytic activity of xylanase by random gene shuffling". In: *Biochemical Journal* 349.2 (2000), pp. 651–656 (cit. on p. 106).

[224] Shan Zhang et al. "Seven N-terminal residues of a thermophilic xylanase are sufficient to confer hyperthermostability on its mesophilic counterpart". In: *PloS one* 9.1 (2014), e87632 (cit. on p. 106).

[225] Yifan Bu et al. "Engineering improved thermostability of the GH11 xylanase from Neocallimastix patriciarum via computational library design". In: *Applied microbiology and biotechnology* 102.8 (2018), pp. 3675–3685 (cit. on p. 106).

[226] Raphael Guerois, Jens Erik Nielsen, and Luis Serrano. "Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations". In: *Journal of molecular biology* 320.2 (2002), pp. 369–387 (cit. on p. 106).

[227] Elizabeth H Kellogg, Andrew Leaver-Fay, and David Baker. "Role of conformational sampling in computing mutation-induced changes in protein structure and stability". In: *Proteins: Structure, Function, and Bioinformatics* 79.3 (2011), pp. 830–838 (cit. on p. 106).

[228] Peng Xiong, Quan Chen, and Haiyan Liu. "Computational protein design under a given backbone structure with the ABACUS statistical energy function". In: *Computational Protein Design*. Springer, 2017, pp. 217–226 (cit. on p. 106).

[229] Samuel H Gellman. "Minimal model systems for β-sheet secondary structure in proteins". In: *Current opinion in chemical biology* 2.6 (1998), pp. 717–725 (cit. on p. 117).

[230] Maria D Crespo et al. "Context-dependent effects of proline residues on the stability and folding pathway of ubiquitin". In: *European journal of biochemistry* 271.22 (2004), pp. 4474–4484 (cit. on p. 117).

[231] Stephen J Eyles et al. "Roles of proline residues in the structure and folding of a β-clam protein". In: *Peptides for the New Millennium*. Springer, 2002, pp. 313–315 (cit. on p. 117).

[232] Georges Köhler and Cesar Milstein. "Continuous cultures of fused cells secreting antibody of predefined specificity". In: *nature* 256.5517 (1975), pp. 495–497 (cit. on p. 123).

[233] Joe J Tjandra, Lanny Ramadi, and Ian FC McKenzie. "Development of human anti-murine antibody (HAMA) response in patients". In: *Immunology and cell biology* 68.6 (1990), pp. 367–376 (cit. on p. 123).

[234] Anna M Wu and Peter D Senter. "Arming antibodies: prospects and challenges for immunoconjugates". In: *Nature biotechnology* 23.9 (2005), pp. 1137–1146 (cit. on p. 123).

[235] Janusz Wesolowski et al. "Single domain antibodies: promising experimental and therapeutic tools in infection and immunity". In: *Medical microbiology and immunology* 198.3 (2009), pp. 157–174 (cit. on p. 123).

[236] Serge Muyldermans and Marc Lauwereys. "Unique single-domain antigen binding fragments derived from naturally occurring camel heavy-chain antibodies". In: *Journal of Molecular Recognition* 12.2 (1999), pp. 131–140 (cit. on p. 123).

[237] S Muyldermans et al. "Camelid immunoglobulins and nanobody technology". In: *Veterinary immunology and immunopathology* 128.1-3 (2009), pp. 178–183 (cit. on p. 123).

[238] CTSG Hamers-Casterman et al. "Naturally occurring antibodies devoid of light chains". In: *Nature* 363.6428 (1993), pp. 446–448 (cit. on p. 124).

[239] K Decanniere, S Muyldermans, and L Wyns. "Canonical antigen-binding loop structures in immunoglobulins: more structures, more canonical classes?" In: *Journal of molecular biology* 300.1 (2000), pp. 83–91 (cit. on p. 125).

[240] Viet Khong Nguyen et al. "Camel heavy-chain antibodies: diverse germline VHH and specific mechanisms enlarge the antigen-binding repertoire". In: *The EMBO journal* 19.5 (2000), pp. 921–930 (cit. on p. 125).

[241] Viet Khong Nguyen, Aline Desmyter, and Serge Muyldermans. "Functional heavy-chain antibodies in Camelidae". In: (2001) (cit. on p. 125).

[242] Laura S Mitchell and Lucy J Colwell. "Comparative analysis of nanobody sequence and structure data". In: *Proteins: Structure, Function, and Bioinformatics* 86.7 (2018), pp. 697–706 (cit. on p. 125).

[243] Erwin De Genst et al. "Molecular basis for the preferential cleft recognition by dromedary heavy-chain antibodies". In: *Proceedings of the National Academy of Sciences* 103.12 (2006), pp. 4586–4591 (cit. on p. 125).

[244] Jessica R Ingram, Florian I Schmidt, and Hidde L Ploegh. "Exploiting nanobodies' singular traits". In: *Annual review of immunology* 36 (2018), pp. 695–715 (cit. on pp. 125, 127).

[245] Cecile Vincke et al. "General strategy to humanize a camelid single-domain antibody and identification of a universal humanized nanobody scaffold". In: *Journal of Biological Chemistry* 284.5 (2009), pp. 3273–3284 (cit. on p. 125).

[246] Julian Davies and Lutz Riechmann. "'Camelising'human antibody fragments: NMR studies on VH domains". In: *FEBS letters* 339.3 (1994), pp. 285–290 (cit. on p. 126).

[247] Joost A Kolkman and Debbie A Law. "Nanobodies–from llamas to therapeutic proteins". In: *Drug discovery today: technologies* 7.2 (2010), e139–e146 (cit. on p. 126).

[248] Martin F Flajnik, Nick Deschacht, and Serge Muyldermans. "A case of convergence: why did a simple alternative to canonical antibodies arise in sharks and camels?" In: *PLoS Biol* 9.8 (2011), e1001120 (cit. on p. 126).

[249] Serge Muyldermans. "Nanobodies: natural single-domain antibodies". In: *Annual review of biochemistry* 82 (2013), pp. 775–797 (cit. on p. 126).

[250] Mireille Dumoulin et al. "Single-domain antibody fragments with high conformational stability". In: *Protein Science* 11.3 (2002), pp. 500–515 (cit. on p. 126).

[251] RHJ Van der Linden et al. "Comparison of physical chemical properties of llama VHH antibody fragments and mouse monoclonal antibodies". In: *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology* 1431.1 (1999), pp. 37–46 (cit. on p. 126).

[252] Sandrine Moutel et al. "NaLi-H1: A universal synthetic library of humanized nanobodies providing highly functional antibodies and intrabodies". In: *Elife* 5 (2016), e16228 (cit. on pp. 126, 127, 130, 134, 140).

[253] Daniel Christ, Kristoffer Famm, and Greg Winter. "Repertoires of aggregation-resistant human antibody domains". In: *Protein Engineering, Design & Selection* 20.8 (2007), pp. 413–416 (cit. on p. 126).

[254] Ole Aalund Mandrup et al. "A novel heavy domain antibody library with functionally optimized complementarity determining regions". In: *PloS one* 8.10 (2013), e76834 (cit. on p. 126).

[255] Ulrich Rothbauer et al. "Targeting and tracing antigens in live cells with fluorescent nanobodies". In: *Nature methods* 3.11 (2006), pp. 887–889 (cit. on pp. 126, 140).

[256] Els Pardon et al. "A general protocol for the generation of Nanobodies for structural biology". In: *Nature protocols* 9.3 (2014), pp. 674–693 (cit. on p. 126).

[257] Jonas Ries et al. "A simple, versatile method for GFP-based super-resolution microscopy via nanobodies". In: *Nature methods* 9.6 (2012), pp. 582–584 (cit. on p. 126).

[258] Rubel Chakravarty, Shreya Goel, and Weibo Cai. "Nanobody: the "magic bullet" for molecular imaging?" In: *Theranostics* 4.4 (2014), p. 386 (cit. on p. 127).

[259] Sabrina Oliveira et al. “Targeting tumors with nanobodies for cancer imaging and therapy”. In: *Journal of Controlled Release* 172.3 (2013), pp. 607–617 (cit. on p. 127).

[260] Isabel Van Audenhove and Jan Gettemans. “Nanobodies as versatile tools to understand, diagnose, visualize and treat cancer”. In: *EBioMedicine* 8 (2016), pp. 40–48 (cit. on p. 127).

[261] Ivana Jovčevska and Serge Muyldermans. “The therapeutic potential of nanobodies”. In: *BioDrugs* 34.1 (2020), pp. 11–26 (cit. on p. 127).

[262] RJ Pantazes and Costas D Maranas. “OptCDR: a general computational method for the design of antibody complementarity determining regions for targeted epitope binding”. In: *Protein Engineering, Design & Selection* 23.11 (2010), pp. 849–858 (cit. on p. 127).

[263] Tong Li, Robert J Pantazes, and Costas D Maranas. “OptMAVEn–a new framework for the de novo design of antibody variable region models targeting specific antigen epitopes”. In: *PloS one* 9.8 (2014), e105954 (cit. on p. 127).

[264] Gideon D Lapidoth et al. “AbDesign: A n algorithm for combinatorial backbone design guided by natural conformations and sequences”. In: *Proteins: Structure, Function, and Bioinformatics* 83.8 (2015), pp. 1385–1406 (cit. on p. 127).

[265] Jared Adolf-Bryfogle et al. “RosettaAntibodyDesign (RAbD): A general framework for computational antibody design”. In: *PLoS computational biology* 14.4 (2018), e1006112 (cit. on p. 127).

[266] Richard A Norman et al. “Computational approaches to therapeutic antibody design: established methods and emerging trends”. In: *Briefings in bioinformatics* 21.5 (2020), pp. 1549–1567 (cit. on p. 127).

[267] Davide Ferrari et al. “A novel nanobody scaffold optimized for bacterial expression and suitable for the construction of ribosome display libraries”. In: *Molecular Biotechnology* 62.1 (2020), pp. 43–55 (cit. on p. 127).

[268] Nicolas Bery et al. “A targeted protein degradation cell-based screening for nanobodies selective toward the cellular RHOB GTP-bound conformation”. In: *Cell chemical biology* 26.11 (2019), pp. 1544–1558 (cit. on pp. 128, 139).

[269] Jonathan CY Tang et al. “Detection and manipulation of live antigen-expressing cells using conditionally stable nanobodies”. In: *Elife* 5 (2016), e15312 (cit. on p. 129).

[270] David Simoncini, Thomas Schiex, and Kam YJ Zhang. “Balancing exploration and exploitation in population-based sampling improves fragment-based de novo protein structure prediction”. In: *Proteins: Structure, Function, and Bioinformatics* 85.5 (2017), pp. 852–858 (cit. on p. 133).

[271] Sandrine Moutel et al. “A multi-Fc-species system for recombinant antibody production”. In: *BMC biotechnology* 9.1 (2009), p. 14 (cit. on pp. 134, 140).

[272] Denis Jullien et al. "Chromatibody, a novel non-invasive molecular tool to explore and manipulate chromatin in living cells". In: *Journal of cell science* 129.13 (2016), pp. 2673–2683 (cit. on p. 140).

[273] Adegoke Ojewole et al. "OSPREY predicts resistance mutations using positive and negative computational protein design". In: *Computational Protein Design*. Springer, 2017, pp. 291–306 (cit. on pp. 146, 152).

[274] Teresa Kaserer and Julian Blagg. "Combining mutational signatures, clonal fitness, and drug affinity to define drug-specific resistance mutations in cancer". In: *Cell chemical biology* 25.11 (2018), pp. 1359–1371 (cit. on pp. 146, 152).

Abstract: Proteins are fundamental components of life. Over the billions of years of evolution, proteins have evolved to perform certain functions better and faster or to achieve new functions in order to pursue the biological needs under diverse and changing conditions. The field of protein engineering is becoming a research domain of great importance. The interest of proteins with new or improved properties is increasing in health, nano/biotechnology and green chemistry.

Computational Protein Design (CPD) plays a critical role in advancing the field of protein engineering and accelerating the delivery of novel proteins displaying high specificity, high efficiency and better stability. The CPD problem can be formalized as an optimization problem. Using an all-atom energy function and a reliable search method, CPD tries to identify amino acid sequences that fold into a target structure and ultimately perform a desired function. The traditional Single State Protein Design (SSD) contrasts with the increasing evidence that proteins do not remain in a unique conformational state but rather sample conformational ensembles. In this thesis we propose a MultiState Design (MSD) method which aims at alleviating SSD limitations by simultaneously considering several conformational states.

In the second part of this thesis, MSD was applied on two projects that led to an experimental characterization and validation. These two projects concern two application domains: health and white biotechnologies. The first one targets GH11 Xylanases. To understand the molecular basis underlying its thermal stability and activity, Molecular Dynamics simulations were used and revealed useful characteristics to design this enzyme. This produced GH11 xylanases with improved thermal stability and catalytic activity. The second project concerns the design of a synthetic humanized nanobody scaffold. The resulting nanobody is highly expressed and shows suitable affinity with different CDR loops.

Keywords: Computational Protein Design, Multistate Design, Molecular Dynamics simulations, GH11 Xylanases, Nanobodies

Résumé: Les protéines sont des composants fondamentaux de la vie. Au cours des milliards d'années d'évolution, elles ont évolué pour mieux remplir leurs fonctions ou pour réaliser de nouvelles fonctions, afin de répondre aux besoins biologiques dans des conditions diverses et changeantes. L'ingénierie des protéines est ainsi un domaine de recherche d'une grande importance. L'intérêt pour les protéines ayant des propriétés nouvelles ou améliorées augmente en santé, en bio/nanotechnologie et en chimie verte.

Le design computationnel de protéines (CPD) joue un rôle essentiel pour faire progresser le domaine de l'ingénierie des protéines et accélérer la conception de nouvelles protéines présentant une haute spécificité, une grande efficacité et une meilleure stabilité. Le problème de CPD peut être formalisé comme un problème d'optimisation. A l'aide d'une fonction d'énergie et d'une méthode de recherche fiable, le CPD tente d'identifier les séquences d'acides aminés qui adoptent une structure cible et qui remplissent une fonction souhaitée. Le modèle classique à état unique (Single State Protein Design - SSD) néglige le fait que les protéines adoptent un ensemble d'états conformationnels. Dans cette thèse, nous proposons une méthode de conception multi-états (MSD) qui vise à atténuer les limitations du SSD en considérant efficacement et simultanément plusieurs états conformationnels.

Dans la deuxième partie de cette thèse, le MSD a été appliqué à deux projets avec une caractérisation et une validation expérimentale. Ces projets concernent deux domaines d'application différents : la santé et les biotechnologies blanches. Le premier concerne les xylanases GH11. Pour comprendre les bases moléculaires qui sous-tendent leur stabilité et leur activité, des simulations de dynamique moléculaire ont révélé des caractéristiques utiles pour la conception de mutants plus thermostables et plus actifs. Le second projet concerne la conception d'un squelette de nano-anticorps humanisés synthétiques. Certains de ces chassis ont montré un haut niveau d'expression et l'affinité attendue avec différentes boucles CDR.

Keywords: Design Computationnel de Protéines, Design Multi-états, Dynamique Moléculaire, Xylanases GH11, Nano-anticorps