Université Paris.Diderot - Paris 7

École Doctorale B3MI - ED 516
Biochimie, Biothérapies, Biologie Moléculaire et Infectiologie

Thèse de Doctorat

Spécialité : Bioinformatique, analyse des génomes et modélization

Maria Kalimeri

Les protéines thermophiles sont-elles rigides ou flexibles ?
Une étude *in silico*.

Thèse dirigée par Fabio Sterpone*, soutenue publiquement le 22 Septembre 2014 devant le jury composé de :*

Fuchs Patrick, Maître de conférences, Université Paris Diderot (Président)
Nilsson Lennart, Professeur, Karolinska Institute, Suéde (Rapporteur)
Senet Patrick, Professeur, Université de Bourgogne (Rapporteur)
Derreumaux Philippe, Professeur, Université Paris Diderot (Examinateur)
Madern Dominique, IRHC, CNRS/IBS Grenoble (Examinateur)
Melchionna Simone, Chercheur, IPCF - CNR, Italie (Invitè)
Sterpone Fabio, CR1, CNRS, Université Paris Diderot, (Directeur de thèse)

Are thermophilic proteins rigid or flexible? An *in silico* investigation

Maria Kalimeri

A thesis supervised by
Dr. Fabio Sterpone

Submitted for the degree of
Doctor of Paris Diderot University

Paris, September 2014

"I think nature's imagination is so much greater than man's,
she's never going to let us relax."

- Richard P. Feynman

Résumé en français

Introduction et motivation

La vie sur terre est adaptée à un large éventail de conditions physico-chimiques[1]. Bien que la vaste majorité des espèces non microbiennes survit confortablement autour de conditions ambiantes standards de température et de pression (SATP)[2], il existe des microorganismes non seulement capables de survivre mais se développant dans des conditions extrêmes. Ceux-ci sont correctement appelés extrêmophiles, du Latin "extrēmus" et du Grecque "φιλῶ"=aimer.

Parmi les extrêmophiles, d'un intérêt particulier pour nous est la sous-classe appelée *thermophiles*, faisant référence aux organismes ayant la capacité de résisté aux hautes températures [2]. Les thermophiles sont essentiellement des bactéries ou archées bien que quelques eucaryotes, notamment des algues et champignons, existent aussi [3]. Ils peuvent être trouvés dans des zones volcaniques et géothermales, sur terre ou dans des évents chauds au fond de l'océan.

Les thermophiles, comme les extrêmophiles en général, s'adaptent à leur environnement en employant différents mécanismes. Dans cette étude, nous nous concentrons sur le niveau moléculaire et en particulier sur les mécanismes physiques et chimiques qui permettent aux protéines thermophiles de supporter la chaleur.

Les protéines sont parmi les plus importants composants des organismes vivants, même les plus simples [4]. Ce sont de grandes molécules organiques constituées d'une ou plusieurs chaînes de composés organiques, appelés acides aminés. Elles ont une grande variété de fonctions comprenant la catalyse de réaction chimiques importantes pour la vie, la réplication de l'ADN, la synthèse protéique, le transport, e.t.c. Généralement, elles sont pleinement fonctionnelles dans leur état appelée natif qui est une complexe structure tridimensionnelle repliée, dont la forme exacte est imposée par leur séquence primaire en acides aminés [3] et les paramètres physico-chimiques de leur environnement [5]. C'est à dire, pour une valeur de

[1] Un organisme vivant, dans ce contexte, est une entité organisée composée d'une ou plusieurs cellules capable de croitre, métaboliser, se reproduire, et répondre à des *stimuli* externes et de s'adapter à des changements environnementaux [1]

[2] SATP: 25 °C (298.15 K) et 0.987 atm (100 kPa)

[3] Les protéines intrinsèquement désordonnées sont une exception.

pH donnée, une séquence primaire spécifique ne peut résulter en deux repliements tridimensionnels –substantiellement– différents. Approximativement, l'état replié des protéines est maintenu intact grâce aux forces cohésives entre les acides aminés la constituant, i.e. les interactions de Van der Waals, l'électrostatique, e.t.c. De ce fait, les protéines repliés sont des systèmes de la matière molle, certaines étant plus flexibles que d'autres.

En fait, il y a une corrélation délicate entre flexibilité, stabilité et fonction [6–9]. Une protéine a besoin de maintenir un repliement stable[19]; mais avec une quantité appropriée de flexibilité afin de faciliter les processus ayant trait à sa fonction, e.g. la liaison de substrats et cofacteurs [10]. Les protéines provenant d'organismes thermophiles constituent un modèle unique qui pourrait permettre de nous aider à comprendre les mécanismes de bases de cette subtile interaction. La sélection naturelle a conduit (ou maintenue) leur stabilité et fonction optimale à des températures qui causeraient le dépliement d'autres protéines provenant d'organismes vivant à des températures plus modérées, c'est à dire les mésophiles. Il se trouve que les protéines thermophiles sont généralement inactives à température ambiante [11, 12]. Cette observation a conduit à l'idée qu'elles possèderaient une matrice plus rigides que leurs homologues mésophiles [13]. Le caractère universel de ce paradigme de rigidité a cependant été questionné à de nombreuses reprises à la fois par les expériences et la simulation [14].

Comprendre les mécanismes de cette meilleure stabilité protéique a des implications médicales et technologiques. Par exemple, les protéines thermostables sont déjà employées pour la biocatalyse industrielle aux températures extrêmes, un processus fondamental dans l'industrie alimentaire, les produits laitiers, le papier, les biocarburants et bien d'autres [15, 16]. Un autre aspect important concerne les maladies à protéines mal repliées comme les maladies d'Alzheimer ou de Creutzfeldt-Jakob, dont la cause est le dépliement ou le dépliement partielle de protéines globulaires et leur agrégation subséquente sous la forme de fibres amyloïdes [17]. L'étude de protéines thermophiles pourrait aidée à comprendre les mécanismes derrière cette amyloïdogénèse [18, 19].

Le but de cette étude est d'identifier, via l'utilisation de simulations de Dynamique Moléculaire, les caractéristiques distinguant les protéines thermophiles des mésophiles, en élaborant en particulier sur le paradigme de rigidité mentionné ci-dessus. De telles caractéristiques pourraient inspirer de nouvelles stratégies pour le dessein de protéines thermostables [20].

Cette thèse est organisé de la manière suivante. Dans le Chapitre 1, nous discutons d'abord les bases de la structure des protéines et subséquemment présentons ce qui est déjà connu des protéines thermophiles en ce qui concerne leur structure et dynamique. Ensuite la base thermodynamique pour la stabilité des protéines

est décrite et le chapitre se termine par une perspective sur l'étude actuelle. Dans le Chapitre 2, nous présentons la Méthodologie utilisée pour attaquer le problème: Dans la première section les principes de la simulation par MD sont présentés, en développant les techniques choisies pour notre approche. La deuxième section traite des méthodes d'échantillonnage augmenté utilisées tandis que dans la dernière section nous présentons les approches de post-traitement de la MD que nous employons. Dans le Chapitre 3, nous étudions les dynamiques à température ambiante de deux protéines globulaires homologues, les domaines G hyperthermophiles et mésophiles extraies des protéines EF-Tu et EF-1α, respectivement. Nos résultats mettent en question le paradigme de rigidité pour cette protéine hyperthermophile et indiquent qu'un partitionnement régulier en parties flexibles et rigides le long de la séquence pourrait empêcher le dépliement de la protéine. Dans le Chapitre 4 nous étendons notre étude des deux homologues en examinant leur stabilité cinétique à de plus hautes températures. En complément des simulations tout-atome du trimère entier de protéines EF, nous individualisons le point faible des protéines mésophiles, étroitement lié à leur fonction enzymatique et donnons quelques perspectives pour un travail futur. En gardant les deux domaines G sous le microscope, le Chapitre 5 traite des aspects thermodynamiques de leur stabilité thermale, tout en testant pour la première fois la capacité d'un champ de force gros grain à distinguer le différent contenu en stabilité thermale de deux protéines. Dans le Chapitre 6, nous étendons notre étude à deux autres homologues, une malate déshydrogénase tétramérique mésophile et thermophile. Nous trouvons une étroite corrélation entre la flexibilité de l'état replié, la stabilité thermale de la protéine et son état d'oligomérisation: l'Oligomérisation rigidifie substantiellement l'homologue thermophile avec un effet potentiel sur son affinité de liaison au substrat. Nous concluons en résumant nos résultats et en posant quelques questions ouvertes et perspectives pour le future.

Conclusions

Tel que discuté ci dedans, la controverse sur la rigidité/flexibilité émane principalement de deux sources. La première est la nature complexe des protéines qui ne permet pas une définition unique de la flexibilité [21]. La seconde, supportée aussi par les résultats de cette étude, est que la Nature n'a pas de panacée contre le stress à haute température. Elle se montre plus inventive, employant différentes stratégies en fonction de la paire d'homologues ou de la famille de protéines.

Donc, sont-elles rigides ou flexibles? Cette étude met sous le microscope deux cas d'études caractéristiques et les disséquant en utilisant la technique de

référence des techniques *in silico*, les simulations par Dynamique Moléculaire. La puissante résolution atomistique de cette technique, avec la possibilité d'explorer différentes échelles de temps, rends l'ambigüité du terme "flexibilité des protéines" immédiatement évidente.

Dans la première étude de cas d'une paire de protéines monomériques homologues, des domaines G hyperthermophile et mésophile, la thermophilie est en corrélation avec la distribution régulière de parties flexibles et rigides le long de la séquence ainsi qu'à des fluctuations conformationnelles plus grandes sur l'échelle de la microseconde. Cependant, la variante mésophile maintient une région particulière qui est plus flexible. Ceci est appelé le switch I et peut exister dans deux conformations très différentes qui impose si l'enzyme est "allumé" ou "éteinte". Cette capacité à se convertir facilement d'une conformation à l'autre nécessite, en dehors d'une activation par une autre enzyme (comportement allostérique), une fixation lâche au corps de la protéine. Cette caractéristique est au final le "tendon d'Achille" de la protéine mésophile: lorsque la température augmente, elle conduit au dépliement.

La seconde étude de cas traite deux protéines tétramériques homologues, des malate déshydrogénases thermophile et mésophile. Ici, les mouvements à l'équilibre de la variante thermophile sont "bloqués" à la fois sur une échelle de taille locale et globale, un comportement résultant des interactions effectives électrostatiques et hydrophobes au niveau de l'interface inter-domaine. La raideur de la matrice protéique thermophile se propage depuis l'interface entre domaines jusqu'au site actif où, en combinaison avec des mutations clefs, la conformation fermée préférée de la boucle du site de liaison gêne l'accessibilité à la poche catalytique. D'un autre côté la protéine mésophile est caractérisé par sa dynamique d'ouverture/fermeture de cette boucle.

De cette étude et d'autres, il devient apparent que pour les protéines mésophiles, comparées à leurs homologues thermophiles, la stabilité thermale est compromise dans l'intérêt d'une activité optimale. Une activité enzymatique efficace aux conditions ambiantes est le résultat d'une organisation subtile de la matrice protéique et de ses fluctuations conformationnelles. Une haute température pourrait aisément perturber cet équilibre en acheminant de l'énergie selon des modes particuliers de la protéine conduisant à son dépliement. Comme exemple caractéristique, un travail récent à trouver qu'une plus haute stabilité pour l'acylphosphatase musculaire humaine peut être atteinte via une augmentation de l'entropie conformationnelle de l'ensemble de l'état replié [22]. Cependant, les mutants stabilisés étaient déficients en activité enzymatique.

En résumé, les protéines thermophiles ne sont pas nécessairement rigides, tel que généralement conçu, bien que cette route puisse être généralement préférée.

Notre étude, encore une fois, suggère que clarifié la relation entre stabilité, flexibilité et fonction nécessite d'individualiser les degrés de libertés clefs et d'explorer toutes les différentes échelles de temps des dynamiques associées [23]. Cette étude démontre aussi comment les méthodes de regroupement et de réseau, utilisées pour la représentation de l'espace conformationnel explorée par les protéines, sont de puissants outils d'analyse pour se renseigner sur la flexibilité protéique. Cette boîte à outils est actuellement employée pour distinguer les différents comportement d'un large ensemble de protéines homologues.

Vers la conception de la résistance thermique. Ciblant les applications industrielles, l'étude présente vérifies une fois encore la relation inextricable entre des interactions électrostatiques clefs et "l'amour pour la chaleur". Dans l'étude des malate déshydrogénases tétramériques, même si la séquence de orthologue thermophile contient moins d'acides aminés chargés, les résidus ioniques sont placés dans des positions stratégiques augmentant la taille des groupes de ponts salins interfaciaux et renforçant la liaison inter-domaine effective. Traditionnellement, les pairs d'ions sont considérées comme des éléments structuraux conférant une rigidité locale à la matrice protéique, cependant, comme nous l'avons illustré en caractérisant nos études de cas, le remaniement de leurs paires d'ions et la flexibilité de leur réseaux peut être source de stabilité ainsi que les fluctuations conformationnelles.

Cette étude soutient aussi une stratégie alternative pour la stabilisation. Une stabilité protéique augmenté peut être atteinte en réglant l'étendue et la distribution de parties flexibles/rigides le long de la séquence primaire. De cette façon l'excitation thermale peut être efficacement contenue et absorbée.

Quelques questions ouvertes et perspectives: Fonctionnalité à haute température. Cette étude pose clairement la question de savoir si la température est le seul paramètre à prendre en compte lors de la description de la fonctionnalité des thermophiles. D'après l'image des "états correspondants", le mécanisme décrivant la fonctionnalité des homologues est identique mais caractérisée par une différence d'énergie d'activation: la flexibilité requise pour la fonctionnalité est la même à la température optimale correspondante de l'organisme hôte. Cependant, les deux cas étudiés ici indiquent que d'autres spécificités viennent avec le fonctionnement à haute température. Par exemple en considérant les domaines G EF-Tu et -1α, nous savons que l'activité de la variante mésophile est associée à un impressionant changement conformationnel localisé au niveau de la région du switch I. Pour autant que nous le sachions, l'existence d'un tel changement conformationnel chez l'hyperthermophile reste inconnu. Nous avons vu que la région du

switch I pour cette variante est caractérisé par l'insertion de motifs structuraux qui pourraient affectés les changements conformationnels associés à l'activité GTPase. Donc la fonction de l'hyperthermophile à haute température n'est pas simplement le résultat d'une activation thermale des mêmes degrés de liberté que pour la variante mésophile, mais pourrait impliquée d'autres chemins conformationnels. Dans le future proche, cet aspect pourrait être étudié en employant de longues simulations de MD, des méthodes d'échantillonnages avancées *ad hoc* ainsi que des modèles de protéines plus simplifiés [24].

Concernant les deux déshydrogénases tétramériques homologues, comme mentionné précédemment, l'accessibilité du site actif thermophile est restreinte à température ambiante en raison de la boucle rigidement fermée à son entrée. Au contraire la boucle respective chez le mésophile est caractérisé par une dynamique d'ouverture/fermeture. Toutefois, une augmentation de température ne suffit pas à activer le mouvement thermophile dans les échelles de temps explorées. Il serait intéressant d'évaluer comment cette boucle accède à l'état ouvert et le rôle potentiel de la proximité du substrat dans la facilitation du changement conformationnel. Ceci pourrait être explorer par de plus longues simulations ou expérimentalement, par exemple, en utilisant des expériences de dispersion-relaxation en RMN [12].

D'un autre côté, si l'idée de correspondance des états concernant la flexibilité est généralement vraie, il serait très intéressant d'explorer plus systématiquement s'il existe une corrélation entre les protéines thermophiles caractérisées comme rigide et leur état oligomérique. Cette suggestion est inspirée par l'effet observé de l'oligomérisation sur la protéine thermophile tétramérique de cette étude, ainsi que par une vue d'ensemble concernant les études réalisées jusqu'à présent (la discussion sur le sujet se trouve Chapitre 1).

Un autre aspect important de la stabilité des protéines concerne les conditions réelles d'encombrement du cytoplasme. Les protéines *in vivo* fonctionnent rarement de manière isolée et, de fait, il a été estimé que les interactions protéine-protéine représente une stabilisation additionnelle de $\sim 2 - 4\, k_B T$ [25]. En même temps il a été suggéré que les protéines thermophiles et mésophiles ont des propriétés de diffusion différentes selon les conditions d'encombrement, ce qui pourrait contribuer substantiellement à leur différente stabilité thermale [26]. Une puissante méthodologie nouvellement développée en interne, combinant le champ de force gros grain OPEP pour les protéines [27] et une représentation mésoscopique du solvant basée sur la méthode de Boltzmann sur réseau [28, 29] est actuellement employée pour répondre à ce problème d'une manière plus systématique [30].

Au niveau suivant, en approchant le sujet de la fonctionnalité protéique de manière plus directe, la question est de savoir comment les différentes conformations explorées par les protéines affectent la réactivité chimique. En modélisation, ce

problème nécessite l'application d'un mélange de méthodologies quantique/classique (QM/MM).

Dans une perpétuelle recherche de tendance, il y a un peu moins d'une décennie une étude suggérait deux mécanismes physiques majeurs pour la stabilisation thermale des protéines en fonction de l'histoire évolutive de l'organisme source, l'un basé sur la structure et l'autre sur la séquence [31]. Les protéines d'organismes provenant d'un environnement chaud (ici les archées) ont une structure bien plus compacte et un cœur hydrophobe. D'un autre côté, les protéines venant d'organismes originellement mésophiles ayant ensuite recolonisé un environnement plus chaud (ici les bactéries) restent structurellement semblable aux homologues mésophiles mais présentent des substitutions de séquence résultant en des interactions clefs dans le repliement final. Cependant, cette classification stricte de l'histoire évolutive dépendant du domaine du vivant auquel l'organisme appartient (archée ou bactérie) n'a pas de base solide. En effet, des études plus récentes montrent que les ancêtre de bactéries étaient aussi thermophiles [32, 33]. En même temps des études structurales, bien qu'elles placent les choses dans une première perspective informative, négligent la dynamique et sont basées sur des structures cristallographiques résolues à basses températures. Il est donc bon de compléter ces études, lorsque possible, par des études expérimentales et computationnelles pertinentes pour expliquer le rôle couplé de la dynamique et de la température.

Abstract

Understanding the relation between protein flexibility, stability and function remains one of the most challenging, open questions in biophysical chemistry. For example, proteins need to be flexible to facilitate substrate binding but locally rigid to sustain substrate specificity. Exemplary cases are enzymes from microorganisms that thrive at elevated temperatures, also referred to as thermophiles. These proteins are stable and functional at the high temperature regime but generally lack activity at ambient conditions. Therefore, their thermal stability has been correlated to enhanced mechanical rigidity through the *corresponding states* paradigm [34]. The generality of this view, however, has been questioned by a number of experimental and computational studies [35, 36].

In the present study, we employ the "gold standard" of computational techniques, namely Molecular Dynamics simulations, in order to identify microscopical characteristics that distinguish thermophilic from mesophilic proteins, elaborating in particular on the rigidity paradigm mentioned above.

We focus on two characteristic study-cases. In the first case, we compare two homologous globular proteins, the G-domains of a hyperthermophilic and a mesophilic elongation factor. Our findings question the rigidity paradigm for this pair and show that the hyperthermophilic variant has comparable if not enhanced flexibility depending on the length-scale. Moreover, at high temperature, while the hyperthermophilic protein is stable, the mesophilic one starts to unfold with its kinetic instability being localized at the switch I region, important to its enzymatic function [37]. This finding supports the view that optimal activity at ambient temperature is fine-tuned at the expense of an enhanced protein stability. A thermodynamic analysis, based on a coarse-grained model, suggests that the thermodynamic mechanism behind the thermal resistance of the hyperthermophilic protein is caused by a smaller heat capacity of unfolding as compared to its mesophilic homologue [38]. Finally, we expand our study to two other homologues, a mesophilic and a thermophilic tetrameric malate dehydrogenase. We find a tight correlation between the folded-state flexibility, the protein's thermal stability and its oligomerization state: Oligomerization substantially rigidifies the thermophilic homologue. Specifically, the binding-site loop is anchored down, hin-

dering the accessibility to the catalytic pocket. Temperature increase does not activate the motion of the loop at the explored timescales, something that opens the question of whether temperature is the only parameter to take into account when describing the functionality of thermophilic proteins.

Finally, some of our findings could inspire new strategies in the design of thermostable proteins.

Publications

Some of the results and ideas presented in this thesis have appeared before in:

1. F. Sterpone, P. Nguyen, M. Kalimeri, and P. Derreumaux. "Importance of the ion-pair interactions in the OPEP coarse-grained force field: Parametrization and validation". *J. Chem. Theory. Comput.* 9 (2013), pp. 4574-4584.

2. M. Kalimeri, O. Rahaman, S. Melchionna and F. Sterpone (2013). "How conformational flexibility stabilizes the hyperthermophilic elongation factor G-domain". *J. Phys. Chem. B* 117.44, pp. 13775-13785

3. F. Sterpone, S. Melchionna, P. Tuffery, S. Pasquali, N. Mousseau, T. Cragnolini, Y. Chebaro, J. St-Pierre, M. Kalimeri, A. Barducci, Y. Laurin, A. Tek, M. Baaden, P. Nguyen, and P. Derreumaux. "The OPEP protein model: From single molecules, amyloid formation, crowding and hydrodynamics to DNA/RNA systems". *Chem. Soc. Rev.* 43.13 (2014), pp. 4871-4893.

4. M. Kalimeri, P. Derreumaux and F. Sterpone (2014). "Are coarse-grained models apt to detect protein thermal stability? The case of OPEP force field". *J. Non-Cryst. Solids.* In press.

Acknowledgements

"How strange is the lot of us mortals! Each of us is here for a brief sojourn; for what purpose he knows not, though he sometimes thinks he senses it. But without deeper reflection one knows from daily life that one exists for other people – first of all for those upon whose smiles and well-being our own happiness is wholly dependent, and then for the many, unknown to us, to whose destinies we are bound by the ties of sympathy. A hundred times every day I remind myself that my inner and outer life are based on the labors of other men, living and dead, and that I must exert myself in order to give in the same measure as I have received and am still receiving..."

- Albert Einstein

Fabio, I could pause here with a full stop, skip three silent pages in a sort of Greek-drama way and then continue with whatever else I have to say. Yet, I have to try to put some of it in words. For it has been three years, most enlightening, most productive and most fun. So, "Fabio, thanks". The completion of this thesis owes to your brilliant and insightful guidance, unreserved support and advice as well as soothing abilities when things seemed desperate. Thank you for the time and energy spent, not only on discussing science, but also on teaching me by example how to be a good human being. I am not sure if you can immediately see the improvement but I can assure you that your important messages will keep echoing in my head, with that necessary and invaluable ingredient of humor. I wish that the fascinating journey in the biophysical world, that started under your guidance and friendship, goes on.

During these last, three, beautiful years, I made quite a few other friends that have contributed directly or indirectly to this work. First of all, I owe a big thanks to Simone Melchionna who shared his expertise in computing and scientific writing even if, unfortunately, due to time and distance restrictions my total number of "flight-hours" next to him are not that many. Next, I want to thank Dominique Madern, for pulling things back into their biological track when they tended to drift away. I learned many beautiful things from him and enjoyed a lot our collaboration along with Eric Girard and Romain Talon from IBS in Grenoble.

I am particularly grateful to Professors Patrick Senet and Lennart Nilsson for accepting the laborious work of reviewing my manuscript as well as to Patrick

Fuchs for accepting to join my thesis' jury.

Coming back to LBT, Tristan Cragnolini, my maximum-intersecting-time officemate, thank you for the brainstorming and the enlightening discussions and, along with the superset of all the others that passed from room 223, thank you for making this office such a funny and loud place to work in. Speaking about loudness, "Francesco, it has been a pleasure... for you of course". You will be truly missed!

A great thanks also to the director of our lab, Philippe Derreumaux but also to Jérôme Hénin, Charles Robert, Antoine Taly and Guillaume Stirnemann for their, sometimes vital, scientific advice, to Obaidur Rahaman for the fruitful collaboration and exchange of ideas, to all the other friends from LBT for making these years unforgettable by usually "forgetting" something sweet in the coffee room but also to Samuel Murail who never lost his courage, or his humor, when it came to asking me to join for lunch. A respectful, deep bow should also go to the people behind the scenes, my dearest friend Victoria Terziyan the bureaucratic Zorro, Geoffrey Letessier the only true master of Hades but also Yves Dapra and Franck Paraskiova whose secret superpowers sometimes magically solved the Gordian knot of cables.

The madmakers Alex Tek and Matthieu Chavent thank you for the occasional madmaking, but also for your technical and/or artistic help. It was Matthieu's award-winning, artistic hand for science that resulted in the mind-blowing representation of electrostatic interactions of Figure 3.1. As for Benoist Laurent, Benjamin Boyer and all the rest, you will always be, rightfully, remembered as "les inévitables" (not inspired by Victor Hugo).

Still, these three years in Paris would not have been as lovely without the deliciously warm soirées franco-arméno-indiennes chez Victoria et Satya. Mes très bons amis, au revoir au Nord ou au Sud!

My friend Kostas, our funny revitalizing coffee breaks at Nouvelle Mairie will remain indelibly marked in my memory. I will miss them as much as the rest of the parisian Greeks, including Giorgos the soul and spirit of Fondation Hellenique.

Oxymoron, but I feel I should pay my dues also to the economical crisis in my home country for motivating me to jump into this new fascinating field and get all the way to this day. Greece is a beautiful place with the best of friends that I miss all the more everyday and I had a hard time leaving.

However, I came back home, primarily to pay my respects and gratitude to the supervisor of my undergraduate thesis, mentor and very good friend Prof. Konstantinos Eftaxias for introducing me to scientific research and the science of complexity. Thank you daskale for wide-opening your door that day and giving me the chance.

I am more than grateful to my master thesis supervisors and collaborators, Prof. Fotis Diakonos, by whom I was also first taught computational physics in the undergraduate years, Vassilios Constantoudis, Harris Papageorgiou, Kostas Karamanos and Constantinos Papadimitriou for sharing with me their excitement on the complexity of natural language. It has been a great pleasure to learn from you and participate in the language group.

As indirect as their contribution to this thesis may be, I am glad to find these lines to say a thank you to four other special professors back in the UoA, Alexandros Karanikas, Paris Sphicas, Theocharis Apostolatos and Petros J. Ioannou for their inspirational, almost cinematographic, teaching that infused within me a great deal of the love I have today for physics.

I was also lucky to meet during my Masters program in NTUA, Prof. Antonios Symvonis and his team who, apart from my first rigorous contact with graph theory, gave me also the chance to learn and teach a practical course of Java in a very systematic way, one of my most precious experiences.

Finally, my most special thanks goes to my family. My beloved Matti, thank you for constantly providing your mathematical insight and physical intuition, for the delightful scientific discussions (thankfully, not so often related to your field) and for your overall support and encouragement. Kiitos, että sinä siedit minua niin kauan 16 neliömetrissä (käytännöllisesti katsoen 14 ilman kaappia)! To my sister, Titika, who has turned into one of the funniest people I know, to my mother, Katerina, who keeps reminding me not to take myself too seriously (I am still working on that) and to my father, Dimitris, having dedicated countless hours of discussion that born inside me the curiosity and love for science, όσο περιττό και αν μοιάζει, σας ευχαριστώ για την συνεχή σας στήριξη και ανιδιοτελή σας αγάπη.

Financial support and computing resources

The research leading to results herein has received funding from the European Research Council under the European Community's Seventh Framework Program (FP7/2007-2013) Grant Agreement no.258748. Part of this work was performed using HPC resources from GENCI [CINES] (Grants 2012 c2012086818 and 2013 x2013076818) and CINECA supercomputing center (ISCRA grant FLEXPROT). We acknowledge the financial support for infrastructures from ANR-11-LABX-0011-01. Figures and calculations were partly done using the R software and packages bio3d and igraph [39–41]. Partial funding for traveling was also provided by the Ecole Doctorale B3MI of Univeristy Paris Diderot.

Contents

List of symbols

$\boldsymbol{a}$	acceleration
α	thermal expansion
β	thermodynamic beta
β_T	isothermal compressibility factor
β_a	apparent compressibility
C_P	isobaric heat capacity
C	transitivity coefficient
γ	friction coefficient
D	diffusion coefficient
δ	Dirac delta function
d	distance (unless denoted otherwise)
E	total energy
$\boldsymbol{F}$	force
f	multi-dimensional arbitrary function
G	Gibbs free energy
$\mathbb{H}$	Hamiltonian
H	Shannon entropy
$\mathcal{H}$	hyperthermophilic protein under study
H	enthalpy
θ	arbitrary angle
k_B	Boltzmann constant

k	force constant
$\mathcal{M}$	mesophilic protein under study
m	mass
μ	chemical potential
N	number of atoms
N_f	number of degrees of freedom
n_t	fraction of native torsion angles
P	pressure
$\mathcal{P}$	probability
$\boldsymbol{p}$	momentum
ρ	density
Q	fraction of native contacts
$\boldsymbol{q}$	generalized coordinate
R_g	radius of gyration
$\boldsymbol{r}$	a point in $\mathbb{R}^3$ space
S	entropy
$\mathcal{T}$	thermophilic protein under study
T	temperature
t	time
τ	characteristic time
U	potential energy
V	volume
$\boldsymbol{V}$	velocity vector
v	velocity
$\boldsymbol{W}$	stochastic force
$\boldsymbol{X}$	fraction of unexchanged peptide hydrogens (in the context of H/D experiment)

X one-dimensional arbitrary collective variable

$\boldsymbol{x}$ a point in $\mathbb{R}^{3N}$ space

Z partition function

$\boldsymbol{z}$ a point in the $\mathbb{R}^{M}$ space (collective variable space)

List of abbreviations

AA	all-atom
a.a.	amino acid
ACF	autocorrelation function
a.k.a	also known as
CG	coarse-grained
CV	collective variable
EF-1α	elongation factor 1α
EF-Tu	elongation factor thermo unstable
FTIR	Fourier transform infrared spectroscopy
HB	hydrogen bond
H/D	hydrogen-deuterium exchange
IR	infrared (spectroscopy)
IP	ion-pair (Salt-bridge)
MCL	Markov clustering algorithm
MD	molecular dynamics
MDH	malate dehydrogenase
NMR	nuclear magnetic resonance
OPEP	Optimized Potential for Efficient Protein Structure Prediction
PDB	Protein Data Bank
REMD	Replica Exchange Molecular Dynamics
$RMSD$	root mean square deviation
$RMSF$	root mean square fluctuations

Introduction

"Frankly, extremophiles would be recruited for the local SWAT team, if they were big enough to carry weapons."

- Seth Shostak, Astronomer, SETI Institute

Life at the extreme

Life on Earth is adapted to a wide range of physicochemical conditions[4]. Although the vast majority of non-microbial species survives comfortably around standard ambient temperature and pressure (SATP)[5], there exist microorganisms not only able to survive but thriving at extreme conditions. These are the rightfully called *extremophiles*, from the Latin "extrēmus" and the Greek "φιλῶ"=to love. Even the variety of their extremity is stunning. For example, *Deinococcus radiodurans* is a bacterium with the ability to withstand supra-lethal ionizing and ultra violet radiation [42, 43]; the eukaryotic micro-algae *Dunaliella salina* manages to survive in the Dead Sea waters with salt concentrations as high as 31% [44]; colonies of photosynthetic cyanobacteria have been isolated from the interior of halite crusts (rock salts) in the Atacama Desert in Chile, the driest place on Earth [45]; not to forget of course the micro-animals tardigrades, that can survive vacuum, extremely high pressure conditions and temperatures ranging from almost absolute zero to above the boiling point of water.

Among extremophiles, of particular interest to us is the subclass known as *thermophiles*, referring to organisms that have the capacity to withstand high temperatures [2]. Thermophiles are mostly bacteria or archaea although some eukaryotes, such as algae or fungi, also exist [3]. They can be found in volcanic and geothermal areas, either on land or in hot deep-ocean vents. Thermophiles are considered to be among the most ancient organisms [33, 46, 47]. The main justification is that, in prehistoric times environmental conditions on the surface of the earlier Earth changed in violent and abrupt ways not allowing enough time for life to develop.

[4] A living organism, in this context, is an organized entity made up of one or more cells that is able to grow, metabolize, reproduce, respond to external *stimuli* and adapt to environmental changes [1]

[5] SATP: 25 °C (298.15 K) and 0.987 atm (100 kPa)

On the other hand, hot vents at the bottom of the oceans haven't substantially changed since the last several millions of years. Therefore, adaptation was not necessary for the survival of organisms living in such hidden corners of the planet. Generally, however, although almost everyone seems to agree that ancestral life sooner or later survived higher temperatures, the hyperthermophilic origin of the last common ancestor is still a subject of debate [32, 48]. Nevertheless, we should keep in mind that thermophiles today have a variable evolutionary history. Some of them originated in an extreme environment while others evolved at ambient temperature but later recolonized a hot place.

It is worth mentioning two thermophiles whose name will come up again in the following. The first one is the archaeon *Sulfolobus solfataricus* with an optimal growth temperature around 75°C that classifies it as a hyperthermophile. It grows in volcanic hot springs with plenty of sulfur which it oxidizes to produce energy and even if it grows in acidic environments, its cytoplasmic pH is close to neutral [49]. The second thermophile is the bacterium *Chloroflexus aurantiacus* usually found in low-sulfur hot springs areas. Due to its photosynthetic ability it can grow in anaerobic (or semiaerobic) conditions. If oxygen is present it can survive in the dark and with an optimal growth temperature between 52 and 60°C is classified as a mild thermophile [50].

Prologue to this work

Thermophiles, as extremophiles in general, adapt to their environment by employing different mechanisms. In this study, we focus on the molecular level and in particular the physical and chemical mechanisms that make thermophilic proteins cope with heat.

Proteins are among the most the important building blocks of even the simplest living organism [4]. They are large organic molecules consisting of one or more chains of organic compounds, called amino acid residues. They have a large variety of functions among which is the catalysis of chemical reactions important for life, DNA replication, protein synthesis, transportation e.t.c. Generally, they are fully functional in their so-called native form which is a complex three-dimensional folded structure, the exact shape of which is dictated by their primary amino acid sequence[6] and the physico-chemical parameters of their environment [5]. That is, for a given pH value, a specific primary sequence cannot result in two – substantially – different three-dimensional folds. Roughly speaking, the folded state of proteins is kept intact as a result of cohesive forces between the constituent amino acids, i.e. van de Waals interactions, electrostatics e.t.c. Thus, folded proteins are

[6] Exceptions are intrinsically disordered proteins.

soft-matter entities, some more flexible than others.

In fact, there is a delicate correlation between flexibility, stability and function [6–9]. A protein needs to maintain a stable fold[19]. Yet, with the appropriate amount of flexibility in order to facilitate processes relevant to function, e.g. substrate and cofactor binding [10]. Proteins from thermophilic organisms constitute a unique template that may help us understand the basic mechanisms of this subtle interplay. Natural selection has driven (or maintained) their stability and optimal function at temperatures that would unfold any other protein from moderate-temperature organisms, a.k.a. mesophiles. As a matter of fact, thermophilic proteins are generally inactive at ambient temperatures [11, 12]. This observation has led to the idea that they possess a more rigid matrix than their mesophilic homologues [13]. The universal character of this rigidity paradigm however has been questioned several times by both experiment and simulation [14].

Understanding the mechanisms of enhanced protein stability has technological and medical implications. For instance, thermostable proteins are already employed for extreme-temperature industrial biocatalysis, a fundamental process in the industries of food, dairy products, paper, biofuel and many others [15, 16]. Another important aspect concerns protein-misfolded diseases such as Alzheimer's and Creutzfeldt-Jakob disease, the cause of which is unfolding or partial unfolding of globular proteins and their subsequent aggregation into the form of amyloid fibrils [17]. The study of thermophilic proteins may help to understand the mechanisms behind amyloidogenesis [18, 19].

The aim of the present study is to identify, through the use of Molecular Dynamics simulations, characteristics that distinguish thermophilic from mesophilic proteins, elaborating in particular on the rigidity paradigm mentioned above. Such characteristics could inspire new strategies in the design of thermostable proteins [20].

This thesis is organized as follows. In Chapter 1, we first discuss the basics of protein structure and subsequently state what is already known about thermophilic proteins as far as both structure and dynamics are concerned. Then the thermodynamic basis for protein stability is described and the chapter closes with an outlook to the current study. In Chapter 2, we present the Methodology used to attack the problem: In the first section the basic principles of MD simulations are presented, elaborating on the techniques chosen for our approach. The second section touches on the enhanced sampling schemes in use while in the last section we present the MD post-processing approaches that we employ. In Chapter 3, we study the ambient-temperature dynamics of two homologous globular proteins, the hyperthermophilic and mesophilic G-domains extracted from EF-Tu and EF-1α proteins, respectively. Our results question the rigidity paradigm for this hyper-

thermophilic protein and indicate that a regular partitioning of flexible and rigid parts along the sequence may prevent protein unfolding. In Chapter 4 we extend the study of the two homologues by examining their kinetic stability at higher temperatures. Complementing with all-atom simulations of the whole trimeric EF proteins, we individuate the weak spot of the mesophilic protein, tightly related to its enzymatic function and set some perspectives for future work. Keeping the two G-domains under the microscope, Chapter 5 deals with the thermodynamic aspects of their thermal stability, while testing for the first time the potentiality of a coarse-grained force-field to distinguish the different thermal stability content of two proteins. In Chapter 6, we expand our study to two other homologues, a mesophilic and a thermophilic tetrameric malate dehydrogenase. We find a tight correlation between the folded-state flexibility, the protein's thermal stability and its oligomerization state: Oligomerization substantially rigidifies the thermophilic homologue with a potential effect on its substrate binding affinity. We conclude by summarizing our findings and stating some open questions and perspectives for the future.

Chapter 1

Thermophilic proteins, "state of the art"

Summary

In this chapter, we begin by setting the basic concepts and nomenclature for proteins in general. We then discuss, in brief, how thermophilic proteins differ from their mesophilic counterparts and present some microscopic characteristics that have been identified as commonly occurring in thermally stable proteins. Subsequently, we address the question of the mechanisms lying at the origin of thermal stabilization. Specifically, we present the "corresponding states" concept, initially introduced by Somero to explain the lack of activity of thermophilic enzymes at ambient temperature. According to this, thermal stability is a consequence of an enhanced protein rigidity that suppresses functionality. We give our current understanding of the issue considering both experimental and computational studies. The thermodynamic perspective is also presented to show that the idea of protein rigidity is only one among the possible strategies of thermal adaptation. We conclude with presenting how modern computational methodologies could provide insight on this intriguing problem with a specific focus on our work.

1.1 Proteins: a look at the structure of life's machinery

Proteins are linear chains of organic compounds called amino acids (a.a.) that are bound together with peptide bonds[1]. An amino acid has a central carbon atom around which are bonded a hydrogen, an amine (-NH2) group, a carboxylic acid (-COOH) group and a characteristic side chain (R). In nature there exist 20

[1] A peptide bond is covalent chemical bond.

amino acids[2] that differ from each other only in the composition of the side chain. They are divided in four categories, nonpolar or hydrophobic, polar uncharged, negatively charged and positively charged (for a more detailed discussion see [51, 52]).

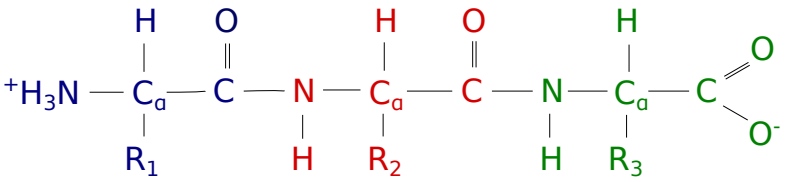

Figure 1.1. Linear sequence of arbitrary amino acids bound together with peptide bonds to form the primary structure of a protein. Each color corresponds to a different amino acid. The different side chains are indicated with R_x whereas the sequence of nitrogen, carbon and oxygen atoms constitute the protein backbone or main chain.

During the cellular process of translation the a.a. are covalently bonded together, as in Fig. 1.1. After their linkage, they are referred to as residues. The order of their sequence, encoded in the DNA, is specific for each protein and constitutes the so-called primary sequence. The primary sequence hasn't acquired yet any structure; rather it looks like a stretched flexible polymer, a.k.a. random coil, like the top left of Fig. 1.2. Shortly after its composition it folds into a complex three-dimensional shape that is in one way unique: a specific primary structure results in a specific fold but different a.a. sequences may result in the same or similar folds.

According to our current understanding, the folding process follows a certain hierarchy of steps (see Fig. 1.2); first comes the formation of local structures that are called secondary. These elements are defined by the backbone hydrogen-bond patterns between neighboring amino acids and are geometrically specific: the ϕ and ψ dihedral angles of the protein backbone (defined in Fig. 1.3) are restricted in a certain range of values for each type of secondary structure. The two main types of secondary structure are α-helices and β-strands (Fig. 1.4 shows a common cartoon representation of secondary structure elements). Right after their formation they acquire a certain configuration with respect to each other resulting that way in the final fold called tertiary structure. Proteins may consist of one primary-sequence

[2] A few other rare amino acids may also occur.

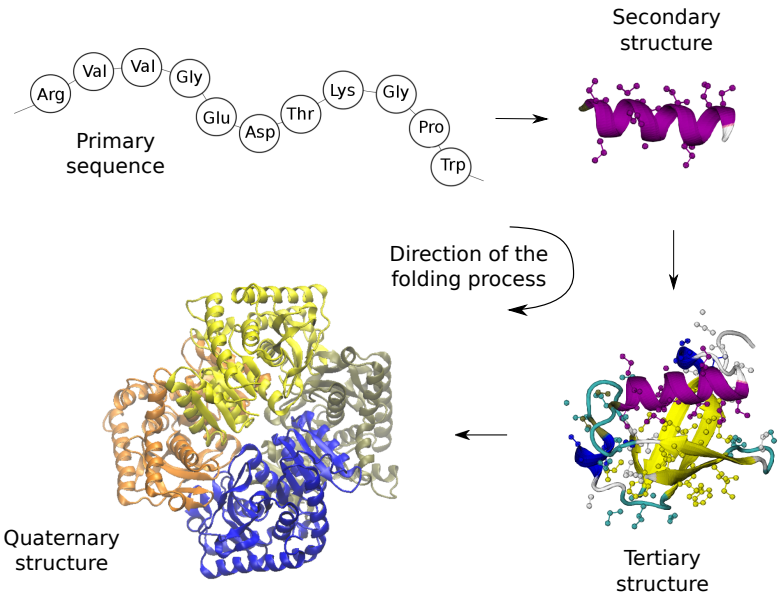

Figure 1.2. The basic levels of hierarchy of a protein structure.

chain (monomeric proteins) or several (oligomeric proteins). In the second case, the tertiary structures of the individual chains are further assembled to a greater quaternary structure. An example of an homotetrameric protein can be seen in the lower left corner of Fig. 1.2.

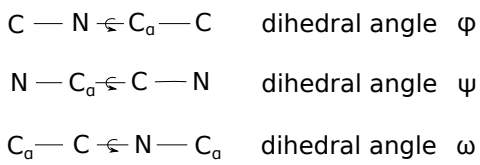

Figure 1.3. The backbone dihedral angles of proteins, ϕ, ψ and ω.

The native folded protein is, obviously, energetically favorable over all the possible unfolded shapes, usually referred to as conformations. The forces that drive the folding of a protein are the same that stabilize the final native fold. These are

- the hydrophobic effect, i.e. the preferential attraction between hydrophobic side-chains immersed in a polar medium [53]. In fact, this has long been considered as the main driving force of protein folding [54] and this is why water is thought to be important in both folding and stability of proteins [55, 56]. From the thermodynamics point of view, the (free) energy of the folded state has both enthalpic and entropic contributions. While being in an unfolded conformation, the protein's hydrophobic residues are exposed to

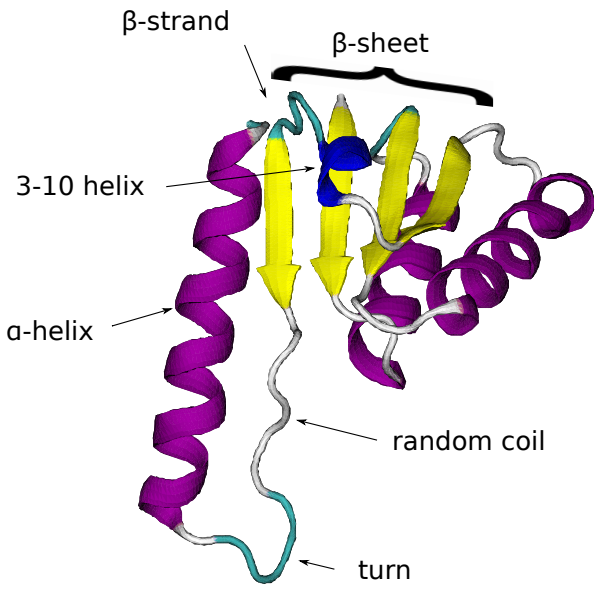

Figure 1.4. Common cartoon representation of the different types of secondary structure. Random coil actually corresponds to the lack of any hydrogen bond pattern. There is one more helix type, other than the two shown above, know as π-helix.

water which in turn is forced to form organized structures around them. The energetic cost of these structures is not compensated by the favorable entropy of the random coil; thus the protein collapses in a globular form in order to protect its hydrophobic parts from water.

- van der Waals interactions between polar residues and electrostatic interactions between oppositely charged residues. These forces usually fine-tune the final fold.

- hydrogen bonds between either side-chain or backbone electronegative atoms, i.e. oxygen and nitrogen. Hydrogen bonds are essentially electrostatic interactions but due to their partially covalent nature, they are traditionally treated separately. The hydrogen-bonds between the side-chains of oppositely charged amino-acids are referred to as salt-bridges or ion-pairs.

At the level of protein stability, these "soft forces", as compared to the strong covalent peptide bonds, are the ones that can be thermally disrupted. Temperature increase does not, practically, break covalent bonds (in numbers, the thermal energy equal to the energy of a typical carbon-carbon covalent bond (85 kcal/mol) yields a temperature of about 40,000 K, $k_B T \approx 80$ kcal/mol). How enthalpic and entropic

forces compete with each other to stabilize the fold is particularly important to us and will be discussed again in the following.

It is worth noting at this point, that the study of protein stability and function has greatly benefitted from the tremendous development of X-ray crystallography that allows to resolve the 3D organization of atoms in crystal proteins. This method gives us also a first idea on the flexibility of the different parts of the protein crystal via the variation of the Debye-Waller factor of atoms, a measure that describes the attenuation of X-ray diffraction. Important information at the structural level comes also from high resolution nuclear magnetic resonance (NMR) of small proteins in solution [51]. Newly resolved protein structures are usually deposited in an accessible online database, the Protein Data Bank (PDB), that way allowing either massive statistical analysis or - as we will specify in the next chapter - to be the starting conformations for computer simulations at atomistic resolution.

1.2 A suitable molecular machinery for high temperatures

The first official discovery of a thermophilic bacterium dates back to P. Miquel in 1879 [57], although a few earlier records also exist [2]. It was not however until the 1940's that a more systematic study of the enzymes of these organisms begun. At the molecular level, the first question to be addressed was whether thermophilicity was achieved by a favorable equilibrium between temperature-induced unfolding of proteins and their fast re-synthesizing. The first studies concluded that thermophilic proteins are, in fact, inherently more stable than their homologues from mesophilic organisms, i.e. they do not unfold at high temperatures [58]. Since then, a whole field of research flourished in the search for the physicochemical mechanism responsible for increasing protein thermal stability. To date, it is widely accepted that there is no such unique or universal mechanism. Systematic analysis between thermophilic and mesophilic homologous proteins suggests that stabilization involves all the levels of hierarchy of a protein structure.

At a primary sequence level, statistical analysis yielded, on average, an excess of charged a.a. in thermophilic sequences suggesting that electrostatic interactions on the final tertiary structure have an important stabilizing effect. This was also supported by analysis of several homologous structures, showing that increased thermostability indeed correlates to an increased number of hydrogen bonds and ion pairs as well as an increased polarity of the protein surface [59, 60]. At the level of secondary and tertiary structure, shorter loops and anchored -C and -N terminals, that are usually observed in thermophilic structures, may prevent water penetration into the protein core and therefore unfolding [61]. Another

recent breakthrough was the discovery of a few unambiguous rules concerning the mapping from specific secondary structure patterns to tertiary structure motifs [62] that follow from minimization of either torsional strain or backbone bendability. These rules tend to be followed by thermophiles, while proteins that were designed based on them exhibited melting temperatures greater than 95 °C. At the tertiary structure level, an optimized packing of the core hydrophobic residues has been advocated as a factor enhancing stability [63, 64], whereas in a recent work the attention was drawn on the quality of hydrophobic contacts, meaning not the mere number of hydrophobic clusters but those that actually result in a low favorable energy [65]. At a final quaternary structure level, earlier studies had noted the possible correlation between thermal stability and a higher oligomeric state [66–68]. However, the ever increasing number of homologous pairs deposited in the Protein Data Bank does not support this hypothesis anymore [69, 70].

Yet, one cannot isolate a single level of protein hierarchy, such as for example the primary sequence, and through comparative analysis unambiguously determine the content of thermophilicity. Despite this, some of the important ingredients mentioned above are already effectively used in protein design [16].

1.3 The rigidity paradigm

Increased stability of the protein matrix at higher temperatures, in whatever way it might be achieved, is a necessary but not sufficient prerequisite to guarantee optimal function at elevated temperatures. Yet, thermophilic proteins function best at the optimal temperature of the host organism. In fact, most of them are largely inactive at ambient conditions [60, 71].

Let's understand how this might be explained considering a simple schematic picture. In several cases, protein activity can be roughly divided in two steps, first the substrate binds to the protein's active site, then the actual chemical step takes place. Substrate binding requires accessibility of the binding site which in turn entails a certain degree of protein flexibility. If a protein is very rigid at ambient temperature, equilibrium fluctuations do not allow for binding of the substrate and function is stalled. An increase in temperature activates the suppressed motion and protein function is recovered.

The above scenario along with Somero's "corresponding states" concept [72], stating that proteins from mesophilic and thermophilic organisms tend to preserve ligand-binding properties and overall efficiency in their corresponding physiological temperatures, led to the common idea that the different homologues have comparable flexibilities at their respective optimal regimes. The assumption behind this idea is that the lack of activity at ambient temperature is indeed rooted in the sup-

pressed conformational fluctuations relevant to activity, e.g. enzyme or cofactor binding.

It is however possible to explain the lack of activity by focusing on the chemical step. According to transition state theory (TST)[3], the catalytic constant of the reaction is given by

$$k_c \propto k_B T \exp(-\Delta G^{\ddagger}/k_B T) \tag{1.1}$$

where $\Delta G^{\ddagger}$ is the Gibbs energy of activation of the chemical reaction, k_B is Boltzmann's constant and T the temperature. Eq. 1.1 dictates that the reaction rate decreases exponentially as the energy barrier increases, while it increases monotonically with the temperature. So, in principle, as was put forward by M. Roca and collaborators [23], thermophilic proteins could have a larger $\Delta G^{\ddagger}$ requiring a higher temperature to activate the chemical reaction.

The first experimental indirect support of the rigidity hypothesis came in 1990 with a hydrogen-deuterium (H/D) exchange experiment by A. Wrba et al. [73] Hydrogen exchange is a very useful technique to experimentally quantify the flexibility and the global stability of the protein matrix. The experiment is based on the fact that in a water-protein solution the amide hydrogens of the protein, that are accessible by water, are exchanging protons with it. Thus, diluting the water solvent with D_2O and subsequently detecting the deuterium atoms that are bound to the protein allows to quantify the fraction of residues that get exposed to water. This serves as a measure of protein flexibility. Detection of deuteriums is done usually by NMR, infrared spectroscopy or mass spectrometry.

In the experiment of A. Wrba et al. [73] done at the ambient temperature of 25 °C, the proteins under study were the D-glyceraldehyde-3-phosphate dehydrogenase (GAPDH) from the hyperthermophilic bacterium *Thermotoga maritima* (optimal T in the range of $80 - 90\,^{\circ}\mathrm{C}$) and its mesophilic homologue from yeast, both tetrameric. The results yielded a smaller number of water-exposed amides for the hyperthermophilic variant suggesting its larger conformational rigidity at ambient temperature.

A more direct support to the rigidity hypothesis came later, in 1998, by R. Jaenicke et al. [74] where the authors performed the same experiment (the mesophilic variant belonged to rabbit in this case) but this time a higher temperature was tested as well. The results can be seen in Fig. 1.5. The plotted quantity $\boldsymbol{X}$, detected by infrared spectroscopy, is the ratio of energy absorbances at the maximum of amide I and amide II bands. The amide I band (wavenumber between 1600 and 1700 cm^{-1}) is mainly associated with the C=O stretching vibration. It is used in the denominator of $\boldsymbol{X}$ as a reference for protein concentration. Amide II

[3] TST, developed by H. Eyring, M.G. Evans, and M. Polanyi, explains the reaction rates of elementary chemical reactions assuming the presence of an activated complex, the transition state, separating the reactant and product states. In short, the kinetics of the reaction depends on the probability to access the transition state.

results mostly from the N–H bending vibration but it has some contribution from the C–N stretching vibration too. The above ratio $\boldsymbol{X}$ is effectively the fraction of unexchanged peptide hydrogen atoms. Therefore the larger the $\boldsymbol{X}$, the more rigid the protein. Figure 1.5(a) shows $\boldsymbol{X}$ versus time for both proteins as monitored at 25 °C (k_0 is a function of pH and temperature and has units of inverse time). As in the previous experiment of the 1990, the hyperthermophilic variant is more rigid. However when the same measurement is done with the hyperthermophile at the higher temperature of 68 °C, its conformational flexibility overlaps with that of its mesophilic homologue at 25 °C (see Fig 1.5(b)). This "mechanism" of thermal adaptation was supported by another H/D experiment of the same year, this time between a mesophilic and a mild thermophilic dimeric isopropylmalate dehydrogenase [75].

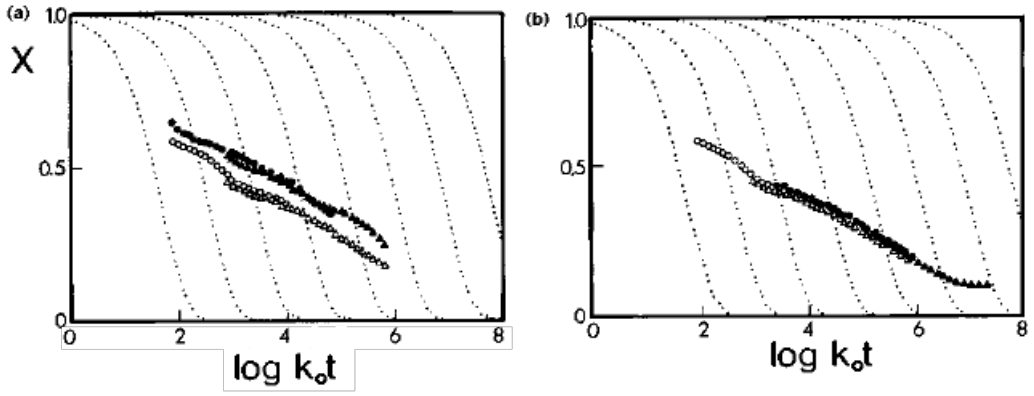

Figure 1.5. Hydrogen-deuterium exchange of GAPDH from Thermotoga maritima (closed symbols) and rabbit (open symbols), measured at pH 6.0 (●, ○) and 7.0 (▲,△) plotted as relaxation spectra. For experimental details see Wrba et al. 1990. (a) Measurements at constant temperature (25 °C); larger $\boldsymbol{X}$ values reflect increased rigidity. (b) Measurements at 68 °C for TmGAPDH and 25 °C for rabbit GAPDH. Coincidence of the curves indicates similar flexibility [Figure taken from R. Jaenicke et al. (1998)].

Only two years later, it was again an amide hydrogen exchange experiment by Hernández et al. [35] that showed that the hypothesis under discussion has no general validity. In this case the protein was the extremely thermostable monomeric rubredoxin from *Pyrococcus furiosus* (its optimal T is 100 °C). Remarkably, solvent access was monitored for this protein within the millisecond timescale for nearly all amide positions at 28 °C. Such hydrogen exchange rates are similar to those of many mesophilic proteins thus not supporting the hypothesis that thermal stability is caused by a higher conformational rigidity. This view found support by one

more infrared spectroscopy H/D experiment of the same year on two homologous α-amylases at the picosecond timescale [36].

The studies that followed returned conflicting results as well. *In vivo* and *in vitro* elastic incoherent neutron scattering (EINS) experiments found the atomistic fluctuations at ambient temperature comparable between thermophilic and mesophilic proteins while noted their weak temperature dependence for the thermophilic homologue [63, 76]. This resilience to temperature increase was later confirmed by MD simulations while the overall atomistic-fluctuation profiles showed a higher flexibility for the thermophile [77]. Another EINS experiment, comparing the distributions of sub-nanosecond atomistic fluctuations between two homologous monomeric dihydrofolate reductases, found a wider distribution for the thermophilic species rendering it more flexible than its mesophilic variant [78]. Very recently, a Trp fluorescence quenching experiment in combination with an MD simulation noted a more rigid thermophilic aqualysin I as compared to its psychrotrophic homologue [79].

Molecular dynamics studies have also been inconclusive. One of the earliest studies somewhat supported the rigidity hypothesis [80]. However, the simulation therein was only 400ps-long and the same system, protein rubredoxin from *Pyrococcus furiosus*, was found later by an H/D exchange experiment to be highly flexible at longer timescales [35] (see discussion above). More recent MD studies along the same lines include Refs. [61, 81]. On the other hand, the first MD study to show that reduced flexibility for thermophiles at ambient T is not a necessary requirement for stability was conducted by G. Colombo and collaborators [82] even before its first experimental support by Hernández et al., while other studies in this direction also followed [83–85].

The rigidity hypothesis has also been investigated with more simplified computational approaches such as constraint protein networks and rigidity theory. For example, in the work by S. Radestock and H. Gohlke, twelve out of twenty homologous pairs of proteins, studied therein, were found to have a more rigid thermophilic counterpart [86, 87]. Using similar approaches, the thermophilic dimeric citrate synthase was also characterized as rigid in contrast to its mesophilic counterenzyme [88].

There are two reasons for the lack of consensus. The first, as will be discussed later on in this Chapter, is that the mechanism of thermal adaptation might be different depending on the pair of homologues. The second reason is that there is no unique measure of protein flexibility. Protein motions range from thermal fluctuations and equilibrium local conformational changes to larger global conformational transitions (e.g. diffusive motions that involve the accumulation and release of stress during and after the binding process) [89]. All these different modes oc-

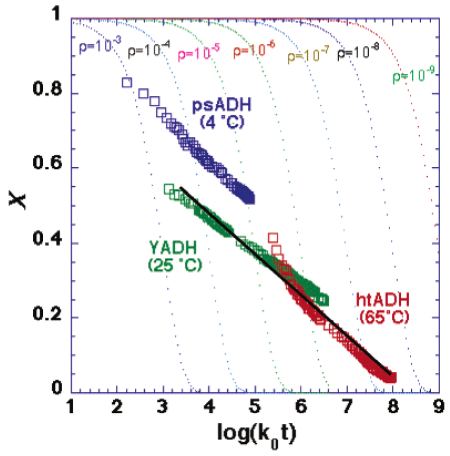

Figure 1.6. H/D exchange probability distribution as presented in the form of relaxation spectra for the psychrophilic (4 °C), mesophilic (25 °C) and thermophilic (65 °C) alcohol dehydrogenases as measured by FTIR spectroscopy. $\boldsymbol{X}$ is the fraction of unexchanged peptide hydrogens, t is time in seconds, and k_0 is the calculated chemical exchange rate constant at the relevant temperature. The dotted lines represent exchange curves for hypothetical polypeptides characterized by an F value, which is defined as the probability of finding a particular hydrogen exposed to the solvent. Coincidence of the curves of mesophilic and thermophilic as indicated by the solid straight line has previously been reported to suggest similar flexibility. [Figure taken from R. Liang et al. (2004)].

cur on largely different timescales. Consequently, a protein that may be rigid in the picosecond timescale (e.g. atomistic or side chain fluctuations) can be more flexible in the nano- or micro- second timescale (e.g. opening/closing dynamics of a loop). A characteristic example of this, is the study by Liang et al. [90]. In the conducted H/D experiment the authors calculated the quantity $\boldsymbol{X}$, as defined above, for three homologues, a psychrophile, a mesophile and a thermophile, each being at its corresponding optimal temperature. The results from the standard H/D exchange showed that the psychrophilic protein is more rigid than the other two (Fig. 1.6). However, local H/D exchange (on protein-derived peptides) showed that the psychrophile exhibits, in fact, enhanced flexibility as compared to its thermophilic homologue but at specific, functionally important regions. The authors noted that as variations in protein flexibility can be local, averaged global H/D exchange measurements may be misleading. Later studies noting the local character of flexibility include the H/D experiments by Oyeyemi et al. [91] where at a global level the thermophilic homologue was found to be more flexible than its mesophilic homologue with the exception of a single region where the opposite is true. This

region is located within the protein interior at the intersection of the cofactor and substrate-binding sites.

From the above, it becomes clear that when discussing the relationship between flexibility, stability and function at different thermodynamic conditions, it is necessary to first clarify the reaction coordinates relevant to the process under study. In other words, one needs to single out the important degrees of freedom, their collective motion and, eventually, the role of their amplitude and kinetics in protein functionality.

1.4 What does thermodynamics have to say?

We now try to approach the subject from a thermodynamics point of view. The assumption is that there are only two distinct states for the protein, the native (F) and the denatured or unfolded state (U, i.e. any other state that substantially deviates from the native). For what follows, we make use of this two state model and reserve a comment on its applicability at the end of this section.

The thermodynamic stability of the folded state is defined as the Gibbs free energy difference between the folded and unfolded states

$$\Delta G_{F\to U} = G_U - G_F \tag{1.2}$$

Given the definition of the Gibbs free energy, ΔG has both an entropic and an enthalpic contribution

$$\Delta G_{F\to U} = \Delta H_{F\to U} - T\Delta S_{F\to U} \tag{1.3}$$

As we will discuss below, one of the thermodynamic mechanisms that thermophilic proteins can adopt in order to increase their melting temperature is increasing $\Delta G_{F\to U}$ (notice that from Eq. 1.2 $\Delta G_{F\to U} > 0$ for a folded protein). This can be achieved by increasing $\Delta H_{F\to U}$, decreasing $\Delta S_{F\to U}$ or both. Increasing $\Delta H_{F\to U}$ practically means increasing the absolute value of H_F with, for example, a larger number of hydrogen bonds or ion-pairs[4]. Decreasing $\Delta S_{F\to U}$ could mean either a more flexible folded protein or residual secondary structure in the unfolded state.

The rigidity argument discussed above sounds reasonable to explain stability at high T. This is because a globally rigid protein may effectively a) be cross-linked by a larger number of hydrogen bonds and ion-pairs and/or b) have a tightly packed hydrophobic core. In the first case, the protein is mainly enthalpically stabilized (electrostatics and van der Waals forces). In the second case, it is stabilized both enthalpically and entropically (van der Waals attraction is also involved between

[4] We note that both the destabilizing character of buried ion pairs as well as ion-pair contribution to stability only at higher T have also been discussed in [92] and [93]

the hydrophobic groups [53] whereas the favorable entropic contribution comes from the water). However, as we saw above, the stability of a protein may in principle be rooted only in $T\Delta S$. That essentially means that a thermostable protein may as well be flexible at ambient temperature. The few subtleties involved in the enthalpic and entropic contribution to ΔG are discussed below, after introducing the PT phase diagram for a two-state protein as proposed by S.A. Hawley [94].

Pressure-temperature phase diagram of proteins by Hawley. The differential form of the Gibbs free energy for constant number of particles is given by

$$d\Delta G = \Delta V dP - \Delta S dT \tag{1.4}$$

where we dropped the subscript $F \to U$ to keep the notation light. We additionally define the quantities

$$\Delta\beta_T = (\partial\Delta V/\partial P)_T, \quad \text{(isothermal compressibility factor)} \tag{1.5}$$
$$\Delta\alpha \ = (\partial\Delta V/\partial T)_P, \quad \text{(thermal expansivity factor)} \tag{1.6}$$
$$\Delta C_P = T(\partial\Delta S/\partial T)_P = (\partial\Delta H/\partial T)_P, \quad \text{(isobaric heat capacity)} \tag{1.7}$$

If we assume that $\Delta\beta_T$, $\Delta\alpha$ and ΔC_P are constant, integration of Eq. 1.4 from an arbitrary chosen reference point P_0, T_0 yields the following expression of ΔG as a function of pressure and temperature

$$\begin{aligned}\Delta G(P,T) =& \frac{\Delta\beta_T}{2}(P-P_0)^2 + \Delta\alpha(P-P_0)(T-T_0) - \Delta C_P\left[T\left(\ln\left(\frac{T}{T_0}\right)-1\right)+T_0\right] \\ &+ \Delta V_0(P-P_0) - \Delta S_0(T-T_0) + \Delta G_0\end{aligned} \tag{1.8}$$

The parameters of Eq. 1.8 are determined experimentally and it turns out that the contours of constant free energy on the PT plane are ellipses [94]. Their elliptical character is reflected on the quadratic nature of the Eq. 1.8 in the vicinity of T_0:

$$T\left(\ln\left(\frac{T}{T_0}\right)-1\right)+T_0 \simeq \frac{(T-T_0)^2}{2T_0} + O(T^3) \tag{1.9}$$

which after substitution to Eq. 1.8 yields a two-dimensional inverted parabola shown in Fig. 1.7. Below the curve the elliptic contour for which $\Delta G = 0$ is drawn. Within this closed ellipse, ΔG is positive and the protein is most of the time folded (or otherwise the population of folded proteins is larger than the population of the unfolded ones). The opposite is true outside the curve. We now focus on temperature, so for constant pressure Eq. 1.8 becomes

$$\Delta G(T) = -\Delta C_P\left[T\left(\ln\left(\frac{T}{T_0}\right)-1\right)+T_0\right] - \Delta S_0(T-T_0) + \Delta G_0 \tag{1.10}$$

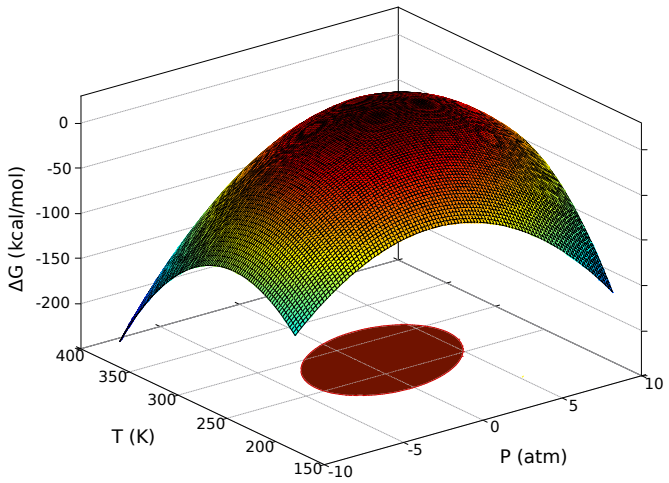

Figure 1.7. Second order approximation of ΔG between the folded and unfolded states as a function of pressure and temperature. The constant ΔG contours on the PT plane are ellipses while the one projected in the figure above corresponds to $\Delta G = 0$. For P and T values within the curve the protein is folded. The exact shape and position of this ellipse is different for different proteins. It is determined by the parameters of Eq. 1.8 which are found by experiment or simulation.

The most common choice for T_0 is the melting temperature of the protein, T_m, where ΔG_0 is zero. Then Eq. 1.10 becomes

$$\begin{aligned}
\Delta G(T) &= -\Delta C_P \left[T\left(\ln\left(\frac{T}{T_\mathrm{m}}\right) - 1\right) + T_\mathrm{m} \right] - \Delta S_\mathrm{m}(T - T_\mathrm{m}) + \Delta G_\mathrm{m} \\
&= -\Delta C_P \left[T\left(\ln\left(\frac{T}{T_\mathrm{m}}\right) - 1\right) + T_\mathrm{m} \right] - \Delta S_\mathrm{m} T + \overbrace{\Delta S_\mathrm{m} T_\mathrm{m} + \Delta G_\mathrm{m}}^{\Delta H_\mathrm{m}} \\
&= -\Delta C_P \left[T\left(\ln\left(\frac{T}{T_\mathrm{m}}\right) - 1\right) + T_\mathrm{m} \right] + \frac{\cancelto{0}{\Delta G_\mathrm{m}} \Delta H_\mathrm{m}}{T_\mathrm{m}} T + \Delta H_\mathrm{m} \\
&= -\Delta C_P \left[T\left(\ln\left(\frac{T}{T_\mathrm{m}}\right) - 1\right) + T_\mathrm{m} \right] + \Delta H_\mathrm{m} \left(\frac{T_\mathrm{m} - T}{T_\mathrm{m}}\right)
\end{aligned} \tag{1.11}$$

In an experiment or simulation, one estimates ΔG for as many temperatures as possible and then fits the data with this model.

The so-called stability curve, $\Delta G(T)$, is plotted in Fig. 1.8 for an actual protein, the mesophilic *E. coli* Pol I polymerase [95]. The parameters for the specific curve, $\Delta C_P = 3.7 \pm 0.8\,\mathrm{kcal/mol{\cdot}K}$, $\Delta H_\mathrm{m} = 128 \pm 12\,\mathrm{kcal/mol}$ and $T_\mathrm{m} = 316 \pm 1.2\,\mathrm{K}$, were determined experimentally via chemical denaturation at different temperatures and monitored by circular dichroism spectroscopy. Figure 1.8 shows also the enthalpic, $\Delta H(T)$, and entropic, $T\Delta S(T)$, contributions to $\Delta G(T)$ with their range of values

on the right. These curves are slightly curved and cross each other at both high and low melting temperatures (in the Figure they are not exact but only qualitatively correct for illustrative purposes).

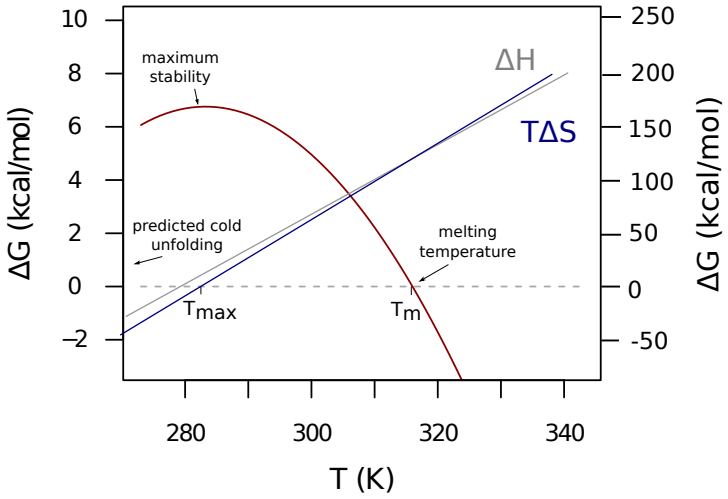

Figure 1.8. The stability curve, $\Delta G(T)$ (axis on the left), for mesophilic *E. coli* Pol I polymerase drawn according to Eq 1.11. The right axis refers to $\Delta H(T)$ and $T\Delta S(T)$. The parameters of the equation were determined experimentally via chemical denaturation with guanidine HCl, monitored by circular dichroism spectroscopy (CD) in 10 mM phosphate buffer at pH 7.5. [C.C. Liu and V.J. LiCata, Proteins, 2013]

The first interesting thing to notice is the range of values for $\Delta H(T)$ and $T\Delta S(T)$ as compared to that of $\Delta G(T)$. The free energy difference between the folded and unfolded states is in general marginal, a small difference between two large opposing contributions (in numbers $\Delta G_{\max}$ is comparable to the energy of a few hydrogen bonds). From the biological point of view, marginal stability can be explained if we think that proteins need to be ready to respond to changes in environmental conditions [96]. A second thing to note is that maximal stability is reached when the entropic contribution cancels out $T\Delta S = 0$, therefore maximal stability is purely enthalpic. Below $T_{\max}$, the enthalpic contribution is unfavorable while the entropic one is favorable (see also Eq. 1.2 and 1.3). The opposite is true for $T > T_{\max}$. Finally, the maximum of the ΔG parabola for a given protein (i.e. maximum stability) is observed at a temperature that is much below the optimal growth temperature (T_{opt}) of the organism. This is true for both mesophilic and

thermophilic organisms [74]. This, in conjunction with the previous observation, means that at the optimal growth temperature the entropic contribution to stability is always a penalty that enthalpic forces have to compensate. In this sense, an "entropically stable" thermophilic protein may only have a reduced such penalty [95].

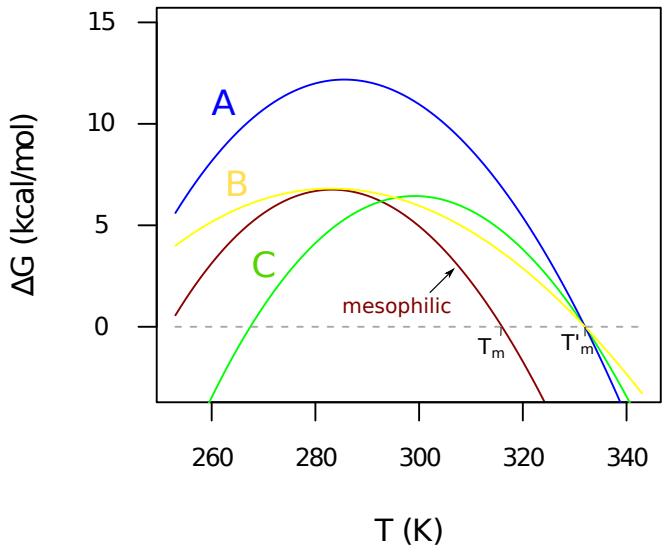

Figure 1.9. The three commonly viewed ways of modifying the stability curve to a higher T_{m}. With respect to the mesophilic stability curve, the thermophilic curve may be up-shifted (A), broadened (B) or right-shifted (C).

There are three commonly proposed modifications of the stability curve in order to get a higher T_{m} [97]. They can be seen in Fig. 1.9. With respect to the mesophilic stability curve, the thermophilic curve may be up-shifted (blue curve), broadened (yellow) or right-shifted (green). There is no single thermodynamic contribution to each of these modifications. For example the up-shift of the stability curve, can be achieved either enthalpically or entropically or both. The same is true for the broadening of the curve that practically means a lower heat capacity of unfolding ($\Delta C_P = -T(\partial^2 \Delta G/\partial T^2)_P$). The right shift of the curve doesn't change either $\Delta G_{\max}$ or ΔC_P. Experimentally, a little less than 20 homologous pairs of mesophilic and thermophilic proteins have been compared in this way [95, 98] and in most cases a combination of broadening and up-shift of the curve was observed (∼80%). For the record, there was at least one case where broadening of the curve with a simultaneous down-shift was observed that although results in a larger T_{m} it gives a smaller $\Delta G_{\max}$ for the thermophile. Notably, in a recent work

by C.C. Liu and V.J. LiCata [95], the authors decomposed the thermodynamic free energy for 18 pairs of homologous proteins in the two competing enthalpic and entropic forces and in almost all the cases they found that increased thermal stability has an entropic nature. That is to say that for the thermophilic homologue the stabilizing enthalpic contribution has to compensate a smaller entropic penalty than for the mesophilic one. Microscopically this could mean one or a combination of the following features, optimized packing of the hydrophobic residues in the folded state, residual secondary structure on the unfolded state or/and a flexible folded state.

At this point, we should make a note on the two-state assumption. For small globular proteins the two-state model is valid [99] while it is not difficult to assess its applicability experimentally for any given system [100]. For some cases however, stability is determined kinetically rather than thermodynamically. For example the cold shock proteins in the work of Jaenicke et al. [74] unfold in a much slower rate that then one in which they fold. Equilibrium thermodynamics are also, generally, not applicable for large systems such as membrane proteins [101], which is why thermodynamic data are not easy to obtain experimentally (or computationally for that matter). Finally, different thermodynamic approaches should be used for oligomeric proteins where the unfolding is complicated by the combination of monomer unfolding and oligomer dissociation [67].

Finally, it might also be useful to interpose a more general comment on the equilibrium nature of both experiments and simulations in the context of this study. In reality a system is never isolated from its environment. For example proteins live and work mostly in the cell, a very dynamic and crowded place. From the physics point of view, the interior of a cell is far from equilibrium (for an insightful comment on this see the introduction of Ref. [96]). However, as far as biochemistry is concerned, the cell reaches a state called dynamic equilibrium, also known as homeostasis. Although reactions do occur, the average composition of a cell does not change. This also means that the average temperature and pressure inside the cell are maintained approximately constant. Of course if we would pick a random instant at the life of a protein, it is possible that we would find a non-uniform pressure in its immediate neighborhood, e.g. an activator molecule disassociated from the protein a few picoseconds ago and is now diffusing away thus causing some turbulence around the binding site. In fact, protein motions can be broadly categorized to equilibrium and non-equilibrium ones [52, 89, 102]. Non-equilibrium *in vivo* relaxations are mostly related to the function of the proteins and they are to a large extend out of the scope of the present work (or at least their direct investigation is). Herein, we focus on equilibrium motions and properties of thermophilic proteins as compared to those from mesophilic homologues. In the future however,

it would be interesting to assess whether different environmental conditions of the cell give contribution to the stability gap between homologues [26, 30].

1.5 An *in silico* inquiry

In summary, neither experiment, theory or computational studies can conclude on whether thermophilic proteins are more rigid than their mesophilic counterparts. From comparison of the mere number of the studies we might say that "they tend to be", without however rigidity being the necessary route to thermal stability. In fact, the currently available data show an intriguing trend, the tendency of the rigid thermophilic proteins to be oligomeric. In most of the experimental and simulation studies mentioned above that support the rigidity paradigm, the system is oligomeric. The only experimental exception is the very recent Trp fluorescence quenching experiment where the rigid thermophilic protein was the monomeric aqualysin. It was, however, being compared to its psychrotrophic and not mesophilic homologue, thus incorporating simultaneously two underlying mechanisms, adaptation to low and high temperatures that in practice may be different. Some rigid monomeric proteins were also found in a computational approach based on constraint protein networks [86] however this technique includes many approximations and is keen to lose both chemical detail and correct thermal dependency. The complimentary trend is also observed, most of the thermophilic flexible proteins tend to be monomeric. Exception is the MD study of Marcos et al. [77] where the flexible thermophilic protein is a tetrameric malate dehydrogenase being compared to its mesophilic lactate dehydrogenase orthologue. However, as will be seen in Chapter 6, for the same timescales another thermophilic malate is found to be more rigid as compared, this time, to a more similar mesophilic homologue. A larger number of future comparative studies should show whether these trends really exist.

In the current study we exploit molecular dynamics (MD) simulations to investigate the mechanisms of thermal stability for two representative pairs of homologues. The thread linking our studies is the question relating protein stability to rigidity but we aim to eventually extract information that will deepen our understanding and inspire the design of new thermostable proteins. At the moment, given the maturity of the available force fields, MD is one of the best *in silico* techniques to address this problem. It can be successfully used to explore local and global protein flexibility in the microsecond time scale and beyond, allowing direct comparison with several experimental techniques. Moreover, when used in combination with simplified models and/or enhanced sampling techniques, it provides insight into the thermodynamic mechanisms underlying thermal stability.

Our first study case, discussed in Chapters 3, 4 and 5, is a pair of homologous G-domains from elongation factor thermo unstable (EF-Tu) and 1α (EF-1α) extracted from the mesophilic bacterium *Escherichia coli* and the hyperthermophilic archaeon *Sulfolobus solfataricus*, respectively. The G-domain constitutes the N-terminal domain of elongation factor that even if isolated, it can still perform GTPase activity and vary its catalytic power upon ribosome binding.

In Chapter 3, anticipating our results, we see that at the level of atomistic fluctuations the striking difference between the two homologues lies, not in the magnitude of fluctuations which is comparable for the two species but, in the distribution of flexible and rigid parts along the sequence. At the same time, a thorough clustering analysis of the conformations sampled in the microsecond timescale shows that the hyperthermophilic protein, in its folded state, is characterized by a larger conformational space than the mesophilic one.

In Chapter 4, we investigate the effect of higher temperature on the two species. The *in silico* stability of the hyperthermophile is verified while we identify the weak spot of the mesophilic protein where the unfolding begins. This region is important to the protein's activity and the equivalent one in the hyperthermophile confers flexibility yet stability to the protein matrix at all temperatures.

Chapter 5 deals with the different thermal stability of the two domains in a thermodynamic context. By exploiting an effective coarse-grained model in combination with the Replica Exchange MD technique, we extract, for the first time, a qualitative estimation of the stability curves for the two homologues and propose that the increase in the melting temperature of the thermophilic protein is a result of the broadening of the stability curve, something suggested already from Chapter 3.

Finally, Chapter 6 puts under the microscope a pair of tetrameric malate dehydrogenases from two bacteria, the mesophilic *Chlorobium vibrioforme* and the thermophilic *Chloroflexus aurantiacus*. We find that the thermophilic protein is more rigid than the mesophilic one both globally and locally around the active site. This result is in accord with previous comparative studies between homologous dehydrogenases. Intriguingly however, when we study the dynamics of the isolated monomers, the rigidity relationship between the two homologues reverses suggesting that increased stiffness of the thermophilic matrix is the result of optimized interfacial interactions propagating along the domains. This finding seems to support a correlation between thermophilic rigidity and oligomeric state.

Chapter 2

Methodology

Summary

The basic method employed in this work is Molecular Dynamics (MD), one of the most widely used computer simulation techniques in material science and biophysics. In short, MD is used to calculate the time evolution of a classical many-body system by numerically integrating Newton's equations of motion. It is currently applied to a large spectrum of systems, from nanomaterials to biomolecules [103, 104]. It has two strong points, it offers atomistic-level resolution and since it doesn't disrupt the kinetics of the system it allows for the calculation of transport properties and relaxation constants. However, when applied to large systems the sampling is limited by the currently available computational resources, e.g. the typical limit for a system of 100,000 atoms is the microsecond timescale. In the first section of this Chapter we give a brief overview of the method in the context of protein simulation. More detailed discussions on the several aspects of MD can be found in classical textbooks [105, 106]. The second section briefly introduces some of the enhanced sampling techniques used in combination with MD to improve sampling and assess thermodynamic quantities. Finally, the last section deals with the post-processing tools used herein to analyze MD trajectories.

2.1 Molecular Dynamics

2.1.1 The basic steps of a Molecular Dynamics simulation

The time evolution of a classical, conservative system obeys Newton's equation of motion [107].

$$m_i \boldsymbol{a}_i = m_i \frac{d^2 \boldsymbol{r}_i}{dt^2} = \boldsymbol{F}_i, \quad \text{with} \tag{2.1}$$

$$\boldsymbol{F}_i = -\boldsymbol{\nabla}_i U \tag{2.2}$$

where m_i is the mass of particle i, $\boldsymbol{a}_i$ its acceleration, $\boldsymbol{r}_i$ its position in the $3D$ Euclidean space, $\boldsymbol{F}_i$ is the force exerted on particle i by all others and U is the potential energy of the system. The idea behind Molecular Dynamics lies in the numerical integration of Eq. 2.1 in conjunction with the ergodic hypothesis, i.e. averaging in time is equivalent to averaging over the ensemble. In the vast majority of situations, the potential energy U of Eq. 2.2 is assumed to be pairwise additive as this reduces computational cost, i.e. the interaction between a particle and all others is the sum of all pairwise interactions[1]. In the context of protein simulations, detailed functional forms of commonly used potentials will be discussed in Subsection 2.1.3.

The core of a molecular dynamics program is comprised of the following steps [105]:

1. *Initialization of coordinates and velocities.*
2. *Calculation of forces.*
3. *Numerical integration of equations of motion.*
4. *Steps 2 and 3 are repeated for the desired length of time.*
5. *At the end, average quantities such as temperature, volume e.t.c. are computed and printed out along with the coordinate trajectories.*

Initialization of coordinates and velocities. We start by fixing the initial coordinates and velocities of the system. For example, in the case where we want to study the equilibrium properties of a folded protein the initial coordinates are usually those of the experimentally resolved structure of the protein, whereas if we want to study equilibrium folding/unfolding we might start from an extended random-coil conformation. The initial velocities are chosen randomly from the Maxwell-Boltzmann distribution that corresponds to the target temperature.

[1] Examples of non-pairwise additive potentials are the so-called polarizable force fields.

Calculation of forces. The next step is to compute all the forces on all the particles and it is the most time-consuming part of the simulation. First of all, in biomolecular simulations, as in molecular simulations in general, we most often use periodic boundary conditions in order to exclude finite size effects and account for more *in vivo* -like environments (e.g. mimic the cell environment inside which proteins act). In this context, every particle interacts only with the closest image of the all other particles in the system. If this so-called minimum image convention is not satisfied, a particle may interact with itself and the results of the simulation are non-physical.

As we will see in a following section describing the basic components of a force field, the interactions between the particles are separated in bonded and non-bonded ones. The slowest part of the calculation concerns the non-bonded interactions since for one unit cell their number scales as N^2, where N is the total number of particles. This time complexity makes the calculation forbiddingly slow and several efficiency techniques are used, for the short-range van der Waals interactions, in order to reduce the scaling down to $N^{3/2}$ or even N. The most common such technique, called Verlet-list [108], uses a cutoff for the force on particle i. That is to say, not all atoms in the box contribute to the energy of particle i but rather its close neighbors. Another approach is the cell-list or linked-list method, where the simulation box is divided into cells with a size equal to some cutoff [109]. Then the particles in each cell interact only with the particles in the same or neighboring cells. A combination of Verlet-list and cell-list can also be used where the latter is used in order to construct the Verlet-list [110].

The above cutoff schemes are applicable for short-range interactions. However, electrostatics is long-range and a simple cutoff scheme cannot be applied. Long-range interactions are thus treated separately with different methods that also reduce time complexity. The most common such method is the Ewald summation [111] were interactions are separated in two terms, a short-range calculated in real space and a long-range term evaluated in reciprocal space. The summation in Fourier space converges quickly and higher order terms may be neglected without loss in accuracy. The modern implementation of the technique, based on particle mesh approaches [112] yields an efficiency of $\mathrm{O}(NlogN)$. Other techniques also exist, like fast multipole expansion schemes, where a group of atoms at a large distance can be considered as one cluster for which the internal pairwise interactions don't need to be calculated [113].

Numerical integration of equations of motion. After all forces have been evaluated, numerical integration of the equations of motion (Eq. 2.1) must be performed. There exist several algorithms for that but care should be taken on the

choice of a good algorithm. Two important criteria that make up a good algorithm are time-reversibility, a fundamental symmetry of the equations of motion, and the symplectic property which ensures a bounded energy drift and stability for large timesteps (the latter is also important so that we need fewer time-consuming force calculations). The most common such algorithm is the Verlet [108]. The idea behind its derivation is very simple. For a lighter notation let's consider a particle in one dimension. We start with the Taylor expansion of its position around time t'.

$$r(t'+\Delta t) = r(t') + v(t')\Delta t + \frac{F\big(r(t')\big)}{2m}\Delta t^2 + \frac{1}{3!}\frac{d^3r(t)}{dt^3}\Big|_{t=t'}\Delta t^3 + \mathrm{O}\big(\Delta t^4\big) \tag{2.3}$$

and

$$r(t'-\Delta t) = r(t') - v(t')\Delta t + \frac{F\big(r(t')\big)}{2m}\Delta t^2 - \frac{1}{3!}\frac{d^3r(t)}{dt^3}\Big|_{t=t'}\Delta t^3 + \mathrm{O}\big(\Delta t^4\big) \tag{2.4}$$

Summation of Equations 2.3 and 2.4 yields

$$r(t'+\Delta t) + r(t'-\Delta t) = 2r(t') + \frac{F\big(r(t')\big)}{m}\Delta t^2 + \mathrm{O}\big(\Delta t^4\big) \tag{2.5}$$

$$r(t'+\Delta t) = 2r(t') - r(t'-\Delta t) + \frac{F\big(r(t')\big)}{m}\Delta t^2 + \mathrm{O}\big(\Delta t^4\big) \tag{2.6}$$

We thus have the position at time $t'+\Delta t$ as a function of the position in the two previous steps and the current force exerted on the particle. The error of this estimation is of the order of Δt^4, where Δt is the chosen timestep in the MD scheme. We notice that the velocities do not enter in the above scheme. In principle, we can proceed with our simulation without any knowledge of the velocities at any step. However this is not particularly useful as velocities are needed for the estimation of the kinetic energy and the temperature and also, due to computational restrains, we almost always need to restart our simulation, for which we need positions *and* velocities. We can evaluate them by subtracting Equation 2.4 from 2.3

$$r(t'+\Delta t) - r(t'-\Delta t) = 2v(t')\Delta t + \mathrm{O}\big(\Delta t^3\big) \tag{2.7}$$

$$v(t') = \frac{r(t'+\Delta t) - r(t'-\Delta t)}{2\Delta t} + \mathrm{O}\big(\Delta t^2\big) \tag{2.8}$$

which gives an estimation of velocities with larger error than for the positions, namely $\mathrm{O}\big(\Delta t^2\big)$. There are several variations of the Verlet algorithm. For example, the Leap Frog algorithm [109] evaluates the velocities at half integer timesteps and uses these velocities to compute new positions. The velocity Verlet algorithm [114] is similar to the Leap Frog, but the velocities are calculated for the same timestep as for the positions. The Leap Frog and velocity Verlet give the same order for the error of the velocity as the basic Verlet. There are however Verlet-like variations with greater accuracy, like the Beeman [115] and the velocity-corrected algorithms [105].

2.1.2 Keeping the temperature and pressure constant

Integration of Newton's equations of motion allow the system to explore the phase-space according to the microcanonical ensemble (NVE), i.e. a hypothetical collection of isolated systems that have the same Hamiltonian $\mathbb{H}(\boldsymbol{q}, \boldsymbol{p})$ but different microscopic states. In NVE the number of particles N, the volume V and the total energy E are constant and the microcanonical ensemble average of an observable A is given by

$$\langle A \rangle = \frac{\int d\boldsymbol{q}\, d\boldsymbol{p}\; A(\boldsymbol{q}, \boldsymbol{p})\, \delta(\mathbb{H}(\boldsymbol{q}, \boldsymbol{p}) - E)}{\int d\boldsymbol{q}\, d\boldsymbol{p}\; \delta(\mathbb{H}(\boldsymbol{q}, \boldsymbol{p}) - E)} \tag{2.9}$$

where $\mathbb{H}$ is the Hamiltonian of the system and $\boldsymbol{q}, \boldsymbol{p}$ the generalized coordinates and momenta. However, NVE yields results not directly comparable to experiment (most experiments are done in conditions of constant temperature and pressure as it is easier). Thus, in practice, the most common ensembles in MD are the canonical with constant NVT, where T is the temperature, the isothermal-isobaric with constant NPT, where P is the pressure, and the grand canonical with constant μVT, where μ is the chemical potential [116].

For the production phase of our simulations we used the isothermal-isobaric ensemble (NPT)[2]. In order to keep P and T constant some *ad hoc* algorithms have been developed. The methods for constraining thermodynamic variables can be categorized in strong and weak coupling formulations, stochastic motivated methods and extended system methods.

Thermostats. In experiments one keeps the temperature of a system constant by coupling it to a large heat bath. The total energy of the system is subject to fluctuations and the probability of finding it in a given energy state E_i is dictated by the Boltzmann distribution

$$\mathcal{P}(E_i) = \frac{e^{-\beta E_i}}{Z}, \quad \beta = \frac{1}{k_B T} \tag{2.10}$$

where Z is a normalizing constant also known as the partition function and k_B is Boltzmann's constant. The kinetic energy is also subject to fluctuations following the Maxwell-Boltzmann distribution, that is the probability of a particle to have momentum equal to $\boldsymbol{p}$ is given by

$$\mathcal{P}(\boldsymbol{p}) = \left(\frac{\beta}{2\pi m}\right)^{3/2} \exp(-\beta \boldsymbol{p}^2 / 2m) \tag{2.11}$$

The simplest approach for temperature control is velocity rescaling [117]. If we consider the, commonly used in simulations, definition of instantaneous tempera-

[2] Information on the average number of particles is not particularly relevant in biomolecular simulations.

ture

$$T(t) = \frac{1}{k_B N_f} \sum_{i=1}^{N_f} m_i v_i^2(t) \tag{2.12}$$

where N_f is the number of degrees of freedom, it is evident that by multiplying the velocity vector $\boldsymbol{V} = (v_1, v_2, ..., v_{N_f})$ with a scaling factor c_t we can reach fast the target temperature T_0

$$\boldsymbol{V'} = c_t \boldsymbol{V}, \quad \text{where} \quad c_t = \sqrt{\frac{T_0}{T}} \tag{2.13}$$

Velocity rescaling belongs to the "less correct" class of approaches in the sense that although it keeps the temperature constant, i.e. yields the desired total kinetic energy, the equilibrium probability distribution of a particle's momentum does not correspond to the Maxwell-Boltzmann one. Thus, even if it is acceptable to use such a method, for example, for effective heating of a system it must be avoided during production phases where we want to measure equilibrium averages which should be subject to the correct fluctuations.

Another velocity-scaling approach is the weak coupling scheme proposed by Berendsen [118]. Here the equations of motion are modified in the following way

$$m_i \frac{d^2 \boldsymbol{r}_i}{dt^2} = \boldsymbol{F}_i - m_i \gamma_t \frac{d\boldsymbol{r}_i}{dt}, \quad \text{where} \quad \gamma_t = \frac{1}{2\tau}\Big(1 - \frac{T_0}{T}\Big) \tag{2.14}$$

where γ_t has units of inverse time, T is the instantaneous kinetic temperature and τ is the time constant of the coupling to the heat bath. Eventually, there is an effective scaling of the velocity vector of Eq. 2.13 by

$$c_t = \sqrt{1 - \frac{\Delta t}{\tau}\Big(1 - \frac{T_0}{T}\Big)} \tag{2.15}$$

The Berendsen thermostat doesn't sample the canonical ensemble but it is roughly correct for large systems (hundreds or thousands of particles). A correction curing its deficiencies has been recently proposed [119].

Stochastic based thermostats are more common and more correct approaches. For example the Andersen thermostat [120] keeps the average temperature constant by coupling the system to a heat bath. This is done by "attempting collisions" with randomly selected particles at regular intervals of time. If a particle is selected to undergo a collision, its velocity is drawn from a Maxwell-Boltzmann distribution that corresponds to the desired temperature. Between stochastic collisions the system evolves at constant energy. The more frequent the collisions the greater the strength of the coupling to the heat bath.

In the present study we have mostly used an alternative stochastic approach, the Langevin dynamics [116, 121]. In Langevin dynamics, the force in the equations

of motion of the system has two extra terms, a friction term and a stochastic term.

$$m_i \frac{d^2 \boldsymbol{r}_i}{dt^2} = \boldsymbol{F}_i - m_i \gamma \frac{d\boldsymbol{r}_i}{dt} + \boldsymbol{W}_i(t) \tag{2.16}$$

where γ is a friction coefficient with units of inverse time, and $\boldsymbol{W}_i$ is a random force that is uncorrelated in time and across particles. The Langevin dynamics sample the canonical ensemble.

Other usual approaches to the constant temperature issue include also the stochastic dissipative particle dynamics thermostat [122], the more elaborate deterministic Nose-Hoover thermostat that adds extra degrees of freedom to the Lagrangian of the system [123, 124] and an extension of it, the Nose-Hoover chains where a thermostat is coupled to a chain of other thermostats [125].

Barostats. In short, just as a real piston would do, the pressure is controlled by allowing the volume to fluctuate.

In the same spirit as for thermostating, a weak coupling Berendsen scheme can be applied also for the pressure control [118] if we rescale both particle positions and total volume. That is, for a cubic box with length L and isotropic pressure

$$L' = c_p L \tag{2.17}$$

$$\boldsymbol{x}' = c_p \boldsymbol{x}, \quad \text{where} \quad c_p = \Big[1 - \frac{\beta \Delta t}{\tau_P}\Big(P_0 - P\Big)\Big]^{1/3} \tag{2.18}$$

where $\boldsymbol{x} \in \mathbb{R}^{3N}$, Δt is the timestep, τ_P is the pressure coupling, P is the instantaneous temperature and P_0 is the target pressure . The larger the τ_P the weaker the coupling.

Usually however, the well established schemes for pressure control are based on the pioneering extended-system approach by Andersen [120]. Here, the system is coupled to an external variable V, the volume of the system, which mimics the action of a piston on a physical system. This V is the coordinate of the piston with a user-defined mass M that quantifies the coupling of the barostat to the system (e.g. infinite mass yields no pressure control and coincides with NVT statistics). The equations of motion incorporating the Andersen barostat sample correctly the isobaric-isoenthalpic ensemble. Relevant schemes include the formulation by Hoover [126] which only approximates the desired distribution or the more accurate formulation by Martyna et al. [127]. In this approach the proposed equations of motion are

$$\frac{d\boldsymbol{r}_i}{dt} = \frac{\boldsymbol{p}_i}{m_i} + \frac{p_\epsilon}{M}\boldsymbol{r}_i \tag{2.19}$$

$$\frac{d\boldsymbol{p}_i}{dt} = \boldsymbol{F}_i - (1 + \frac{3}{N_f})\frac{p_\epsilon}{M}\boldsymbol{p}_i - \frac{p_\xi}{Q}\boldsymbol{p}_i \tag{2.20}$$

$$\frac{dV}{dt} = \frac{3Vp_\epsilon}{M} \tag{2.21}$$

$$\frac{dp_\epsilon}{dt} = 3V(P - P_0) + \frac{3}{N_f}\sum_{i=1}^{N}\frac{\boldsymbol{p}_i{}^2}{m_i} - \frac{p_\xi}{Q}p_\epsilon \tag{2.22}$$

$$\frac{d\xi}{dt} = \frac{p_\xi}{Q} \tag{2.23}$$

$$\frac{dp_\xi}{dt} = \sum_{i=1}^{N}\frac{\boldsymbol{p}_i{}^2}{m_i} + \frac{p_\epsilon^2}{M} - (N_f + 1)k_BT \tag{2.24}$$

where number 3 comes from the dimensions of the system. In the above scheme there is also a thermostat introduced via the variables ξ, p_ξ and Q that is similar to the NVT version of the Nose-Hoover chain algorithm [125] (not described herein). The barostat enters with the variables ϵ, p_ϵ and M where $\epsilon = ln(V/V_0)$, with V_0 being the initial volume, p_ϵ the momentum conjugate to ϵ and M the mass parameter associated to ϵ (mass of the "piston"). In the present study we have used the so called Nose-Hoover Langevin piston [128] which is a combination of the above Nose-Hoover method, with piston fluctuation control implemented using Langevin dynamics [129]. That essentially means that two stochastic terms are added to the system, a noise term W is added to the atom's Eq. 2.20 and a term W_e to the piston's Eq. 2.22. The user specifies the desired pressure P_0, the decay of the piston and its oscillation period τ_P entering via its mass $M = 3N\tau_P^2 k_BT$.

2.1.3 Protein force fields

We now briefly discuss the empirical potentials (a.k.a. force fields) used for Molecular Dynamics simulations of proteins. Force fields are broadly categorized in all-atom (AA), that as the name suggests, represent each atom explicitly and coarse-grained (CG) that represent clusters of atoms collapsed in one center of force (technically referred to as *bead*) in order to improve computational efficiency.

According to a chemical-mechanical representation of a molecule the "fundamental", classical interactions governing its motion can be separated in two main terms, bonded and non-bonded.

$$U = U_{\text{bonded}} + U_{\text{non-bonded}} \tag{2.25}$$

The bonded term accounts for 1-2, 1-3 and 1-4 pairwise interactions, corresponding to bond stretching, angle bending and dihedral (torsional) angle rotation, respectively (see Fig. 2.1). These three interactions are incorporated even in the simplest protein models. Bond stretching and angle bending are represented by harmonic potentials that model the oscillations around equilibrium values. Dihedral angle rotation around a central bond accounts for the possible *cis*/*trans* iso-

mers[3] and, as a multi-minima term, it is usually represented by a *cosine* function. Out-of plane deviations, a.k.a improper dihedrals, are generally described by an extra harmonic term. Reference values of bond lengths and angles are measured either experimentally (X-ray structures) or via computational *ab initio* methods and along with the associated force constants, they represent the set of bonded parameters of the model.

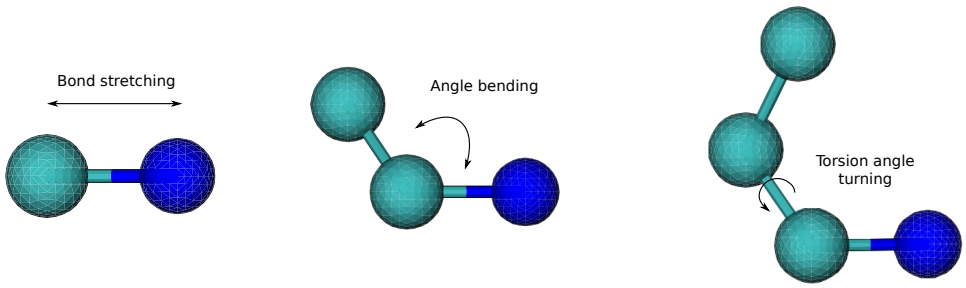

Figure 2.1. Basic molecular modes. Bond stretching (left), angle bending (middle) and dihedral angle turning (right).

The non-bonded term of the potential energy includes the dispersive van der Waals interactions, generally represented by Lennard-Jones -like potentials, and electrostatics stemming from the partial charges associated to each atom. Formal descriptions of these terms, in the context of specific force fields, are provided in the following paragraph.

The most common, all-atom, macromolecular force fields used at the moment are CHARMM (Chemistry at HARvard Molecular Mechanics) [130, 131], AMBER (Assisted Model Building and Energy Refinement) [132] and GROMOS (GROningen MOlecular Simulation package) [133]. In the all atom approaches of this work (Chapters 3, 4 and 6) we have used almost entirely the CHARMM22/CMAP [134] force field but some complementary results where also acquired using AMBER99sb [135] as well as the later version CHARMM36 [136]. A concentrating table with the simulation details of the present study can be seen at the Appendix of this Chapter.

CHARMM all-atom force field. The complete functional form of CHARMM is given below

$$U_{CHARMM} = \overbrace{U_{\text{bond}} + U_{\text{angle}} + U_{\text{UB}} + U_{\text{dihedral}} + U_{\text{improper}} + U_{\text{CMAP}}}^{U_{\text{bonded}}} + \overbrace{U_{\text{LJ}} + U_{\text{elec}}}^{U_{\text{non-bonded}}}$$

[3] cis/trans isomerism describes the relative orientation of functional groups

where

$$
\begin{aligned}
U_{bond} &= \sum_{bonds} K_b(b - b_0)^2, \\
U_{angle} &= \sum_{angles} K_\theta(\theta - \theta_0)^2, \\
U_{UB} &= \sum_{Urey-Bradley} K_{UB}(b^{1-3} - b_0^{1-3})^2, \\
U_{dihedral} &= \sum_{dihedrals} K_\varphi((1 + \cos(n\varphi - \delta)), \\
U_{improper} &= \sum_{impropers} K_\omega(\omega - \omega_0)^2, \\
U_{CMAP} &= \sum_{residues} u_{CMAP}(\Phi, \Psi), \\
U_{LJ} &= \sum_{nonb.pairs} \varepsilon_{ij} \left[\left(\frac{r_{ij}^{min}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^{min}}{r_{ij}} \right)^{6} \right], \\
U_{elec} &= \sum_{nonb.pairs} \frac{q_i q_j}{\epsilon r_{ij}}
\end{aligned}
$$

The term U_{UB} (Urey-Bradley) was added by the force field developers to complement the bond bending term with an extra harmonic bias between the distance of the 1st and 3rd atoms in an angle and was applied from case to case during the final optimization stage of the vibrational spectra. The term U_{CMAP} is a recent correction for cross terms of backbone dihedral angles.

Coarse-grained force fields, unlike all-atom ones, don't retain full resolution for the biomolecule but rather merge together groups of atoms in order to speed up simulation time and improve sampling. The exact resolution depends on the specific model and its purpose (i.e. predict structure, reproduce thermodynamic properties, e.t.c.). Below we mention the two coarse-grained models used herein.

OPEP coarse-grained force field. The OPEP (Optimized Potential for Efficient Protein Structure Prediction) coarse-grained force field, used in Chapter 5, is designed to fold peptides and small proteins and has been used successfully in a wide variety of cases [27, 30, 137–139]. The model represents each amino acid by six centers of force: the side-chain is represented by a unique bead, while atomistic resolution is reserved for the backbone that includes N, H_N, C_α, C, O atoms. Exception is proline whose side-chain is represented by all heavy atoms (see Fig. 2.2). The water effect is accounted for implicitly via the parameters of the potential. As in most force fields, the Hamiltonian consists of short-range and long-range

non-bonded interactions with the addition of an extra term, U_{HB}, that accounts explicitly for the backbone hydrogen-bond propensity.

$$U = U_{\text{local}} + U_{\text{non-local}} + U_{\mathrm{HB}} \tag{2.26}$$

The local contributions are given by

$$\begin{aligned} U_{\text{local}} = {} & w_b \sum_{bonds} K_b(b-b_0)^2 + w_\theta \sum_{angles} K_\theta(\theta-\theta_0)^2 \\ & + w_\varphi \sum_{dihedrals} K_\varphi((1+\cos(n\varphi-\varphi_0)) + w_\omega \sum_{impropers} K_\omega(\omega-\omega_0)^2 \\ & + w_{\phi\psi}(\sum_{\phi} K_{\phi\psi}(\phi-\phi_0)^2 + \sum_{\psi} K_{\phi\psi}(\phi-\psi_0))^2 \end{aligned} \tag{2.27}$$

where each w_x represents an optimizable weighting factor. The first four terms are bond stretching, angle bending, dihedral twisting and improper dihedral. The last two terms are additional harmonic biases confining the dihedrals within the Ramachandran limits, where $\phi_0 = \phi$ if the torsion is found in the interval $[\phi_{\text{lower}}, \phi_{\text{upper}}]$ otherwise $\phi_0 = min(\phi - \phi_{\text{lower}}, \phi - \phi_{\text{upper}})$. The limits for the ϕ angle are $\phi_{\text{lower}} = -160°$ and $\phi_{\text{upper}} = -60°$ while for the ψ angle are $\psi_{\text{lower}} = -60°$ and $\psi_{\text{upper}} = 160°$.

The non-bonded interactions consist of the terms

$$\begin{aligned} U_{\text{non-local}} = {} & w_{1,4} \sum_{1,4} U^1_{\mathrm{VdW}} \\ & + w_{1>4}\Big(\sum_{\mathrm{M'},\mathrm{M'}} U^1_{\mathrm{VdW}} + \sum_{\mathrm{M'},\mathrm{C}_\alpha} U^1_{\mathrm{VdW}} + \sum_{\mathrm{M'},\mathrm{S}_c} U^1_{\mathrm{VdW}} + \sum_{\mathrm{C}_\alpha,\mathrm{S}_c} U^1_{\mathrm{VdW}}\Big) \\ & + w_{C_\alpha,C_\alpha} \sum_{C_\alpha,C_\alpha} U^2_{\mathrm{VdW}} + w_{S_c,S_c} \sum_{S_c,S_c} U^2_{\mathrm{VdW}} \end{aligned} \tag{2.28}$$

where M′ denotes all the backbone atoms except for C_α and S_c is the side-chain bead. The first short-range term accounts for all 1-4 interactions along each torsional degree of freedom. The second block of terms is long-range with respect to sequence position and involves (M′,M′) interactions and mixed interactions (M′,S_c),(M′,C_α) and (C_α,S_c) that share the same weighting factor $w_{1>4}$. The last two terms specifically account for (C_α,C_α) $j > i+3$ interactions and (S_c,S_c) $j > i+1$ interactions, where i and j are two amino acids in the sequence. The VdW potential in OPEP has two distinct forms denoted with VdW^1 and VdW^2 that depend on the center of forces involved. Their form is described in detail elsewhere [27, 137, 138]. The energy scale of the model is set by both the optimized weighting factors of each term of the Hamiltonian and a global scaling factor. In numbers: as reference we consider the minimum of the non-bonded Ile/Ile interaction that is generally set to 3.89 kcal/mol. For some applications and tests a weaker value 3.49

kcal/mol has also been used. OPEPv5 [138] was also recently introduced, where the ionic interactions Lys-Asp, Lys-Glu, Arg-Asp and Arg-Glu are not described anymore by the VdW[2] long-range interaction – as they do in the previous versions of the model – but instead are replaced by *ad hoc* pair potentials $V^{eff}(r)$ derived using the iterative Boltzmann inversion (IBI) procedure and targeting the radial distribution functions between the center of mass of the side chains obtained by extended atomistic MD simulations of the ionic pairs in explicit solvent. In Chapter 5, however, we employ the version 4 of the force field.

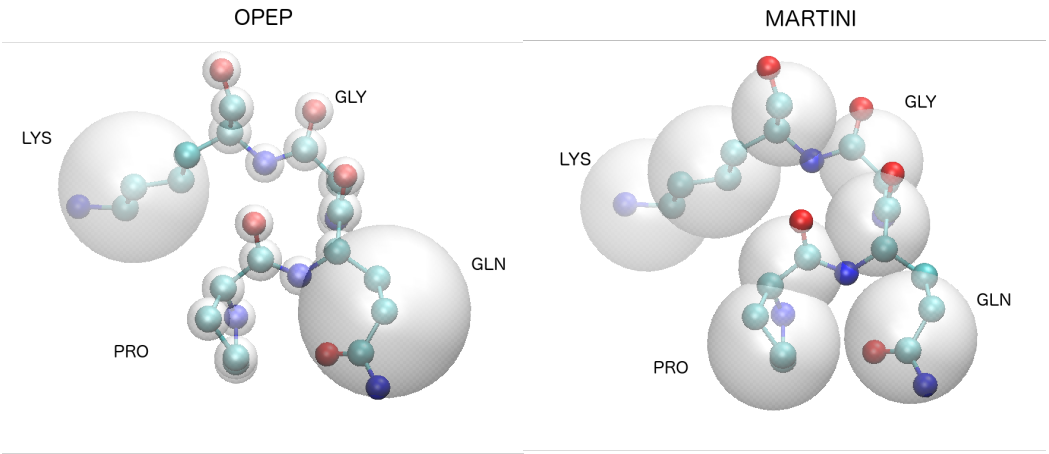

Figure 2.2. Resolution of OPEP (left) and MARTINI (right) CG force fields. With solid color is the all-atom representation whereas the larger transparent grains correspond to the CG beads.

MARTINI coarse-grained force field. The MARTINI coarse-grained force field was initially developed for MD simulations of lipids [140] and was later extended to proteins and other biomolecules [141]. Since it imposes secondary structure constraints, its purpose is fundamentally different from that of OPEPs. It has been parametrized focusing on non-bonded interactions and aiming to reproduce thermodynamic properties of the systems under study. Among the model's applications are lipid aggregation, membrane protein-lipid interaction, self-assembly of soluble peptides and proteins, membrane protein oligomerization e.t.c. [142]. The model maps on average four heavy atoms to a single bead with the exception of ring-like fragments (e.g. the side-chain of proline e.t.c.) where a higher resolution is used (see Fig. 2.2). Martini uses explicit water where, again, four real water molecules are mapped to a CG water bead, while ions are represented by one bead. The functional forms for the force field can be found in Ref. [141]. In the last part of Chapter 6, we use MARTINI v2.1 with a recently introduced polarizable water

model [143].

2.2 Enhanced sampling techniques

As mentioned in the first section of this Chapter, despite the powerful atomistic resolution of MD, many processes of biological relevance are not accessible with the current computational power. That is, high-probability stable and metastable states of the conformational space are separated from each other by low-probability regions, also known as energetic barriers, the transition of which is a rare event in the timescales explored by MD. One example of what this practically means is that it is impossible to observe, with a single serial MD simulation (a.k.a. brute force MD), the folding event of any protein other than small peptides and ultrafast folders. Notably, only the massively parallel MD-designed Anton computer has been able to observe multiple repeated folding/unfolding events in a single all-atom MD simulation, and the proteins under study were fast folding ones [144]. But the problem is not only restricted to protein folding. Sufficient and correct exploration of the conformational space can help us understand also the function of proteins since the tight relationship between flexibility and function is in most cases far from understood [7, 89]. Thus, we might be talking about equilibrium simulations but brute force MD, especially for biological systems, is practically far from ergodicity.

In the context of protein stability, brute force MD can give in certain cases, as we will see in Chapter 4, insightful information on the kinetic stability of proteins. However, as already discussed in Chapter 1, thermodynamic stability is defined as the free energy difference between the folded and the unfolded states. Since the free energy of a state is closely related to the probability of this state, it is obvious that we can not estimate free energies via brute force MD. There exist several techniques that combined with MD can tackle the problem [145, 146]. We have already mentioned coarse-grained models that improve sampling of the conformational space by reducing the system's degrees of freedom at the cost of less accuracy. These models stretch the timescale of simulation but their simplified potential energy doesn't provide precise estimates of thermodynamic quantities. At the same time there is a large number of currently used techniques which are based on improved sampling. These techniques either use an external bias that forces the system to sample the low probability regions (*collective variable biasing* techniques) or speed up barrier crossing by making use of the temperature effect (*tempering* techniques). The field is vast, however below we briefly mention the basic ideas behind the techniques used in this work.

2.2.1 Collective variable biasing

A "collective variable" (CV) is, in general, a multidimensional function $f : \mathbb{R}^{3N} \rightarrow \mathbb{R}^M$ mapping a point of the $\mathbb{R}^{3N}$ configurational space of a system with N atoms to a point in $\mathbb{R}^M$ CV space where $M << 3N$. The new trajectory in the CV space visits the point $\boldsymbol{z} \in \mathbb{R}^M$ with a given frequency that, after sufficiently long time, converges to a probability $\mathcal{P}(\boldsymbol{z}) = \langle \delta\big(f(\boldsymbol{x}) - \boldsymbol{z}\big) \rangle$, where $\boldsymbol{x} \in \mathbb{R}^{3N}$ and the Dirac delta function picks only the initial configurations that map, via f, to the point $\boldsymbol{z} \in \mathbb{R}^M$. This probability can be expressed as a free energy [147]

$$\mathcal{F}(\boldsymbol{z}) = -k_B T ln\big(\mathcal{P}(\boldsymbol{z})\big) \tag{2.29}$$

where k_B is Boltzmann's constant and T is the temperature. Our problem usually boils down to estimating the free energy difference between two points of interest of the CV space.

One approach that allows the estimation of free energy differences by biasing a collective variable is *thermodynamic integration* and somehow follows an experimental "treatment": makes use of the fact that one can calculate the derivative of the free energy $-(\partial\mathcal{F}(\boldsymbol{z})/\partial\boldsymbol{z})$, a.k.a. the "mean force", with respect to the CV at the chosen points of the CV space. Then, the free energy difference between two states $\boldsymbol{z}_A$ and $\boldsymbol{z}_B$ is given by the line integral

$$\Delta\mathcal{F}_{\boldsymbol{z_A} \rightarrow \boldsymbol{z_B}} = \int_{\boldsymbol{z}_A}^{\boldsymbol{z}_B} d\boldsymbol{z} \cdot \Big(\frac{\partial\mathcal{F}(\boldsymbol{z})}{\partial\boldsymbol{z}}\Big) \tag{2.30}$$

Several techniques can be used to compute the local free-energy gradients along the chosen path, e.g. by applying a mechanical constraint (e.g. blue moon sampling), a harmonic restraint or a time dependent bias (e.g. adaptive biasing force) to the system's Hamiltonian

$$\mathbb{H}_{\text{total}}(\boldsymbol{x}) = \mathbb{H}_{\text{system}}(\boldsymbol{x}) + U_{\text{external}}(f(\boldsymbol{x})) \tag{2.31}$$

The strongest point of *thermodynamic integration* is that the path doesn't need to be a real physicochemical process, it can as well be an alchemical process such as the transformation of one element to another [148]. In such a case the initial potential energy of the system is different from the final one while the path from the initial to the final state is via a coupling parameter λ that goes from zero to one

$$U(\lambda) = (1 - \lambda)U_{\text{I}} + \lambda U_{\text{II}} \tag{2.32}$$

Then, assuming Boltzmann statistics

$$\Delta\mathcal{F}_{\lambda=1\rightarrow\lambda=0} = \int_{\lambda=0}^{\lambda=1} d\lambda \Big(\frac{\partial\mathcal{F}(\lambda)}{\partial\lambda}\Big) \tag{2.33}$$

$$= -\int_{\lambda=0}^{\lambda=1} d\lambda \frac{\partial\Big(k_B T ln\big(Z(\lambda)\big)\Big)}{\partial\lambda} \tag{2.34}$$

$$= -\int_{\lambda=0}^{\lambda=1} d\lambda \frac{k_B T}{Z(\lambda)} \frac{\partial Z(\lambda)}{\partial\lambda} \tag{2.35}$$

$$= \int_{\lambda=0}^{\lambda=1} d\lambda \frac{\int d\boldsymbol{x}^{3N} \frac{\partial U(\lambda)}{\partial\lambda} \exp(-U(\lambda)/k_B T)}{\int d\boldsymbol{x}^{3N} \exp(-U(\lambda)/k_B T)} \tag{2.36}$$

$$= \int_{\lambda=0}^{\lambda=1} d\lambda \left\langle \frac{\partial U(\lambda)}{\partial\lambda} \right\rangle_{\lambda} \tag{2.37}$$

where in this case the free energy is classically defined as $\mathcal{F} = -k_B T ln\big(Z(\lambda)\big)$ with $Z(\lambda) = \int d\boldsymbol{x}^{3N} \exp(-U(\lambda)/k_B T)$ being the partition function. The $\langle ... \rangle_{\lambda}$ denotes an ensemble average performed for a fixed value of λ. The final result is important because it expresses a free energy difference as an ensemble average that can be calculated directly in a simulation [105].

A similar alternative to *thermodynamic integration* is *free energy perturbation* [149] where the free energy is given by

$$\Delta\mathcal{F}_{\boldsymbol{z}_A\rightarrow\boldsymbol{z}_B} = -k_B T \ln \left\langle \exp\left(-\frac{\mathbb{H}(\boldsymbol{z}_B) - \mathbb{H}(\boldsymbol{z}_A)}{k_B T}\right)\right\rangle_{\boldsymbol{z}_A} \tag{2.38}$$

In this approach we use the information collected in the initial state A to obtain the free energy difference between A and B. This generally requires that the method is applied for overlapping, consequent intervals along the reaction path.

Another approach applying an *a priori* known external potential to the Hamiltonian is *umbrella sampling*, used in Chapter 6 [150, 151]. In this case, we make use of the fact that by construction we know the biased potential, in order to recover the unbiased probability of the system for the specific region of the CV space as a function of the biased one. The method introduces a weighting function $w(\boldsymbol{x})$, which in the vast majority of the cases is an explicit function of the CV coordinates $w(\boldsymbol{x}) = W[f(\boldsymbol{x})]$, and which is effectively as if adding the bias potential

$$U_{\text{external}}(\boldsymbol{x}) = -k_B T \ln w(\boldsymbol{x}) \tag{2.39}$$

The usual implementation is to run several independent MD simulations each one "weighting" a different region of the CV space. At the end, with the restriction that there is a good overlap between the sampling of each region, the statistics can be combined using the weighted-histogram analysis method (WHAM) [152, 153]. In Chapter 6, we combine the *umbrella sampling* with the CG model Martini to estimate the inter-domain binding free energies of two oligomeric homologues.

2.2.2 Temperature Replica Exchange Molecular Dynamics

Another way to enhance the sampling in MD simulations is the temperature Replica Exchange Molecular dynamics (REMD) technique [154]. Here, one runs in parallel independent simulations of the exact same system (replicas) at different temperatures. Every n_t steps, an exchange between – usually neighboring – configurations is attempted with an acceptance ratio that guarantees that the final (equilibrium) distribution for each temperature corresponds to the canonical ensemble. The probability of acceptance is

$$\mathcal{P} = \min\left(1, e^{(E(\boldsymbol{x}_j)-E(\boldsymbol{x}_i))\left(\frac{1}{k_B T_i}-\frac{1}{k_B T_j}\right)}\right) \tag{2.40}$$

where $E(\boldsymbol{x}_{i,j})$ and $T_{i,j}$ are the energies and temperatures of replicas i and j the timestep before the attempt. Eq. 2.40 is essentially the Metropolis criterion used in Monte-Carlo simulations to attempt jumps from one configuration to another. Thus, REMD is a combination of MD with a random walk in temperature space. A relevant scheme of the REMD method is given in Fig. 2.3. In Chapter 5 we employ REMD in conjunction with the OPEP CG force field to estimate the stability curves and melting temperatures of two homologues under study.

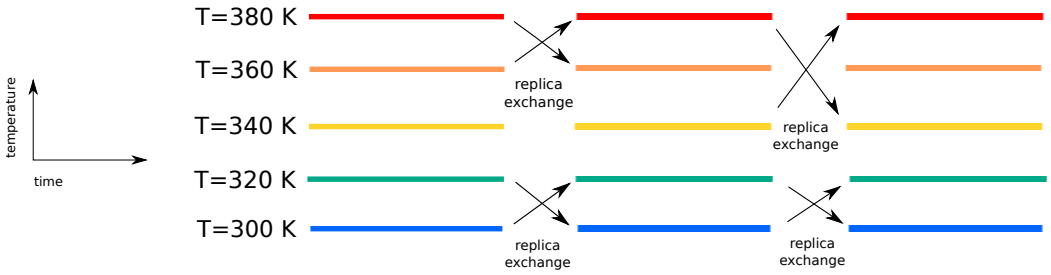

Figure 2.3. Schematic representation of the Replica Exchange Molecular Dynamics (REMD) enhanced sampling techinque.

2.3 Analysis of molecular dynamics' trajectories

At the end of a simulation run, one is left with an intimidating amount of information to interpret: the evolution in time of the x, y and z coordinates of every particle in the system. Given that proteins range from hundreds to hundreds of thousands of particles (excluding the number of water molecules for the case of explicit solvent) the problem hardly ends after the simulation does. MD post-processing tools range from the definition of collective variables (CV) to clustering techniques, network analysis and others, where in almost all cases the goal

is to reduce the dimensionality of the system for a more effective description of its behavior. Below we first introduce the CVs used in this work, chosen for their capability to effectively distinguish either folded-state fluctuations (e.g. $RMSD$) or between folded and unfolded states (e.g. radius of gyration). Then we present additional measures and concepts used in the following Chapters to quantify the flexibility of a protein's folded state; these are the harmonic diffusive model for the description of protein internal dynamics and cluster analysis in combination with network approaches.

2.3.1 Collective variables

The radius of gyration provides information on the size of the system and is given by

$$R_g(t) = \sqrt{\frac{1}{N_{BB}} \sum_{i=1}^{N_{BB}} (\boldsymbol{r}_i(t) - \langle \boldsymbol{r}(t) \rangle)^2} \tag{2.41}$$

where the summation is over all the heavy backbone atoms, that is N_{BB}=C, C_α, N and O, $\boldsymbol{r}_i(t) \in \mathbb{R}^3$ is the position of the i-th atom at time t and $\langle \boldsymbol{r}(t) \rangle$ is the average position over all backbone atoms at time t. We note that this definition differs from the standard definition of the radius of gyration that is averaged over time and depends on atomic masses.

The Root Mean Square Displacement (RMSD) is used to estimate the distance between two different conformations. It is computed via the following expression

$$RMSD(t) = \sqrt{\frac{1}{N_{\text{atom}}} \sum_{i=1}^{N_{atom}} (\boldsymbol{r}_i(t) - \boldsymbol{r}'_i)^2} \tag{2.42}$$

where N_{atom} is the number of atoms used, each time, for the calculation. We usually choose either C_α atoms or backbone atoms (C, C_α, N and O) or, very rarely, all heavy atoms, i.e. all proteins atoms except for hydrogen. Again $\boldsymbol{r}_i(t)$ is the position of the i-th atom at time t and $\boldsymbol{r}'_i$ is its reference position in a reference structure, which in the majority of the case is the crystal structure of the protein. For the calculation of $RMSD$ we remove rigid body motions by super-imposing the set of chosen atoms in their configuration at time t on those of the reference structure. In Chapter 5, where we make use of the CG model OPEP, we compute the $RMSD$ for the rigid-core of the proteins, i.e. the stretches of the protein that correspond to well defined secondary structure in the crystal structure (i.e. anything but coil or loop). The reason for this is that due to the absence of viscous aqueous medium (OPEP represents water implicitly), turns and coils, located at the surface of the protein move in a much noisier way.

Another useful CV is the fraction of native contacts that quantifies the deviation from the native folded state by looking at the proximal content of atoms.

Specifically here, the number of native contacts l'_i for a given C_α site is the number of all other C_α atoms located within a cut-off distance of 8 Å from the site in the crystallographic structure. Thus, the fraction of native contacts, referring to the whole chain, is defined as

$$Q(t) = \frac{1}{N_{C_\alpha}} \sum_{i=1}^{N_{C_\alpha}} \frac{l_i(t)}{l'_i} \tag{2.43}$$

where N_{C_α} is the number of C_α atoms, having l'_i native C_α contacts in the reference state and $l_i(t)$ of them appearing also at time t ($l_i(t) \leq l'_i$).

In the same spirit, the fraction of native torsion angles is calculated as follows

$$n_t(t) = \frac{1}{N_\theta} \sum_{i=1}^{N_\theta} \exp\left[-\frac{(\theta_i(t) - \theta'_i)^2}{\sigma^2}\right], \quad \text{where} \quad |\theta_i(t) - \theta'_i| < 180^\circ \tag{2.44}$$

where N_θ is the number of torsion angles θ, having values θ'_i in the equilibrated structure and values $\theta_i(t)$ at time t and $\sigma = 60^o$. For our calculations we considered both ϕ and ψ dihedrals.

In this work we have also used the fraction of secondary structure, which corresponds to the number of residues belonging to a well-defined secondary structure, namely α-helix or β-strand (codes G, H, I, E or B) as calculated by the DSSP algorithm [155] divided by total number of residues in the sequence.

The Root Mean Square Fluctuations ($RMSF$) of an atom, although technically is not a CV as defined above, is a useful measure to quantify atomistic fluctuations, and for a given atom i it is computed via the following expression

$$RMSF_i = \left\langle \sqrt{\langle (\boldsymbol{r}_i(t) - \overline{\boldsymbol{r}}_i)^2 \rangle_w} \right\rangle \tag{2.45}$$

where the two brackets denote a double time average. The inner average $\langle ... \rangle_w$ is calculated over a small timescale window denoted by w. Specifically in Chapter 3 this window was 350 ps ensuring the unimodal distribution of the atomic fluctuations. Larger windows were also used in Chapter 6 in order to examine the dependence of $RMSF$ on the timescale. The outer average $\langle ... \rangle$ is calculated along the trajectory over all blocks of w ps. This averaging improves statistics given that no large scale conformational changes have occurred in the proteins. We note that both R_g and $RMSF$ are quantities directly comparable with experiment.

2.3.2 Harmonic diffusion model for the folded state

The concept of diffusion along a single reaction coordinate dates back to Kramers' theory introduced in 1940 [156]. In the original paper a trapped Brownian particle was used to model the kinetics of chemical reactions. Ever since diffusive models

have been extensively used to describe the internal friction of biomolecules, especially in the theory of protein folding. For example, McCammon and Karplus used a diffusive model to theoretically estimate characteristic times of the internal motions of IgG-class antibodies [157] and later to characterize the rotational motions of rings in BPTI [158]. Subsequently, a protein-inspired theoretical framework for one-dimensional diffusion in a rough potential was presented by Zwanzig [159]. Another characteristic application was that of Woolf and Roux [160] where they used a harmonic approach (with umbrella sampling potentials) to quantify the diffusivity of the dihedral angles of two polar phospholipid head groups in both the absence and presence of explicit water. The same harmonic approach was used also by Hummer [161] for the estimation of diffusion coefficients of the dynamics of a peptide in explicit water. Finally, diffusion is a key concept in the theory of protein folding with the notable work by Socci, Onuchic and Wolynes where a single CV was enough to explain the folding kinetics on lattice models in very good agreement with experiment [162].

In theory, if we consider diffusion in a one-dimensional harmonic potential

$$U(X) = \frac{1}{2}kX^2 \tag{2.46}$$

– which in practice is the potential that the one-dimensional CV experiences – the corresponding Smoluchowski diffusion equation describing the time evolution of the probability density function of a particle's position

$$\partial_t \mathcal{P}(X,t|X_0,t_0) = D\left(\partial_X^2 + \beta\, k\, \partial_X X\right)\mathcal{P}(X,t|X_0,t_0) \tag{2.47}$$

where D is the diffusion coefficient and X_0, t_0 initial conditions, can be solved analytically. With the prerequisite that the stationary solution has to be the Boltzmann distribution, the diffusion coefficient with respect to the CV X is given by [163]

$$D = \frac{\langle \delta X^2 \rangle}{\tau_{corr}} \tag{2.48}$$

where $\delta X = X - \langle X \rangle$ is the instantaneous fluctuation of the collective variable and τ_{corr} its correlation time, being defined as

$$\tau_{corr} = \int_\tau R(\tau)d\tau, \quad \text{where} \quad R(\tau) = \frac{\langle \delta X(t) \cdot \delta X(t+\tau) \rangle}{\langle \delta X^2 \rangle} \tag{2.49}$$

In some sort of analogy to the work by Woolf and Roux [160] where they artificially imposed harmonic constraints along the reaction coordinate in order to get the value of D at different positions, here we employ the definition 2.48 for the diffusion coefficient assuming that the folded state is itself to some extent harmonic.

Depending on the system and the chosen CV this approximation is less or more valid. In practice, we check its validity by verifying that the autocorrelation

function $R(\tau)$ in Eq. 2.49 decays exponentially after an initial fast and short transient time ($\sim 300\,\text{ps}$). In this work we use an exponential fit to estimate τ_{corr} on long stationary stretches of the trajectory. As an example, in Chapter 3 the harmonic approximation is valid for the n_t and Q variables, but for the case of the $RMSD$ of the two systems under study, there is a rather unsteady behavior and the stationary intervals are not sufficiently long to correctly estimate τ_{corr}. There, we perform additional simulations with a harmonic restraint on the $RMSD$ variable, biasing its motion around its average value. The diffusion constant D is again evaluated as above mainly to verify the results obtained from the other CV but we acknowledge that alone, it only provides a very local information due to the applied restraint.

2.3.3 Cluster analysis

Cluster analysis is the process of grouping together multidimensional data based on some similarity measure [164, 165] and it can be particularly useful for pattern extraction from MD data. The basic steps of a clustering process are [165]

1. *Choose the desired data representation.*
2. *Choose the distance measure.*
3. *Perform the actual clustering or grouping.*
4. *Assess the output.*

Choose the desired data representation. Specifically for the case of MD trajectories there are countless options depending on the system and the question asked. The most frequent one is to use all, or some of, the x, y and z coordinates of the system. Although this is somehow standard, it is not always useful. In order to perform the following up pairwise comparison, rigid body motions must be removed which is not always straightforward (if even possible in cases such as protein unfolding or peptide aggregation). In this study, since the main question concerns the flexibility of the folded state, our clustering analysis was done each time on the ambient-temperature trajectories with the proteins being fully stable. However, except for spatial coordinates, we also applied two other representations to examine the robustness of our results against the chosen observable. These were the fraction of native contacts and the fraction of native torsion angles (described below).

Choose the distance measure. Although in general this also depends on the question, for the case of the spatial coordinates of a protein the obvious choice is

RMSD as defined in Eq. 2.42, which is also what we use in this work. For the case of the fraction of native contacts Q each configuration is represented by a vector of length equal to the number of C_α atoms (N_{C_α}) with its $i-$th component being the quantity $l_i(t)/l'_i$ of Eq. 2.43. Then, the distance between two configurations is defined as

$$d(t,t') = \sqrt{\frac{1}{N_{C_\alpha}} \sum_{i=1}^{N_{C_\alpha}} (L_i(t) - L_i(t'))^2}, \quad \text{where} \quad L_i(t) = \frac{l_i(t)}{l'_i} \tag{2.50}$$

Equivalently for the fraction of native ϕ and ψ angles, each configuration is represented by a vector of length equal to the number of ϕ and ψ angles along the sequence ($N_\theta = N_\phi + N_\psi$) and where the $i-$th component of the vector is given by the quantity $\exp\left[-(\theta_i(t) - \theta'_i)^2/\sigma^2\right]$ of Eq. 2.44. Then the distance between two configurations is defined as

$$d(t,t') = \sqrt{\frac{1}{N_\theta} \sum_{i=1}^{N_\theta} [\Theta_i(t) - \Theta_i(t')]^2}, \quad \text{where} \quad \Theta_i(t) = \exp\left[-\frac{(\theta_i(t) - \theta'_i)^2}{\sigma^2}\right]$$
$$\text{and} \quad |\theta_i(t) - \theta'_i| < 180^\circ \tag{2.51}$$

Perform the actual clustering or grouping. The proposed algorithms, up to date, can be divided in two basic categories, partitional and hierarchical ones. Partitional algorithms produce one single series of clusters whereas hierarchical algorithms produce several nested series of clusters. The two most popular partitional algorithms are the k-means [166] and the leader/follower [167] that have one basic difference. K-means depends on *a priori* knowledge of the number of clusters that one gets at the end, i.e. the number of clusters is a parameter of the algorithm. Leader/follower, on the other hand, takes as a parameter the cutoff distance between the cluster centroids, so the number of clusters is one of the outputs of the algorithm. At the other end, hierarchical clustering does not depend on a priori knowledge of either the number of clusters or the cutoff distance but unfortunately it requires to save the $N(N-1)/2$ distance values simultaneously and has a time complexity larger than $O(N^2)$ that practically forbid its use for large data sets (as is usually the case for MD data). Here clustering was used as an additional measure of protein flexibility through the resulting number of clusters. To that purpose we employed the leader/follower scheme which is described below. In some cases, we verified our results using an hierarchical agglomerative scheme on a smaller strode trajectory or making use of k-means for an extra refinement of the cluster memberships.

The leader/follower algorithm. It was first proposed by Hartigan in 1975 [167] and it is the fastest algorithm in terms of time complexity $O(Nk)$, where k is

some constant and N is the size of the initial data set. One of its drawbacks is that it is order dependent, that is the partitioning may be different depending on the order with which the data are fed to the algorithm. However this problem is not very relevant to us. First, since the MD simulation preserves the kinetics of the system feeding a trajectory to the algorithm with any order other than the chronological is meaningless. Secondly, the number and size of the generated clusters, which is what we are actually interested in, depends almost entirely on the chosen distance threshold. In the applications of the following chapters, when comparing the clusters between two systems we always use the same cutoff and several cutoff values are tested for the sake of robustness. Another important point should be made here. For a random polymer the possible number of clusters grows exponentially with the length of the chain. This is why we should in principle compare chains of equal size. At the same time, let's keep in mind that folded proteins are far from random and in their compact fold this dependency is much weaker.

Assess the output. In the framework of our approach, we begin by plotting the number of newly created clusters versus time in order to assess the available conformational space for different systems. This plot can also be used to estimate the "goodness" of the local sampling: given that the barrier-heights separating the different clusters of folded conformations are comparable, the number of newly created clusters saturates exponentially fast to a constant value N_∞, following the simple model $N = N_\infty(1 - e^{-t/\tau})$. If the fit is not possible or a saturation is not evident then one should in principle extend the simulation. Of course there is a subjectivity entering with the choice of the cutoff, which is why one should test the robustness of any conclusions by trying several cutoffs.

Continuing with the assessment of the clustering output, we also make use of network approaches. We represent the clusters as nodes of a network, the size of which is proportional to their occupancy. The nodes are connected by edges weighted with the number of interconversions between two adjacent nodes. Then the network is drawn with the use of a force-based algorithm, effectively projecting on a $2D$-plane the topology of the underlying conformational landscape. Force-directed algorithms are one of the most widely used graph-drawing techniques where, in order to obtain the final network representation, forces are introduced between the nodes of the network and the algorithm searches for an equilibrium with the force on each node being zero. The algorithm used herein is comprised of

- an attractive, spring-like force between connected nodes with a spring constant proportional to the weight of the adjacent edge
- a general charge-like repulsive force acting upon all the nodes

- a gravitational force
- and a bias towards placing the hubs of the network at more central positions

In the following, whenever comparisons are made between different networks the parameters where set to equal values.

We, also, use measures from complex networks and information theory to complement this representation. Such are the transitivity coefficient of a network that quantifies the tendency of the nodes to cluster together. It is defined as the probability that two nodes adjacent to a third are also connected to each other [168]. Finally, in some cases, we also evaluate the Shannon entropy [169] for the obtained networks $\mathrm{H} = -\sum_{i=1}^{N} p_i \ln(p_i)$, where the summation runs over all the different clusters N and p_i is the relative occupation of the i_{th} cluster.

Kinetic clustering. We then take the network analysis presented above one step further. The original clustering is coarse grained via an iterative procedure based on the "Markov clustering algorithm" (MCL) by S. van Dongen [170, 171] which separates the states depending on their effective kinetic barriers. MCL is based on a random walk on a network and works as follows.

(i) From the original network of clusters a transition matrix A_{ij} is constructed, where element A_{ij} is the number of transitions from cluster i to j. This matrix is then transformed into a stochastic one by normalizing its columns to one, $A'_{ij} = A_{ij} / \sum_{j=n}^{N} A_{ij}$, with N the number of substates.

(ii) Then the matrix is squared, $M_{ij} = \sum_{k=1}^{N} A'_{ik} A'_{kj}$ to yield the probabilities of transition from node j to node i in a two-step path. This operation is called expansion.

(iii) The elements of the squared matrix are raised to the power p, with $p > 1$, in order to promote the most probable paths at the cost of the less probable ones and the resulting matrix is again column-normalized, $M'_{ij} = M^p_{ij} / \sum_{j=n}^{N} M^p_{ij}$. This operation is the so-called inflation.

(iv) $A'_{ij} = M'_{ij}$ and steps (ii), (iii) and (iv) are repeated until convergence where the matrix is invariant upon the two operations. Recurrent application of the expansion and inflation operators results to a matrix with exactly one non-zero entry per column and isolated paths. Each of these paths includes the most connected substates with respect to a certain granularity defined by the exponent p. Within the free energy landscape picture, the granularity parameter determines the minimum height of the kinetic barriers detected by the algorithm that confine the walkers in separate parts of the network. Small values of p result in a very coarse-grained final representation while

large values of p result in a more detailed one. In the following discussions the testing of several exponents yields robustness to our results. Figure 2.4 gives a schematic representation of the algorithm.

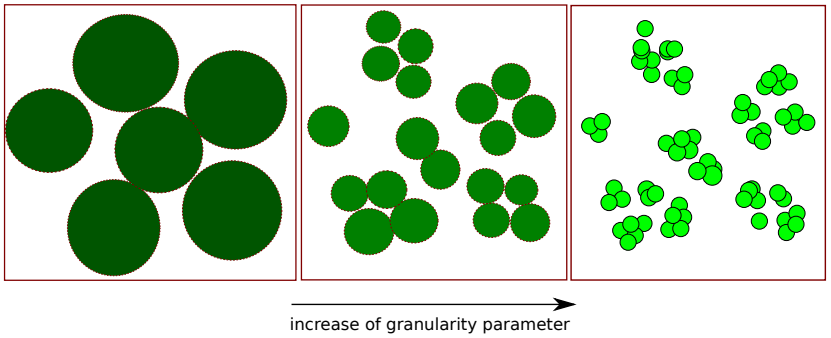

Figure 2.4. Schematic representation of the MCL algorithm. As the granularity parameter of the algorithm increases, random walkers on a network get confined in smaller and smaller regions.

Appendix

Table 2.1. Simulations performed in this study using different force fields and methodologies.

System	Force Field	Duration (ns)	T (K)	Run type	Chapter
1EFC ($\mathcal{M}$ G-domain)	CHARMM22/CMAP	600	300	brute force MD	3
	CHARMM22/CMAP	250	330, 360	brute force MD	3
	CHARMM22/CMAP	50	310-370 by 10	brute force MD	3
	CHARMM22/CMAP	1500	360	brute force MD	4
	CHARMM22/CMAP	7 runs of 300-500	360	brute force MD	4
	CHARMM22/CMAP	220	390	brute force MD	4
	AMBER99SB	550	360	brute force MD	4
	CHARMM22/CMAP	500 (larger box)	360	brute force MD	4
	CHARMM22/CMAP	500 (larger box)	400	brute force MD	4
	CHARMM36	600 (larger box)	360	brute force MD	4
	CHARMM22/CMAP	10	300	R_g biasing MD	4
	OPEPv4	100	300, 325, 350	brute force MD	5
	OPEPv4	230	260-582	REMD	5
1EFC ($\mathcal{M}$ EF-Tu)	CHARMM22/CMAP	450	360	brute force MD	4
1OB2 ($\mathcal{M}$ G-domain)	CHARMM22/CMAP	600	360	brute force MD	4
	CHARMM22/CMAP	10	300	R_g biasing MD	4
1SKQ ($\mathcal{H}$ G-domain)	CHARMM22/CMAP	600	300	brute force MD	3
	CHARMM22/CMAP	250	330, 360	brute force MD	3
	CHARMM22/CMAP	50	310-370 by 10	brute force MD	3
	CHARMM22/CMAP	1000	360	brute force MD	4
	CHARMM22/CMAP	220	390	brute force MD	4
	CHARMM22/CMAP	10	300	R_g biasing MD	4
	OPEPv4	100	300, 325, 350	brute force MD	5
	OPEPv4	230	260-582	REMD	5
1SKQ ($\mathcal{H}$ EF-Tu)	CHARMM22/CMAP	200	360	brute force MD	4
1GV1 ($\mathcal{M}$ tetramer)	CHARMM22/CMAP	200	300, 360	brute force MD	6
1GV1 ($\mathcal{M}$ dimers)	Martini v2.1	30 (per window)	300	umbrella sampling	6
1GV1 ($\mathcal{M}$ monomer)	CHARMM22/CMAP	200	300, 360	brute force MD	6
4CL3 ($\mathcal{T}$ tetramer)	CHARMM22/CMAP	200	300, 360	brute force MD	6
4CL3 ($\mathcal{T}$ dimers)	Martini v2.1	30 (per window)	300	umbrella sampling	6
4CL3 ($\mathcal{T}$ monomer)	CHARMM22/CMAP	200	300, 360	brute force MD	6

Chapter 3

How conformational flexibility stabilizes the hyperthermophilic elongation factor G-domain.[1]

Summary

We first focus on two homologous, monomeric, globular proteins and try to understand the correlation between thermal stability and flexibility. We consider the dynamics of the systems, primarily, at 300 K since the main concept under question is the alleged rigidity of thermophilic proteins at ambient conditions. We do, however, also examine the impact of higher temperatures in the flexibility of the folded state. Anticipating our results, at the atomistic level, while the magnitude of fluctuations is comparable for the two species, the distribution of flexible and rigid stretches of amino-acids is more regular in the hyperthermophilic protein causing a cage-like correlation of amplitudes along the sequence. This caging effect is suggested to favor stability at high T by confining mechanical excitation. Moreover, when looking at the molecular length-scale, the hyperthermophilic protein visits a higher number of conformational substates than the mesophilic homologue. The entropy associated with the occupation of the different substates and the thermal resilience of the protein's intrinsic compressibility provide a qualitative insight on the thermal stability of the thermophilic protein as compared to its mesophilic homologue. In short, in the microsecond timescale the hyperthermophilic protein shows comparable or even enhanced flexibility as compared to its mesophilic counterpart depending on the length-scale.

[1] M. Kalimeri, O. Rahaman, S. Melchionna and F. Sterpone (2013). "How conformational flexibility stabilizes the hyperthermophilic elongation factor G-domain". *J. Phys. Chem. B* 117.44, pp. 13775-13785

3.1 Prologue

System description. The systems under study are the G-domains of elongation factor -thermo unstable (EF-Tu) and -1α (EF-1α). Elongation factors play an important role in the biosynthesis process. Archaeal and eukaryal EFs are collectively designated as EF-1α because they are more similar to each other (sequence identities in the range of 50-60%) than to their bacterial analogues, called EF-Tu. The mesophilic protein studied here (PDB code 1EFC [172]) belongs to the bacterium *Escherichia coli* while the hyperthermophilic one (PDB code 1SKQ [173]) belongs to the archaeon *Sulfolobus solfataricus*. The G-domain is the catalytic core of the, otherwise, heterotrimeric EF and sets a "basic" level of thermostability for the whole protein [174] (see Fig. 3.1*A*). When isolated it is still able to perform GTPase activity and vary its catalytic power upon ribosome binding. The domains from the two homologues studied here share high structural homology (see Fig. 3.1*B*) and 34% of sequence identity. They are a good study-case mostly for three reasons; their thermal stabilities are separated by a large gap – optimal substrate binding activity at 308 K for the mesophilic versus 357 K for the hyperthermophilic – their fold contains both α and β structures and the hyperthermophilic is enriched in charged amino acids as commonly observed in thermophilic proteins.

Simulation setup. The G-domain corresponds to the N-terminal part of the protein, and in our simulations the mesophilic homologue was cut at the level of the residues T8-E203 while the hyperthermophilic G-domain encompasses the stretch of residues K4-V229. In the remainder of the text the residue numbering refers to our simulated systems and a shift of 7 and 3 residues is needed in order to match the numbering in the 1EFC and 1SKQ crystallographic structures, respectively (see Fig. 3.1). Molecular Dynamics simulations (MD) were performed using the CHARMM22/CMAP Force Field for proteins [134] and TIP3P-CHARMM model for water. The mesophilic domain (196 amino acids), denoted from now on by $\mathcal{M}$, was solvated with 7440 water molecules and the hyperthermophilic one (226 amino acids), denoted from now on by $\mathcal{H}$, with 10673. Counter-ions were added to neutralize the systems. Details of the systems preparation can be found in Ref. [175, 176]. The systems were simulated in the temperature range $T = 300 - 360$ K with a variable simulation length depending on T. To sample the long timescale behavior in the folded state at ambient condition ($T = 300$ K), the proteins were simulated for 0.6 μs. At the higher temperatures of 330 K and 360 K the simulations were carried out for about 250 ns. In order to study the temperature dependence of the protein compressibility, a set of independent trajectories of about 50 ns each were produced at intermediate temperatures separated from one another by 10 K

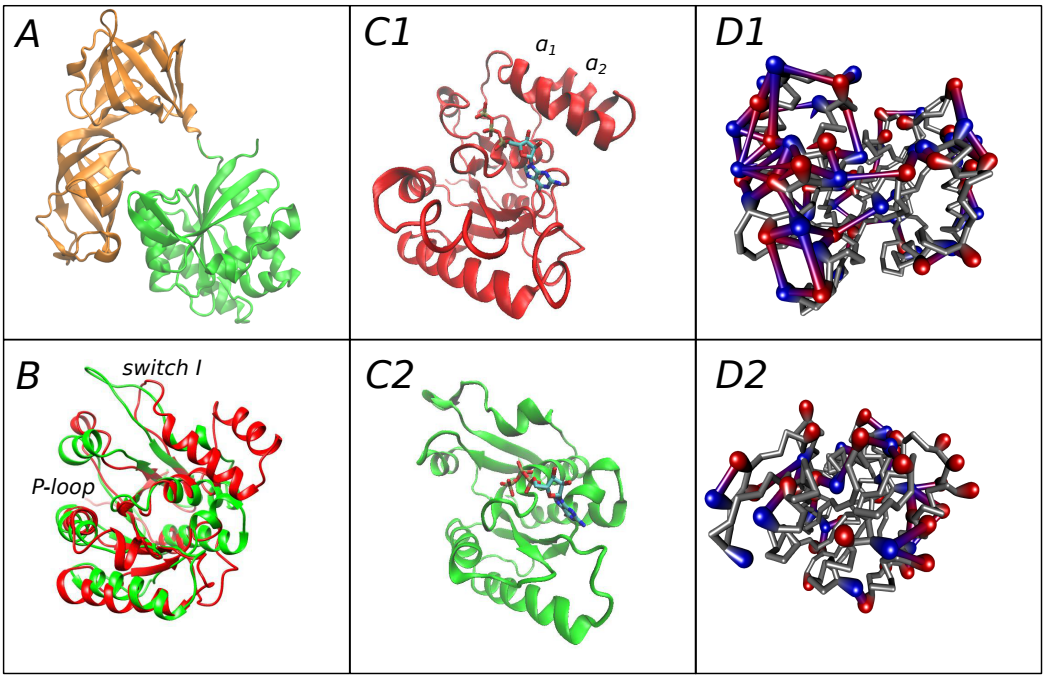

Figure 3.1. Overview of the two homologous G-domains. (A) Elongation factor Tu from *E. coli*. The G-domain of the protein is colored in green. (B) Overlap of EF-Tu (mesophilic) and EF-1α (hyperthermophilic) G-domains. Panels C1 and C2 show the hyperthermophilic ($\mathcal{H}$, red) and mesophilic ($\mathcal{M}$, green) G-domains respectively bound to GTP. Panels D1 and D2 depict the time-averaged salt-bridge networks for $\mathcal{H}$ and the $\mathcal{M}$, respectively. The positively charged residues are colored in red while the negatively charged ones in blue. The ion-pairs are represented as a colored bond with thickness proportional to the probability of being formed. The rendering of these panels was done with the Hyperball software[†].

and ranging from 310 to 370 K.

All simulations were performed using the NAMD software package [178, 179]. The equations of motion were integrated using a timestep of 2 fs, with all bonds treated as flexible except for those involving hydrogen atoms which were kept rigid. Temperature and pressure were kept constant using the Langevin thermostat (with friction coefficient $\gamma = 5\,\mathrm{ps}^{-1}$) and barostat (with oscillation period of $\tau_P = 100\,\mathrm{fs}$), respectively. Electrostatics in a periodic simulation box was solved via the Ewald summation method and handled by the PME algorithm with a grid spacing of 1 Å. The production phases were, each, preceded by a nanosecond-long equilibration phase. Volumetric properties were calculated using the program *trjVoronoi* [180, 181]. As will be seen in Chapter 4, where we deal with the kinetic stability of the

[†]Special thanks to Matthieu Chavent for realizing the figures with Hyperball [177].

two systems at high temperature, the mesophilic protein above 360 K and within a couple of hundreds of nanoseconds starts to unfold. Therefore, as far as the present discussion is concerned, the calculations of the $RMSF$ and the volumetric properties at high temperatures were restrained to the part of the trajectory where the mesophilic protein maintained a folded structure.

3.2 Results and discussion

3.2.1 Local fluctuations

As a preliminary step, we verify that the two homologues maintain their fold stable at 300 K and within the explored timescales. Figure 3.2 shows the root mean square deviation (RMSD) for the two systems in the course of the 600 ns simulation.

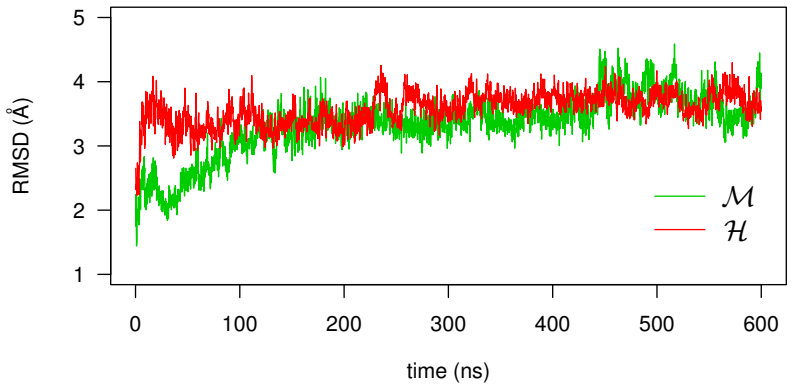

Figure 3.2. $RMSD$ of $\mathcal{M}$ (green) and $\mathcal{H}$ (red) for two 600-nanosecond simulations at ambient temperature (T=300 K) depicting the stable dynamics of the two systems.

We then monitor the atomistic motion occurring on the sub-nanosecond timescale by analyzing the root mean square fluctuation (RMSF) of the backbone C_α atoms. The fluctuations are evaluated in a 350 ps simulation window, ensuring the unimodal distribution of the atomic positions [182]. The average $RMSF$ at $T = 300\,K$ is reported in Fig. 3.3A along with its min/max variability (dotted bars) evaluated by block analysis.

The magnitude of the larger fluctuations as well as the average value along the sequence are comparable for both proteins. However, a striking difference emerges when the autocorrelation of the $RMSF$ values is taken along the protein sequence (see Fig. 3.3B). Namely the $\mathcal{H}$ protein shows a remarkable anti-correlating behavior

in the $RMSF$ between groups of neighboring residues with a characteristic sequence length ξ of about 10 residues. This behavior persists at higher temperatures (see Fig. 3.8 in the Appendix of this Chapter) and, borrowing a concept from the theory of liquids, features a caging effect along the sequence. For the $\mathcal{H}$ protein, flexible and rigid parts of the sequence alternate more regularly than in the mesophilic homologue. This finding suggests that a regular distribution of rigid/flexible fragments can possibly suppress the propagation of mechanical stress along the protein matrix, thus preventing progressive unfolding at high temperature. If verified for other pairs of homologous proteins, this feature could inspire new procedures to design thermostable proteins by tuning the local rigidity/flexibility pattern.

The different correlation profiles of the atomistic flexibility of the two proteins should be traced back to the structural differences distinguishing the two homologues. A sketch of the secondary structure for the two systems is drawn in Fig. 3.3*A*, top panel. From a quantitative point of view, the number of amino acids belonging to flexible motifs, i.e. turn or coil, are more frequent in $\mathcal{H}$ (44%) than in $\mathcal{M}$ (25%). Moreover, they are more uniformly distributed along the sequence (see Appendix of this Chapter and Fig. 3.9 therein). On closer inspection, the pairs of residues (i, i+15) in $\mathcal{H}$ that mostly contribute to the anticorrelation of $RMSF$ (Fig. 3.3*B*) span uniformly the first part of the sequence (from residue 1 to residue 120, see Fig. 3.4). For all those pairs but one, when one member belongs to a well defined secondary structure (α-helix or β-strand) its partner belongs to either a turn or a coil. Only exception is the stretch K36-K43 that belongs to a short α-helix but exhibits higher $RMSF$ values. This stretch is part of a key region of the $\mathcal{H}$ protein formed by two helices (α^1 [E32-L45] and α^2 [E48-E63]) that although preserves well its secondary structure, shows high mobility and intermittent unwinding of its terminal 3 residues. This is due to its large concentration of charged residues that results in frequent partner exchange of the ion-pairings. Typical example is the E40 that exchanges partners between K36 and K44 in nanosecond timescales. On the contrary, in the $\mathcal{M}$ protein the region G33-A45 shows large fluctuations and gives the small positive correlation at residue-separation 15-17 with other highly fluctuating amino-acids (see Fig. 3.3*B* and Fig. 3.4). This region, as we will see in the next Chapter, constitutes the protein's weak spot where the high-temperature unfolding begins. Single-point or more extended mutations can be designed to confirm the stabilizing effect of a regular alternation of rigid and flexible fragments in this region of the protein. It would also be intriguing to examine via pump-probe experiments whether the local caging effect that we report here has an impact on the energy transfer processes along the protein backbone; this could be achieved, for example, by monitoring the vibration of labeled, photo-excited carbonyl C=O bonds [183] or during progressive thermal unfolding monitored via multimode 2D

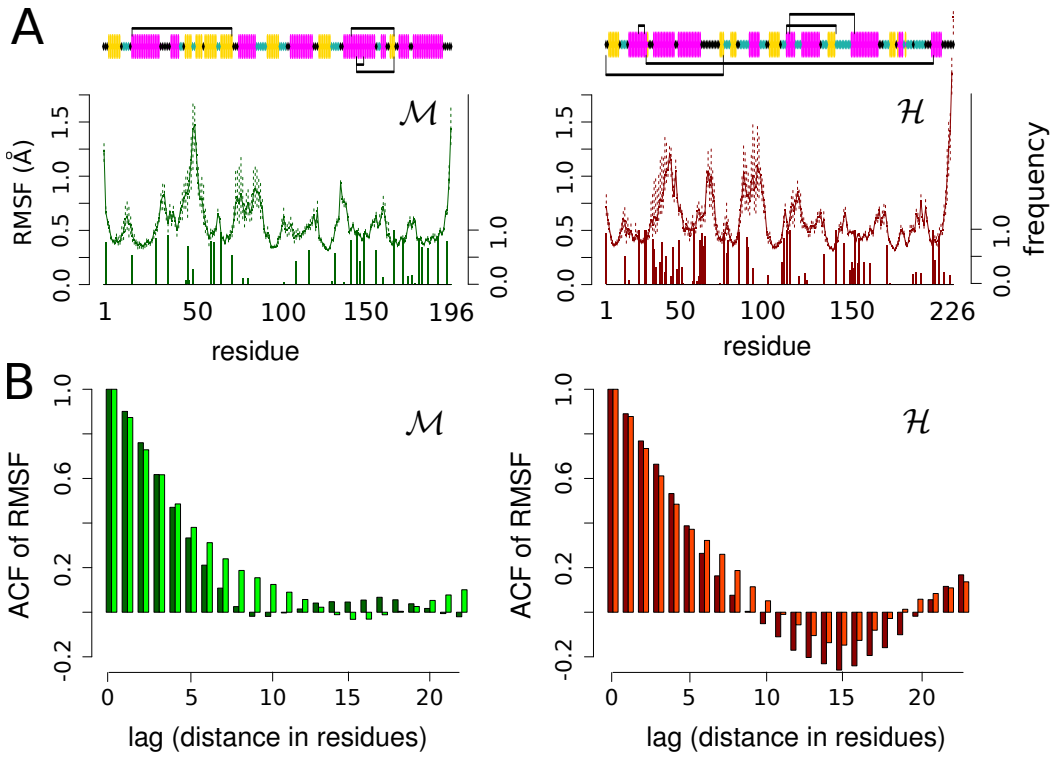

Figure 3.3. Atomistic fluctuations: mesophilic protein (left column) and hyperthermophilic (right column). (A) The top of the figure displays the secondary structure of each protein as well as the most stable salt-bridges in the form of connecting black lines. In the secondary structure bar β-strands and α-helical secondary structures are represented as yellow and magenta bars while turn and coil regions are represented as thinner cyan and black lines, respectively. Right below, the continuous lines show the mean atomistic C_α $RMSF$ for $\mathcal{M}$ and $\mathcal{H}$ at $T = 300\,K$ while the dotted lines correspond to its min/max variation (axis on the left). The bars below the $RMSF$ indicate the relative frequency of formation of a salt-bridge (axis on the right). (B) Autocorrelation of the atomistic $RMSF$ as a function of the residue-lag along the sequence. The darker colors correspond to the mean $RMSF$ autocorrelation while the lighter colors to the min/max variation.

IR spectroscopy [184].

3.2.2 Electrostatic interactions

The local flexibility of the protein matrix was probed by investigating the network of electrostatic interactions, namely hydrogen-bonds (HB) and ion-pairs (IP). At room temperature, the number of HB per unit of volume is comparable for the two proteins, being $\rho_{HB} = \frac{n_{HB}}{V_p} = 0.022\,\text{Å}^{-3}$ with fluctuations of the order of 10%.

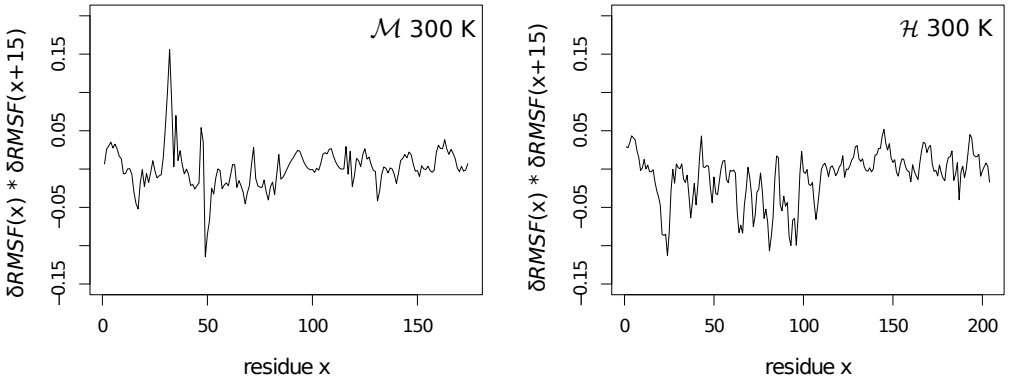

Figure 3.4. Product of the fluctuation of $RMSF$ with itself at positions x and $x + 15$, for both $\mathcal{M}$ and $\mathcal{H}$ at $T = 300\,\mathrm{K}$. The pairs of residues in $\mathcal{H}$ (right) that mostly contribute to the anticorrelation of $RMSF$ (see Fig. 3.3 and 3.8) span uniformly the first part of the sequence (from residue 1 to residue 120). On the contrary, the observed small positive correlation for $\mathcal{M}$ at distances 15-17 residues has contribution only from the region G33-A45 (left).

In n_{HB} we considered both intraprotein and protein-solvent HBs, and the protein volume V_p was evaluated by Voronoi tessellation [185]. However, as a consequence of its chemical composition, the $\mathcal{H}$ protein is cross-linked by a higher number of instantaneous IPs (∼16) than the $\mathcal{M}$ species (∼7). Figure 3.3*A* illustrates the frequency of IP formation for each residue in the form of bars underneath the $RMSF$ profile. There is a clear correlation between the low-RMSF fragments of the sequence and a high probability for IP formation – with the exception of the K36-K43 region of $\mathcal{H}$ – suggesting that the IPs act as structural clamps. The high density of IPs localized at the level of the $\mathcal{H}$ switch I region (L29-P73), as mentioned above, although resulting in a higher mobility of this region, confers a long term resistance to the secondary structure of this key stretch of amino-acids. A few IPs are long-lived and particularly stable, as schematically shown in Fig. 3.3*A* as black straight lines above and below the secondary structure inset.

It is worth noting that for $\mathcal{H}$, the interaction E77-K1 keeps the N-terminus closely packed to the body of the protein for the whole range of temperatures, while the stable interactions D217-K79 and D217-K80 ensure the effective packing of the C-terminus. On the contrary, for the $\mathcal{M}$ protein, the interactions D63-K2 and E194-R67 that link the terminals to the protein body at room temperature, gradually weaken with increasing temperature. The stable anchorage of C- and N-terminals is a recognized structural peculiarity of thermophilic proteins [71], and

in the context of our investigation, the link of the N-terminal to the switch I region is individuated as the key stabilizing interaction for the G-domain of the EF-Tu from the thermophilic *Bacillus stearothermophilus* bacterium [186].

In the $\mathcal{H}$ protein the larger number of charged amino acids favors a variety of IP patterns at the protein surface. At any given instant, the maximum number of residues involved in a IP network, is of $4-5$ for both proteins; but when the IP formation is averaged over time, the surface of the $\mathcal{H}$ protein is covered by extended IP networks while at the $\mathcal{M}$ surface small IP clusters remain disconnected (see Fig. 3.1 , panels D1 and D2). In other words, the side-chain flexibility of the ionic amino-acids – that we have verified to be comparable among the two proteins by computing their conformational entropy as in Ref. [187] – favors IP partner exchange in the $\mathcal{H}$ protein. Conversely, in the $\mathcal{M}$ protein the breakage of a IP is not compensated by the formation of a new ion-pairing. It was previously suggested, by studying a coil-coiled system [187], that the fluctuations of the ion-pair patterns at the interfaces of three aggregated helices help to accommodate the high-temperature entropy and therefore maintain the system stable.

In conclusion, we observe that ion-pairs contribute to the enhanced stability of the $\mathcal{H}$ protein, not only via short-range permanent links, but also forming, as a result of side-chain flexibility, dynamic extended network of electrostatic interactions.

3.2.3 Compressibility

The number and spatial distribution of charged groups are generally correlated to the heat capacity of unfolding [188, 189] and protein compressibility [190]. Fluctuations of the protein packing are therefore a natural candidate to monitor protein flexibility and its contribution to stability [191, 192]. For the homologous G-domains at room temperature we obtain a comparable and quite high apparent compressibility, as estimated by the method introduced in Ref. [193], β_a=8.7−9.2(±2) 10^{-5} Mpa^{-1}, that correlates to the presence of large water-filled cavities in the protein structures. The weighted fluctuations of the protein volume, or intrinsic protein compressibility, $\beta_T = \frac{\langle \delta V_p^2 \rangle}{k_B T \langle V_p \rangle}$, slightly increase with temperature for both proteins, as reported in Fig. 3.5, but for the $\mathcal{H}$ protein we observe a milder T dependence than for the $\mathcal{M}$ protein. A link between protein compressibility and the enthalpy of unfolding has been previously pointed out [190, 192] such that proteins with a lower compressibility are enthalpically more stable. Our finding suggests that while at ambient temperature the two proteins have a similar enthalpy of unfolding, as temperature increases ΔH increases faster for $\mathcal{H}$ than for $\mathcal{M}$. The $\mathcal{H}$ protein is characterized by a better atomic packing at all temperatures [175], with the volume per atom being about 2% smaller than in the $\mathcal{M}$ protein

and a thermal expansion of $\alpha \sim 4{\times}10^{-3}\text{Å}^3\,\text{K}^{-1}$.

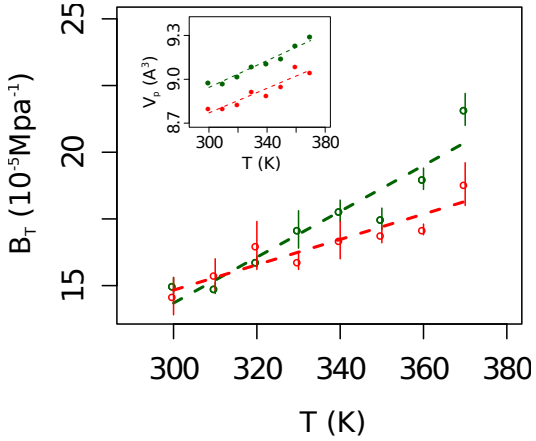

Figure 3.5. Intrinsic protein compressibility β_T versus temperature for the $\mathcal{M}$ (green) and the $\mathcal{H}$ (red) protein. Inset: specific volume per atom versus temperature.

3.2.4 Conformational states

The long-time fluctuations of the two protein conformations are now compared. A clustering analysis performed on trajectories of equal duration (0.6 μs) returns a clear result: the conformational landscape of the $\mathcal{H}$ protein is characterized by a larger number of substates than that of the $\mathcal{M}$ counterpart (Fig. 3.6*A* and 3.6*B*). This finding is robust for different cutoffs, different tested algorithms (see Cluster analysis in Chapter 2) and for several order parameters used for the clustering ($RMSD$, fraction of native contacts Q and fraction of native torsion angles n_t). We have also verified that the number of clusters saturates within a few hundreds of nanoseconds according to a simple exponential evolution, $N = N_\infty(1 - e^{-\frac{t}{\tau}})$. From the data of Fig. 3.6*A*, the timescale τ is in the range of 170-350 ns. The cutoff values used for the plotted data are 2.5 Å for $RMSD$, 0.20 for Q and 0.22 for n_t.

The observed differences in the conformational landscapes mark the structural properties of the two homologues. The $\mathcal{H}$ protein has a larger number of turn and coil residues whose fluctuations result in visiting a larger number of conformational states.

Figure 3.6*B* illustrates for the native contact parameter Q, a representative example of the clustering analysis. The clusters are depicted as interconnected nodes in the network. The size and color intensity are proportional to their occupancy. The edges represent the transitions from one substate to another and are weighted

with the respective transition frequencies computed along the trajectories. The conformational landscape of the $\mathcal{M}$ protein, projected on the network with the use of a force-based algorithm [194], is compact with the majority of substates organized around a main local basin. On the contrary, for the $\mathcal{H}$ protein, more basins of attraction are visible. The two types of organization can be distinguished by computing the transitivity coefficient C of the network, quantifying the tendency of the nodes to cluster together. It is defined as the probability that two nodes adjacent to a third are also connected to each other [168]. The results are reported in Table 3.1. For the two networks the coefficient C is found to be higher for $\mathcal{M}$ than for $\mathcal{H}$. We have also evaluated the Shannon entropy [169] for the obtained networks $\mathrm{H} = -\sum_i^N p_i \ln(p_i)$, where the summation runs over all the different clusters N and p_i is the relative occupation of the i_{th} cluster. The normalized quantity $\mathrm{H}/\mathrm{H}_{max}$, where H_{max} corresponds to a uniform distribution of occupancies, is reported in Table 3.1 for both systems and for the three order parameters. Systematically higher values are associated to the $\mathcal{H}$ protein indicating a more uniform distribution of substate occupancies as a consequence of the larger number of attractive basins.

Table 3.1. Transitivity coefficient of conformational networks and Shannon entropy of the node memberships

System	CV	Transitivity C	$\mathrm{H}/\mathrm{H}_{max}$
$\mathcal{M}$	$RMSD$	0.65	0.83
	Q	0.73	0.74
	n_t	0.79	0.79
$\mathcal{H}$	$RMSD$	0.62	0.91
	Q	0.52	0.85
	n_t	0.73	0.87

For the calculation of the transitivity coefficient the weights of the edges (i.e. frequencies of transition) were taken into account.

One point should be made concerning the small difference in the sequence length of the two proteins, 196 ($\mathcal{M}$) versus 226 ($\mathcal{H}$) residues. Although for a random polymer the possible number of clusters should grow exponentially with the length of the chain, in the compact fold this dependency is expected to be much weaker. However, in order to observe possible size effects, we performed additional validations. At first we excluded the last 3 highly fluctuating residues in the C-terminal of the $\mathcal{H}$ protein and compared with the original clustering of the $\mathcal{M}$ protein; next we clustered several equal-length residue-stretches for both systems. Both tests confirmed the main finding: the landscape of the $\mathcal{H}$ protein is characterized by a

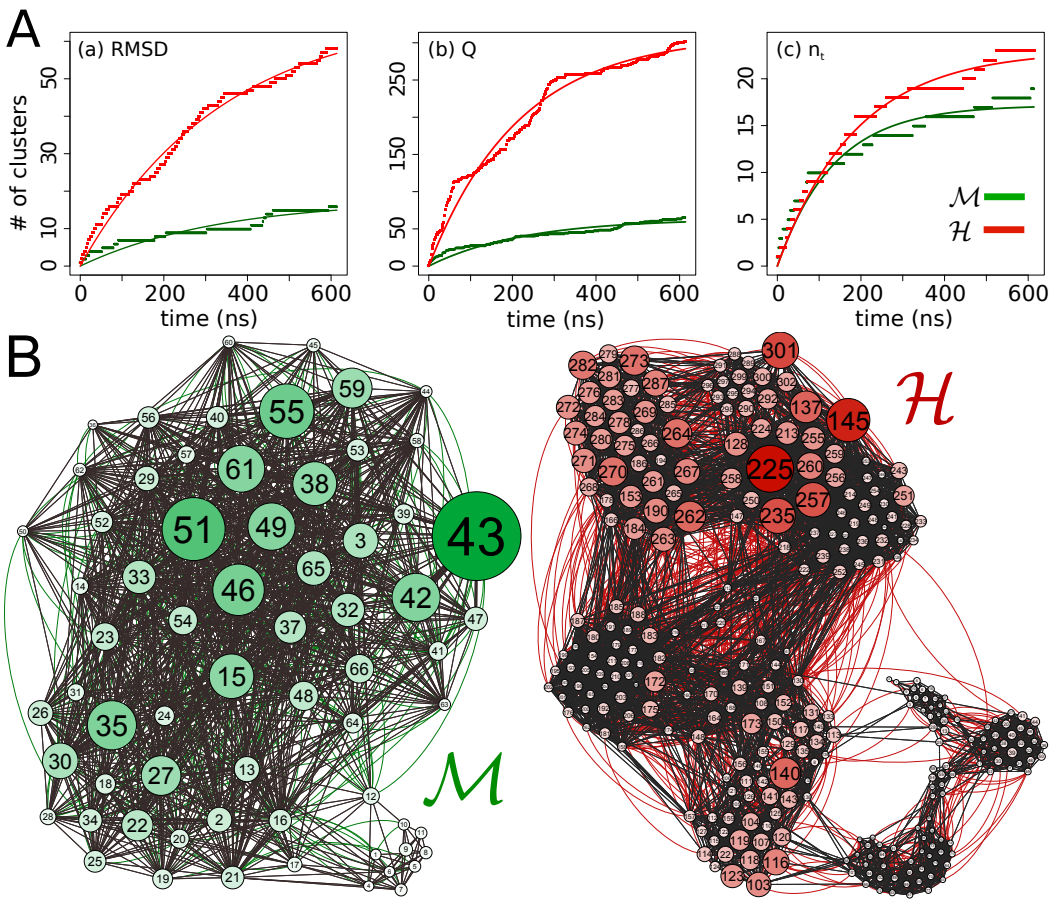

Figure 3.6. Conformational substates. (A) Number of clusters versus time (dotted lines) as obtained with the leader/follower algorithm at $T = 300\,K$. The clustering variables are (a) $RMSD$, (b) fraction of native contacts Q and (c) fraction of native torsion angles n. The solid lines correspond to the exponential fitting of the form $N = N_\infty(1 - e^{-t/\tau})$. (B) Network representations for the $\mathcal{M}$ and $\mathcal{H}$ proteins after clustering with the Q variable. The two networks are projected on 2D space using a force-based algorithm with its parameters set the same for both systems. The networks were drawn with the Gephi software [194].

larger number of conformational states.

As has been discussed in Chapter 1, from the thermodynamic point of view, thermophiles can achieve high thermal stability via three possible strategies or a combination thereof. In a simple two-state model the unfolding free energy as a function of temperature, $\Delta G(T)$, resembles an inverted parabola with its maximal value ($\Delta G_{\max}$) at the temperature of maximun stability, $T_{\max}$ [98]. The zero of the parabola at high temperature corresponds to the melting temperature T_{m} at which unfolded and folded states have equal probabilities. Thermophiles increase the value of T_{m} by either i) increasing the value of $\Delta G_{\max}$, ii) shifting $T_{\max}$ to higher

values or iii) reducing the curvature of $\Delta G(T)$ [97, 98]. It was found that a large class of thermophiles adopt the latter strategy and therefore manifest a smaller heat capacity of unfolding, ΔC_P, with respect to the mesophilic counterparts [98]. Several microscopic mechanisms might cause a smaller ΔC_P of unfolding. For example, a study of the Ribonuclease H protein proposed that thermophiles could preserve partial secondary structures in the unfolded state that shield hydrophobic groups from water [195]; on the other hand it was shown that the higher content of polar groups generally found in thermophiles suffices to explain the observed lower ΔC_P [188]. Finally, and more related to the analysis presented above, on the basis of NMR experiments probing the backbone bonds orientation [196, 197] it was proposed that a broadening of the stability curve could be related to high conformational fluctuations in the folded state, as discussed in Ref. [83, 84, 198]. The larger number of conformational states along with its chemical composition suggest that ΔC_P of unfolding for the $\mathcal{H}$ G-domain should be smaller than that for the $\mathcal{M}$ protein. Unfortunately, to the best of our knowledge, a systematic, experimental comparison of calorimetric data for the isolated G-domains is still lacking. The few available calorimetric studies [199–201] focused on the role of inter-domain interactions, flexibility and ligand binding on the EF proteins' thermal stability but ΔC_P values are not available. For the entire hyperthermophilic EF-1α protein the presence of secondary structures was observed in the denatured ensemble [201] but it is not clear if those are located in the G-domain or elsewhere. It is also worth mentioning that, from the experimental point of view, an accurate estimate of ΔC_P is possibly compromised by the irreversible nature of the thermal unfolding process of the EFs proteins [200, 201]. Finally, even if experimental data are missing, the thermodynamic mechanism suggested here is in accord with the form of the stability curves for the two homologues, as extracted with the use of REMD and the coarse-grained model OPEP in Chapter 5.

3.2.5 Diffusion in the folded landscape

The network analysis presented above is a powerful comparative tool in order to gain information on the conformational fluctuations and the free energy landscape for homologous proteins. However, it is important to verify in a more robust way if the landscape is representative of the kinetic separation between substates [202]. To this end, for each observable and protein, the original clustering is coarse grained via an iterative procedure based on a Markov clustering algorithm [170, 171](see Chapter 2, Methodology) which separates the states depending on their effective kinetic barriers. The clusters with fast interconversions are merged and represented as a unique node, as shown in Fig. 3.7. The obtained results confirm for all the CVs that the landscape of $\mathcal{M}$ constitutes a unique main basin while at the same

level of resolution the $\mathcal{H}$ protein has many kinetically separated states.

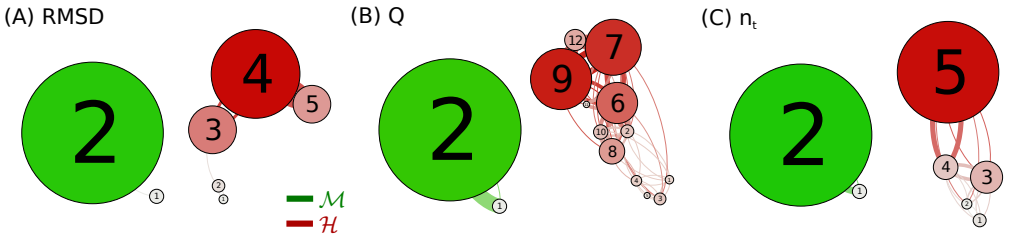

Figure 3.7. Coarse-grained network representations of the folded substates. The above networks were generated using the Markov Clustering Algorithm (MCL)[50] on the respective collective variable clustering (see Chapter 2, Methodology). Thus they represent a coarse grained description of the initial clustering where the more kinetically relevant substates have been merged into one node. In particular, depicted in (C) are the coarse grained representations of the two networks in Fig. 3.6*B*. The single control parameter for the MCL algorithm, namely the granularity, was set equal to 1.3, 2.0 and 1.7 for $RMSD$, Q and n_t respectively. These values correspond to the lowest granularity that results in two nodes for $\mathcal{M}$. Note that for each variable and for the same granularity, the hyperthermophilic system is always decomposed in more substates.

The number of substates and the distribution of the separating kinetic barriers impact the diffusivity in the free energy landscape pertaining to the folded state. The motion with respect to a given collective variable is here associated to a diffusion coefficient D [161] (see Chapter 2, Methodology). Within the harmonic approximation, D is given by the fluctuations of the CV divided by its characteristic decorrelation time, $D = \langle \delta X^2 \rangle / \tau_{corr}$. For the Q and n_t variables this approximation is valid and the time-correlation function of their fluctuations decays exponentially, $R(\tau) = \frac{\langle \delta X(t) \cdot \delta X(t+\tau) \rangle}{\langle \delta X^2 \rangle} \simeq e^{-t/\tau_{corr}}$, where X indicates either n_t or Q (see Fig. 3.10 in Appendix). While the fluctuations of the CVs are rather comparable between the two systems, the decorrelation times are systematically longer for the $\mathcal{H}$ G-domain mirroring a higher internal protein friction (see Table 3.2). For the $RMSD$ variable the fluctuations are largely anharmonic, and it is somewhat arbitrary to individuate a time-window of a stationary behavior. Thus, an external harmonic bias is applied to restrain the $RMSD$ around the respective, ambient-temperature averages for $\mathcal{M}$ and $\mathcal{H}$; the resulting τ_{corr} and D reproduce the trend observed for n_t and Q CVs.

The intramolecular diffusion over a conformational landscape is a key parameter in the theory of protein folding [162]. Experiments based on atomic force microscopy [203] and Föster resonance energy transfer [204] have provided estimates for the coefficient D for unfolded and transition state configurations along

the end-to-end distance reaction coordinate. By using atomic force spectroscopy it would be of great interest to observe if thermophilic proteins exhibit a slower motion along the pulling direction than their mesophilic homologues.

Table 3.2. Diffusion constant for the protein conformational motion

System	CV	$\langle \delta X^2 \rangle$	τ_{corr} (ns)	D
$\mathcal{M}$	$RMSD^*$	5.0×10^{-3}	0.6	8.4×10^{-3}
	Q	25.5×10^{-5}	1.8	143×10^{-6}
	n_t	5.0×10^{-5}	3.2	15.6×10^{-6}
$\mathcal{H}$	$RMSD^*$	4.0×10^{-3}	7.1	6.0×10^{-4}
	Q	19.8×10^{-5}	4.3	46×10^{-6}
	n_t	10.0×10^{-5}	23.1	3.9×10^{-6}

* The $RMSD$ fluctuations were computed by applying an harmonic restraint (force constant k_{RMSD}=100 kcal/mol·Å^2) around the values of $RMSD_0$=3.3 Å and $RMSD_0$=3.7 Å for the $\mathcal{M}$ and $\mathcal{H}$ proteins, respectively.

3.3 Concluding remarks

By using extensive simulations and several different indicators we questioned the common view according to which thermophilic proteins are more rigid than their mesophilic homologues at ambient conditions. In this view, an enhanced rigidity of thermophiles confers resistance to thermal stress and is the cause for the lack of activity of this class of proteins at ambient temperature.

When focusing on the behavior of the folded state at ambient condition, we noted that the hyperthermophilic protein shows comparable or even enhanced flexibility with respect to the mesophilic protein, depending on the time- and length-scale considered. Average atomistic fluctuations were comparable in magnitude among the two proteins but rigid and flexible stretches of amino acids are differently partitioned over the matrix, with the $\mathcal{H}$ species being characterized by an alternation of atomistic flexibility, recalling a *caging* effect along the sequence. This alternation is an effect of the specific structural motives α^1 and α^2 that differentiate the $\mathcal{H}$ G-domain and, as discussed above, contribute to stabilizing this region. The *caging* attitude of the investigated thermophile suggests a new strategy to enhance the stability of proteins by tuning the extension and distribution of flexible/rigid parts along the primary sequence.

When analyzing the global protein dynamics, we represented the conformational landscapes of the two proteins as a network of substates. Such an approach catered a direct evidence that the hyperthermophilic protein is characterized by

a larger number of conformational substates kinetically separated; on the contrary the $\mathcal{M}$ protein fluctuates in an isolated conformational basin. The observed enhanced conformational flexibility of the $\mathcal{H}$ protein should cause a smaller heat capacity of unfolding [197] with respect to the $\mathcal{M}$ species and therefore a reduced curvature of the stability curve along with a higher melting temperature. Unfortunately, up to now the thermodynamic mechanism [98] underlying the higher stability of the *Solfolobus solfataricus* G-domain with respect to that from *E. coli* has not been experimentally probed. Notably, however, our expectation is verified by the computational thermodynamic approach of Chapter 5.

Appendix

Block entropy for secondary structure

In order to quantify the uniformity of the distribution of turns (T) and coils (C) along the sequence of the two chains, we used the so-called block entropies [169, 205], a generalization of the Shannon entropy to blocks with m symbols.

We first mapped the crystallographic secondary sequence (as calculated by the algorithm STRIDE) in a sequence of $0s$ and $1s$ in the following way; if the residue is found to belong to either a T or a C stretch it is mapped to 1, otherwise to 0. An example is given below,

$$\text{CCCEEEEEEECTTTT} \rightarrow 111000000011111$$

Then, by moving along the binary sequence, we consider the (overlapping) blocks of length m and calculate the following Shannon-like entropy.

$$S_m = -\frac{1}{K}\sum_{j=1}^{K} \ln(p(m_i)), \tag{3.1}$$

where j runs from 1 to $K = N - m + 1$, the total number of blocks of length m along the sequence, and i runs over all unique blocks of length m. Thus $p(m_i)$ is the relative frequency of the unique $i-$th block. This quantity is maximized when all possible blocks of length m appear with equal probability. The values of $S_m/S_{m,max}$ w.r.t. m are shown in Figure 3.9 (A). In Fig. 3.9 (B) we followed the same procedure as described above but where *only* the residues belonging to T stretches were mapped to 1; in Fig. 3.9 (C) *only* the residues that belong to C stretches were mapped to 1. In all three cases, for the smaller values of m, the quantity $S_m/S_{m,max}$ is slightly larger for the hyperthermophilic sequence than for the mesophilic one. This reflects a more uniform distribution of turns and coils along the sequence of the hyperthermophilic system, which notably, is more pronounced when only turn residues are taken under consideration (Figure 3.9 (B)).

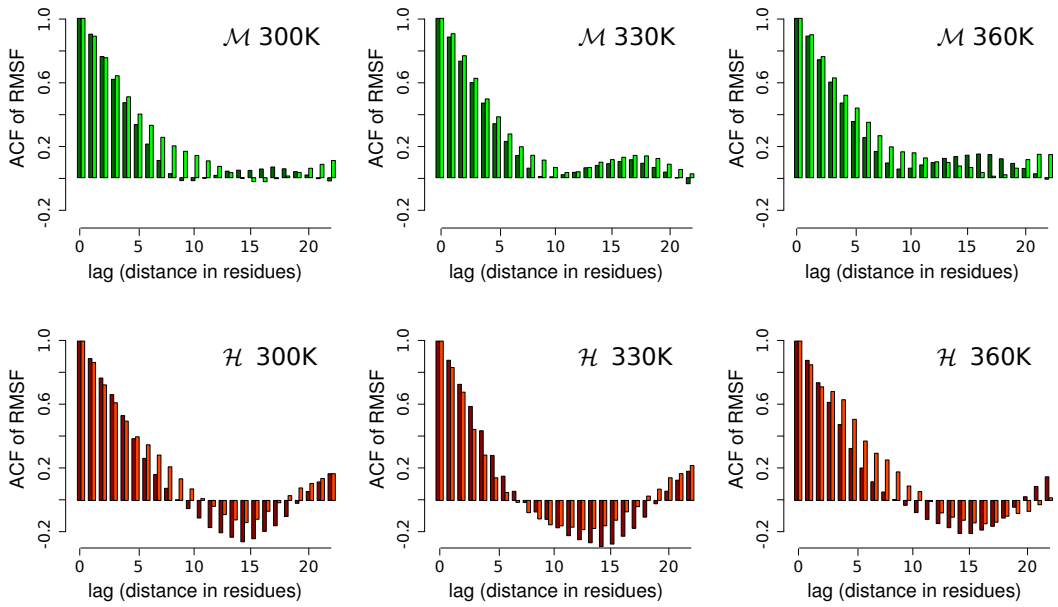

Figure 3.8. Autocorrelation of the atomistic $RMSF$ as a function of the distance between residues in the sequence. This figure is an extension of Figure 2 of the main text, depicting the behavior for the ACF of $RMSF$ for higher temperatures (including $T = 300\,\mathrm{K}$). As in the main text, the darker colors correspond to the ACF of its mean $RMSF$ while the lighter colors to the ACF of the min/max variation. Data for the mesophilic system are colored with green color (above) and those for the hyperthermophilic system with red color (below)

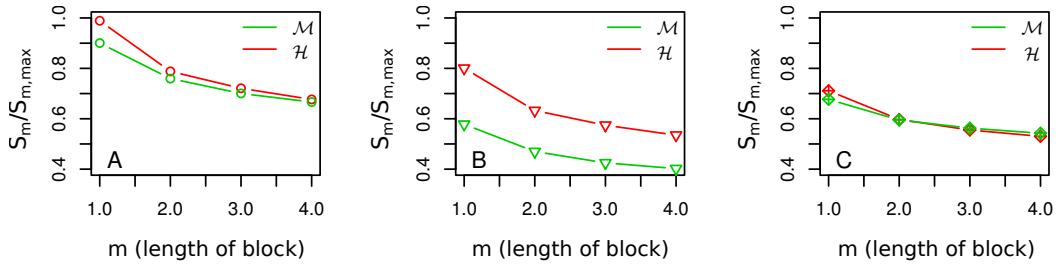

Figure 3.9. Normalized block entropies versus the length of block quantifying the uniformity of the distribution of turns (T) and coils (C) along the protein-chains of the two systems. (A) Both T and C residues have been taken into account, (B) only T residues are considered and (C) only C residues are considered.

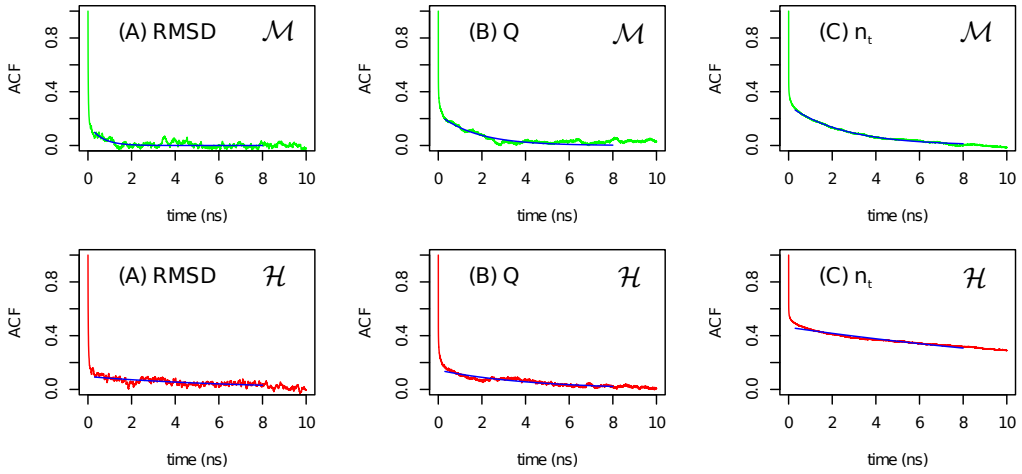

Figure 3.10. Normalized time correlation function, $R(\tau) = \frac{\langle \delta X(t) \cdot \delta X(t+\tau) \rangle}{\langle \delta X^2 \rangle}$, computed for the three CVs $RMSD$, Q and n_t. The exponential fit is represented as a solid blue line. The top panels refer to the $\mathcal{M}$ proteins and the bottom panels to the $\mathcal{H}$ proteins.

Chapter 4

Stability of elongation factor at high temperatures[1].

Summary

In this chapter, we extend the study of the two G-domains by probing their kinetic stability at high temperatures in the microsecond timescale. We first verify that at its high temperature regime (360 K) the hyperthermophilic G-domain is stable while its mesophilic homologue starts to unfold. The unfolding event is localized in a key region for the protein activity, referred to as switch I. This region is known to undergo an important conformational change during the GTPase activity. Therefore, a high flexibility of this stretch of amino acids, while makes the enzymatic turnover efficient at ambient conditions, it weakens the mesophilic protein at high temperature. The similar region in the hyperthermophile is overall stable. Different conformational changes are probably required for this protein's functionality at high temperatures. The effect of inter-domain interactions on the stability of the two proteins and the flexibility of the switch I region is also investigated by considering the dynamics of the whole heterotrimeric protein.

4.1 Results and discussion

4.1.1 G-domain of EF-Tu: Stability versus early step of unfolding

Here, by extending the initial 360 K simulations up to the microsecond scale we verify that the hyperthermophilic G-domain is more stable – in a kinetic sense –

[1] M. Kalimeri, O. Rahaman, S. Melchionna and F. Sterpone (2013). "How conformational flexibility stabilizes the hyperthermophilic elongation factor G-domain". *J. Phys. Chem. B* 117.44, pp. 13775-13785

in its high functional temperature regime ($T = 360\,K$) than the mesophilic homologue[2]. Within one microsecond the fold of the $\mathcal{H}$ protein remains close to its crystallographic structure and shows a steady root mean square displacement of its C_α atoms (C_α-RMSD) with an average around 3.8 Å (see Fig. 4.1*A*). At the same temperature, the mesophilic protein exhibits the early onset of unfolding: the time-evolution of the C_α-RMSD is first marked by a sequence of jumps until a further highly fluctuating drift toward larger C_α-RMSD values, ~ 10 Å. Fig. 4.1*B* reports the conformational "free energy" landscape sampled by the two proteins and projected on two CVs, the fraction of native contacts, Q, and the fraction of native torsion angles, n_t [206, 207].

The kinetic instability of the mesophilic protein is reproduced in 7 shorter extra runs of duration $300-500$ ns at 360 K as well as in a 220 ns long simulation at 390 K (see Fig. 4.2*A*). From those, the characteristic timescale of the early unfolding event is estimated to be $140-400$ ns. At 360 K a run using the AMBER99SB Force Field [135] is also in agreement with this timescale (see Fig. 4.2*C*).

The early steps of unfolding of $\mathcal{M}$ take place in the protein's "Achilles' heel": the stretch G33-A45. In our long simulation, this region unpacks at approximately 150 ns, rapidly followed by the disruption of the F39-D43 helix and a gradual unpacking of the residues A45-C74. Reaching 1 μs of simulation, the stretch G33-C74 loses its secondary structure motifs and expands in the form of a random-coil with the first moment of the protein's gyration tensor being 40% larger than at previous times (see the snapshot at the top of Fig. 4.1). The finally disrupted sequence includes a key region for the protein activity, the so-called switch I region (in our numbering G33-I55, see Fig. 4.3), that is known to undergo a large conformational change in bacterial EF-Tu proteins during GTPase activity [208–210]. As discussed in Chapter 3, the equivalent region in $\mathcal{H}$ comprises two helices (α^1 [E32-L45] and α^2 [E48-E63]) packed in a very different way and although their secondary structure is well preserved at high temperature, they are in fact – mainly as a rigid body – rather flexible. Figure 4.3 shows the crystal structures of the two homologues highlighting the location of this region. To the best of our knowledge, it is not yet elucidated whether in eukaryal and archaeal EF-1α the same region is subject to a conformational change during GTPase activity.

Mesophilic G-domain: Why does the unfolding stop? By extending the 360 K simulation of the mesophilic G-domain for another 0.5 μs we are unable to monitor further unfolding. Figure 4.4 shows the $RMSD$ of $\mathcal{M}$ for a total simulation time of 1.5 μs reaching a plateau after 1.2 μs. The unfolded switch I region – see the three-dimensional representation on top of the $RMSD$ in Fig. 4.1 – collapses

[2] The system setup and simulation protocol of Chapter 3 applies also here.

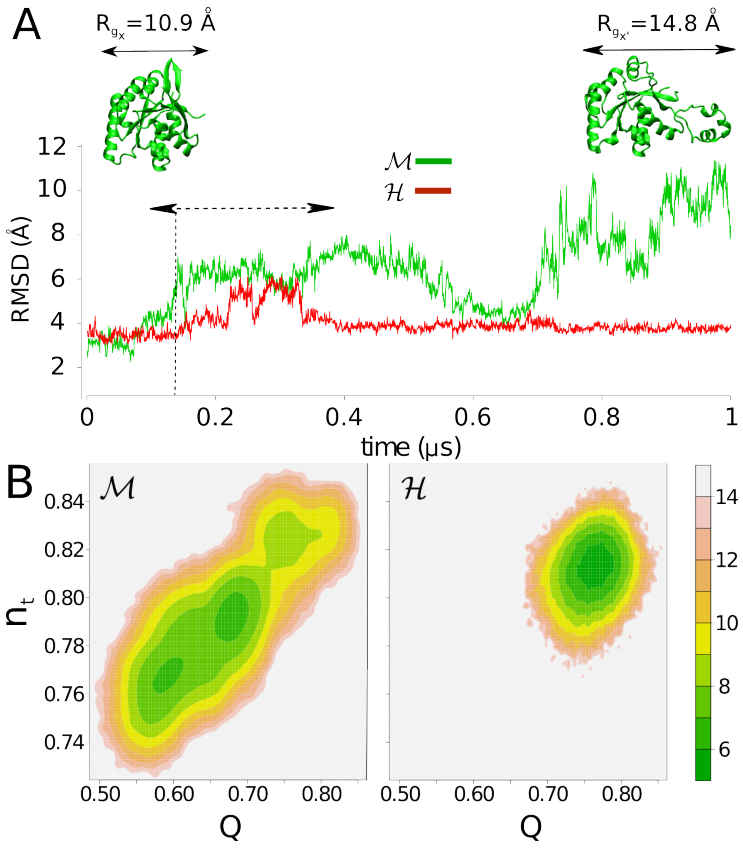

Figure 4.1. Microsecond simulations at $T = 360\,K$: (A) C_α-RMSD for $\mathcal{M}$ and $\mathcal{H}$ proteins. The horizontal dotted arrow indicates the time window over which the first $RMSD$ "jump" for $\mathcal{M}$ takes place in a collection of 7 independent simulations (see Fig. 4.2). Above the respective $RMSD$ values, two snapshots of the folded (left) and partially unfolded (right) $\mathcal{M}$ protein are shown with the indication of the square root of the first principal moment of their gyration tensor. (B) Dimensionless 2D free energy landscape at $T = 360\,K$ for the two systems with CVs the fraction of native torsion angles, n, against the fraction of native contacts, Q.

back on the protein fold. The packing, however, occurs in the opposite direction of its initial configuration while its secondary structure acquires the form of two small helices (see Fig. 4.5A).

We first examine whether the collapse is caused by the size of the box which, at this stage, restrains further unfolding. However, resolvating the extended conformation of the molecule in a much larger box containing 45307 water molecules (as compared to the initial box containing 7440 molecules) and simulating for another 0.5 μs at 360 K results in approximately the same misfolded configuration as before (see Fig. 4.5A and 4.6). Only a subsequent restart at the higher temperature of 400 K, after 0.5 μs, yields an unpacking of the misfolded stretch with a simulata-

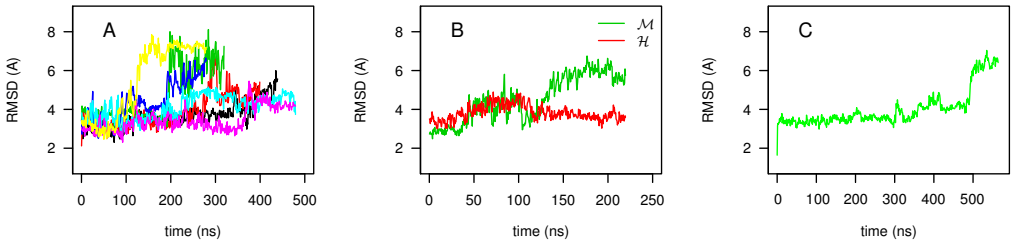

Figure 4.2. Mesophilic protein, early step of unfolding. (A) $RMSD$ of 7 independent trajectories at $T = 360\,\mathrm{K}$ for the $\mathcal{M}$ protein. The first sudden "up-drift" of the $RMSD$ occurs within the time window $140 - 500\,\mathrm{ns}$. (B) $RMSD$ of $\mathcal{M}$ and $\mathcal{H}$ at $T = 390\,K$ and (C) $RMSD$ of an independent simulation of $\mathcal{M}$ at $T = 360\,\mathrm{K}$ as obtained with the AMBER99sb Force Field [135].

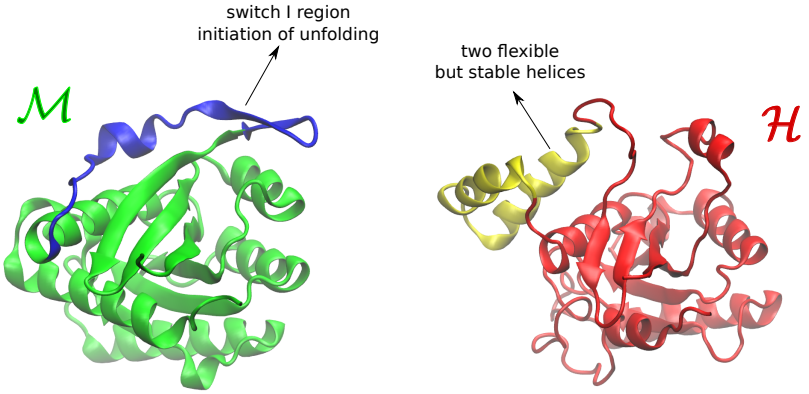

Figure 4.3. The three-dimensional structure of the two G-domains, mesophilic $\mathcal{M}$ and hyperthermophilic $\mathcal{H}$, highlighting the important for activity switch I region in blue and yellow color, respectively.

neous loss of its helical secondary structure without however proceeding to any further unfolding (see Fig. 4.6). Another attempt is to restart the simulation of the extended conformation in the large box using a very recent re-parametrization of the employed force field, the CHARMM36 [136], that corrects the small propensity for α-helix of the previous CHARMM22/CMAP version. This new $0.6\,\mu$s long simulation, just like the higher temperature simulation of the previous f.f. version, results in keeping the particular stretch mobile and in a random coil conformation without however unfolding any further either (Figures 4.5B and 4.6). The above tests suggest that the next kinetic barrier along the unfolding pathway has a different timescale of crossing.

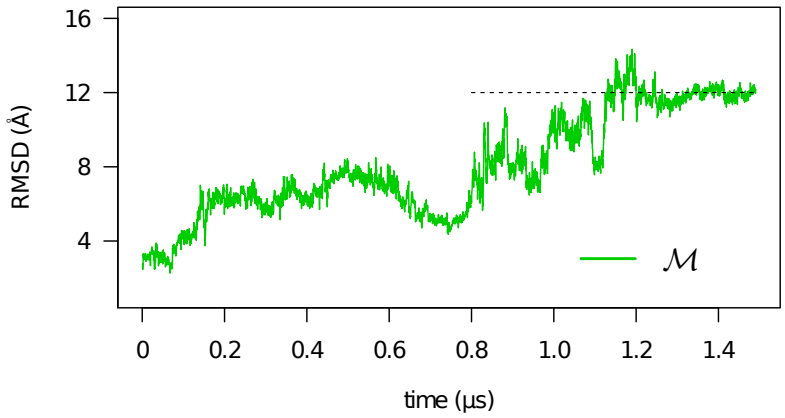

Figure 4.4. Unfolding stops: $RMSD$ of $\mathcal{M}$ for a 1.5 μs-long simulation showing a plateau after the 1.2 μs.

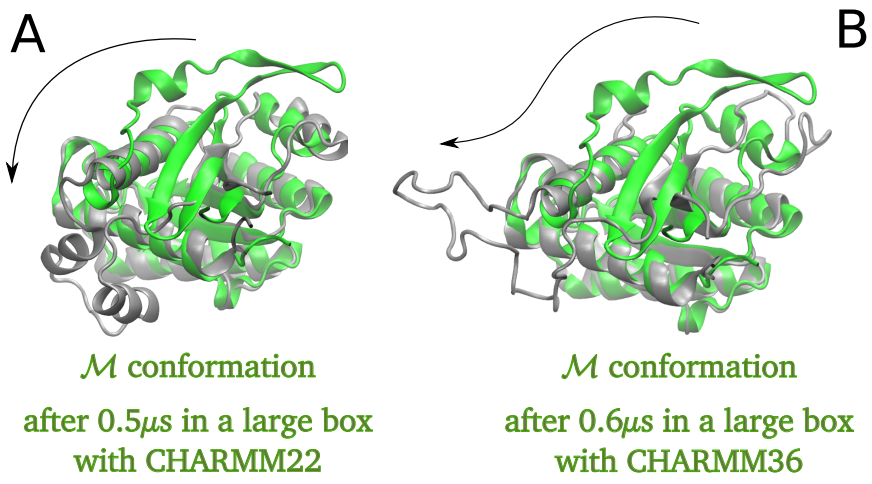

Figure 4.5. Unfolding of the mesophilic G-domain. (A) The initial crystal structure of $\mathcal{M}$ (green) overlapped with the final conformation of a 0.5 μs-long simulation at 360 K in a large solvation box (CHARMM22/CMAP). The same misfolded conformation occurs also at the end of the initial 1.5 μs-long simulation. (B) The crystal structure of $\mathcal{M}$ (green) overlapped with the final conformation of a 0.6 μs long simulation at 360 K with CHARMM36 Force Field.

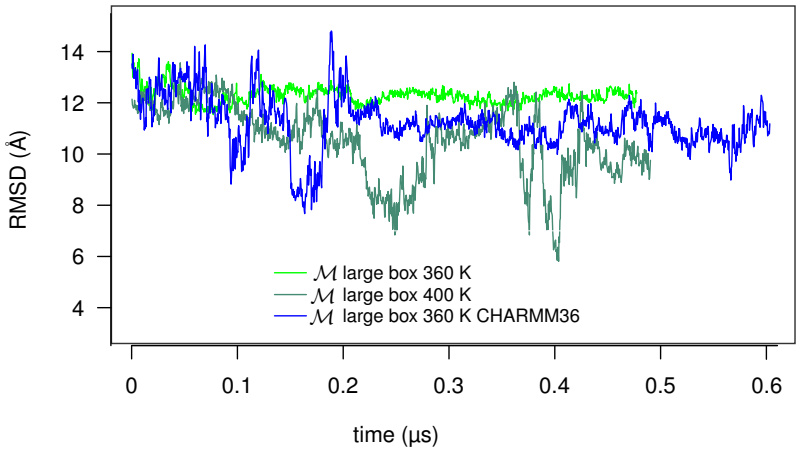

Figure 4.6. Unfolding of the mesophilic G-domain. *RMSD* timelines of the high-temperature simulations of the mesophilic G-domain after resolvating in a larger box with CHARMM22/CMAP (green and petrol colors) and with CHARMM36 (blue color).

4.1.2 Multimeric apo† EF-Tu: A β to α conformational drift

Activity of G-proteins. We mentioned briefly in the previous section that the switch I region of the mesophilic protein, as with all bacterial EF-Tu proteins, undergoes a large conformational change during activity. In fact, considering the protein as a whole, this conformational change is only one of the several structural rearrangements that can happen during the elongation factor's lifetime [172, 208, 211–217]. The activation cycle of heterotrimeric GTPase proteins deserves here an overview: EF-Tu was the first protein found to be regulated by the binding of GDP and GTP molecules [211, 218]. In its basic state, the EF-Tu/EF-1α – more generically G-protein – is bound to GDP and is "turned-off". The whole complex is itself bound to the membrane. The guanine nucleotide exchange factor, the G-protein's activator, in its inactive state is also separately bound to the membrane. Upon the latter's activation, it becomes highly affine for the G-protein-GDP complex and when the two bind, the GDP is released. The now GDP-free complex becomes highly affine for GTP. Upon GTP binding, the factor dissociates followed by a large inter-domain rearrangement. The G-protein-GTP complex is now "turned-on". It can thus form a ternary complex with tRNA anticodon domain and transport it to the ribosomal A site where it associates with the mRNA codon domain and

† *Apo* is a shortcut of the term *apoenzyme* or *apoprotein* referring to enzymes that require a cofactor but do not have one bound. In contrast to that, *holo* or *holoenzyme* refers to the enzyme in its active, bound form.

proceeds to protein synthesis. While at the ribosome, the hydrolysis from GTP to GDP occurs with mRNA. The scheme of this process can be seen in Fig. 4.7

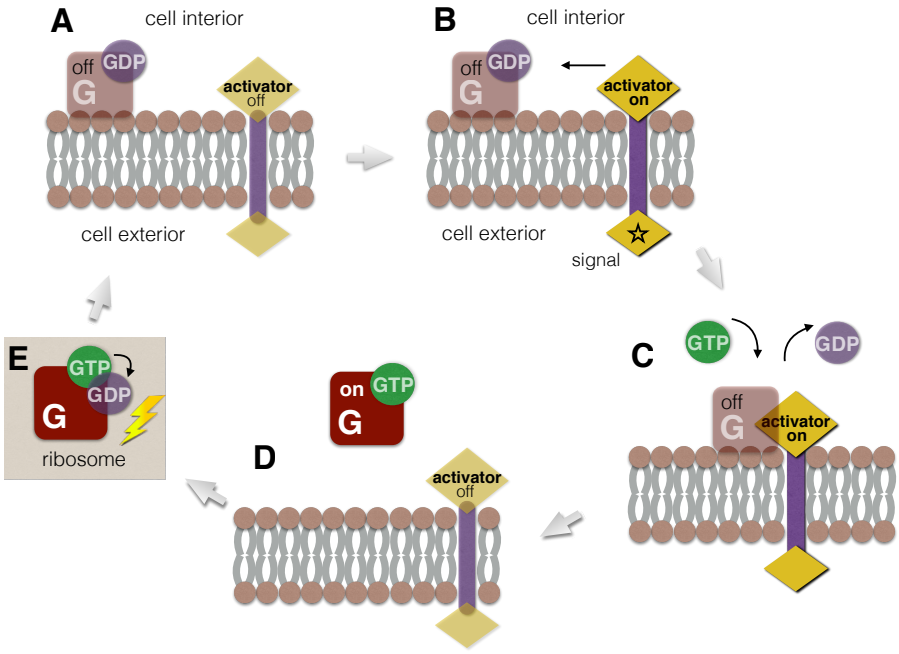

Figure 4.7. Schematic representation of GTPase protein's activation cycle. (A) The G-protein-GDP complex as well as its activator are turned off and bound independently on the membrane. (B) Signaling molecules from outside the cell turn on the G-protein activator, i.e. the guanine nucleotide exchange factor, that binds to the G-protein-GDP complex. (C) GDP is released and the new complex becomes highly affine for GTP. (D) GTP binds on the G-protein and turns it on, allowing it to bind to tRNA and transfer it to the ribosome. (E) While at the ribosome the hydrolysis from GTP to GDP occurs followed by a large conformational change of the G-protein.

Specifically for the bacterial elongation factors, EF-Tu, the conformational change of the G-domain's switch I region involves the conversion of a small α-helix in the GTP-bound form to a double-stranded antiparallel β-sheet in the GDP-bound form. An overlap of the two conformers of the isolated mesophilic G-domain can be seen on the left side of Fig. 4.11.

Dynamics of the whole heterotrimeric EF. We now address an aspect of the behavior of the whole multimeric proteins even if in their *apo* forms. In order to probe the stability of the whole EFs, we simulated both mesophilic and hyperthermophilic systems at 360 K for 0.45 μs and 0.2 μs, respectively. Although these

timescales are effectively small for the size of the trimeric EFs (393 a.a. for the mesophilic and 435 a.a. for the hyperthermophilic EFs) we do manage to observe a striking conformational change in the mesophilic protein. The simulation having started with the EF-Tu in its beta conformer (Fig. 4.8(left)), a little after 250ns the structure acquires a conformation closer to the alpha conformer (Fig. 4.8(right)). This is further quantified in Fig. 4.9, where two $RMSD$ timelines, one with respect to the crystal structure of the beta conformer and another with respect to the crystal structure of the alpha conformer are shown.

The evolution of this high-temperature simulation suggests that the α conformer in the apoprotein form has a lower free energy than the β conformer with a kinetic barrier that – at 360 K – is accessible in a 100-nanosecond timescale. Although we lack the amount of data needed to actually evaluate that free energy, this finding is consistent with the activation cycle of EF-Tu; after the release of GDP, the apo-EF-Tu adopts a conformation that is highly affine to GTP. We should also note that a secondary structure prediction program also predicted an α conformation for this stretch [208].

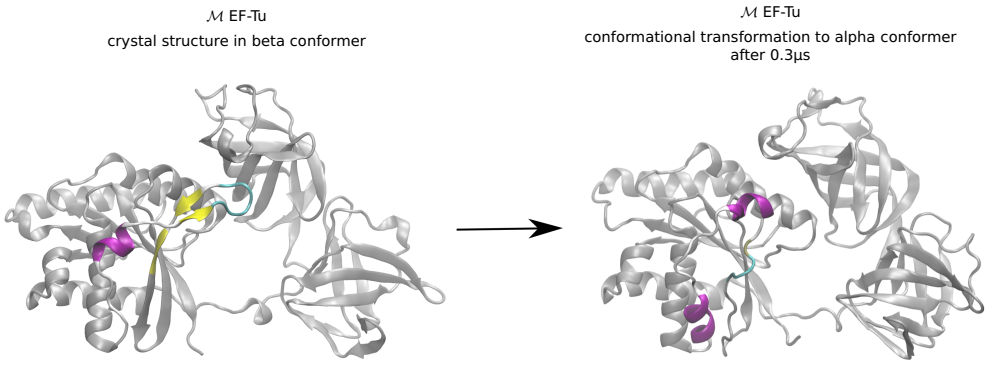

Figure 4.8. β to α conformational transition of EF-Tu. Snaphots of mesophilic EF-Tu at the beginning (left) and at the end (right) of a 0.45 μs simulation at 360 K.

Kinetic stability of the G-domain alpha conformer. The above results call for a final analysis, the investigation of the high-temperature kinetic stability of the α conformer for the isolated G-domain. A 0.6 μs-long simulation at $T = 360$ K shows that the stretch G33-C74 doesn't extend as effectively as with the high-temperature simulation of the β conformer. The C_α $RMSD$ for this simulation fluctuates around an average value of ~ 3.5 Å indicating the α to β direction of the transition might have a higher kinetic barrier (see Fig. 4.10). Given the previous discussion and observations, this behavior is not surprising. Besides, the α to β

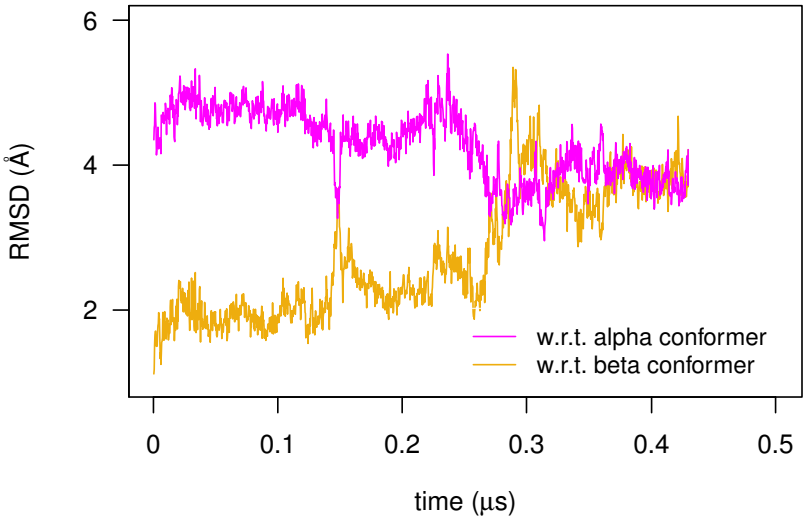

Figure 4.9. β to α conformational transition of mesophilic EF-Tu at 360 K. C_α-RMSD with respect to the crystallographic α conformer (magenta) and C_α-RMSD with respect to the crystallographic β conformer (orange). Although the corresponding simulation is that of the whole EF-Tu, the two $RMSD$ timelines refer only to the G-domain in order to exclude the noise coming from the rigid body rotation of domains 2 and 3.

conformational change happens during the catalytic step and not spontaneously. Yet, at the end of this simulation the stretch G33-C74 is substantially disrupted with loss of secondary structure (see the snapshot on the right of Fig. 4.11) indicating that the weak spot of the alpha conformer is located at the same switch I region. Indeed, further unfolding simulations, biasing the radius of gyration of the protein to gradually increasing values, verify that the unfolding always initiates from the same region independently of the starting conformer (see middle and bottom panels of Fig. 4.12). Interestingly, the respective region unfolds first also for the $\mathcal{H}$ G-domain (see top panel of Fig. 4.12). More rigorous free energy calculations, quantifying the height of the barrier for each protein and for each conformer of the $\mathcal{M}$ homologue are reserved for a future investigation.

4.2 Concluding remarks

In this chapter we verified that the two homologous proteins exhibit different thermal stabilities *in silico*, with the mesophilic protein signalling the onset of unfolding at a high, but physical, melting temperature ($T = 360\,K$) while the hyperthermophilic domain preserves its native structure.

The weak spot of the mesophilic protein, i.e. the region where unfolding initi-

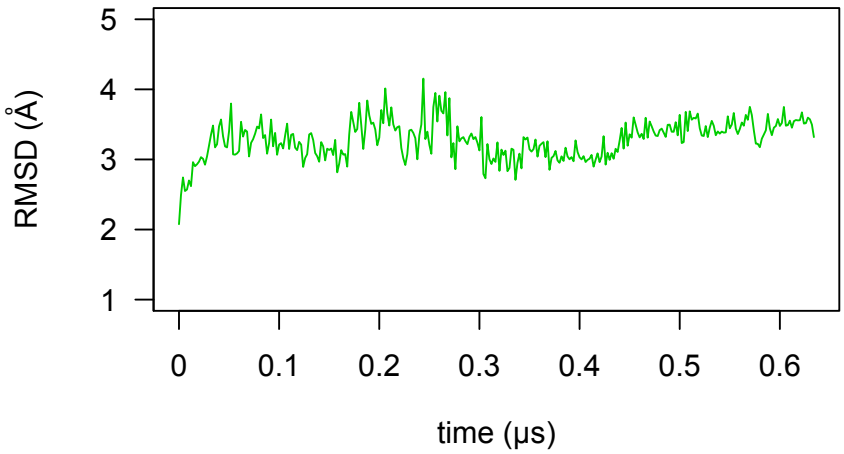

Figure 4.10. C_α *RMSD* of a 0.6 μs-long simulation at 360 K of the isolated mesophilic G-domain in its α conformer.

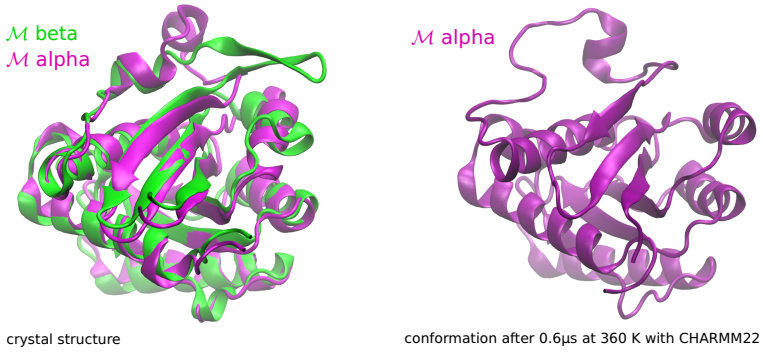

Figure 4.11. β and α conformers of mesophilic EF-Tu. (Left) Overlap of the two conformers, the α GTP-bound in magenta (PDB code 1OB2) over the β GDP-bound in green (PDB code 1EFC). (Right) Snapshot of the α conformer after 0.6 μs at $T = 360$ K.

ates, involves the so-called switch I region, a key motif for both the catalytic GTPase catalysis and the long range allosteric conformational displacement occurring upon ribosome binding [208–210]. The equivalent region in the hyperthemophilic protein is structurally stabilized by the insertion of two small helices (α^1 [E32-L45] and α^2 [E48-E63], see Fig. 4.3) and by frequent ion-pairing between charged amino-acids (see Discussion of the previous Chapter) [172, 175]. Complimentary simulations of the whole EF-Tu reveal a spontaneous β to α conformational change at the switch I stretch. This first result suggests that the flexibility and structure of the switch I region is optimized in the mesophilic protein for an enhanced catalytic activity at room temperature. When temperature is raised, the early unfolding of this region undermines the pre-organisation of the active site [219]. In the hyper-

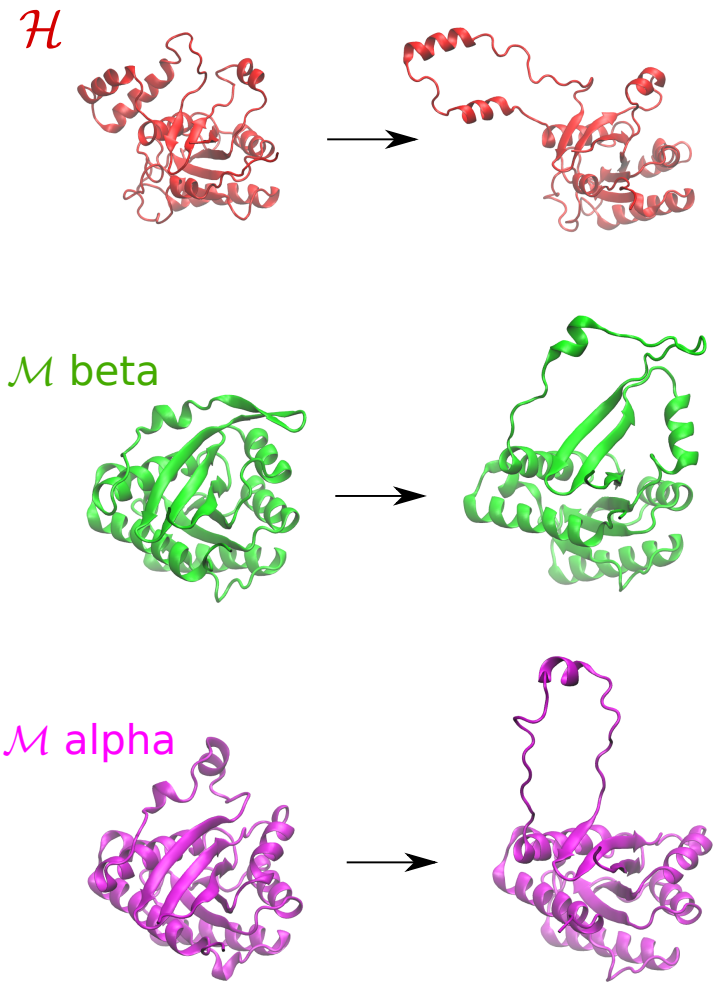

Figure 4.12. Unfolding simulations via R_g biasing. The simulation of each system, $\mathcal{H}$, $\mathcal{M}$ beta and $\mathcal{M}$ alpha had a total duration of 10 ns where every 1 ns the target R_g was augmented by 0.5 Å. For all three systems, the unfolding is initiated from the same stretch, that is the switch I region of $\mathcal{M}$ (middle and bottom panels) and its equivalent in $\mathcal{H}$ (top panel).

thermophilic protein, the switch I region is more resistant to temperature increase, thus its potential contribution to optimal catalysis is preserved at high temperatures. To the best of our knowledge, it is still an open question what conformational changes happen to this region during catalytic activity of $\mathcal{H}$.

Chapter 5

Are coarse-grained models apt to detect protein thermal stability? The case of OPEP force field[1].

Summary

Both kinetic and thermodynamic stability of the two homologous G-domains is now examined using the coarse-grained model OPEP. Free of external constraints, OPEP (see Chapter 2, Methodology) is able to maintain stable the fold of these relatively large proteins within the hundrend-nanosecond timescale. This makes possible to characterize anew, with a description of low resolution, the conformational landscape of the folded proteins as well as to explore their unfolding. In agreement with all-atom simulations used as a reference, we show that the conformational landscape of the thermophilic protein is characterized by a larger number of substates with slower dynamics on the network of states and more resilient to temperature increase. We verify the stability gap between the two proteins using replica-exchange simulations and estimate a difference between the melting temperatures of about 23 K, in fair agreement with experiment. The detailed investigation of the unfolding thermodynamics, gives insight into the mechanism underlying the enhanced stability of the thermophile relating it to a smaller heat capacity of unfolding.

[1] M. Kalimeri, P. Derreumaux and F. Sterpone (2014). "Are coarse-grained models apt to detect protein thermal stability? The case of OPEP force field". *J. Non-Cryst. Solids.* In press.

5.1 Prologue

As we've seen so far, the use of AA MD simulations allows to study, with accuracy, microscopic properties of the folded state such as HB and IP networks and their short-timescale dynamics, atomistic fluctuations or even, in some cases, larger conformational changes attainable within the microsecond timescale (see Chapters 3 and 4). Moreover, additional cluster and network analysis can provide important information on the structure of the free energy landscape pertaining to the folded state (see Chapter 3). As a consequence, an AA comparative study between thermophilic and mesophilic homologues may allow to speculate on the underlying thermodynamic mechanisms of thermal resistance. For example in Chapter 3, the observed enhanced conformational flexibility of $\mathcal{H}$ is expected to cause a smaller heat capacity of unfolding as compared to $\mathcal{M}$, therefore a reduced curvature of the stability curve. Generally, however, the use of AA models with the currently available computational resources make unfeasible a more direct calculation of macroscopic properties (e.g. melting temperature, heat capacity e.t.c.). Here is where coarse-grained models in combination with enhanced-sampling techniques enter the game [30]. In this Chapter we exploit the OPEP CG force-field, described in Chapter 2 with a simulation protocol as described below.

Simulation setup The coarse-grained MD simulations are performed using the OPEP v4 [27] model for proteins with implicit solvent. The simulations run using an *in-house* developed code implementing the OPEP Hamiltonian. The trajectory is evolved using a timestep of 1.5 fs, and the temperature of the system is kept constant by applying the Berendsen thermostat ($\tau_B = 0.1$ ps). Before the production phase at temperatures $T = 300$ K, $T = 325$ K and $T = 350$ K, the systems were progressively equilibrated at lower temperatures, $T = 250$ K and $T = 275$ K for about 50 ns. The replica exchange molecular dynamics is done using 24 parallel replicas and an exponential temperature distribution in the range of 260-582 K. Exchanges between two neighboring replicas are attempted every 7.5 ps. Each replica is extended for 230 ns. Specific heat curves, $C_V(T)$, and free energy profiles are computed using the PTwham algorithm [220].

The simulation protocol for the ambient temperature (300 K) all-atom simulation can be seen in Chapter 3.

5.2 Results and discussion

5.2.1 Stability on a long timescale

Up to date, the OPEP force field has been extensively applied to study small peptides, and it was only recently tested on mid-size proteins with generally less than 80 amino-acids [27, 176, 221]. Therefore, given their size, the mesophilic ($\mathcal{M}$) and hyperthermophilic ($\mathcal{H}$) G-domains, are a challenging study-case. The capability of the force field to maintain the fold of the two proteins in MD simulations extending up to 100 ns is first probed and discussed below.

After an equilibration phase at low temperatures (250 K and 275 K), three independent trajectories, of length 100 ns each, were generated at temperatures 300 K, 325 K and 350 K in order to monitor the kinetic stability of the two systems and characterize the dynamics of their folded state.

Figure 5.1 shows, for the MD simulations at $T = 300$ K, the time evolution of four CVs that monitor conformational properties, namely the radius of gyration R_g, the root mean square deviation ($RMSD$) calculated with respect to the equilibrated configuration, the fraction of native torsion angles n_t and the fraction of secondary structure. The average values of these CVs are reported in Table 5.1 along with the respective values from AA simulations. Overall, the fold of the two proteins is stable during the simulation time; the $\mathcal{M}$ and $\mathcal{H}$ domains remain folded in a compact globular state of radius R_g= 16.0 Å and R_g=17.3 Å, respectively. We note that these values are about 3-5% smaller than those obtained by all atoms simulations. This stronger cohesive packing is caused by several factors characteristic of the CG model, such as the lack of "adhesive" interactions with the solvent treated as implicit, the preferential filling of the space due to the spherical graining of the amino-acid side-chains, the specific weight of hydrophobic mimicking potentials and the lack of repulsive electrostatic interactions. While the fraction of native torsion angles n_t and the percentage of secondary structure show a steady behavior, when looking locally, we observe several instabilities. For either protein, two helices and two small β-strands located around the middle and the end of each sequence are not well preserved. For $\mathcal{M}$ these stretches are lost during the equilibration phase at low temperature (250 K). For $\mathcal{H}$, one small helix is similarly lost during the equilibration and a second short peripheral helical stretch unwinds at the beginning of the 300 K simulation (Q128-G140). It is however noteworthy that at a later time, part of this helix refolds and that even the short helix lost during the equilibration phase (E118-M123) is also recovered. The main contribution to the initial $RMSD$ jump of $\mathcal{H}$ comes from a specific region of the $\mathcal{H}$ domain, two helices located at the switch I region of the protein, α^1 [E32-L45] and α^2 [E48-E63] (see Fig. 3.1(*C1*) or Fig. 4.3). These two helices maintain very well their sec-

ondary structure; however, they do move in a rather flexible way as a rigid body. By removing the contribution from this region the $RMSD$ shifts down and follows the behavior of the $\mathcal{M}$ protein as shown by the orange curve in Fig. 5.1 (b). The functional role of the switch I region and how this region contributes to the different stabilities of the $\mathcal{M}$ and the $\mathcal{H}$ proteins is discussed in Chapters 3 and 4. The timeline of the secondary structure for both systems is reported in Fig. 5.2.

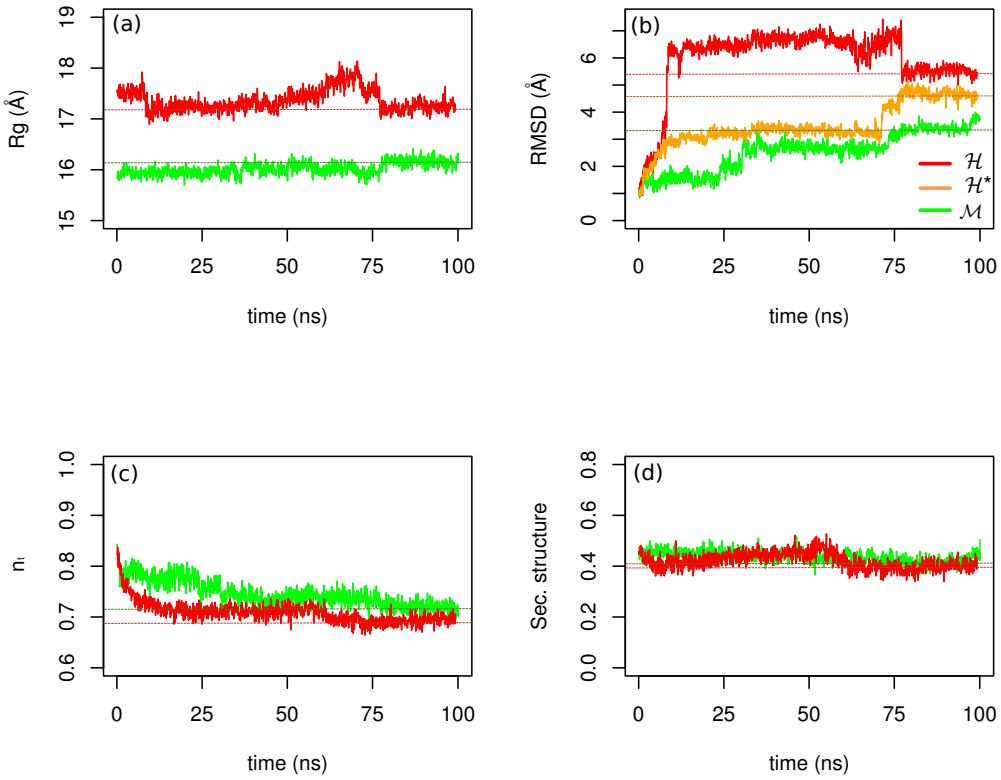

Figure 5.1. Timeline of (a) radius of gyration R_g (b) rigid-core C_α $RMSD$, (c) fraction of native torsion angles n_t and (d) fraction of secondary structure along the sequence for the OPEP MD simulations of the two systems at 300 K. Data in red refer to the $\mathcal{H}$ protein and data in green to the $\mathcal{M}$ protein. In panel (b) the orange curve $\mathcal{H}^*$ refers to a calculation performed after removing the contribution from helices α^1 and α^2 of the switch I region (see Fig. 4.3 of Chapter 4).

The examined CVs attest that the overall structure of $\mathcal{M}$ is more rigid and better maintained than that of $\mathcal{H}$. The somewhat floppy dynamics of the $\mathcal{H}$ domain is caused by the larger number of flexible regions, turns and coils (see also discussion in Section 3.2.1 and Appendix of Chapter 3), located at the surface of the protein and which, as a consequence of the absence of the viscous aqueous medium, move

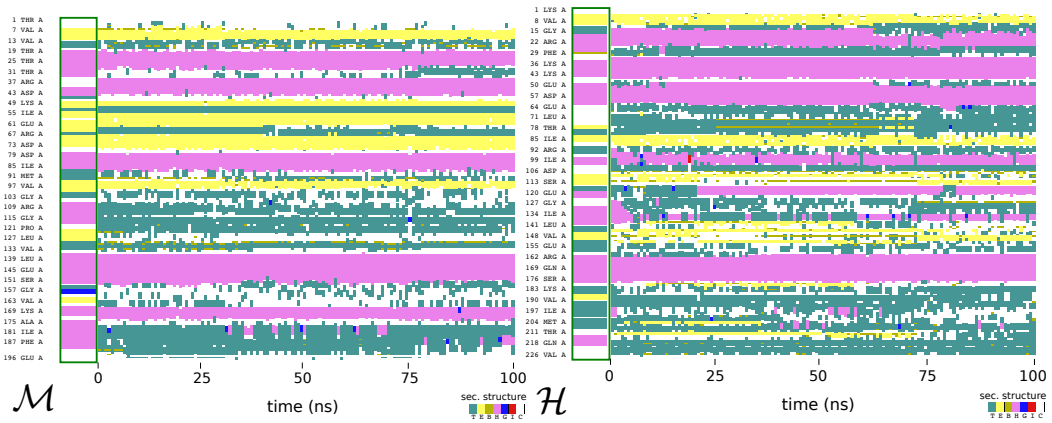

Figure 5.2. Secondary structure timeline for the 100 ns OPEP simulation at ambient temperatures. The mesophilic system is on the left and the thermophilic one on the right. At the beginning of each timeline, in a green-bordered box the secondary structure of the crystal structure is also reported.

more freely. Such behavior is also expected to impact the rigid core since α-helices and β-strands are on average shorter and more frequently interrupted by coils and loops.

The average values of the four CVs for the two simulations at the higher temperatures of 325 K and 350 K are given in Table 5.1 whereas a timeline is also shown in Fig. 5.7 at the end of this Chapter. At these timescales, we do not observe any temperature-driven kinetic instability for both systems, something also verified by two simulations at the higher temperatures of $T = 375$ K and $T = 400$ K (data not shown). We recall that in the all-atom simulations the early steps of unfolding are observed in a well localized region of the $\mathcal{M}$ protein and occur between 200 and 500 ns at $T = 360$ K (see Chapter 4). As we discuss later in detail, the OPEP force field like other CG models in general, as compared to the all-atom simulations, impacts both the effective timescale of relaxation processes (kinetics) as well as their energy scales (thermodynamics). Therefore it is important to quantify this shift when discussing the relative stability of proteins and its time and temperature dependence.

5.2.2 Exploring the folded-state dynamics

As has been extensively discussed in Chapter 1, according to several experimental studies, thermophilic proteins have been considered more rigid at ambient conditions than their mesophilic homologues [71]; in fact the mechanical rigidity was postulated as the main source of thermal stability as well as the cause of the lack of activity of thermophiles at ambient temperature [198]. Only at the optimal

Table 5.1. Average values of radius of gyration R_g, rigid-core C_α $RMSD$, fraction of native torsion angles n_t and fraction of secondary structure along the sequence for the OPEP MD simulations at 300 K, 325 K and 350 K

	CV	$T = 300\,\mathrm{K}$	$T = 325\,\mathrm{K}$	$T = 350\,\mathrm{K}$
$\mathcal{M}$	R_g (Å)	16.0 ± 0.1 (16.3 ± 0.1)	16.0 ± 0.2	16.3 ± 0.1
	$RMSD$ (Å)	2.5 ± 0.7 (3.0 ± 0.5)	4.4 ± 0.6	4.4 ± 0.3
	n_t	0.70 ± 0.03 (0.85 ± 0.01)	0.71 ± 0.01	0.70 ± 0.02
	sec. structure	0.44 ± 0.02 (0.65 ± 0.01)	0.44 ± 0.02	0.41 ± 0.02
$\mathcal{H}$	R_g (Å)	17.4 ± 0.1 (18.0 ± 0.1)	17.3 ± 0.1	17.4 ± 0.1
	$RMSD$ (Å)	5.9 ± 0.7 (3.4 ± 0.3)	5.8 ± 0.3	5.8 ± 0.2
	n_t	0.71 ± 0.02 (0.86 ± 0.01)	0.69 ± 0.01	0.69 ± 0.01
	sec. structure	0.42 ± 0.03 (0.52 ± 0.02)	0.39 ± 0.02	0.40 ± 0.02

For $T = 300\,\mathrm{K}$ the reported data in parenthesis correspond to the AA simulations. Errors correspond to standard deviation.

growth temperature of the host organism, thermophilic proteins function efficiently because the thermal excitation activates conformational motions that eventually match those of mesophiles at ambient condition [14]. This corresponding-states picture for mechanical rigidity has been however questioned recently, both experimentally [35, 36] and theoretically [37, 84]. Indeed, depending on the time and length scales considered, thermophilic proteins show, at ambient conditions, comparable if not enhanced flexibility with respect to their mesophilic variants. In other words, the entropic route to stability also exists [95]. Moreover quality matters. As put forward in Chapter 3 the spatial distribution of flexible and rigid stretches differs substantially between the $\mathcal{H}$ and $\mathcal{M}$ domains; the $\mathcal{H}$ homologue gains kinetic stability via a more uniform alternation of flexible and rigid structural patterns.

In the lines of our all-atom investigation (see Chapter 3), we now focus on the dynamics at ambient temperature in order to characterize the conformational landscape of the proteins. This is achieved by performing a conformational clustering on the trajectories, where we use the C_α $RMSD$ as a distance between two configurations and a cut-off of 2.5 Å to separate the clusters. The results are shown in Fig. 5.3. In the top panels, the cluster growth versus time is plotted for the OPEP CG simulations (Fig. 5.3(a)) in comparison to the results from all-atom simulations of equal length (Fig. 5.3(b)). Bearing in mind that the clustering cut-off for both systems and models is the same we observe a striking difference in the final number of clusters between the two models. As expected, the sampling of the available conformational space is much more effective using a CG model over an AA one. Interestingly, for this CV, we also note that in the OPEP simulations the

$\mathcal{H}$ protein visits a larger number of sub-states than $\mathcal{M}$, in agreement with the reference all-atom result, albeit the effective timescales in CG and all-atom simulations are different [30] as we will discuss later on.

In the all-atom simulations, the number of clusters visited as a function of time can be successfully fitted using a simple exponential model, $N = N_\infty(1 - e^{-t/\tau})$, obtaining $N_\infty = 5$ and $\tau = 24\,\text{ns}$ for $\mathcal{M}$ and $N_\infty = 13$ and $\tau = 46\,\text{ns}$ for $\mathcal{H}$. However, the same model cannot be fitted on the totality of the CG data that actually show sudden jumps.

To elaborate more on the above, in the bottom panel of Fig. 5.3, a network representation of the CG clustering is drawn using a force-based algorithm, with the size of the nodes and edges being proportional to their occupancy and number of interconversions, respectively (the edge weights have been normalized so that all edges that exit one node sum up to one). With the use of the Markov clustering algorithm mentioned previously [170] we identified the substates where random-walks representing protein motion in the network of states get confined. These clusters are grouped together in larger bundles and outlined by the larger ellipsoidal lines in Fig. 5.3 (bottom). We call those *basins of attraction* and, for the same granularity parameter of the algorithm that set the height of the kinetic barrier confining the walkers, we identified 5 and 4 of them for $\mathcal{M}$ and $\mathcal{H}$ respectively. We then went back to the cluster growth of the CG model and tried to fit the exponential model to the parts of the trajectory that correspond to each of the basins. As can be seen in Fig. 5.3(a) for three cases in $\mathcal{M}$ and one in $\mathcal{H}$ the fit was not possible suggesting that the trajectory was only transiting those substates. Similar fast transitions have been observed previously in all-atom simulations of a protein in crystal environment and could be associated to Lévy flight motion [222].

Our findings show that with respect to AA, the CG dynamics not only is faster in exploring the conformational space as shown by the larger number of clusters visited in simulations of numerically equal length, but it also explores more efficiently the hierarchical organisation of the conformational states [89, 102, 223] as demonstrated by the sudden jumps in the clustered trajectories. The larger number of clusters identified by the OPEP simulations suggests a speed-up of at least 5-6 times with respect to the AA model. This finding agrees with previous validations of the model [30] as well as by monitoring relaxation processes as discussed below.

Although the OPEP CG force field allows us to explore more efficiently the conformational landscape, it is at this point not safe to compare the global flexibility of the two systems since it is clear from Fig. 5.3(a) that longer simulations are needed to achieve a convergence. However, we can try to quantify the diffusivity of the systems along the conformational landscape, a property that relates to the local disorder of the surface [159] and represents a key parameter in the theory of protein

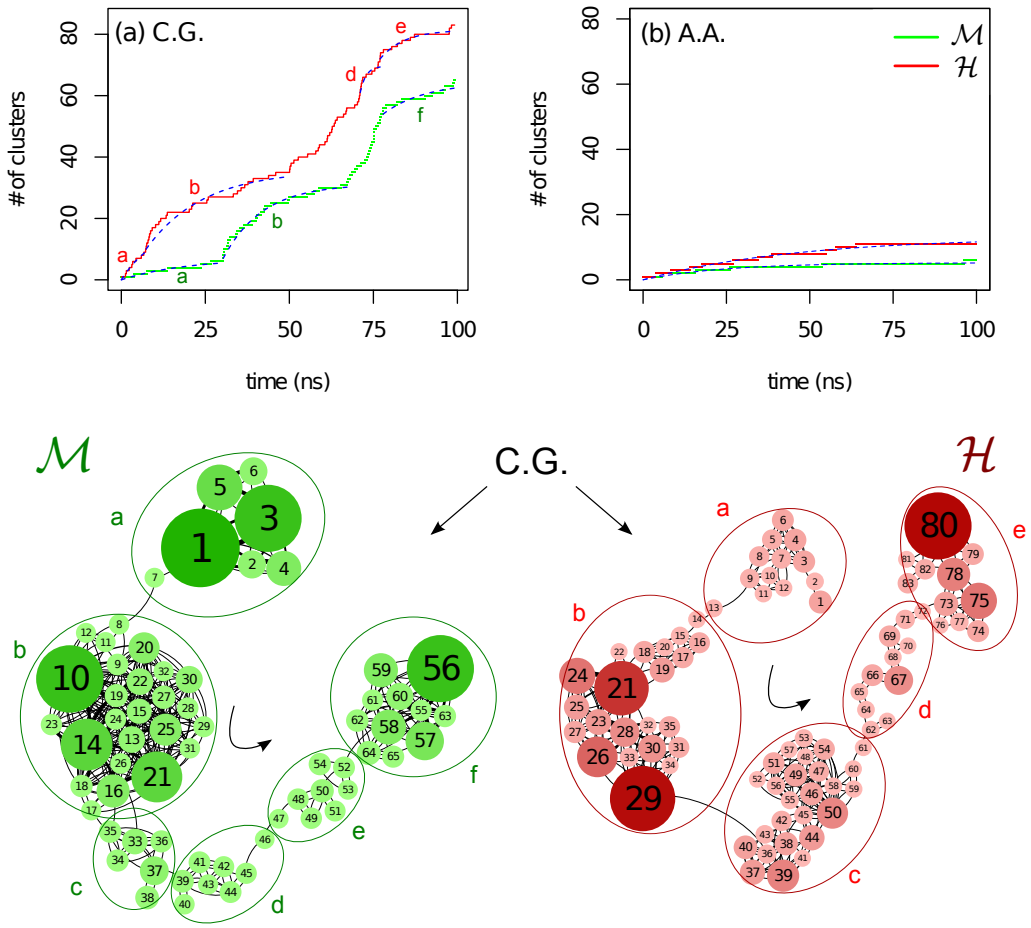

Figure 5.3. Conformational substates at 300 K. Top: Number of clusters versus time for (a) the CG and (b) the AA MD simulations at 300 K. The dashed blue lines correspond to an exponential fit on the function $N = N_\infty(1-e^{-t/\tau})$. Bottom: Network representations [194] of the coarse-grained MD simulations clusters shown in (a) drawn with a force-based algorithm for $\mathcal{M}$ on the left and $\mathcal{H}$ on the right. The larger ellipsoidal borderlines outline the basins of attraction extracted using a Markov clustering algorithm with a granularity parameter equal to 1.2 [170].

folding [162]. In this context, the motion of a protein with respect to a given CV X is associated to a diffusion constant D. Within the harmonic approximation, D is given by $D = \langle \delta X^2 \rangle / \tau_{corr}$ (see Chapter 2, Methodology). This approximation is valid for the fraction of native torsion angles n_t for which the autocorrelation of fluctuations decays exponentially, $c(t) = \langle \delta n_t(t) \cdot \delta n_t(0) \rangle \simeq e^{-t/\tau_c}$. We only stress that when the CV does not show a harmonic behavior, more sophisticated

approaches are needed to estimate the local diffusivity on the projected landscape [207, 224] such as, for example, the CV constraints used in Chapter 3.

The obtained results for n_t are shown in Fig. 5.4. The main plot shows, in solid lines, the diffusion coefficient D with respect to the temperature for the two systems whereas in dashed lines we report the value of D as estimated in our previous all-atom approach for the n_t at $T = 300\,\mathrm{K}$ (see Chapter 3). The values are systematically higher for $\mathcal{M}$, meaning that the internal motion of the $\mathcal{H}$ domain is slowed down by a distribution of higher barriers separating substates [159]. At ambient temperature, the value of D for both systems is shifted, for OPEP, to slightly larger values as compared to the AA force field. This verifies, as we expected, that dynamics is more diffusive in the CG model. The ratio of the computed diffusion constants, $\mathrm{D}^{CG}/\mathrm{D}^{AA} \sim 2-7$ provides a supplemental estimate of the effective timescale characterizing the OPEP internal protein dynamics with respect to all-atom simulations.

Moreover, as the temperature increases the values of D increase as well with a notable resilience for $\mathcal{H}$ as compared to $\mathcal{M}$. Since the dynamics depends on temperature, as the temperature increases it allows for more frequent transitions among substates; thus, the above picture reveals again the presence of higher kinetic barriers for $\mathcal{H}$ than for $\mathcal{M}$.

5.2.3 Towards thermodynamics

The OPEP force field has been routinely used in combination with a variety of simulation techniques [30, 225–227], among which is replica exchange molecular dynamics (REMD) [154] that enhances the sampling of the protein folding/unfolding process yielding correct thermodynamic properties of the systems under study.

A first attempt to probe the different thermal stabilities of the homologous G-domains was already reported for a weak energy scaling factor of the force field [30]. It was indeed verified that the curve of the specific heat $\mathrm{C}_V(\mathrm{T})$ of the $\mathcal{H}$ protein was systematically shifted at higher temperatures with respect to that of $\mathcal{M}$, mirroring the enhanced stability of the hyperthermophile. Here, we report the results from REMD simulations using a higher energy scaling factor (the same used for the MD simulations discussed previously). The curves of the specific heat are plotted in Fig. 5.5. Likewise, we observe for this set up, a systematic shift towards higher temperatures for the $\mathcal{H}$ domain as compared to $\mathcal{M}$. We also note the presence of several peaks that signal the onset of unfolding of different secondary structures as well as unpacking. In particular, the $\mathcal{M}$ protein is characterized by two peaks, a minor one at $T = 366\,\mathrm{K}$ and a large one at $T = 420\,\mathrm{K}$. The hyperthermophilic protein also presents a peak at $T = 420\,\mathrm{K}$, but also a series of others at the higher temperatures of $T = 460\,\mathrm{K}$ and $500\,\mathrm{K}$. While the absolute value of the temperatures

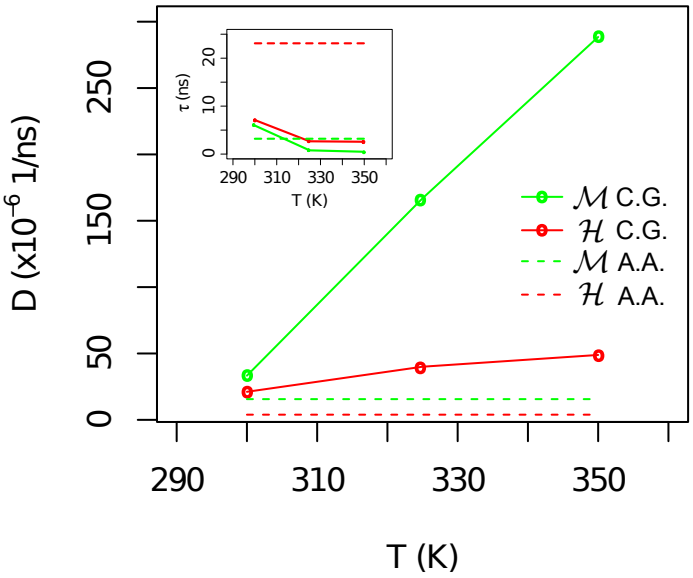

Figure 5.4. Diffusion in a harmonic basin of attraction. Main plot: In solid lines we report the diffusion coefficients versus temperature for n_t as estimated for the OPEP simulations. In dashed lines we report the value of D as estimated in our previous all-atom approach for n_t at $T = 300\,\mathrm{K}$ (see Chapter 3). Inset: Characteristic decorrelation time of δn_t with respect to the temperature. The decorrelation time is always smaller for $\mathcal{M}$ in agreement with our previous AA results.

at the C_V peaks is generally too high as compared to experimental data [174, 201] the stability gap between the two proteins is actually within the correct range of $40 - 50\,\mathrm{K}$.

Subsequently, from the same simulation data, we attempt to extract the stability curves for the two proteins. We use the gyration radius as an order parameter to distinguish between the folded and unfolded states. From the trajectories of all the replicas we reconstruct the free energy landscape projected on the reaction coordinate R_g, $G(R_g) = -k_B T ln\, P(R_g)$. This is done by calculating the probability distribution of the variable $P(R_g)$ using the PTwham unbiasing technique [220]. The probability distributions $P(R_g)$ display a well defined bimodal profile, thus allowing for a clear a separation of the folded and unfolded states at all temperatures and for both systems. This can be appreciated by looking at the top panel of Fig. 5.6 where the free energy profiles for different temperatures are shown with respect to the radius of gyration for $\mathcal{M}$ (left) and $\mathcal{H}$ (right) proteins. The dividing

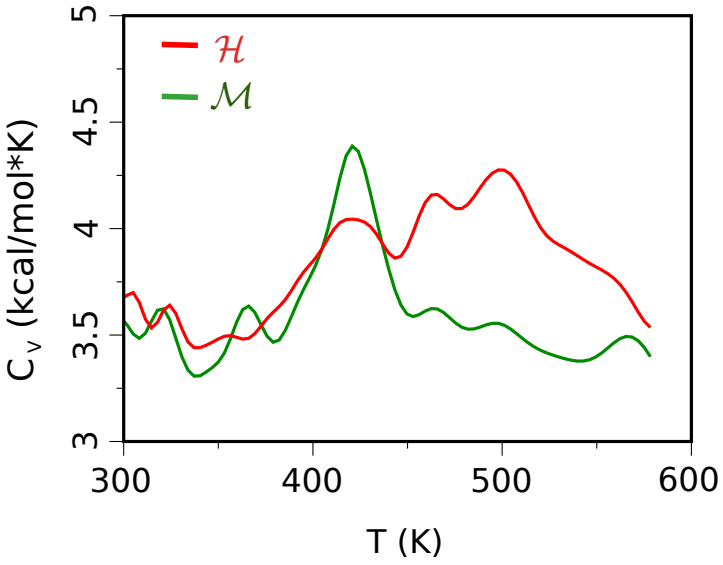

Figure 5.5. Specific heat C_V for $\mathcal{M}$ (green) and $\mathcal{H}$ (red) domains of the EF-Tu and 1α proteins, respectively, calculated from OPEP-REMD simulations. The presence of multiple peaks in the C_V profile is caused by the unfolding events of different secondary structure motifs as well as progressive unpacking. In a simple two-state model, the C_V is expected to show a single peak at the melting temperature T_m where the populations of the folded (p_f) and unfolded (p_u) states are equal.

value between folded and unfolded states is indicated with a vertical dashed line, being 16.6 Å for $\mathcal{M}$ and 18.6 Å for $\mathcal{H}$. As temperature increases the population of unfolded proteins (p_u) increases at the expense of the folded ones (p_f). In the bottom panel of Fig. 5.6, the free energy difference between folded and unfolded states, $\Delta G = -k_B T ln\frac{p_u}{p_f}$ is calculated as a function of temperature. This is the celebrated stability curve, intersecting the x-axis at the melting temperature T_m. The obtained data were fitted to the Gibbs-Helmholtz equation given by

$$\Delta G_{f\to u}(T) = \Delta H_\mathrm{m}\big[(T_\mathrm{m}-T)/T_\mathrm{m}\big] - \Delta C_P\big[T_\mathrm{m} - T\big(1 - ln(T/T_\mathrm{m})\big)\big]$$

estimating the values of $T_\mathrm{m} = 388 \pm 2\,\mathrm{K}$, $\Delta C_P = 0.103 \pm 0.005\,\mathrm{kcal/(mol{\cdot}K)}$ and $\Delta H_\mathrm{m} = 6.2 \pm 0.3\,\mathrm{kcal/mol}$ for $\mathcal{M}$ and $T_\mathrm{m} = 411 \pm 3\,\mathrm{K}$, $\Delta C_P = 0.020 \pm 0.004\,\mathrm{kcal/(mol{\cdot}K)}$ and $\Delta H_\mathrm{m} = 3.5 \pm 0.2\,\mathrm{kcal/mol}$ for $\mathcal{H}$.

With the use of a single reaction coordinate, the simple OPEP model allows to detect a thermal stability gap of about 25 K between the two proteins. This is not far from the experimentally reported difference of the optimal enzymatic activity for the two domains (40 K). Of course, as already discussed above and reported in previous investigations, the OPEP force field is prone to a systematic deviation

of $\pm 23\,\mathrm{K}$ with respect to experiments [30, 138, 228]. At the same time we should stress that also all-atom force fields, although nowadays well-refined and capable to follow folding/unfolding events at long timescales [144], generally fail to estimate the exact temperature dependence of thermodynamic properties [144, 229].

Moreover, the values obtained for the heat capacity of unfolding ΔC_P and enthalpy of unfolding at the melting temperature ΔH_{m}, are out of scale. There are several causes to that starting from the absence of explicit water that is considered to give an important contribution to ΔC_P. Additionally, given the coarse-grained nature of system, the separation among entropic and enthalpic contribution is compromised. Finally, the lack of explicit electrostatic interactions should also be added to the list [230].

Up to date, no calorimetric data are available for the two G-domains. However, two detailed experimental studies on the thermal stability of the whole elongation factors -Tu and -1α, using circular dichroism and fluorescence, have been carried out determining the melting temperatures of the trimeric proteins at $320\,\mathrm{K}$ and $365\,\mathrm{K}$ respectively as well as showing that the G-domains set up a "basic" level of the thermostability for the whole proteins [174, 201]. Here, we verify in a qualitative manner that indeed the different thermal stability content of the two proteins is also reflected when the G-domains are taken isolated. More importantly, we observe the broadening of the hyperthermophilic's stability curve over that of its mesophilic homologue suggesting that the thermodynamic mechanism behind the increase of its thermal stability is that of a smaller heat capacity of unfolding [97, 98]. This is coherent with what deduced from AA simulations considering the conformational fluctuations in the folded state as well as the change of compressibility with temperature (see Chapter 3).

5.3 Concluding remarks

This is the first time that the thermal stability of two homologous thermophilic and mesophilic proteins is examined using the OPEP force field. In fact, to the best of our knowledge, this has never been attempted with any other coarse-grained force field of the same nature for such large systems consisting of about 200 residues. First, we show the capability of the force field to preserve the native state at the timescale of a hundred of nanoseconds by MD without the need of external constrains. This enables the characterization of the conformational landscapes of the homologues. More specifically, we show that the qualitative description of the conformational landscapes of the two proteins matches that from all-atom simulations. The space visited by the hyperthermophilic protein is characterized by a large number of sub-states and its dynamics is slower and more resilient to

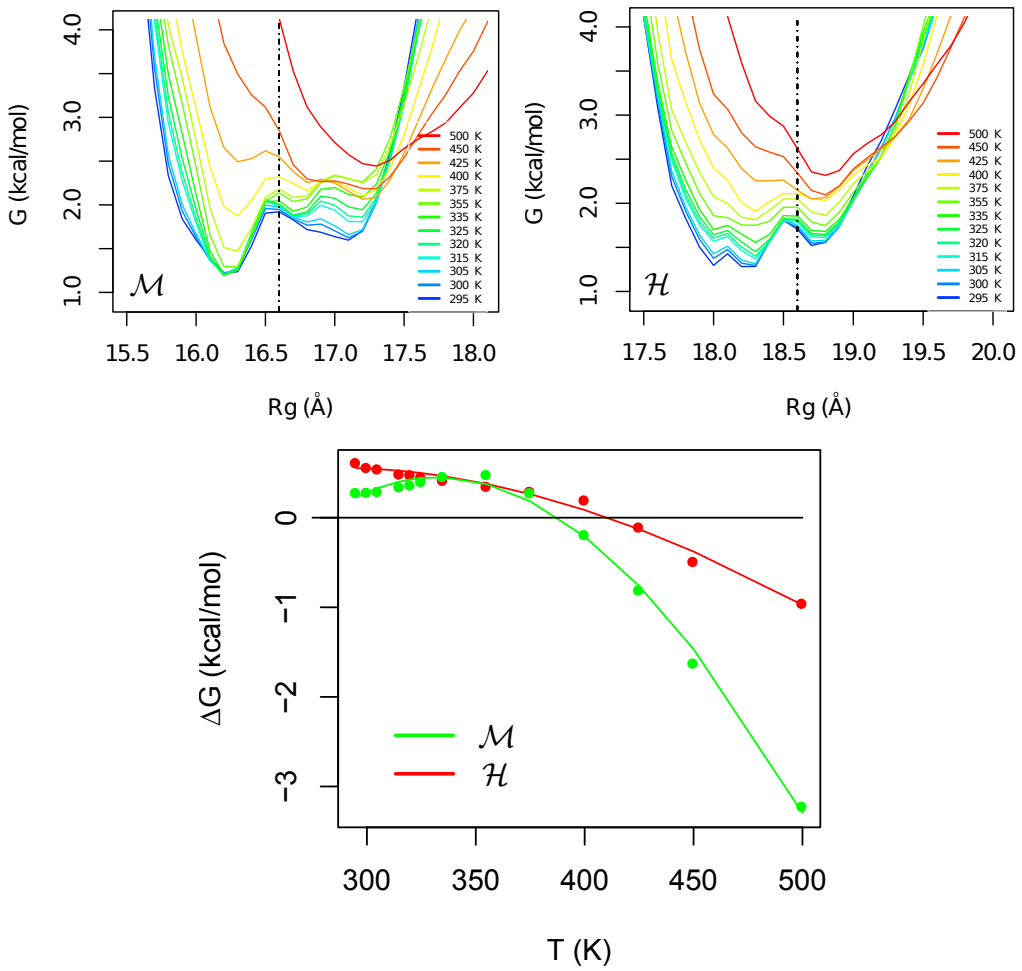

Figure 5.6. Stability curves. Top: Free energy profiles for different temperatures w.r.t. the radius of gyration for the mesophilic (left) and hyperthermophilic (right) proteins. As temperature increases the population of unfolded proteins increases at the expense of the folded population. The dividing value between folded and unfolded states is indicated with a vertical dashed line, being 16.6 Å for $\mathcal{M}$ and 18.6 Å for $\mathcal{H}$. Bottom: Free energy difference, a.k.a. stability curves, as extracted from the data of the top panel. The melting temperature is where the stability curve intersects the x-axis, estimated at 388 K and 411 K for $\mathcal{M}$ and $\mathcal{H}$, respectively.

temperature increase than that of the mesophilic variant. By using enhanced-sampling REMD, we also probe more explicitly the different thermal stabilities of the two proteins computing a stability gap of about 23 K in fair agreement with experiment. The shape of the extracted stability curve suggests that a smaller specific heat of unfolding for the hyperthermophilic protein is key for increasing its melting temperature.

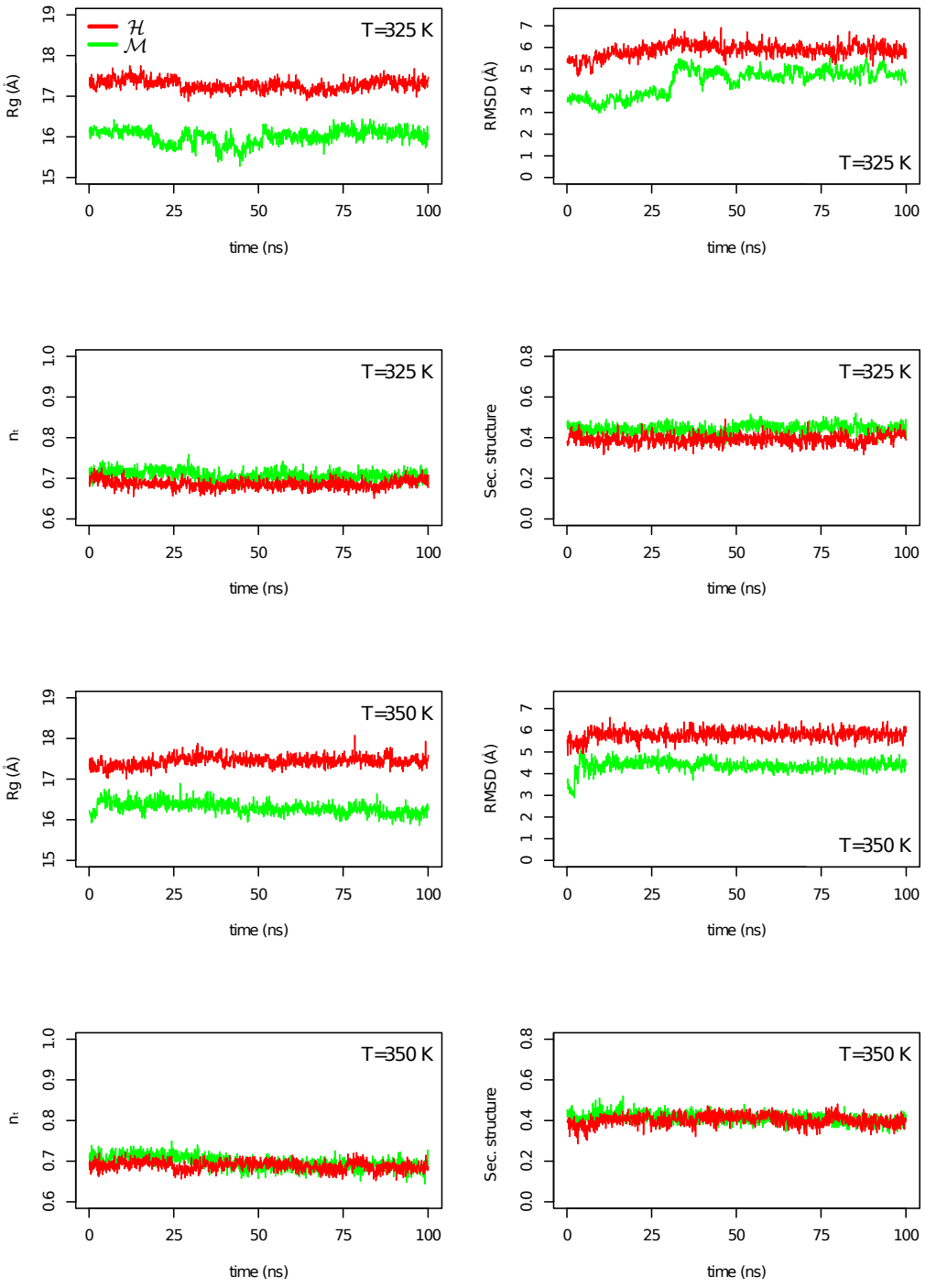

Figure 5.7. Timeline of radius of gyration R_g, rigid-core C_α RMSD, fraction of native torsion angles n_t and fraction of secondary structure for the two independent OPEP simulations at 325 K (top) and 350 K (bottom).

Chapter 6

Interface matters: The stiffness route to stability of a thermophilic tetrameric malate dehydrogenase[1]

Summary

In this final Chapter we turn our attention to the folded state of two orthologous bacterial proteins, a mesophilic and a thermophilic tetrameric malate dehydrogenase (MDH). Using the toolbox applied previously to explore the conformational space of folded proteins, we show that at the molecular length-scale of these orthologues, the thermophilic variant is indeed more rigid that the mesophilic one. Moreover, we show that the rigidification is the result of efficient inter-domain interactions. In fact, when considered isolated, the thermophilic domain is more flexible than the respective mesophilic one. Upon oligomerization, the induced stiffening of the thermophilic protein propagates from the interface to the active site, with a direct impact on the expected mechanism of cofactor and substrate binding [231, 232]. Our results open questions on the similarities of the binding processes in different homologues. Furthermore, our finding, when compared to literature, seems to indicate that the rigidity paradigm is more pertinent to oligomeric proteins.

6.1 Prologue

System description. MDH catalyzes the reversible oxidation of malate to oxaloacetate using the NAD+ coenzyme. Historically, the concept of "corresponding states" was introduced by studying a lactate dehydrogenase (LDH), a close homo-

[1] M. Kalimeri, E. Girard, D. Madern and F. Sterpone, (2014) "Interface matters: The stiffness route to stability of a thermophilic tetrameric malate dehydrogenase". In preparation.

logue of the protein studied here, the malate dehydrogenase (MDH) [72]. In fact, the protein family of LDH and MDH has been a study-case for thermal adaptation since the last 50 years [233] and several crystal structures have been resolved for a large number of organisms living at various temperatures [234–238]. For instance, neutron scattering experiments found the small-scale atomistic fluctuations of the tetrameric MDH from *Methanococcus jannaschii* to be more resilient to temperature increase as compared to its homologous mesophilic lactate dehydrogenase from *Oryctolagus cunniculus*, something that was later confirmed by simulations for larger time-scales [77]. Another recent aspect that was put forward is that the different thermal behavior of mesophilic and thermophilic MDHs may have an important contribution from their diffusion properties under crowding conditions [26]. From the structural point of view, a study of two moderate thermophilic MDH from the phototropic bacteria *Chlorobium tepidum* and *Chloroflexus aurantiacus* (CaMDH) versus their mesophilic variant from *Chlorobium vibrioforme* (CvMDH) showed, in agreement with the general trend, that thermal stability is correlated to an increased number of salt-bridges and hydrogen bonds as well as aromatic interactions across the domain interfaces. For the more thermostable CaMDH, a reduced flexibility was also forecasted on the basis of a proline and alanine surplus [235].

The two orthologue proteins herein are the CaMDH and CvMDH mentioned just above. In the rest of the text we refer to them as $\mathcal{T}$ and $\mathcal{M}$ for the thermophilic (PDB code 4CL3 [236]) and the mesophilic (PDB code 1GV1 [235]), respectively. The two orthologues have a 74% sequence similarity (52.2% identity) with very similar structures; excluding the flexible loop at the binding site region, the subunits are superimposed with an C_α-$RMSD$ of 1.0 Å. Figure 6.1 shows an overlap of the two structures. The thermophilic species has 309 amino acids (a.a.) and the mesophilic one 310 a.a., per chain. The optimal growth temperatures of the organisms are 328 K (55 °C) and 305 K (32 °C) for $\mathcal{T}$ and $\mathcal{M}$, respectively. Based on the characteristics of the different monomer-monomer interfaces and on the fact that the mesophilic homologue has been found to exist also in a dimeric form, the two tetramers are best described as a dimer of dimers [235]. In Figure 6.2, the subunits A+B and C+D constitute each of the two dimers, which in turn interact with each other to form the tetramer.

Simulation setup. All-atom Molecular Dynamics (MD) simulations were performed using the CHARMM22/CMAP Force Field for proteins [134] and TIP3P-CHARMM model for water. The two systems were simulated in both their tetrameric and monomeric forms (i.e. simulation of an isolated domain). The tetrameric $\mathcal{T}$ and $\mathcal{M}$ proteins were solvated respectively with 35422 and 38259 water molecules, while

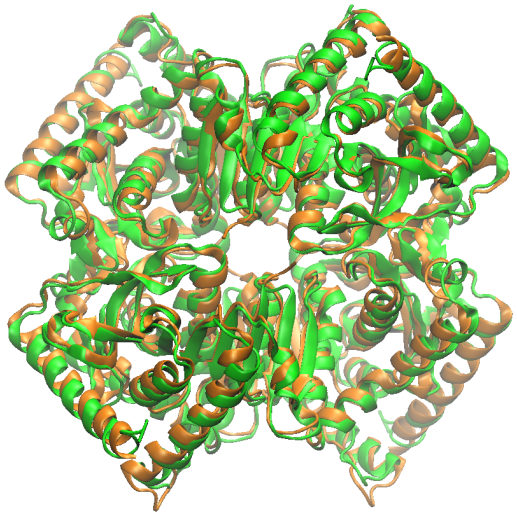

Figure 6.1. Overlap of the thermophilic (orange) and the mesophilic (green) malate dehydrogenases. PDB codes 4CL3 and 1GV1, respectively.

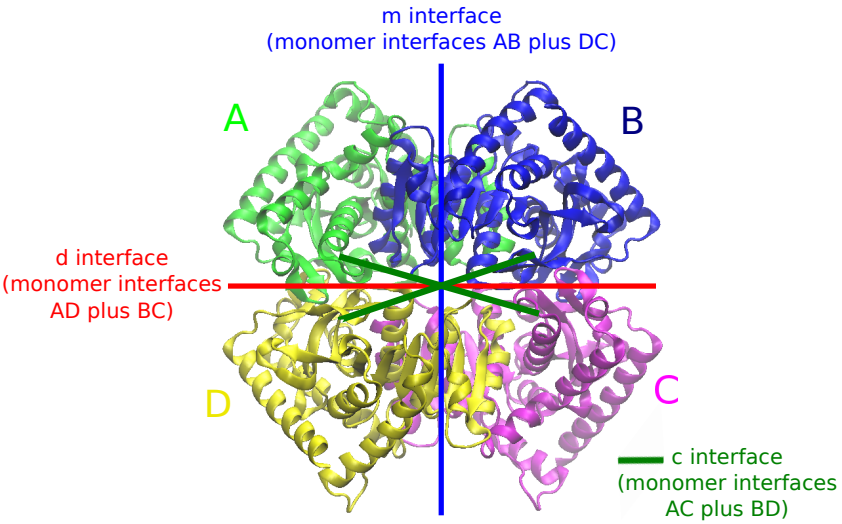

Figure 6.2. Structure of the MDH tetramer. The above structure belongs to the thermophilic ($\mathcal{T}$) homologue but the chain and interface nomenclatures in the figure apply also for the mesophilic one ($\mathcal{M}$). The tetramer is best described as a dimer of dimers. In the figure, the subunits A+B and C+D constitute each of the two dimers, which in turn interact with each other to form the tetramer. In order to study the interfacial interactions in a rather efficient way we have decomposed them into three different kinds comprising interface m, d and c the definition of which can be seen above.

the two monomers with 7992 and 9009 water molecules. Counter-ions were added to neutralize the systems. All four systems were simulated both at 300 K and 360 K,

for 200 ns for each temperature.

All simulations were performed using the NAMD software package [178]. The equations of motion were integrated using a timestep of 2 fs, with all bonds treated as flexible except for those involving hydrogen atoms which were kept rigid. Temperature and pressure were kept constant using the Langevin thermostat (with a friction coefficient $\gamma = 5\,\mathrm{ps}^{-1}$) and barostat (with an oscillation period of $\tau_P = 100\,\mathrm{fs}$), respectively. Electrostatics in a periodic simulation box was solved via the Ewald summation method and handled by the PME algorithm with a grid spacing of 1 Å. The production phases were preceded by 2 ns of equilibration. The trajectories were dumped with a frequency of 4 ps.

The potential of mean force (PMF) calculations were performed using the coarse-grained force field MARTINI v2.1 with polarizable water [143] and using the simulation package Gromacs 4.6.3 [239]. The PMFs were calculated for the separation of only two bound domains each time. Namely we separately considered the separation of A from B, A from D and A from C for both $\mathcal{M}$ and $\mathcal{T}$. The starting conformations were as in the crystal structures. Each dimer (AB, AD and AC) was first solvated in a rectangular box of dimensions 75 x 80 x 260 Å and with the axis that connects the center of masses of the two domains being parallel to the z-axis. Our calculations followed the same protocol as in [153]. After an equilibration phase, the domains were forced apart. This procedure was necessary to generate the initial configuration for the umbrella sampling simulations [240]. The umbrella sampling was based on 60 to 70 windows depending on the system, separated one from the other by 0.5 Å. In each window the simulation run for 30 ns proceeded by a 10 ns equilibration phase. For the final profiles the Weighted Histogram Analysis Method (WHAM) was used [152] as implemented in Gromacs 4.6.3. The errors were estimated with bootstrap analysis.

6.2 Results

6.2.1 Conformational dynamics: insight into stability and function

Protein stability. We first point out that the two tetrameric systems are stable within the explored timescales and temperatures (see Fig. 6.3). In fact, at ambient temperature the two systems fluctuate tightly around their crystallographic structures with a very low average $RMSD, \sim 1.7\,\text{Å}$. At the higher temperature, the average value of the $RMSD$ is slightly excited but remains lower than 3 Å for both systems. Things differ when the isolated monomers are considered. At ambient temperature, after a first drift that occurs for both systems within 40 to 80 ns, the $RMSD$s show a steady behavior around the values 3.0 Å and 3.3 Å for $\mathcal{T}$ and $\mathcal{M}$, respectively. This conformational departure with respect to the X-ray

Table 6.1. Radius of gyration, volume and intrinsic compressibility

		T = 300 K		
System		R_g (Å)	V (Å^3)	β_T (10^{-5}MPa^{-1})
tetramer	$\mathcal{M}$	30.9 ± 0.1	8.96 ± 0.02	11.9 ± 0.1
	$\mathcal{T}$	30.9 ± 0.1	8.98 ± 0.02	11.5 ± 0.1
monomer	$\mathcal{M}$	19.3 ± 0.1	8.90 ± 0.04	12.9 ± 0.3
	$\mathcal{T}$	19.4 ± 0.1	8.92 ± 0.04	13.4 ± 0.3

		T = 360 K		
System		R_g (Å)	V (Å^3)	β_T (10^{-5}MPa^{-1})
tetramer	$\mathcal{M}$	31.0 ± 0.1	9.19 ± 0.03	14.2 ± 0.1
	$\mathcal{T}$	31.1 ± 0.1	9.20 ± 0.02	14.4 ± 0.1
monomer	$\mathcal{M}$	19.5 ± 0.2	9.12 ± 0.04	14.9 ± 0.3
	$\mathcal{T}$	19.4 ± 0.1	9.14 ± 0.05	15.3 ± 0.3

The errors correspond to standard deviation. The values of R_g for chains in the tetramers are identical to those calculated in the isolated monomers.

structure is not surprising since it measures the lack of the packing/confinement of the tetrameric state. Interestingly, at $T = 360\,K$ the mesophilic monomer shows signs of instability ($RMSD > 4.0\,$Å) localized at the curved helix stretch in the proximity of the active site (α1G-α2G). On the contrary the thermophilic monomer remains stable even at this high temperature.

Table 6.1 reports the radius of gyration, the volume per atom and the intrinsic compressibility data for all four systems and for the two temperatures. Within the error, the radius of gyration and the atomic volume are the same between the different orthologues, in either monomeric or tetrameric form. Thus, the enhanced thermal stability of $\mathcal{T}$ does not correlate to an improved atomic packing [198]. What we do note, however, is an important difference in the compressibility values. As noted for $RMSD$, the monomers behave differently with respect to the tetramers. Indeed the intrinsic compressibility of the monomers, β_T, is higher than that of the tetramers as a signature of larger "breathing" modes and possibly a decreased stability [190]. Moreover, this difference is larger for the $\mathcal{T}$ system and, as we will discuss widely later in the text, this indicates a strong, specific effect of the assembling in the tetrameric state of this species.

Rigidity at atomistic length scales. We start investigating protein flexibility at the atomistic length scale. Neutron scattering experiments by M. Tehei et al. [63], probing the atomistic diffusion at small time scale (150 ps) of a hyperther-

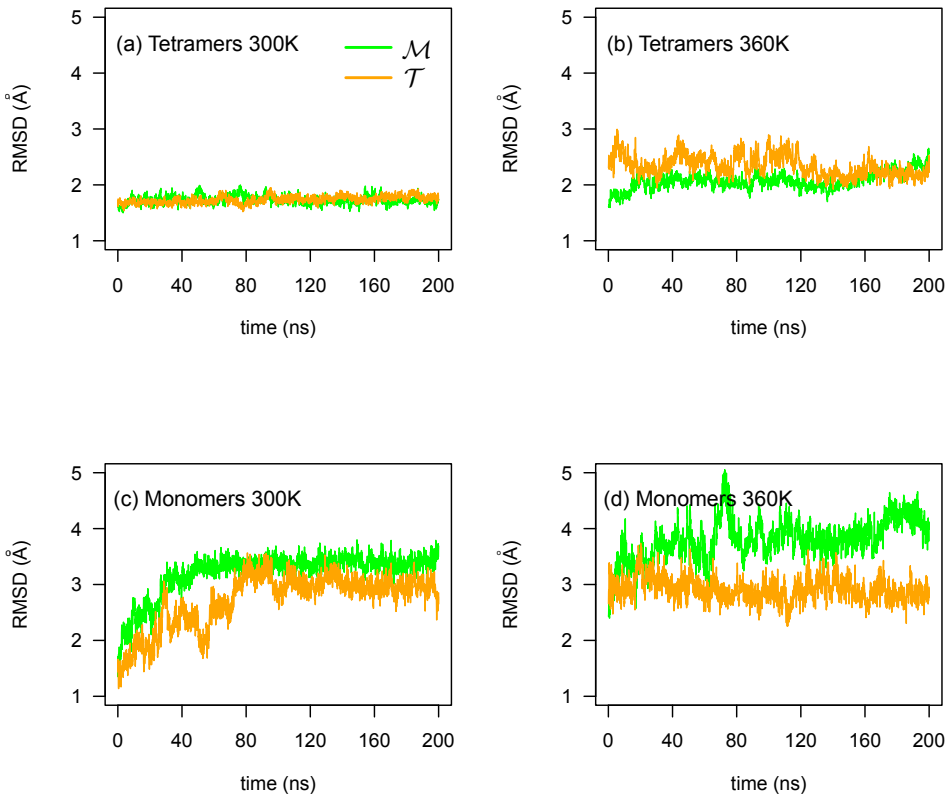

Figure 6.3. Root mean square deviation of C_α atoms at $T = 300\,K$ and $T = 360\,K$. (a) and (b) correspond to the two tetrameric MDH homologues while (c) and (d) show the respective timelines for the monomers simulated in an isolated form.

mophilic malate (from *Methanococcus jannaschii*) as compared to a homologous mesophilic lactate dehydrogenase (from *Oryctolagus cunniculus*), suggested for the former a lower temperature dependence of atomic flexibilities. This behavior was also confirmed in silico by larger-scale simulations for the exact same pair of proteins [77].

Here, we use the root mean square fluctuations of C_α atoms or $RMSF$ (see Methods), to examine to what extend our two $\mathcal{M}$ and $\mathcal{T}$ orthologues, that have a much larger sequence and structure identity than the pair mentioned above, comply to the previous observations. First, as opposed to what found in [77], the average $RMSF$ over all C_α atoms on short and long time-scales (up to 10 ns) shows that the tetrameric $\mathcal{M}$ protein is more flexible than the tetrameric $\mathcal{T}$ independently of the temperature. When we look at the isolated monomers, at ambient temperature the relation reverses, and $\mathcal{M}$ is now more rigid than $\mathcal{T}$. At the higher temperature of T=360 K, the $RMSF$ values of the $\mathcal{M}$ monomer become now larger than those for $\mathcal{T}$, being this an extra indication of its kinetic instability (Fig. 6.4). More

importantly, by considering the shift due to the temperature increase, we also probe that our systems respond similarly (see bottom panel of Fig. 6.4). In other words, the excitation of the atomistic fluctuations in the folded state (as sampled by our simulations) does not mirror the different thermal stabilities of our tetramers. A similar response to temperature increase was also reported for other homologues [78].

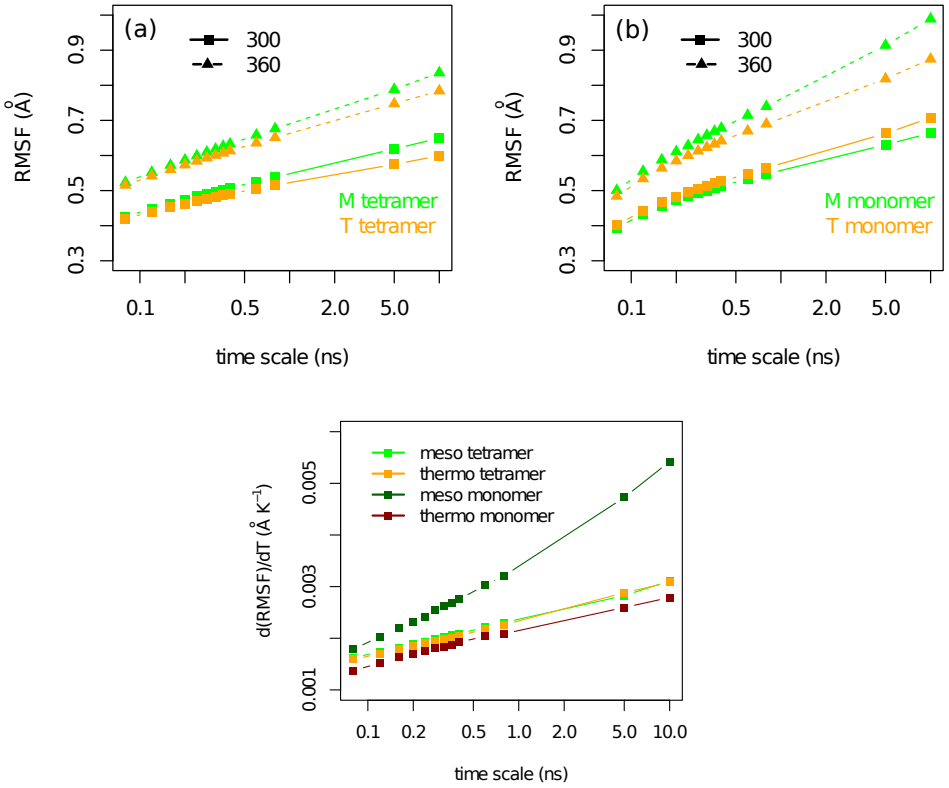

Figure 6.4. Flexibility at different time-scales. (Top panel) Average $RMSF$ for $\mathcal{M}$ and $\mathcal{T}$ (a) tetramers and (b) monomers for both 300 K and 360 K. (Bottom panel) Temperature dependence of RMSF. Derivative of the average RMSF w.r.t. temperature.

Tetramer rigidity. As a further step, we inquire into the rigidity of the proteins at a molecular scale by describing the conformational landscape explored by the systems at ambient temperature. The conformational states visited by the proteins are individuated by using a clustering procedure (see Methods and [37]) based on the all-heavy-atom $RMSD$. The total number of visited clusters versus time is extracted. The results are shown in the top panel of Fig. 6.5. When the monomers are isolated we notice that the $\mathcal{T}$ protein visits a larger number of conformational states than the $\mathcal{M}$ variant (see the right panel (c) of Fig. 6.5). Quite

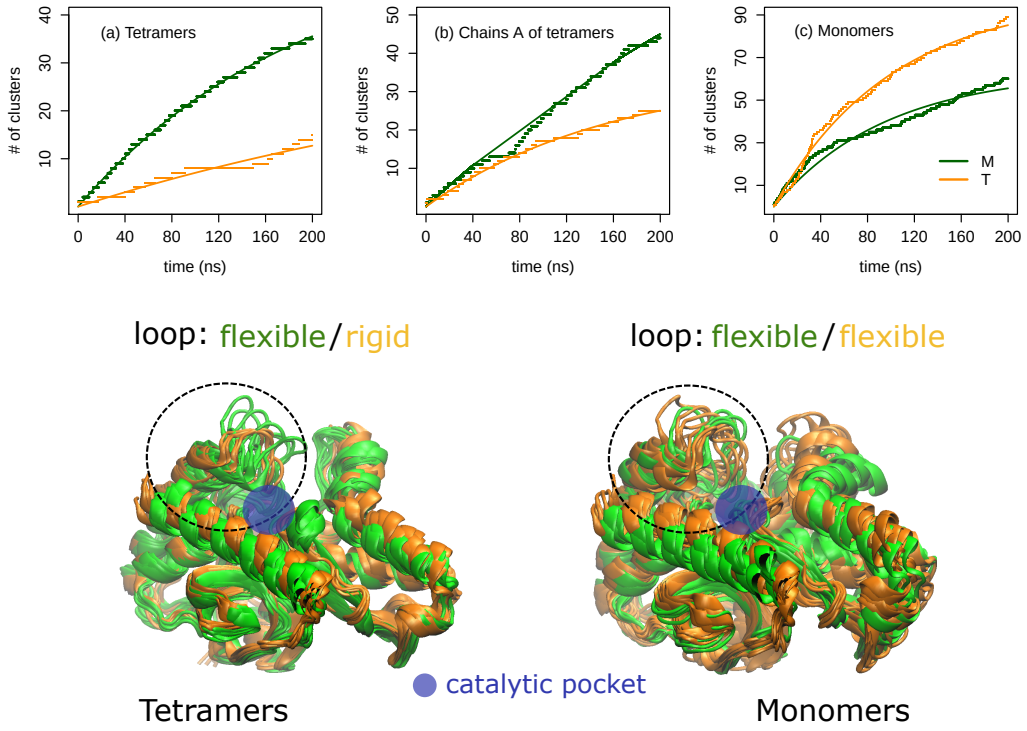

Figure 6.5. Conformational substates. (Top panel) Dotted lines correspond to the number of clusters at ambient temperature versus the simulation time for the (a) two $\mathcal{M}$ and $\mathcal{T}$ tetramers, (b) chains A of the $\mathcal{M}$ and $\mathcal{T}$ tetramers and (c) $\mathcal{M}$ and $\mathcal{T}$ isolated monomers. Straight lines correspond to an exponential evolution fit of the form $N = N_\infty(1 - e^{-t/\tau})$. (Bottom panel) Thermophilic protein is in orange and mesophilic in green color. (Left) Cluster leaders for chain A of the tetramers' clustering and (right) cluster leaders of the monomer's clustering. Notice the anchoring of the loop at the active site of tetrameric $\mathcal{T}$ on the left with respect to the flexible loop of its isolated monomer on the right.

surprisingly the situation reverses when the simulations of the tetrameric systems are considered, with $\mathcal{T}$ being significantly more rigid and exploring a smaller number of conformational states (see left panel (a) of Fig. 6.5). In order to quantify the effect of rigidification upon oligomerization, we have performed the clustering along the trajectories of the tetrameric systems but considering only one chain in the calculation. The results for chain A of $\mathcal{M}$ and $\mathcal{T}$ tetramers are given as an example in Fig. 6.5 (b). For the $\mathcal{T}$ species the effect is quite important, indeed when in the tetrameric assembly the number of accessible states of the single chain is reduced by a factor of three.

To quantify our finding we report in Table 6.2 the estimated maximum number

of clusters and the characteristic time of their saturation by fitting our data to an evolution function $N = N_\infty(1 - e^{-t/\tau})$.

Given that at ambient temperature the proteins are stable, the reported differences quantify the relative flexibility of the proteins in their folded states. Comparing Figures 6.5(b) and (c), it is clear that when the four monomers of each system come together the interfacial interactions between them have a rigidification effect on the protein matrix by reducing the number of accessible conformations. This effect is especially pronounced for the thermophilic variant $\mathcal{T}$.

The characteristic saturation times reported in Table 6.2 signal also the different kinetics of the proteins across the network of states. In a previous study of two homologous G-domain proteins [37], we found that the collective motion of the hyperthermophilic variant has a highly frictional character, i.e the native state is composed of multiple local minima separated by higher kinetic barriers that result in a slow internal diffusion with respect to that of its mesophilic counterpart. To quantify this diffusivity, the motion of the proteins with respect to a given collective variable (CV) was associated to a diffusion coefficient D. Within the harmonic approximation, D is given by the fluctuations of the CV divided by its characteristic decorrelation time, $D = \langle \delta X^2 \rangle / \tau_{corr}$. Interestingly, we herein agree with our previous findings, in fact for the tetramers the internal dynamics of $\mathcal{T}$ is about 20% slower than that of $\mathcal{M}$. The data are shown in Table 6.2 for two CVs, namely the fraction of native contacts Q and the fraction of native torsion angles n_t (see Methods). Just as for the G-domains, the fluctuations of the CVs are comparable between the two tetrameric systems but the decorrelation times are systematically larger for $\mathcal{T}$ which reflects higher kinetic traps for this system. Again, the situation reverses for the case of the isolated monomers. The decorrelation time becomes now small and the diffusion coefficient larger for $\mathcal{T}$. We note that two other tested variables, namely radius of gyration and $RMSD$ follow the same trend (data not shown).

Rigidity and ion-pair networks. The regions of the single domain that get mostly stiffened upon assembling in the tetrameric state are shown in the bottom panel of Fig. 6.5 where the clustered conformations from the simulations at ambient temperature are presented. Conformations visited by chains A of the tetramers are shown on the left of the figure while the conformations visited by the isolated monomers are represented on the right. The mesophilic $\mathcal{M}$ and the thermophilic $\mathcal{T}$ structures are overlapped and represented in green and orange, respectively. The largest effect is localized at the level of the loop at the entrance of the catalytic pocket. In fact, in its tetrameric form the $\mathcal{T}$ homologue maintains this loop in a closed state during all the simulation time, while in the isolated $\mathcal{T}$ monomer

Table 6.2. Clustering: maximum number of clusters N_∞ and their characteristic saturation time τ. Diffusion: magnitude of CV fluctuations $\langle \delta X^2 \rangle$, CV decorrelation time τ_{corr} and the resulting diffusion coefficient D.

		Clustering		Diffusion					
		$RMSD$		Q			n_t		
System		N_∞	τ (ns)	$\langle \delta X^2 \rangle$	τ_{corr} (ns)	D	$\langle \delta X^2 \rangle$	τ_{corr} (ns)	D
tetramer	$\mathcal{M}$	55	190	7	3.3	21	12	4.7	26
	$\mathcal{T}$	36	454	6	9.2	6	15	7.2	21
monomer	$\mathcal{M}$	63	96	26	6.4	40	12	13.3	9
	$\mathcal{T}$	100	101	27	4.7	57	14	7.2	20

the same loop is significantly more flexible. The respective region in $\mathcal{M}$ is equally flexible in either monomeric or tetrameric form. We note that this behavior applies for the loops of all chains of the two tetrameric homologues. This finding, as we will discuss later, might be important to dissect functional conformational changes occurring at the optimal working temperature of the thermophile. In fact, even if our tetrameric MDHs, as other member of the LDH-like family, have never been crystallized in the presence of substrate analogues, the crystal structures of *apo* and *holo* LDH proteins [232] suggest critical conformational changes at level of this biding-site loop.

At this point, the first question that arises is why in the $\mathcal{T}$ tetramer the loop is anchored down. The answer is found in the network of ionic interactions formed between this stretch of amino acids and the inner part of the catalytic pocket. First we note that in both homologues the loop hosts threes basic amino acids, namely Arg81, Lys82 and Arg87 in $\mathcal{M}$ and Arg82, Lys83 and Arg88 in $\mathcal{T}$ (see Fig. 6.6). As can be seen in Fig. 6.6 these residues can form several ion pairs with the acidic residues located inside the pocket. The fine differences between the sequences of the two proteins highlight two important features: first, residue Glu178 in $\mathcal{T}$ doesn't have an acidic analogue in $\mathcal{M}$ since at this position we find a hydrophobic amino acid (Ala177) and secondly, the salt-bridge Arg87-Glu300 in $\mathcal{M}$ doesn't exist in $\mathcal{T}$ (upon structural alignment of the two homologues, Glu300 is replaced by Ala301 in $\mathcal{T}$ while at position 300 we find a positively charged arginine). These two factors are responsible for i) a reduced mobility of the loop in the $\mathcal{T}$ tetramer where the extra salt-bridge with Glu178 rigidifies the region and ii) for the increased flexibility of the loop in $\mathcal{M}$ where the loop motion correlates to an alternating dynamics of ion-pairing of Arg87 with the partners Asp122 or Asp176 and Glu300. It is also worth noting that the arginine in position 81 ($\mathcal{M}$) and 82 ($\mathcal{T}$) is conserved in all

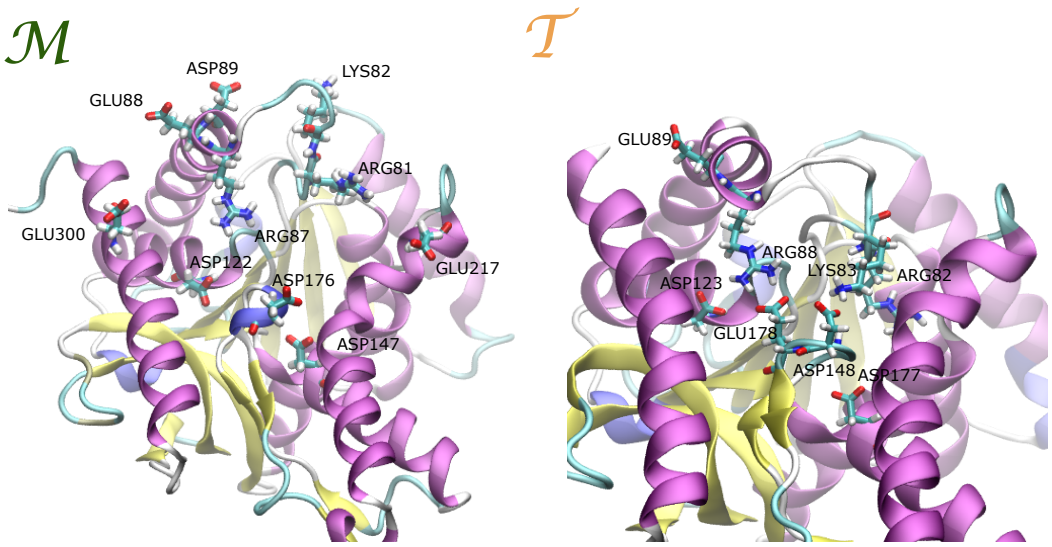

Figure 6.6. Salt-bridges at the external loop of the catalytic pocket. (Left) Mesophilic tetramer. The ion pairs (IPs) that form during the 200ns simulation are Arg81-Glu217, Arg81-Asp147 and Arg81-Asp176, Lys82-Asp89 and Arg87-Asp176, Arg87-Asp122, Arg87-Glu300 and Arg87-Glu88. (Right) Thermophilic tetramer. The IPs that form during the 200 ns simulation are Arg82-Asp148, Arg82-Asp177 and Arg82-Glu178, Lys83-Asp177, Lys83-Glu178 and Lys83-Asp123 and Arg88-Glu178, Arg88-Asp123 and Arg88-Glu89.

MDHs, and its role during the enzymatic activity is well documented [234, 241]. In fact, this basic amino acid binds one of the carboxylates extremities of the substrate. Therefore, during the binding process, the ion-pairs formed by Arg81 (Arg82) must be replaced by the functional interactions with the substrate.

The second question that arises by looking at Fig. 6.5 is why for the thermophile $\mathcal{T}$ the loop is rigid in the tetramer and flexible in the monomer. At the molecular level this is due to a conformational funnel that constrains the residue Arg88 to closer distances with the partners Asp123, Glu178 and Asp177 (see Fig. 6.6). This locked state is caused by an acquired global rigidity of the protein matrix upon oligomerization. In fact, we verified that even by removing the motion of the loop, the number of conformational substates visited by the thermophilic tetramer is always smaller than for its mesophilic variant and the isolated monomer.

Concluding, we have verified that the packing of the interface causes a global rigidification of the $\mathcal{T}$ tetramer resulting in the anchoring of the binding site loop. The consequence of this locked state on the protein-substrate binding process will be addressed in the Discussion.

6.2.2 Forces at the interfaces

Electrostatics and hydrophobicity. We now focus on the cause of the stiffening of the protein matrices by dissecting the energetics of the interdomain interfaces. In order to most effectively study the interfacial interactions we have decomposed the interfaces into three different types comprising m, d and c the definition of which can be seen in Figure 1. Each one of them is the sum of two different monomer-monomer interfaces. For example the interface m is the sum of the interfaces between chains A and B as well as D and C.

The first three columns of Table 6.3 report the fraction of surface area of hydrophobic - hydrophobic, hydrophilic - hydrophilic and hydrophobic - hydrophilic (mixed) contacts along each of the three interfaces as estimated via Voronoi tessellation of the space [193] [2]. The first observation is that m and d interfaces are favored hydrophobically for $\mathcal{T}$ while they are favored hydrophilically for $\mathcal{M}$. This is in line with certain structural facts; $\mathcal{T}$ has, per chain, 10 more hydrophobic a.a. than $\mathcal{M}$ and in particular 1.5 times more along each of the m and d interfaces within a distance of 4.5 Å from hydrophobic a.a. of the opposite side. While both systems are slightly negatively charged the mesophilic has, per chain, 9 more charged a.a. than the thermophilic one. However, looking back in Table 6.3, the frustration (percentage of mixed surface) along the m interface is, for both systems, comparable with the sum of hydrophobic and hydrophilic surfaces. That roughly means that quantity doesn't matter there; it is rather the quality and specificity of interactions along this interface that result to a favourable free energy for the bound state of either system.

In this regard, the last two columns of Table 6.3, report the number of interdomain hydrogen bonds (HB) and ion-pairs (IP). In the thermophile, interfaces d and c have large numbers of either HBs or IPs, even if the number of charged a.a. is less than in $\mathcal{M}$. It is worth to note that along the three interfaces of the thermophilic variant we have a rather similar number of IPs and HBs, see Fig. 6.7 where the number of IPs is plotted as a function of time for the two systems at T=300 K and 360 K. This uniform interfacial strain could contribute cooperatively to the global stiffening of the $\mathcal{T}$ domains in the tetramer. We also stress that by increasing temperature, while the number of interfacial IPs tends to decrease in the $\mathcal{M}$ protein, it increases in $\mathcal{T}$ (see also Table 6.3 and Fig. 6.7). The higher connectivity in $\mathcal{T}$ can be explained by the fact that the higher temperature facilitates small energy-barrier crossing events that favor new partnerships, while at the same time the two systems as a whole remain kinetically stable at the explored timescale. The possibility that IP large networking supports conformational changes across

[2] Since the total interfacial area is not exactly the same for the two systems, to facilitate the comparison the surface has been normalized for each of the m, d and c interfaces, i.e. philic-philic, phobic-phobic and mixed add up to one (an additional table with the absolute values in $Å^2$ can be seen in the Appendix of this Chapter)

Table 6.3. Electrostatics and hydrophobicity at the interface

		T = 300 K				
	System	S_{phobic} (%)	S_{philic} (%)	S_{mixed} (%)	H-bonds	N_{IP}
$\mathcal{M}$	*m* interface	22 ± 1	28 ± 1	50 ± 2	44± 3	9.8 ± 1.1
	d interface	27 ± 2	48 ± 2	25 ± 2	29± 2	4.0 ± 0.1
	c interface	13 ± 2	43 ± 4	44 ± 4	7± 2	4.4 ± 0.8
$\mathcal{T}$	*m* interface	27 ± 1	23 ± 1	50 ± 1	33± 3	7.6 ± 0.9
	d interface	35 ± 1	38 ± 2	27 ± 2	27± 3	7.4 ± 1.3
	c interface	7 ± 4	64 ± 7	29 ± 3	20± 3	8.2 ± 1.0
		T = 360 K				
$\mathcal{M}$	*m* interface	26 ± 1	28 ± 1	46 ± 2	44 ± 3	8.8 ± 1.2
	d interface	31 ± 2	41 ± 3	28 ± 2	24 ± 3	4.0 ± 0.2
	c interface	16 ± 3	46 ± 4	38 ± 4	9 ± 2	5 ± 0.9
$\mathcal{T}$	*m* interface	28 ± 1	25 ± 1	47 ± 1	37 ± 3	10.4 ± 0.9
	d interface	34 ± 3	38 ± 3	28 ± 3	27 ± 3	8.8 ± 1.2
	c interface	10 ± 5	56 ± 9	34 ± 5	19 ± 2	10 ± 1.2

Errors correspond to standard deviation.

the interfaces during the enzymatic activity at high temperature is an appealing hypothesis to be investigated in future work.

Free energy of domain separation We, additionally, estimated the strength of the interfacial matches by performing free energy calculations. Namely, we computed the work needed to separate the different domains of the two systems (see Methods). For computational reasons, we only considered the separation of two domains each time, and for our calculations we employed a coarse-grained model (MARTINI v2.1 with polarizable water [143]). The results are shown in Fig. 6.8. The leftmost panel (Fig. 6.8(a)) shows the potential of mean force to separate domain A from domain B. There is, clearly, a larger binding free energy for $\mathcal{T}$ ($\Delta G_{\mathcal{T}} = 18 \pm 2$ kcal/mol) as compared to $\mathcal{M}$ ($\Delta G_{\mathcal{M}} = 12 \pm 2$ kcal/mol). On the other hand, the binding free energy for domains A and D are, within error, comparable for the two systems with a slightly larger value for the $\mathcal{M}$ homologue, $\Delta G_{\mathcal{T}} = 12 \pm 2$ kcal/mol and $\Delta G_{\mathcal{M}} = 10 \pm 1$ kcal/mol (Fig. 6.8(b)). Finally, the binding energy of domains AC are $\Delta G_{\mathcal{T}} = 3.5 \pm 1$ kcal/mol and $\Delta G_{\mathcal{M}} = 2.3 \pm 0.7$ kcal/mol. Unfortunately, it is computationally very expensive to get a well converged potential of mean force for the dimer-dimer separation, that is the separation of dimer (A+B) from (C+D). However, our preliminary results indicate

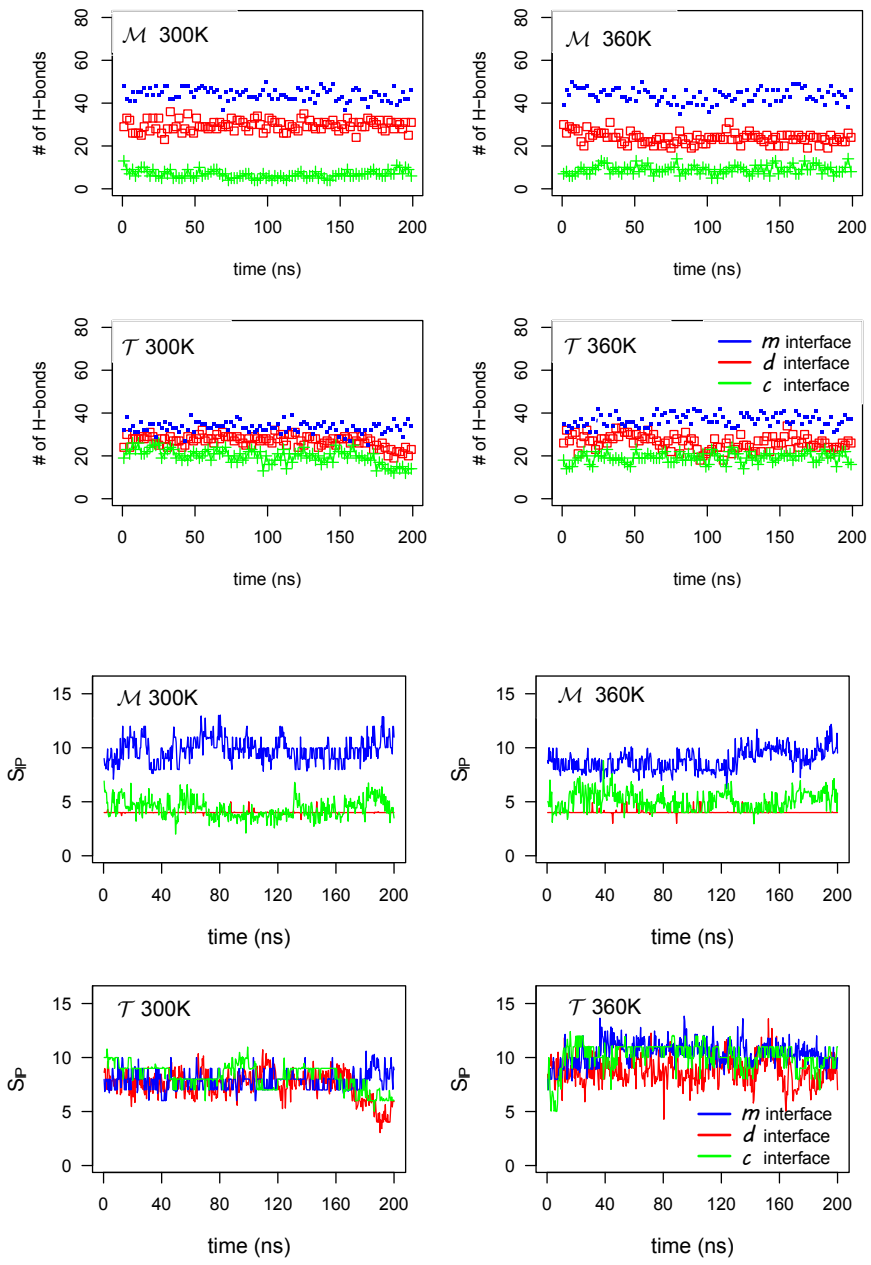

Figure 6.7. Electrostatics at the interface. (Top panel) Number of interdomain protein-protein hydrogen bonds at the 3 different interfaces of the tetramers. (Bottom panel) Number of interdomain protein-protein ion-pairs on the 3 different interfaces of the tetramers at 300 K and 360 K.

a larger binding free energy for $\mathcal{T}$ than that for $\mathcal{M}$ which is in line with previous experimental indications [235].

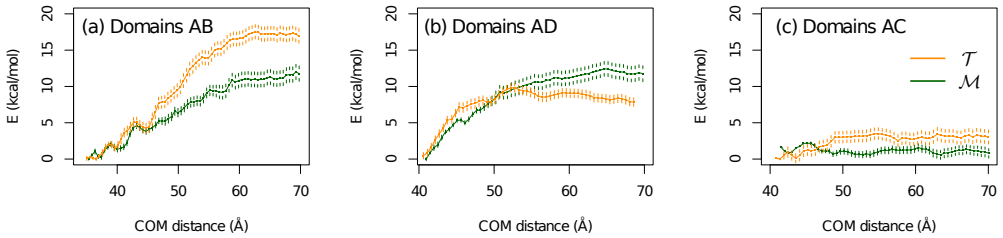

Figure 6.8. Potential of mean force profiles of domain separation. (a) Separation of domain A from B (see also Fig. 6.2), (b) Separation of domain A from D and (c) Separation of domain A from C

6.3 Discussion

The melting temperatures of the mesophilic MDH from *Chlorobium vibrioforme* and the thermophilic MDH from *Chloroflexus aurantiacus* are 52.6 °C (325.75 K) and 67.8 °C (340.95 K), respectively [235]. Here, although the highest temperature in our simulations (360 K) is above both, we do not observe any signs of kinetic instability for the tetramers in the explored time scale. Interestingly though, at this temperature the isolated mesophilic monomer is less stable than its thermophilic homologue as revealed by higher conformational fluctuations (see Fig. 6.3). Moreover, the region where this instability is localized is the curved helix (α1G-α2G), an important portion of which is part of the catalytic pocket. This finding indicates that the isolated domains set possibly a baseline in the tetramers' thermal resistance but extra stability is gained by domain-domain interactions [174].

Nevertheless, the important differences in the dynamics of the two homologues are revealed upon oligomerization. The main finding of our work is that in the tetrameric state the protein domains are systematically more rigid than in the isolated monomeric state and more importantly the rigidification process is very pronounced for the thermophilic variant. This was probed at both the atomistic and the molecular length-scales as well as considering volumetric properties. The $\mathcal{T}$ tetramer appears to be less compressible than $\mathcal{M}$, a relation that reverses when the monomers are considered isolated. In agreement to that, the internal motion of the $\mathcal{T}$ tetramer is slowed down with respect to $\mathcal{M}$ as effect of higher kinetic traps in the conformational landscape, a relation that also reverses for the isolated monomers. This picture is complimented with our cluster analysis of the explored conformational space; upon oligomerization both systems get stiffer yet the $\mathcal{T}$ tetramer is confined in a much smaller conformational space than that of $\mathcal{M}$.

By analyzing the X-ray structure, Dalhus and coworkers [235] forecasted a reduced flexibility for $\mathcal{T}$ on the basis of the observed surplus of proline, however our finding points the attention to a more cooperative effect due to the interfacial

packing. How the surplus of proline amino acids would contribute to enhance the domain rigidity upon oligomerization is an open question and relates to how interfacial packing transmits rigidity across the protein matrix. This will be the focus of a forthcoming work.

A structural comparison identified a few other factors as responsible for the increased thermal stability of $\mathcal{T}$, for example the increased number of alanine and aromatic residues on the m interface [235]. This finding in conjunction with our estimated gap between the domain binding free energies of $\mathcal{T}$ versus $\mathcal{M}$ (Fig. 6.8(a)) reveals the importance of hydrophobic interactions along this interface [65].

For the other two types of interfaces, namely d and c, the thermophilic tetramer $\mathcal{T}$, even if depleted in charged amino-acids with respect to the $\mathcal{M}$ homologue, presents an higher number of both ion-pairs and hydrogen-bonds (see Table 6.3 and Fig. 6.7,). However, for these interfaces, the free energy calculations do not mark any meaningful stability gaps between the two homologues. The role of ionic groups at the interfaces of *Ca* MDH was investigated by single point mutations obtaining different results depending on the targeted amino acids [242, 243]. It was shown that when residue Glu25 that is located at the c interface was mutated to both a lysine and a glutamine the thermal stability at pH 7.5 is only slightly decreased, on the other hand when the Glu165, that belongs to the same network of ionic interactions (see Fig. 6.9), is mutated in a similar way, the thermal stability of the protein increases of about 25 degrees without compromising the catalytic activity. Our simulations showed that at the c and d interfaces $\mathcal{T}$ not only has a higher number of IPs and HBs but also a higher degree of connectivity. The charged residues in $\mathcal{T}$ are placed along the interfaces in such a way so that they are topologically able to interact with multiple partners belonging to different domains. The patterns of this connectivity, absent in $\mathcal{M}$, are represented in Fig. 6.9 and could play a role on the protein functionality by controlling long range motion and domain communication during the protein activity. Because of the extension of these ionic interactions, an adequate computational method and model should be used to obtain a more precise estimate of the ionic contribution to the protein stability [92].

The overall rigidity of the thermophilic protein in combination with local sequence specificities have an important consequence on the binding site dynamics. Namely, we refer to the external loop that upon formation of the enzyme-coenzyme-substrate ternary complex (MDH/NAD/NADH) closes to act as a screening gate to the catalytic vacuole. Resent simulations of the dimeric MDH from *Thermus thermophilus* with NAD showed that the loop, having started from an open conformation, closes during the simulation in order to bring key residues in contact with the co-substrate [237]. For the thermophilic malate under study, both the crystal

structure and its 200-nanosecond dynamics are characterized by a constantly closed loop although the protein is coenzyme- and substrate- free. On the other hand, the respective loop in the mesophilic protein undergoes several openings and closings during the course of the simulation.

To quantify this observation, we used as reference a third orthologous protein, a lactate dehydrogenase (LDH), whose crystal structure has been fully resolved in both holoenzyme and apoenzyme conformation [232]. For the simulated trajectories we calculated the $RMSD$ of the backbone atoms of the residues that form the catalytic pocket for $\mathcal{M}$ and $\mathcal{T}$ with respect to both the *apo* and *holo* form of LDH. The results can be seen in Fig. 6.10. While the mesophilic tetramer switches intermittently between conformations close to the LDH *apo* form and LDH *holo* form, the thermophilic tetramer remains rigidly around a conformation that mostly resembles the *holo* form of LDH. The stiffness of such a region might explain the reduced activity of the thermophilic protein at ambient temperature. However, the observed behavior of the loop does not depend on temperature on the explored time scale. The key question is then, how the loop behaves at the working temperature of the $\mathcal{T}$ MDH? According to the "corresponding states" view, at high temperature one would expect the loop to acquire the necessary flexibility to facilitate the binding process. Clearly, a precise characterization of this gating requires to evaluate the kinetic barrier separating the open and close states as well as to evaluate the temperature effect on the transition path. Moreover, it is also possible that the opening of the loop requires a cooperative role for the coenzyme, whose charged groups could trade the stability of the IP network that anchors down this region for an optimized co-enzyme substrate.

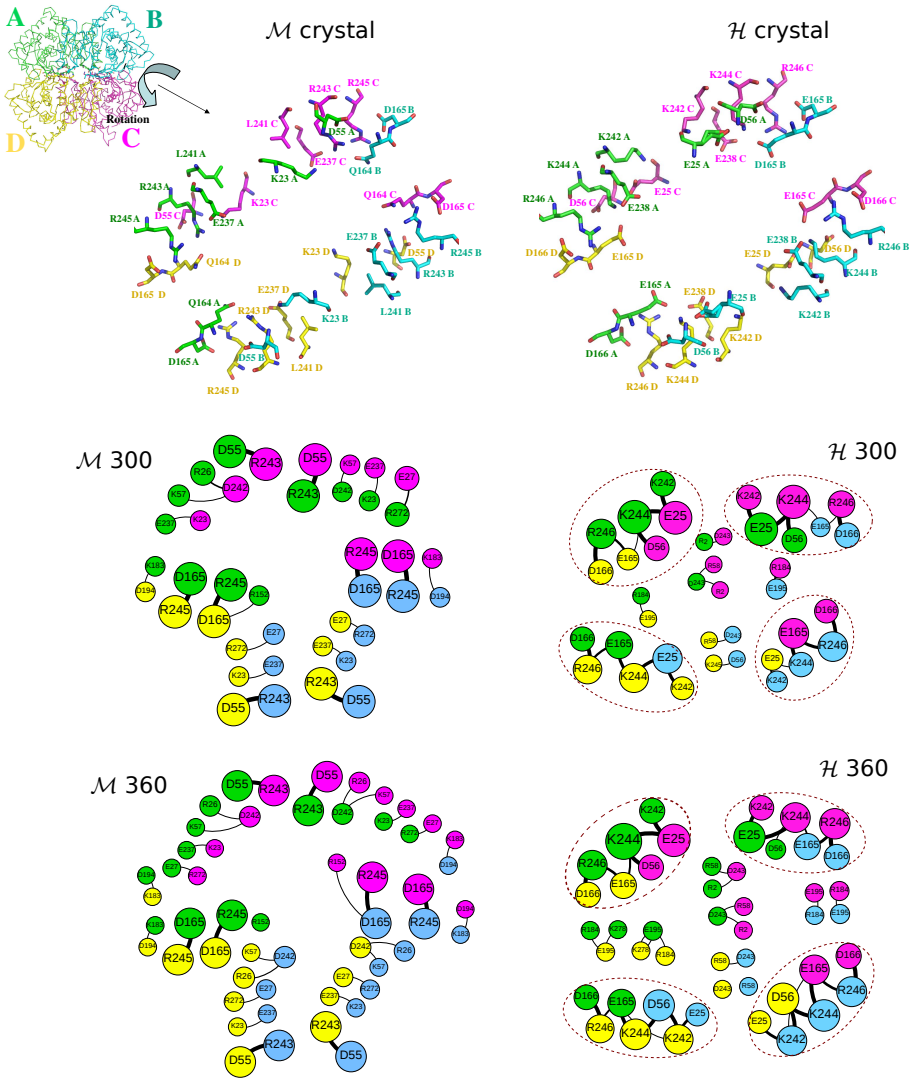

Figure 6.9. Ionic interactions at the d and c interfaces. (Top panel) Molecular representations of the charged residues among all domains. (Middle panel) Network representation of only d and c interfacial IPs for $\mathcal{M}$ (left) and $\mathcal{T}$ (right) at 300 K. The nodes represent charged a.a. that form salt-bridges between different domains. The node-size is proportional to the time the a.a. formed a salt-bridge with any other a.a. (the largest size is equal to 100% of the time). The links between the nodes represent salt-bridge formation with thickness proportional to the time the salt-bridge was formed (the largest size is equal 100% of the time). The coloring code refers to the four different domains, green for domain A, blue for B, magenta for C and yellow for D. (Bottom panel) Network of IPs as above but for 360 K. Notice the connectivity between the clusters of $\mathcal{T}$, absent in $\mathcal{M}$.

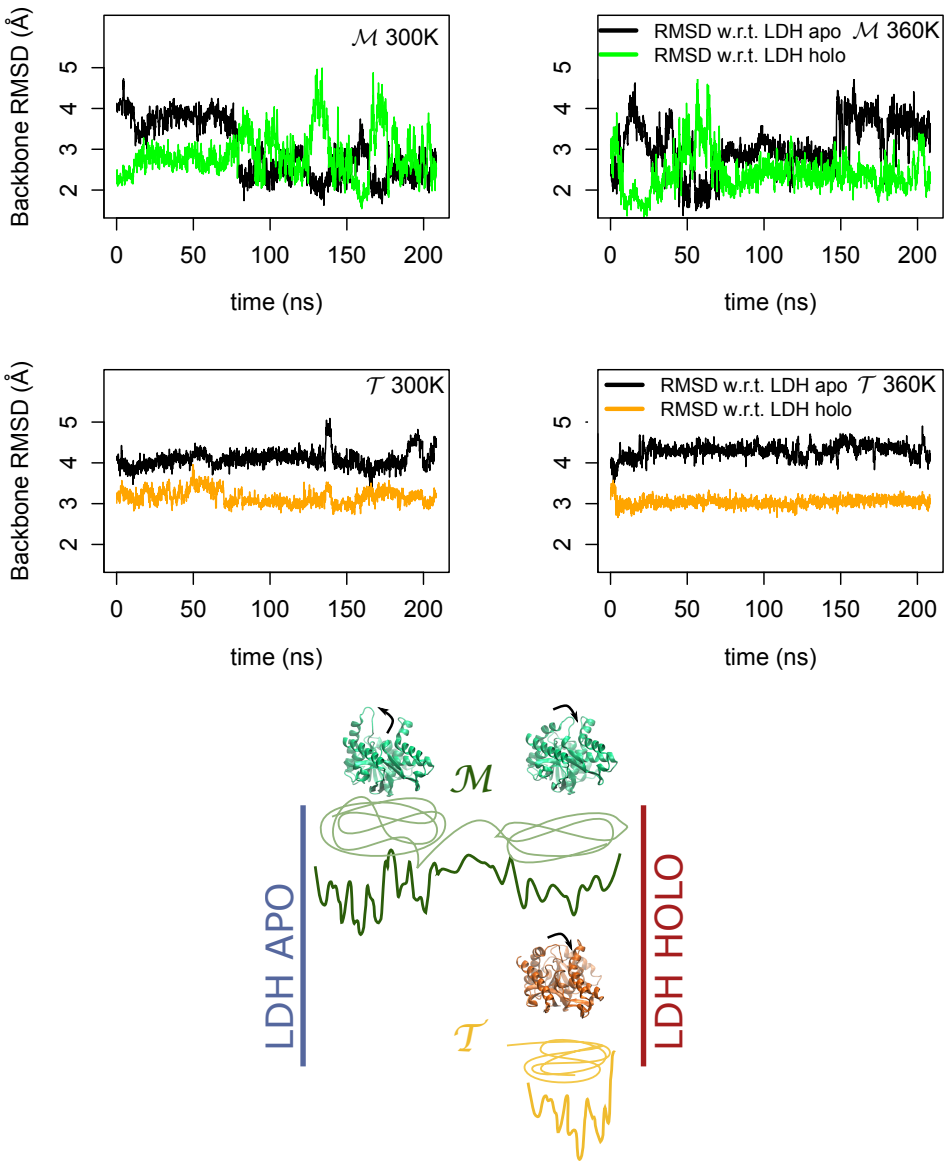

Figure 6.10. Catalytic pocket dynamics. In the upper part of the figure we report the $RMSD$ computed using the backbone heavy atoms of the residues that form the catalytic pocket of $\mathcal{M}$ (upper graphs) and $\mathcal{T}$ (bottom graphs) w.r.t. those of the *apo* and *holo* forms of lactate dehydrogenase at 300 K (left) and 360 K (right). In the lower part of the figure we present a pictorial representation of the conformational states accessible by the proteins when considered as function of the distance with respect to LDH *apo* and *holo* conformers. In the green we sketched the two states visited by the $\mathcal{M}$ MDH, one characterized by an open conformation of the binding site loop and one associated to a close state. The $\mathcal{T}$ MDH (orange) is instead tightly confined in a closed state.

Appendix

Table 6.4. Electrostatics and hydrophobicity at the interface. This Table presents the values of Table 6.3 in absolute values (Å^2).

	$T = 300\,\mathrm{K}$				
System	S_{phobic} (Å^2)	S_{philic} (Å^2)	S_{mixed} (Å^2)	H-bonds	N_{IP}
m interface	629 ± 37	780 ± 31	1391 ± 43	44± 3	9.0 ± 1.2
$\mathcal{M}$ d interface	431 ± 27	683 ± 25	384 ± 22	29± 2	4.0 ± 0.2
c interface	50 ± 6	155 ± 19	168 ± 15	7± 2	1.4 ± 0.7
m interface	724 ± 29	644 ± 28	1340 ± 39	33± 3	7.5 ± 0.9
$\mathcal{T}$ d interface	577 ± 25	634 ± 27	458 ± 29	27± 3	5.9 ± 1.0
c interface	63 ± 30	223 ± 40	140 ± 43	20± 3	7.8 ± 1.0
	$T = 360\,\mathrm{K}$				
System	S_{phobic} (Å^2)	S_{philic} (Å^2)	S_{mixed} (Å^2)	H-bonds	N_{IP}
m interface	744 ± 41	825 ± 42	1378 ± 50	44 ± 3	7.3 ± 1.0
$\mathcal{M}$ d interface	462 ± 32	609 ± 38	416 ± 28	24 ± 3	4.0 ± 0.3
c interface	65 ± 10	193 ± 21	163 ± 22	9 ± 2	3.5 ± 0.9
m interface	782 ± 34	675 ± 36	1285 ± 47	37 ± 3	9.2 ± 0.9
$\mathcal{T}$ d interface	531 ± 45	606 ± 49	481 ± 34	27 ± 3	7.3 ± 1.0
c interface	38 ± 22	198 ± 23	122 ± 37	19 ± 2	9.6 ± 1.2

Errors correspond to standard deviation.

Structure and energetics of the hydration shell

Our study of the thermophilic protein (CaMDH) is based on a recently, *de novo* resolved structure at 1.7Å resolution, in its dimeric form (chains A and D in Fig. 6.2) [236]. The crystallographic structure is resolved along with 945 water molecules enveloping the protein up to a distance of 8.5 Å. Amongst them, 892 (94%) are within the first hydration shell, $d < 4.5$ Å. As already discussed in Ref. [236], approximately one forth of these waters are clustered together in pentameric and hexameric structures. The hydration content and the peculiar presence of closed structures therein were suggested to correlate with the thermophilic character of the protein. In this short discussion, we present preliminary results of a systematic investigation of this issue. First, we extend the definition of polygons to closed structures with vertices from 3 to 7 (i.e. from trigons to heptagons) and compare them with polygon formation in the hydration shell of the protein during the Molec-

ular Dynamics (MD) simulations. Herein, we limit ourselves to the thermophilic system due to the smaller amount of crystal waters and polygons in the crystal structure of the mesophilic homologue that doesn't allow a analogous comparison between simulation and experiment.

In Fig. 6.11*A* we give the number of polygons found in the crystal for each polygon type. Smaller polygons contained in bigger ones are not taken into account and the distance between two neighboring oxygens has been set to 3.0 Å. We observe the substantial lead of pentagons over the rest of the closed structures. Then follow trigons, that have been reported in the past as the most common crystal-water polygon [244], hexagons, tetragons and heptagons. A detail check of the angular dependency of the hydrogen bonds confirms that most of the detected pentagons are almost planar, something not surprising as the angle of a regular pentagon - 108° - is very close to the angle HOH of a water molecule, 104.45°. At the same time the vast majority of these pentagons form a cap on top of hydrophobic sites of the protein. We also point out the presence of larger structures than heptagons, with the notable case of pentagon and hexagon rings fused together on top of the helices α1G-α1G and αH.

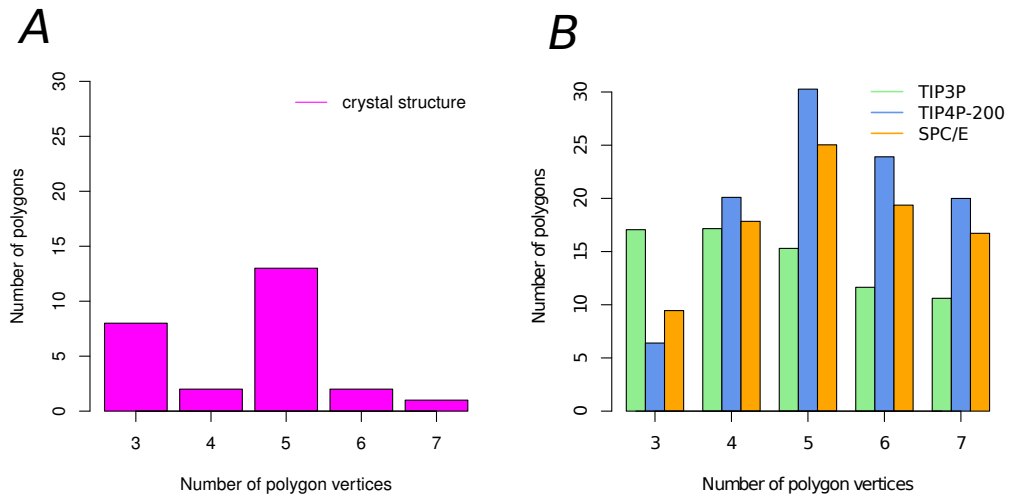

Figure 6.11. Number of polygons in the hydration shell. (*A*) Number of closed polygons within 4.5 Å from the protein in the crystal structure of the dimeric *Ca* MDH. (*B*) Average in time number of closed polygons within 4.5 Å from the protein in the MD simulation of the tetrameric *Ca* MDH ($\mathcal{T}$) for 3 different water models, TIP3P, SPC/E and TIP4P.

In Fig. 6.11*B* we report the average number of closed structures found during MD within the first hydration shell of $\mathcal{T}$ ($d < 4.5$ Å) for three different water

models, namely TIP3P, SPC/E and TIP4P. The polygons were individuated using the same criteria used for the X-ray system: we used only a distance restriction of 3.0 Å between water oxygens without double-counting smaller polygons contained in larger ones. It should be reminded at this point that the results reported for the crystal structure correspond to the dimer whereas the results for the MD refer to the simulated tetramer. Even so, in the MD we find 2.7 to 4 times – depending on the water model – more closed polygons than in the crystal structure. At the same time, the total number of water molecules in the hydration shell is 892 in the crystal versus ∼2000 in the simulation, independently of the water model.

Both the number of hydration water molecules and the number of polygons are in good agreement between experiment and simulation if we take into consideration the fact that in the tetramer 3 more interfaces are present (versus 1 in the dimer) that can accommodate slow water molecules and favor the formation of closed water structures. Interestingly, the two figures agree qualitatively, as well as quantitatively, on the pentagon structure as mostly preferred over the others. That is as long as we consider either the SPC/E or TIP4P water models. Not surprisingly, the TIP3P model doesn't favor larger than 4 structures, probably as a consequence of a weaker HB propensity [245, 246].

Water pentagons around proteins have been reported several times in the past in either crystal structures or simulation [244, 247, 248]. Dating back to 1958, M. Klotz first proposed that water molecules would be organized in pentagonal rings on top of hydrophobic sites [249]. What remains however questionable is the stability of those structures at ambient temperature or above.

M. Teeter pointed out in an earlier experimental work [248], that if most of the oxygens of any closed polygon are not pinned down via hydrogen bonds to the protein they would end up being very labile and contribute unsubstantially to the X-ray scattering. In the crystal structure of *Ca* MDH, 20 out of the 26 polygons discussed here, have more than one oxygens within 3.2 Å from either a protein donor or acceptor and only one of those polygons has no pinning sites at all. This observation also applies if we consider pentagons alone. In the simulations, the polygons of the hydration shell are also, on average, well hydrogen-bonded to the protein with a good agreement between the three water models (see Fig. 6.12). But what is the average lifetime of those structures at ambient temperature, what does their lifetime depend on and do they have any contribution to the stability of the protein?

In the following, we attempt to answer the first two questions reserving a more rigorous approach to the third for a future work. We focus on three of the regions of the protein where pentagons are localized in the crystal structure. Namely, two atoms from the catalytic pocket (region I), two atoms from the helix αH where a

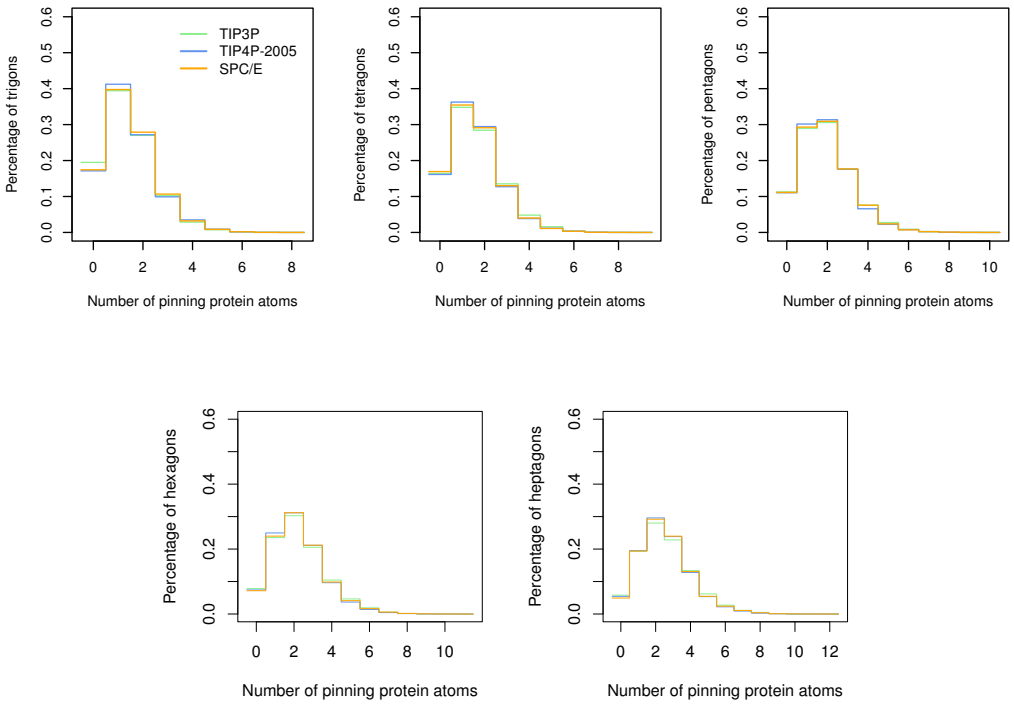

Figure 6.12. Percentage of polygons with a certain number of pinning sites for the simulations with the 3 different water models, TIP3P, TIP4P-2005 and SPC/E. The results are averaged over the total simulation time.

large network of pentagons is anchored (region II) and the oxygen and nitrogen of the Q105 side-chain where a pentagon is present with its plane perpendicular to the line segment connecting the CB atom of A101 and the CD1 atom of I34 (region III). The temperature decrease might render spontaneous the formation of such pentagon rings which separate hydrated aliphatic groups in a single-water layer distance from each other [248]. In Table 6.5 we report the characteristic lifetime[3] of a pentagon structure attached to an atom and the percentage of time that this atom was found to be in contact with a water pentagon.

The first observation is that only the polygons of the catalytic pocket (region I) show long lifetimes, about 20-50 ps, depending on the water model. This high stability is probably caused by the weak exposure of the water molecules to the bulk solution. On the contrary, the other two regions are on the surface of the protein and formation of instantaneous pentagons is favored even if the structure is not stable in time. When focusing on the three water models, the TIP3P shows again its weekness in forming stable hydrogen bonds between water molecules even inside

[3]For the lifetime estimation, a pentagon structure is considered destroyed when at least one water molecule has either detached or been replaced by another.

Table 6.5. Pentagon lifetime

	Characteristic survival time of closed pentagon structures		
Residue/atom	TIP3P	SPC/E	TIP4P
(I) Q143/O	18.2 ± 0.2	55.4 ± 0.7	29.6 ± 0.5
(I) N119/O	18.4 ± 0.1	44.9 ± 0.6	30.4 ± 0.5
(II) A203/O	1.56 ± 0.01	4.3 ± 0.2	5.6 ± 0.2
(II) Q204/OE1	1.40 ± 0.02	1.5 ± 0.1	2.0 ± 0.1
(III) Q105/OE1	1.27 ± 0.03	3.0 ± 0.1	3.2 ± 0.1
(III) Q105/NE2	1.02 ± 0.02	3.2 ± 0.1	3.4 ± 0.2

For the three different water models and for six different atoms (two for the each of the three regions mentioned in the text) we report the characteristic lifetime of a pentagon attached to the atom. The lifetimes are averages over five 4 ns-long stretches extracted from a total trajectory length of 200 ns for the TIP3P and the TIP4P-2005 simulations and 50 ns for the SPC/E simulation.

the catalytic pocket. It is worth noting at this point that the pentagons detected in region III, have a characteristic surviving time of only $3.0 - 3.3$ ps (see for example the SPC/E and TIP4P model) or even less. This low stability is associated to the high mobility of the aliphatic residue I34 located on a turn of the sequence. Thus the hydrophobic chain does not provide a sufficient topological constraint for pentagon survival.

Therefore, the pentagons' lifetime depends on their location with respect to the protein matrix as well as on the temperature. Pentameric rings found in crystal structures that are capping hydrophobic sites on the protein surface, being well exposed to the bulk water, have generally a small lifetime at ambient temperature and are expected to have even smaller lifetimes at higher temperatures. Thus, their contribution to stability is expected to be negligible and their formation on the surface of the protein could just be related to the low temperature at which the crystal is obtained. However, isolated cases of pentagon, or other polygon, structures within internal pockets of the protein with a high lifetime expectancy, such as the structures in the catalytic pocket of *Ca* MDH, may have a substantial contribution to stability. We reserve a more rigorous approach to this issue for a future study.

It is worth noting at this point that ordered waters formed and stabilized by protein hydrophobic groups have been very recently reported for the case of an the antifreeze protein Maxi [247]. The four-helical bundle of the protein has minimal protein-protein interactions but is mainly glued together via an extended sheet of water pentagons in its middle. However, antifreeze proteins work at temperatures close to, and below, the freezing point of water, where such a structured phase of

water can survive for longer times even at a large network scale.

Conclusions

Investigating whether thermophilic proteins are rigid or flexible aims to identify microscopic factors responsible for thermal stability and ultimately to understand the underlying mechanisms sustaining protein functionality at high temperature. As discussed herein, the rigidity/flexibility dispute emanates mainly from two sources. The first is the complex nature of proteins that doesn't allow for a unique definition of flexibility [21]. The second, supported also by the results of this study, is that nature has no panacea to high temperature stress. She is rather more inventive, employing different strategies depending on the pair of homologues or the protein family.

So, are they rigid or flexible? This study puts under the microscope two characteristic study-cases dissecting them by using the gold standard of *in silico* techniques, Molecular Dynamics simulations. The powerful atomistic detail of this technique, along with the possibility to explore different timescales, makes the ambiguity of the term "protein flexibility" immediately obvious.

In the first study-case of a pair of homologous monomeric proteins, a hyperthermophilic and a mesophilic G-domain, thermophilicity correlates to the regular distribution of flexible and rigid parts along the sequence as well as to larger conformational fluctuations at the microsecond timescale. Yet, the mesophilic variant maintains one special region that is more flexible. This is referred to as switch I and it can exist in two very different conformations that dictate whether the enzyme is "on" or "off". Its potentiality to convert easily from one conformation to the other requires, except for activation by another enzyme (allosteric behavior), a loose attachment to the protein body. This feature is ultimately the "Achilles' heel" of the mesophilic protein: when temperature raises, it drives the unfolding.

The second study-case deals with two homologous tetrameric proteins, a thermophilic and a mesophilic malate dehydrogenase. Here, equilibrium motions of the thermophilic variant are "locked" in both global and local length-scale, a behavior resulting from effective electrostatic and hydrophobic interactions on the inter-domain interfaces. The stiffness of the thermophilic protein matrix propagates from the domains' interface to the active site where, in combination with a

few key mutations, the preferential closed conformation of the binding-site loop hinders the accessibility to the catalytic pocket. On the other hand the mesophilic protein is characterized by open/close dynamics of this loop.

From this and other studies, it becomes apparent that for mesophilic proteins, as compared to their thermophilic homologues, thermal stability is compromised for the sake of optimal activity. Efficient enzymatic activity at ambient conditions results from a subtle organization of the protein matrix and its conformational fluctuations. High temperature could easily disrupt this equilibrium by funneling energy on particular protein modes, ultimately leading to unfolding. As a characteristic example, a recent work found that enhanced stability of the human muscle acylphosphatase can be attained via an increase in conformational entropy of the folded state ensemble [22]. However, the stabilized mutants were deficient in enzymatic activity.

In summary, thermophilic proteins are not necessarily rigid, as commonly viewed, although this route might be usually preferred. Our investigation, once again, suggests that clarifying the relationship between stability, flexibility and function requires individuating the key degrees of freedom and exploring all the different time-scales of the associated dynamics [23]. This study also demonstrates how clustering and network approaches, used for the representation of the conformational space explored by proteins, is a powerful analysis tool to inquire into protein flexibility. This toolkit is currently applied to discern the different behavior of a large set of homologous proteins.

Toward the design of thermal resistance. Targeting industrial applications, the current study verifies once more the inextricable relationship between key electrostatic interactions and the "love for heat". In the study of the tetrameric malate dehydrogenases, even if the sequence of the thermophilic orthologue is depleted in charged amino acids, the ionic residues are placed in strategic positions increasing the size of interfacial salt-bridge clusters and strengthening the effective interdomain binding. Traditionally, ion-pairs are considered a structural element that confers local rigidity to the protein matrix, however, as we illustrated by characterizing our study-cases, the reshuffling of ion-pairing and the flexibility of their networking can be source of stability as well as conformational fluctuations.

This investigation supports also an alternative strategy for stabilization. Enhance protein stability can be attained by tuning the extension and distribution of flexible/rigid parts along the primary sequence. That way thermal excitation can be effectively caged and absorbed.

Some open questions and perspectives

Functionality at high temperature. This study clearly opens the question of whether temperature is the only parameter to take into account when describing the functionality of thermophiles. According to the "corresponding states" picture, the mechanism describing the functionality of homologues is identical but characterized by a different activation energy: the flexibility needed for functionality is the same at the corresponding optimal temperature of the hosting organism. However, both cases studied here indicate that other specificities come along with the high temperature functionality. For example considering the EF-Tu and -1α G-domains, we know that the activity of the mesophilic variant is associated to an impressive conformational change localized at level of the switch I region. To the best of our knowledge, it remains unknown whether such a conformational change happens in the hyperthermophile too. We saw that the switch I region for this variant is characterized by the insertion of structural motifs that should affect the conformational changes associated to GTPase activity. So the function of the hyperthermophile at high T is not merely the result of thermal activation of the same degrees of freedom as for the mesophilic variant, but may involve different conformational paths. In the near future, this aspect could be investigated by applying long brute force MD simulations, *ad hoc* enhanced sampling techniques as well as more simplified protein models [24].

Concerning the two homologous tetrameric dehydrogenases, as mentioned before, accessibility of the thermophilic active site is restrained at ambient temperature due to the rigidly closed loop at its entrance. On the contrary the respective loop in the mesophile is characterized by open/close dynamics. However, temperature increase does not manage to activate the thermophilic motion within the explored timescales. It would be interesting to assess how this loop accesses the open state and the potential role of substrate proximity in facilitating the conformational shift. This could be explored either by longer simulations or experimentally, for example, using NMR relaxation dispersion experiments [12].

On the other hand, if the corresponding-states view of flexibility is not generally true it would be very interesting to explore in a more systematic way whether there is a correlation between the thermophilic proteins characterized as rigid and their oligomeric state. This suggestion is inspired from the observed effect of oligomerization on the tetrameric thermophilic protein of this study, as well as from an overview of studies so far (the relevant discussion can be found in Chapter 1).

Another important aspect of protein stability concerns the actual crowding conditions of the cytoplasm. Proteins *in vivo* rarely function isolated and, in fact, protein-protein interactions have been estimated to account for a $\sim 2-4\, k_B T$ of ad-

ditional stabilization [25]. At the same time it has been suggested that thermophilic and mesophilic proteins have different diffusional properties under crowding conditions, something that might substantially contribute to their different thermal stability [26]. An *in house*, newly developed, powerful methodology that combines the OPEP coarse-grained force field for proteins [27] and a mesoscopic solvent representation based on the Lattice Boltzmann method [28, 29] is currently employed in order to address this issue in a more systematic way [30].

At a next level, approaching the subject of protein functionality in a more direct way, the question is how the different conformations explored by proteins affect chemical reactivity. Computationally, this issue would require the application of mixed quantum/classical methodologies (QM/MM).

In a perpetual search for trends, a little less than a decade ago a study suggested that there are two major physical mechanisms for protein thermal stabilization depending on the evolutionary history of the source organism, a "structure-based" and a "sequence-based" one [31]. Proteins from organisms that originated in hot environments (therein archaea) have a much more compact structure and hydrophobic core. On the other hand, proteins from organisms that started as mesophiles and later recolonized a hotter environment (therein bacteria) remain structurally similar to mesophilic homologues but present some sequence substitutions that result in a few key interactions in the final fold. However, this strict classification of evolutionary history depending on the domain of life to which an organism belongs (archaea or bacteria) has no solid ground. In fact, more recent studies showed that the ancestors of bacteria were also thermophiles [32, 33]. At the same time structural studies, although they put things in a first informative perspective, they neglect dynamics and are based on X-ray structures resolved at low temperatures. It is therefore good to complement such investigations, whenever possible, with experimental and computational studies keen to account for the coupled role of dynamics and temperature.

Bibliography

[1] *The American Heritage Dictionary of the English Language*. 4th ed. Houghton Mifflin Harcourt, 2000.

[2] E. Gaughran. "The thermophilic microorganisms". *Bacteriol. Rev.* 3.11 (1947), pp. 189–225.

[3] R. Castenholz. "Thermophilic blue-green algae and the thermal environment". *Bacteriol. Rev.* 33.4 (1969), pp. 476–504. DOI: 10.1007/978-1-4612-6284-8_8.

[4] C. Fraser, J. Gocayne, O. White, M. Adams, R. Clayton, R. Fleischmann, C. Bult, A. Kerlavage, G. Sutton, J. Kelley, J. Fritchman, J. Weidman, K. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T. Utterback, D. Saudek, and C. Phillips. "The minimal gene complement of Mycoplasma genitalium". *Science* 270.5235 (1995), pp. 397–403. DOI: 10.1126/science.270.5235.397.

[5] C. Anfinsen. "The formation and stabilization of protein structure". *Biochem. J.* 128.4 (1972), pp. 737–749.

[6] P. Fields. "Review: Protein function at thermal extremes: Balancing stability and flexibility". *Comp. Biochem. Phys. A* 129.2-3 (2001), pp. 417–431. DOI: 10.1016/S1095-6433(00)00359-7.

[7] K. Teilum, J. Olsen, and B. Kragelund. "Protein stability, flexibility and function". *Biochim. Biophys. Acta* 1814.8 (2011), pp. 969–976. DOI: 10.1016/j.bbapap.2010.11.005.

[8] R. Jaenicke. "Protein structure and function at low temperatures". *Phil. Trans. R. Soc. B* 326.1237 (1990), pp. 535–551. DOI: 10.1098/rstb.1990.0030.

[9] P. Agarwal. "Enzymes: An integrated view of structure, dynamics and function". *Microb. Cell Fact.* 5 (2006). DOI: 10.1186/1475-2859-5-2.

[10] S. Teague. "Implications of protein flexibility for drug discovery". *Nat. Rev. Drug. Discov.* 2.7 (2003), pp. 527–541. DOI: 10.1038/nrd1129.

[11] S. D'Amico, J.-C. Marx, C. Gerday, and G. Feller. "Activity-stability relationships in extremophilic enzymes". *J. Biol. Chem.* 278.10 (2003), pp. 7891–7896. DOI: 10.1074/jbc.M212508200.

[12] M. Wolf-Watz, V. Thai, K. Henzler-Wildman, G. Hadjipavlou, E. Eisenmesser, and D. Kern. "Linkage between dynamics and catalysis in a thermophilic - mesophilic enzyme pair". *Nat. Struct. Mol. Biol.* 11.10 (2004), pp. 945–949. DOI: 10.1038/nsmb821.

[13] G. Somero. "Proteins and temperature". *Annu. Rev. Physiol.* 57 (1995), pp. 43–68. DOI: 10.1146/annurev.ph.57.030195.000355.

[14] R. Jaenicke. "Do ultrastable proteins from hyperthermophiles have high or low conformational rigidity?" *Proc. Natl. Acad. Sci. U.S.A.* 97.7 (2000), pp. 2962–2964. DOI: 10.1073/pnas.97.7.2962.

[15] J. Woodley. "Protein engineering of enzymes for process applications". *Curr. Opin. Chem. Biol.* 17.2 (2013), pp. 310–316. DOI: 10.1016/j.cbpa.2013.03.017.

[16] G. D. Haki and S. K. Rakshit. "Developments in industrially important thermostable enzymes: A review". *Bioresource Technol.* 89.1 (2003), pp. 17–34. DOI: 10.1016/S0960-8524(03)00033-6.

[17] C. M. Dobson. "Protein-misfolding diseases: Getting out of shape". *Nature* 418.6899 (2002), pp. 729–730. DOI: 10.1038/418729a.

[18] A. Thangakani, S. Kumar, D. Velmurugan, and M. Gromiha. "How do thermophilic proteins resist aggregation?" *Proteins* 80.4 (2012), pp. 1003–1015. DOI: 10.1002/prot.24002.

[19] L. Cruzeiro. "Why are proteins with glutamine- and asparagine-rich regions associated with protein misfolding diseases?" *J. Phys.: Condens. Matter* 17.50 (2005), pp. 7833–7844. DOI: 10.1088/0953-8984/17/50/005.

[20] H. Wijma, R. Floor, and D. Janssen. "Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability". *Curr. Opin. Struct. Biol.* 23.4 (2013), pp. 588–594. DOI: 10.1016/j.sbi.2013.04.008.

[21] T. Kamerzell and C. Middaugh. "The complex inter-relationships between protein flexibility and stability". *J. Pharm. Sci.* 97.9 (2008), pp. 3494–3517. DOI: 10.1002/jps.21269.

[22] S. Dagan, T. Hagai, Y. Gavrilov, R. Kapon, Y. Levy, and Z. Reich. "Stabilization of a protein conferred by an increase in folded state entropy". *Proc. Natl. Acad. Sci. U.S.A.* 110.26 (2013), pp. 10628–10633. DOI: 10.1073/pnas.1302284110.

[23] M. Roca, H. Liu, B. Messer, and A. Warshel. "On the relationship between thermal stability and catalytic power of enzymes". *Biochemistry* 46.51 (2007), pp. 15076–88. DOI: 10.1021/bi701732a.

[24] A. Nicolaï, P. Senet, P. Delarue, and D. Ripoll. "Human inducible Hsp70: Structures, dynamics, and interdomain communication from all-atom molecular dynamics simulations". *J. Chem. Theory Comput.* 6.8 (2010), pp. 2501–2519. DOI: 10.1021/ct1002169.

[25] P. Dixit and S. Maslov. "Evolutionary capacitance and control of protein stability in protein-protein interaction networks". *PLoS Comput. Biol.* 9.4 (2013). DOI: 10.1371/journal.pcbi.1003023.

[26] E. Marcos, P. Mestres, and R. Crehuet. "Crowding induces differences in the diffusion of thermophilic and mesophilic proteins: A new look at neutron scattering results". *Biophys. J* 101.11 (2011), pp. 2782–2789. DOI: 10.1016/j.bpj.2011.09.033.

[27] Y. Chebaro, S. Pasquali, and P. Derreumaux. "The coarse-grained OPEP force field for non-amyloid and amyloid proteins". *J. Phys. Chem. B* 116.30 (2012), pp. 8741–8752. DOI: 10.1021/jp301665f.

[28] M. Fyta, E. Kaxiras, S. Melchionna, and S. Succi. "Multiscale simulation of nanobiological flows". *Comput. Sci. Eng.* 10 (2008), pp. 10–19. DOI: 10.1109/MCSE.2008.100.

[29] M. Bernaschi, S. Melchionna, S. Succi, M. Fyta, E. Kaxiras, and J. Sircar. "MUPHY: A parallel MUlti PHYsics/scale code for high performance bio-fluidic simulations". *Comput. Phys. Commun.* 180.9 (2009), pp. 1495–1502. DOI: 10.1016/j.cpc.2009.04.001.

[30] F. Sterpone, S. Melchionna, P. Tuffery, S. Pasquali, N. Mousseau, T. Cragnolini, Y. Chebaro, J. St-Pierre, M. Kalimeri, A. Barducci, Y. Laurin, A. Tek, M. Baaden, P. Nguyen, and P. Derreumaux. "The OPEP protein model: From single molecules, amyloid formation, crowding and hydrodynamics to DNA/RNA systems". *Chem. Soc. Rev.* 43.13 (2014), pp. 4871–4893. DOI: 10.1039/C4CS00048J.

[31] I. Berezovsky and E. Shakhnovich. "Physics and evolution of thermophilic adaptation". *Proc. Natl. Acad. Sci. U.S.A.* 102.36 (2005), pp. 12742–12747. DOI: 10.1073/pnas.0503890102.

[32] B. Boussau, S. Blanquart, A. Necsulea, N. Lartillot, and M. Gouy. "Parallel adaptations to high temperatures in the Archaean eon". *Nature* 456.7224 (2008), pp. 942–945. DOI: 10.1038/nature07393.

[33] S. Akanuma, Y. Nakajima, S. Yokobori, M. Kimura, N. Nemoto, T. Mase, K. Miyazono, M. Tanokura, and A. Yamagishi. "Experimental evidence for the thermophilicity of ancestral life". *Proc. Natl. Acad. Sci. U.S.A.* 110.27 (2013), pp. 11067–11072. DOI: 10.1073/pnas.1308215110.

[34] P. Závodszky, J. Kardos, A. Svingor, and G. A. Petsko. "Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins". *Proc. Natl. Acad. Sci. U.S.A.* 95 (1998), pp. 7406–7411.

[35] G. Hernandez, F. E. Jenney, M. Adams, and D. M. LeMaster. "Millisecond time scale conformational flexibility in a hyperthermophile protein at ambient temperature". *Proc. Natl. Acad. Sci. U.S.A.* 97.7 (2000), pp. 3166–3170. DOI: 10.1073/pnas.97.7.3166.

[36] J. Fitter and J. Heberle. "Structural equilibrium fluctuations in mesophilic and thermophilic a-amylase". *Biophys. J* 79.3 (2000), pp. 1629–1636. DOI: 10.1016/S0006-3495(00)76413-7.

[37] M. Kalimeri, O. Rahaman, S. Melchionna, and F. Sterpone. "How conformational flexibility stabilizes the hyperthermophilic elongation factor G-domain". *J. Phys. Chem. B* 117.44 (2013), pp. 13775–13785. DOI: 10.1021/jp407078z.

[38] M. Kalimeri, P. Derreumaux, and F. Sterpone. "Are coarse-grained models apt to detect protein thermal stability? The case of OPEP force field". *J. Non-Cryst. Solids* (2014). Article in press. DOI: 10.1016/j.jnoncrysol.2014.07.005.

[39] R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria, 2014.

[40] B. J. Grant, A. P. C. Rodrigues, K. M. Elsawy, J. A. McCammon, and L. S. D. Caves. "Bio3d: An R package for the comparative analysis of protein structures". *Bioinformatics* 22.21 (2006), pp. 2695–2696. DOI: 10.1093/bioinformatics/btl461.

[41] G. Csardi and T. Nepusz. "The igraph software package for complex network research". *InterJournal* Complex Systems (2006), p. 1695.

[42] M. Yuan, W. Zhang, S. Dai, J. Wu, Y. Wang, T. Tao, M. Chen, and M. Lin. "*Deinococcus gobiensis* sp. nov., an extremely radiation-resistant bacterium". *Int. J. Syst. Evol. Microbiol.* 59.6 (2009), pp. 1513–1517. DOI: 10.1099/ijs.0.004523-0.

[43] O. Singh and P. Gabani. "Extremophiles: Radiation resistance microbial reserves and therapeutic implications". *J. Appl. Microbiol.* 110.4 (2011), pp. 851–861. DOI: 10.1111/j.1365-2672.2011.04971.x.

[44] S. Emeish. "Production of natural β-carotene from *Dunaliella* living in the dead sea". *J. J. Ear. Envir. Sci.* 4.2 (2012), pp. 23–27.

[45] P. Vítek, H. Edwards, J. Jehlička, C. Ascaso, A. De Los Ríos, S. Valea, S. Jorge-Villar, A. Davila, and J. Wierzochos. "Microbial colonization of halite from the hyper-arid Atacama desert studied by Raman spectroscopy". *Phil. Trans. R. Soc. A* 368.1922 (2010), pp. 3205–3221. DOI: 10.1098/rsta.2010.0059.

[46] C. Woese. "The universal ancestor". *Proc. Natl. Acad. Sci. U.S.A.* 95.12 (1998), pp. 6854–6859. DOI: 10.1073/pnas.95.12.6854.

[47] E. Gaucher, J. Thomson, M. Burgan, and S. Benner. "Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins". *Nature* 425.6955 (2003), pp. 285–288. DOI: 10.1038/nature01977.

[48] N. Galtier, N. Tourasse, and M. Gouy. "A nonhyperthermophilic common ancestor to extant life forms". *Science* 283.5399 (1999), pp. 220–221. DOI: 10.1126/science.283.5399.220.

[49] T. Brock, K. Brock, R. Belly, and R. Weiss. "*Sulfolobus*: A new genus of sulfur-oxidizing bacteria living at low ph and high temperature". *Archiv für Mikrobiologie* 84.1 (1972), pp. 54–68. DOI: 10.1007/BF00408082.

[50] B. Pierson and R. Castenholz. "A phototrophic gliding filamentous bacterium of hot springs, *Chloroflexus aurantiacus*, gen. and sp. nov." *Arch. Microbiol.* 100.1 (1974), pp. 5–24. DOI: 10.1007/BF00446302.

[51] T. E. Creighton. *Proteins: Structures and molecular properties*. Second Edition. W. H. Freeman & Company, 1993.

[52] H. Frauenfelder. *The physics of proteins: An introduction to biological physics and molecular biophysics*. Springer, 2010.

[53] C. van Oss, D. Absolom, and A. Neumann. "The 'hydrophobic effect': Essentially a van der Waals interaction". *Colloid. Polym. Sci.* 258.4 (1980), pp. 424–427. DOI: 10.1007/BF01480835.

[54] K. Dill. "Dominant forces in protein folding". *Biochemistry* 29.31 (1990), pp. 7133–7155. DOI: 10.1021/bi00483a001.

[55] F. Sterpone, C. Bertonati, and S. Melchionna. "Water around thermophilic proteins: The role of charged and apolar atoms". *J. Phys.: Condens. Matter* 22.28 (2010), p. 284113. DOI: 10.1088/0953-8984/22/28/284113.

[56] Y. M. Rhee, E. J. Sorin, G. Jayachandran, E. Lindahl, and V. S. Pande. "Simulations of the role of water in the protein-folding mechanism". *Proc. Natl. Acad. Sci. U.S.A.* 101.17 (2004), pp. 6456–6461. DOI: 10.1073/pnas.0307898101.

[57] P. Miquel. "Monographie d'un bacille vivant au delà de 70 centigrades". *Ann. Micrographie* 1 (1888), pp. 3–10.

[58] S. Friedman. "Protein-synthesizing machinery of thermophilic bacteria". *Bacteriol. Rev.* 32.1 (1968), pp. 27–38.

[59] G. Vogt and P. Argos. "Protein thermal stability: Hydrogen bonds or internal packing?" *Fold. Des.* 2.4 (1997), S40–S46. DOI: 10.1016/S1359-0278(97)00062-X.

[60] G. Feller. "Protein stability and enzyme activity at extreme biological temperatures". *J. Phys.: Condens Matter* 22.32 (2010), p. 323101. DOI: 10.1088/0953-8984/22/32/323101.

[61] T. Mamonova, A. Glyakina, M. Kurnikova, and O. Galzitskaya. "Flexibility and mobility in mesophilic and thermophilic homologous proteins from molecular dynamics and foldunfold method". *J. Bioinf. Comput. Biol.* 8.3 (2010), pp. 377–394. DOI: 10.1142/S0219720010004690.

[62] N. Koga, R. Tatsumi-Koga, G. Liu, R. Xiao, T. Acton, G. Montelione, and D. Baker. "Principles for designing ideal protein structures". *Nature* 491 (2012), pp. 222–227. DOI: 10.1038/nature11600.

[63] M. Tehei, D. Madern, B. Franzetti, and G. Zaccai. "Neutron scattering reveals the dynamic basis of protein adaptation to extreme temperature". *J. Biol. Chem.* 280.49 (2005), pp. 40974–40979. DOI: 10.1074/jbc.M508417200.

[64] M. Vijayabaskar and S. Vishveshwara. "Comparative analysis of thermophilic and mesophilic proteins using protein energy networks". *BMC Bioinformatics* 11.SUPPLL.1 (2010). DOI: 10.1186/1471-2105-11-S1-S49.

[65] P. Rathi, H. Höffken, and H. Gohlke. "Quality matters: extension of clusters of residues with good hydrophobic contacts stabilize (hyper)thermophilic proteins". *J. Chem. Inf. Model* 54.2 (2014), pp. 355–361. DOI: 10.1021/ci400568c.

[66] V. Villeret, B. Clantin, C. Tricot, C. Legrain, M. Roovers, V. Stalon, N. Glansdorff, and J. van Beeumen. "The crystal structure of *Pyrococcus furiosus* ornithine carbamoyltransferase reveals a key role for oligomerization in enzyme stability at extremely high temperatures". *Proc. Natl. Acad. Sci. U.S.A.* 95.6 (1998), pp. 2801–2806. DOI: 10.1073/pnas.95.6.2801.

[67] J. Backmann and G. Schäfer. "Thermodynamic analysis of hyperthermostable oligomeric proteins". *Hyperthermophilic enzymes, Part C*. Ed. by M. Adams and R. Kelly. Vol. 334. Methods in Enzymology. Academic Press, 2001, pp. 328–342.

[68] J. Backmann, G. Schäfer, L. Wyns, and H. Bönisch. "Thermodynamics and kinetics of unfolding of the thermostable trimeric adenylate kinase from the archaeon *Sulfolobus acidocaldarius*". *J. Mol. Biol.* 284.3 (1998), pp. 817–833. DOI: 10.1006/jmbi.1998.2216.

[69] P. Gallego, R. Planell, J. Benach, E. Querol, J. Perez-Pons, and D. Reverter. "Structural characterization of the enzymes composing the arginine deiminase pathway in Mycoplasma penetrans". *PLoS ONE* 7.10 (2012). DOI: 10.1371/journal.pone.0047886.

[70] A. Criswell, E. Bae, B. Stec, J. Konisky, and G. Phillips Jr. "Structures of thermophilic and mesophilic adenylate kinases from the genus *Methanococcus*". *J. Mol. Biol.* 330.5 (2003), pp. 1087–1099. DOI: 10.1016/S0022-2836(03)00655-7.

[71] C. Vieille and G. Zeikus. "Hyperthermophilic enzymes: Sources, uses, and molecular mechanisms for thermostability". *Microbiol. Mol. Biol. Rev.* 65.1 (2001), pp. 1–43. DOI: 10.1128/MMBR.65.1.1-43.2001.

[72] G. N. Somero. "Temperature adaptation of enzymes: Biological optimization through structure-function compromises". *Ann. Rev. Ecol. Syst.* 9 (1978), pp. 1–29. DOI: 10.1146/annurev.es.09.110178.000245.

[73] A. Wrba, A. Schweiger, V. Schultes, R. Jaenicke, and P. Zavodszky. "Extremely thermostable d-glyceraldehyde-3-phosphate dehydrogenase from the eubacterium Thermotoga maritima". *Biochemistry* 29.33 (1990), pp. 7584–7592. DOI: 10.1021/bi00485a007.

[74] R. Jaenicke and G. Böhm. "The stability of proteins in extreme environments". *Curr. Opin. Struct. Biol.* 8.6 (1998), pp. 738–748. DOI: 10.1016/S0959-440X(98)80094-8.

[75] P. Závodszky, J. Kardos, A. Svingor, and G. Petsko. "Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins". *Proc. Natl. Acad. Sci. U.S.A.* 95.13 (1998), pp. 7406–7411. DOI: 10.1073/pnas.95.13.7406.

[76] M. Tehei, B. Franzetti, D. Madern, M. Ginzburg, B. Ginzburg, M.-T. Giudici-Orticoni, M. Bruschi, and G. Zaccai. "Adaptation to extreme environments: Macromolecular dynamics in bacteria compared *in vivo* by neutron scattering". *EMBO Reports* 5.1 (2004), pp. 66–70. DOI: 10.1038/sj.embor.7400049.

[77] E. Marcos, A. Jiménez, and R. Crehuet. "Dynamic fingerprints of protein thermostability revealed by long molecular dynamics". *J. Chem. Theory. Comput.* 8.3 (2012), pp. 1129–1142. DOI: 10.1021/ct200877z.

[78] L. Meinhold, D. Clement, M. Tehei, R. Daniel, J. L. Finney, and J. C. Smith. "Protein dynamics and stability: the distribution of atomic fluctuations in thermophilic and mesophilic dihydrofolate reductase derived using elastic incoherent neutron scattering". *Biophys. J* 94.12 (2008), pp. 4812–4818. DOI: 10.1529/biophysj.107.121418.

[79] A. Sigtryggsdóttir, E. Papaleo, S. Thorbjarnardóttir, and M. Kristjánsson. "Flexibility of cold- and heat-adapted subtilisin-like serine proteinases evaluated with fluorescence quenching and molecular dynamics". *Biochim. Biophys. Acta* 1844.4 (2014), pp. 705–712. DOI: 10.1016/j.bbapap.2014.02.009.

[80] T. Lazaridis, I. Lee, and M. Karplus. "Dynamics and unfolding pathways of a hyperthermophilic and a mesophilic rubredoxin". *Prot. Sci.* 6.12 (1997), pp. 2589–2605. DOI: 10.1002/pro.5560061211.

[81] K. Manjunath and K. Sekar. "Molecular dynamics perspective on the protein thermal stability: A case study using SAICAR synthetase". *J. Chem. Inf. Model.* 53.9 (2013), pp. 2448–2461. DOI: 10.1021/ci400306m.

[82] G. Colombo and K. Merz Jr. "Stability and activity of mesophilic subtilisin E and its thermophilic homolog: Insights from molecular dynamics simulations". *J. Am. Chem. Soc.* 121.29 (1999), pp. 6895–6903. DOI: 10.1021/ja990420s.

[83] P. L. Wintrode, D. Zhang, N. Vaidehi, F. H. Arnold, and W. A. Goddard III. "Protein dynamics in a family of laboratory evolved thermophilic enzymes". *J. Mol. Biol.* 327.3 (2003), pp. 745–757. DOI: 10.1016/S0022-2836(03)00147-5.

[84] E. D. Merkley, W. W. Parson, and V. Daggett. "Temperature dependence of the flexibility of thermophilic and mesophilic flavoenzymes of the nitroreductase fold". *Protein Eng. Des. Sel.* 23.5 (2010), pp. 327–36. DOI: 10.1093/protein/gzp090.

[85] X.-L. Feng, X. Zhao, H. Yu, T.-D. Sun, and X.-R. Huang. "Molecular dynamics simulations of the thermal stability of tteRBP and ecRBP". *J. Biomol. Struct. Dyn.* 31.10 (2013), pp. 1086–1100. DOI: 10.1080/07391102.2012.721497.

[86] S. Radestock and H. Gohlke. "Exploiting the link between protein rigidity and thermostability for data-driven protein engineering". *Eng. Life Sci.* 8.5 (2008), pp. 507–522. DOI: 10.1002/elsc.200800043.

[87] S. Radestock and H. Gohlke. "Protein rigidity and thermophilic adaptation". *Proteins* 79.4 (2011), pp. 1089–1108. DOI: 10.1002/prot.22946.

[88] S. Wells, S. Crennell, and M. Danson. "Structures of mesophilic and extremophilic citrate synthases reveal rigidity and flexibility for function". *Proteins: Structure, Function and Bioinformatics* (2014). Article in press. DOI: 10.1002/prot.24630.

[89] H. Frauenfelder, S. Sligar, and P. Wolynes. "The energy landscapes and motions of proteins". *Science* 254.5038 (1991), pp. 1598–1603. DOI: 10.1126/science.1749933.

[90] Z.-X. Liang, I. Tsigos, T. Lee, V. Bouriotis, K. Resing, N. Ahn, and J. Klinman. "Evidence for increased local flexibility in psychrophilic alcohol dehydrogenase relative to its thermophilic homologue". *Biochemistry* 43.46 (2004), pp. 14676–14683. DOI: 10.1021/bi049004x.

[91] O. Oyeyemi, K. Sours, T. Lee, A. Kohen, K. Resing, N. Ahn, and J. Klinman. "Comparative hydrogen-deuterium exchange for a mesophilic vs thermophilic dihydrofolate reductase at 25 °C: Identification of a single active site region with enhanced flexibility in the mesophilic protein". *Biochemistry* 50.38 (2011), pp. 8251–8260. DOI: 10.1021/bi200640s.

[92] L. Xiao and B. Honig. "Electrostatic contributions to the stability of hyperthermophilic proteins". *J. Mol. Biol.* 289.5 (1999), pp. 1435–44. DOI: 10.1006/jmbi.1999.2810.

[93] C. Danciulescu, R. Ladenstein, and L. Nilsson. "Dynamic arrangement of ion pairs and individual contributions to the thermal stability of the cofactor-binding domain of glutamate dehydrogenase from thermotoga maritima". *Biochemistry* 46.29 (2007), pp. 8537–8549. DOI: 10.1021/bi7004398.

[94] S. Hawley. "Reversible pressure-temperature denaturation of chymotrypsinogen". *Biochemistry* 10.13 (1971), pp. 2436–2442. DOI: 10.1021/bi00789a002.

[95] C. Liu and V. LiCata. "The stability of *Taq* DNA polymerase results from a reduced entropic folding penalty; identification of other thermophilic proteins with similar folding thermodynamics". *Proteins* 82.5 (2014), pp. 785–793. DOI: 10.1002/prot.24458.

[96] T. Lazaridis and M. Karplus. "Thermodynamics of protein folding: A microscopic view". *Biophys. Chem.* 100.1-3 (2003), pp. 367–395. DOI: 10.1016/S0301-4622(02)00293-4.

[97] H. Nojima, A. Ikai, T. Oshima, and H. Noda. "Reversible thermal unfolding of thermostable phosphoglycerate kinase. Thermostability associated with mean zero enthalpy change". *J. Mol. Biol.* 116.3 (1977), pp. 429–442. DOI: 10.1016/0022-2836(77)90078-X.

[98] A. Razvi and J. M. Scholtz. "Lessons in stability from thermophilic proteins". *Protein Sci.* 15.7 (2006), pp. 1569–1578. DOI: 10.1110/ps.062130306.

[99] N. Go. "Theoretical studies of protein folding". *Annu. Rev. Biophys. Bio.* 12 (1983), pp. 183–210. DOI: 10.1146/annurev.bb.12.060183.001151.

[100] L. Marky and K. Breslauer. "Calculating thermodynamic data for transitions of any molecularity from equilibrium melting curves". *Biopolymers* 26.9 (1987), pp. 1601–1620. DOI: 10.1002/bip.360260911.

[101] T. Haltia and E. Freire. "Forces and factors that contribute to the structural stability of membrane proteins". *Biochim. Biophys. Acta* 1241.2 (1995), pp. 295–322. DOI: 10.1016/j.bbamem.2011.11.006.

[102] A. Ansari, J. Berendzen, S. Bowne, H. Frauenfelder, I. Iben, T. Sauke, E. Shyamsunder, and R. Young. "Protein states and proteinquakes". *Proc. Natl. Acad. Sci. U.S.A.* 82.15 (1985), pp. 5000–5004. DOI: 10.1073/pnas.82.15.5000.

[103] P. Balbuena and J. Seminario, eds. *Nanomaterials: Design and simulation.* Vol. 18. Theoretical and Computational Chemistry. Elsevier, 2007.

[104] M. Karplus and J. McCammon. "Molecular dynamics simulations of biomolecules". *Nature Struct. Biol.* 9.9 (2002), pp. 646–652. DOI: 10.1038/nsb0902-646.

[105] D. Frenkel and B. Smit. *Understanding molecular simulation, second edition: From algorithms to applications (computational science).* 2nd ed. Academic Press, 7, 2001.

[106] M. Tuckerman. *Statistical mechanics : Theory and molecular simulation.* Oxford graduate texts. Oxford: Oxford University Press, 2010.

[107] I. Newton. *Philosophiæ naturalis principia mathematica.* Ed. by A. Koyré and B. Cohen. Third. Cambridge, MA, 1972, Harvard UP, 1726.

[108] L. Verlet. "Computer 'experiments' on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules". *Phys. Rev.* 159.1 (1967), pp. 98–103. DOI: 10.1103/PhysRev.159.98.

[109] R. Hockney and J. Eastwood, eds. *Computer simulations using particles.* McGraw-Hill, New York, 1981.

[110] D. Auerbach, W. Paul, A. Bakker, C. Lutz, W. Rudge, and F. Abraham. "A special purpose parallel computer for molecular dynamics: Motivation, design, implementation, and application". *J. Phys. Chem.* 91.19 (1987), pp. 4881–4890. DOI: 10.1021/j100303a004.

[111] P. Ewald. "Die Berechnung optischer und elektrostatischer Gitterpotentiale". *Ann. Phys.* 369.3 (1921), pp. 253–287. DOI: 10.1002/andp.19213690304.

[112] T. Darden, D. York, and L. Pedersen. "Particle mesh ewald: An N·log(N) method for ewald sums in large systems". *J. Chem. Phys.* 98.12 (1993), pp. 10089–10092. DOI: 10.1063/1.464397.

[113] A. Appel. "An efficient program for many-body simulation". *SIAM Journal on Scientific and Statistical Computing* 6.1 (1985), pp. 85–103. DOI: 10.1137/0906008.

[114] W. Swope, H. Andersen, P. Berens, and K. Wilson. "A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters". *J. Chem. Phys.* 76.1 (1982), pp. 637–649. DOI: 10.1063/1.442716.

[115] D. Beeman. "Some multistep methods for use in molecular dynamics calculations". *J. Comput. Phys.* 20.2 (1976), pp. 130–139. DOI: 10.1016/0021-9991(76)90059-0.

[116] T. Schlick. *Molecular modeling and simulation: An interdisciplinary guide.* Secaucus, NJ, U.S.A.: Springer-Verlag New York, Inc., 2002.

[117] L. Woodcock. "Isothermal molecular dynamics calculations for liquid salts". *Chem. Phys. Lett.* 10.3 (1971), pp. 257–261. DOI: 10.1016/0009-2614(71)80281-6.

[118] H. Berendsen, J. Postma, W. van Gunsteren, A. Dinola, and J. Haak. "Molecular dynamics with coupling to an external bath". *J. Chem. Phys.* 81.8 (1984), pp. 3684–3690. DOI: 10.1063/1.448118.

[119] G. Bussi, D. Donadio, and M. Parrinello. "Canonical sampling through velocity rescaling". *J. Chem. Phys.* 126.1 (2007). DOI: 10.1063/1.2408420.

[120] H. Andersen. "Molecular dynamics simulations at constant pressure and/or temperature". *J. Chem. Phys.* 72.4 (1980), pp. 2384–2393. DOI: 10.1063/1.439486.

[121] G. Grest and K. Kremer. "Molecular dynamics simulation for polymers in the presence of a heat bath". *Phys. Rev. A* 33.5 (1986), pp. 3628–3631. DOI: 10.1103/PhysRevA.33.3628.

[122] P. Nikunen, M. Karttunen, and I. Vattulainen. "How would you integrate the equations of motion in dissipative particle dynamics simulations?" *Comput. Phys. Commun.* 153.3 (2003), pp. 407–423. DOI: 10.1016/S0010-4655(03)00202-9.

[123] S. Nosé. "A unified formulation of the constant temperature molecular dynamics methods". *J. Chem. Phys.* 81.1 (1984), pp. 511–519. DOI: 10.1063/1.447334.

[124] W. Hoover. "Canonical dynamics: Equilibrium phase-space distributions". *Phys. Rev. A* 31.3 (1985), pp. 1695–1697. DOI: 10.1103/PhysRevA.31.1695.

[125] G. Martyna, M. Klein, and M. Tuckerman. "Nosé-Hoover chains: The canonical ensemble via continuous dynamics". *J. Chem. Phys.* 97.4 (1992), pp. 2635–2643. DOI: 10.1063/1.463940.

[126] W. Hoover. "Constant-pressure equations of motion". *Phys. Rev. A* 34.3 (1986), pp. 2499–2500. DOI: 10.1103/PhysRevA.34.2499.

[127] G. Martyna, D. Tobias, and M. Klein. "Constant pressure molecular dynamics algorithms". *J. Chem. Phys.* 101.5 (1994), pp. 4177–4189. DOI: 10.1063/1.467468.

[128] *NAMD user's guide*. 2012. eprint: http://www.ks.uiuc.edu/Research/namd/2.9/ug.pdf.

[129] S. Feller, Y. Zhang, R. Pastor, and B. Brooks. "Constant pressure molecular dynamics simulation: The Langevin piston method". *J. Chem. Phys.* 103.11 (1995), pp. 4613–4621. DOI: 10.1063/1.470648.

[130] B. Brooks, R. Bruccoleri, D. Olafson, D. States, S. Swaminathan, and M. Karplus. "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations". *J. Comput. Chem.* 4 (1983), pp. 187–217. DOI: 10.1002/jcc.540040211.

[131] A. MacKerel Jr., C. Brooks III, L. Nilsson, B. Roux, Y. Won, and M. Karplus. "CHARMM: The energy function and its parameterization with an overview of the program". Ed. by P. v. R. Schleyer et al. Vol. 1. The Encyclopedia of Computational Chemistry. John Wiley & Sons: Chichester, 1998, pp. 271–277.

[132] W. Cornell, P. Cieplak, C. Bayly, I. Gould, K. Merz, D. Ferguson, D. Spellmeyer, T. Fox, J. Caldwell, and P. Kollman. "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules". *J. Am. Chem. Soc.* 117.19 (1995), pp. 5179–5197. DOI: 10.1021/ja00124a002.

[133] W. F. van Gunsteren and H. Berendsen. *Groningen Molecular Simulation (GROMOS) Library Manual*. 1987.

[134] A. D. MacKerell, M. Feig, and C. L. Brooks(III). "Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations". *J. Comput. Chem.* 25 (2004), pp. 1400–1415. DOI: 10.1002/jcc.20065.

[135] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. "Comparison of multiple Amber force fields and development of improved protein backbone parameters". *Proteins* 15 (2006), pp. 712–725. DOI: 10.1002/prot.21123.

[136] R. Best, X. Zhu, J. Shim, P. Lopes, J. Mittal, M. Feig, and A. MacKerell Jr. "Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain χ_1 and χ_2 dihedral angles". *J. Chem. Theory Comput.* 8.9 (2012), pp. 3257–3273. DOI: 10.1021/ct300400x.

[137] J. Maupetit, P. Tuffery, and P. Derreumaux. "A coarse-grained protein force field for folding and structure prediction". *Proteins: Struct., Funct., Genet.* 69.2 (2007), pp. 394–408. DOI: 10.1002/prot.21505.

[138] F. Sterpone, P. Nguyen, M. Kalimeri, and P. Derreumaux. "Importance of the ion-pair interactions in the OPEP coarse-grained force field: Parametrization and validation". *J. Chem. Theory. Comput.* 9 (2013), pp. 4574–4584. DOI: 10.1021/ct4003493.

[139] P. Nguyen and P. Derreumaux. "Understanding amyloid fibril nucleation and $\alpha\beta$ oligomer/drug interactions from computer simulations". *Acc. Chem. Res.* 47 (2014), pp. 603–611. DOI: 10.1021/ar4002075.

[140] S. Marrink, A. De Vries, and A. Mark. "Coarse grained model for semiquantitative lipid simulations". *J. Phys. Chem. B* 108.2 (2004), pp. 750–760. DOI: 10.1021/jp036508g.

[141] L. Monticelli, S. Kandasamy, X. Periole, R. Larson, D. Tieleman, and S. Marrink. "The MARTINI coarse-grained force field: Extension to proteins". *J. Chem. Theory Comput.* 4.5 (2008), pp. 819–834. DOI: 10.1021/ct700324x.

[142] S. Marrink and D. Tieleman. "Perspective on the Martini model". *Chem. Soc. Rev.* 42.16 (2013), pp. 6801–6822. DOI: 10.1039/C3CS60093A.

[143] S. Yesylevskyy, L. Schäfer, D. Sengupta, and S. Marrink. "Polarizable water model for the coarse-grained MARTINI force field". *PLoS Comput. Biol.* 6.6 (2010), e1000810. DOI: 10.1371/journal.pcbi.1000810.

[144] K. Lindorff-Larsen, S. Piana, R. Dror, and D. Shaw. "How fast-folding proteins fold". *Science* 334.6055 (2011), pp. 517–520. DOI: 10.1126/science.1208351.

[145] H. Lei and Y. Duan. "Improved sampling methods for molecular simulation". *Curr. Opin. Struct. Biol.* 17.2 (2007), pp. 187–191. DOI: 10.1016/j.sbi.2007.03.003.

[146] T. Schlick. "Molecular dynamics-based approaches for enhanced sampling of long-time, large-scale conformational changes in biomolecules". *Curr. Opin. Struct. Biol.* 51.1 (2009). DOI: 10.3410/B1-51.

[147] C. Abrams and G. Bussi. "Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration". *Entropy* 16.1 (2013), pp. 163–199. DOI: 10.3390/e16010163.

[148] J. Kirkwood. "Statistical mechanics of fluid mixtures". *J. Chem. Phys.* 3.5 (1935), pp. 300–313. DOI: 10.1063/1.1749657.

[149] R. Zwanzig. "High-temperature equation of state by a perturbation method. I. Nonpolar gases". *J. Chem. Phys.* (1954), pp. 1420–1426. DOI: 10.1063/1.1740409.

[150] G. Torrie and J. Valleau. "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling". *J. Comput. Phys.* 23.2 (1977), pp. 187–199. DOI: 10.1016/0021-9991(77)90121-8.

[151] T. Beutler and W. van Gunsteren. "The computation of a potential of mean force: Choice of the biasing potential in the umbrella sampling technique". *J. Chem. Phys.* 100.2 (1994), pp. 1492–1497. DOI: 10.1063/1.466628.

[152] S. Kumar, J. Rosenberg, D. Bouzida, R. Swendsen, and P. Kollman. "The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method". *J. Comput. Chem.* 13.8 (1992), pp. 1011–1021. DOI: 10.1002/jcc.540130812.

[153] J. Lemkul and D. Bevan. "Assessing the stability of Alzheimer's amyloid protofibrils using molecular dynamics". *J. Phys. Chem. B* 114.4 (7, 2010), pp. 1652–1660. DOI: 10.1021/jp9110794.

[154] Y. Sugita and Y. Okamoto. "Replica-exchange molecular dynamics method for protein folding". *Chem. Phys. Lett.* 314.1-2 (1999), pp. 141–151. DOI: 10.1016/S0009-2614(99)01123-9.

[155] W. Kabsch and C. Sander. "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features". *Biopolymers* 22.12 (1983), pp. 2577–2637. DOI: 10.1002/bip.360221211.

[156] H. Kramers. "Brownian motion in a field of force and the diffusion model of chemical reactions". *Physica* 7.4 (1940), pp. 284–304. DOI: 10.1016/S0031-8914(40)90098-2.

[157] J. McCammon and M. Karplus. "Internal motions of antibody molecules". *Nature* 268.5622 (1977), pp. 765–766. DOI: 10.1038/268765a0.

[158] M. Karplus and J. McCammon. "The internal dynamics of globular proteins". *CRC Cr. Rev. Bioch. Mol.* 9.4 (1981), pp. 293–349. DOI: 10.3109/10409238109105437.

[159] R. Zwanzig. "Diffusion in a rough potential". *Proc. Natl. Acad. Sci. U.S.A.* 85.7 (1988), pp. 2029–2030. DOI: 10.1073/pnas.85.7.2029.

[160] T. Woolf and B. Roux. "Conformational flexibility of o-phosphorylcholine and o-phosphorylethanolamine: A molecular dynamics study of solvation effects". *J. Am. Chem. Soc.* 116.13 (1994), pp. 5916–5926. DOI: 10.1021/ja00092a048.

[161] G. Hummer. "Position-dependent diffusion coefficients and free energies from bayesian analysis of equilibrium and replica molecular dynamics simulations". *New J. Phys.* 7.1 (2005), p. 34. DOI: 10.1088/1367-2630/7/1/034.

[162] N. D. Socci, J. N. Onuchic, and P. G. Wolynes. "Diffusive dynamics of the reaction coordinate for protein folding funnels". *J. Chem. Phys.* 104 (1996), p. 5860. DOI: 10.1063/1.471317.

[163] K. Schulten and I. Kosztin. *Lectures in Theoretical Biophysics*. Department of Physics and Beckman Institute, University of Illinois. 2000.

[164] A. Jain, M. Murty, and P. Flynn. "Data clustering: A review". *ACM Computing Surveys* 31.3 (1999), pp. 264–323. DOI: 10.1145/331499.331504.

[165] L. Parsons, E. Haque, and H. Liu. "Subspace clustering for high dimensional data: A review". *SIGKDD Exploration* 6.1 (2004), pp. 90–105. DOI: 10.1145/1007730.1007731.

[166] S. Lloyd. "Least squares quantization in PCM". *IEEE Trans. Inf. Theory* 28 (1982), pp. 129–137. DOI: 10.1109/TIT.1982.1056489.

[167] J. Hartigan. *Clustering algorithms*. New York:Wiley, 1975.

[168] S. Wasserman. *Social network analysis: methods and applications*. Cambridge University Press, 1994.

[169] C. E. Shannon. "Prediction and entropy of printed english". *Bell Syst. Tech. J.* (1951), pp. 50–64. DOI: 10.1002/j.1538-7305.1951.tb01366.x.

[170] S. M. van Dongen. "Graph clustering by flow simulation". PhD thesis. University of Utrecht, The Netherlands, 2000.

[171] D. Gfeller, P. De Los Rios, A. Caflisch, and F. Rao. "Complex network analysis of free-energy landscapes". *Proc. Natl. Acad. Sci. U.S.A.* 104.6 (6, 2007), pp. 1817–1822. DOI: 10.1073/pnas.0608099104.

[172] L. Vitagliano, A. Ruggiero, M. Masullo, P. Cantiello, P. Arcari, and A. Zagari. "The crystal structure of *Sulfolobus solfataricus* elongation factor 1a in complex with magnesium and GDP". *Biochemistry* 43.21 (2004), pp. 6630–6636. DOI: 10.1021/bi0363331.

[173] H. Song, M. Parsons, S. Rowsell, G. Leonard, and S. Phillips. "Crystal structure of intact elongation factor EF-Tu from *Escherichia coli* in GDP conformation at 2.05 Å resolution". *J. Mol. Biol.* 285.3 (1999), pp. 1245–1256. DOI: 10.1006/jmbi.1998.2387.

[174] H. Šanderová, M. Hůlková, P. Maloň, M. Kepková, and J. Jonák. "Thermostability of multidomain proteins: Elongation factors EF-Tu from *Escherichia coli* and *Bacillus stearothermophilus* and their chimeric forms". *Prot. Sci.* 13.1 (2004), pp. 89–99. DOI: 10.1110/ps.03272504.

[175] F. Sterpone, C. Bertonati, G. Briganti, and S. Melchionna. "Key role of proximal water in regulating thermostable proteins". *J. Phys. Chem. B* 113.1 (2009), pp. 131–7. DOI: 10.1021/jp805199c.

[176] O. Rahaman, S. Melchionna, D. Laage, and F. Sterpone. "The effect of protein composition on hydration dynamics". *Phys. Chem. Chem. Phys.* 15 (10 2013), pp. 3570–3576. DOI: 10.1039/C3CP44582H.

[177] M. Chavent, A. Vanel, A. Tek, B. Levy, S. Robert, B. Raffin, and M. Baaden. "GPU-accelerated atom and dynamic bond visualization using hyperballs: A unified algorithm for balls, sticks and hyperboloids". *J. Comput. Chem.* 32 (2011), p. 2924. DOI: 10.1002/jcc.21861.

[178] J. C. Phillips, R. Braunand, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten. "Scalable molecular dynamics with NAMD". *J. Comp. Chem.* 26.16 (2005), pp. 1781–1802. DOI: 10.1002/jcc.20289.

[179] G. Fiorin, M. Klein, and J. Hénin. "Using collective variables to drive molecular dynamics simulations". *Mol. Phys.* 111.22–23 (2013), pp. 3345–3362. DOI: 10.1080/00268976.2013.813594.

[180] S. Abel, F. Dupradeau, and M. Marchi. "Molecular dynamics simulations of a characteristic DPC micelle in water". *J. Chem. Theory. Comput.* 8.11 (2012), pp. 4610–4623. DOI: 10.1021/ct3003207.

[181] C. Rycroft. "Voro++: A three-dimensional Voronoi cell library in C++". *Chaos* 19 (2009), p. 041111. DOI: 10.1063/1.3215722.

[182] L. Maragliano, G. Cottone, L. Cordone, and G. Ciccotti. "Atomic mean-square displacements in proteins by molecular dynamics: A case for analysis of variance". *Biophys. J.* 88 (2004), pp. 2765–2772. DOI: 10.1016/S0006-3495(04)74330-1.

[183] V. Botan, E. H. G. Backus, R. Pfister, A. Moretto, M. Crisma, C. Toniolo, P. H. Nguyen, G. Stock, and P. Hamm. "Energy transport in peptide helices". *Proc. Natl. Acad. Sci. U.S.A.* 104 (2007), pp. 12749–12754. DOI: 10.1073/pnas.0701762104.

[184] Z. Ganim, H. S. Chung, A. W. Smith, L. P. DeFlores, K. C. Jones, and A. Tokmakoff. "Amide I two-dimensional infrared spectroscopy of proteins". *Acc. Chem. Res.* 41.3 (2008), pp. 432–441. DOI: 10.1021/ar700188n.

[185] G. F. Voronoi. "Nouvelles applications des paramètres continus à la théorie des formes quadratiques". *J. Reine Angew. Math.* 134.198-287 (1908). DOI: 10.1515/crll.1908.134.198.

[186] H. Šanderová, H. Tišerová, I. Barvík, L. Sojka, J. Jonák, and L. Krásný. "The N-terminal region is crucial for the thermostability of the G-domain of *Bacillus*". *Biochim. Biophys. Acta.* 1804 (2010), pp. 147–155. DOI: 10.1016/j.bbapap.2009.09.024.

[187] J. H. Missimer, M. O. Steinmetz, R. Baron, F. K. Winkler, R. A. Kammerer, X. Daura, and W. F. van Gunsteren. "Configurational entropy elucidates the role of salt-bridge networks in protein thermostability". *Protein Sci.* 16.7 (2007), pp. 1349–59. DOI: 10.1110/ps.062542907.

[188] H.-X. Zhou. "Toward the physical basis of thermophilic proteins: Linking of enriched polar interactions and reduced heat capacity of unfolding". *Biophys. J.* 83.6 (2002), pp. 3126–33. DOI: 10.1016/S0006-3495(02)75316-2.

[189] V. P. Ninad and K. A. Sharp. "Heat capacity in proteins". *Annu. Rev. Phys. Chem.* 56 (2005), pp. 521–548. DOI: 10.1146/annurev.physchem.56.092503.141202.

[190] V. M. Dadarlat and C. B. Post. "Adhesive-cohesive model for protein compressibility: An alternative perspective on stability". *Proc. Natl. Acad. Sci. U.S.A.* 100.25 (2003), pp. 14778–83. DOI: 10.1073/pnas.2434157100.

[191] C. Lopez, R. Darst, and P. Rossky. "Mechanistic elements of protein cold denaturation". *J. Phys. Chem. B* 112.19 (2008), pp. 5961–7. DOI: 10.1021/jp075928t.

[192] D. Phelps and C. Post. "A novel basis for capsid stabilization by antiviral compounds". *J. Mol. Biol.* 254.4 (1995), pp. 544–551. DOI: 10.1006/jmbi.1995.0637.

[193] M. Marchi. "Compressibility of cavities and biological water from Voronoi volumes in hydrated proteins". *J. Phys. Chem. B* 107.27 (2003), pp. 6598–6602. DOI: 10.1021/jp0342935.

[194] M. Bastian, S. Heymann, and M. Jacomy. "Gephi: An open source software for exploring and manipulating networks". 2009.

[195] S. Robic, M. Guzman-Casado, J. M. Sanchez-Ruiz, and S. Marqusee. "Role of residual structure in the unfolded state of a thermophilic protein". *Proc. Natl. Acad. Sci. U.S.A.* 100.20 (2003), pp. 11345–9. DOI: 10.1073/pnas.1635051100.

[196] M. J. Stone, S. Gupta, N. Snyder, and L. Regan. "Comparison of protein backbone entropy and β-sheet stability: NMR-derived dynamics of protein G B1 domain mutants". *J. Am. Chem. Soc.* 123.1 (2001), pp. 185–186. DOI: 10.1021/ja0030941.

[197] M. Stone. "NMR relaxation studies of the role of conformational entropy in protein stability and ligand binding". *Acc. Chem. Res.* 34.5 (2001), pp. 379–388. DOI: 10.1021/ar000079c.

[198] F. Sterpone and S. Melchionna. "Thermophilic proteins: Insight and perspective from *in silico* experiments". *Chem. Soc. Rev.* 41 (5 2012), pp. 1665–1676. DOI: 10.1039/c1cs15199a.

[199] T. V. Budkevich, A. A. Timchenko, E. I. Tiktopulo, B. S. Negrutskii, V. F. Shalak, Z. M. Petrushenko, V. L. Aksenov, R. Willumeit, J. Kohlbrecher, I. N. Serdyuk, and A. V. El'skaya. "Extended conformation of mammalian translation elongation factor 1a in solution". *Biochemistry* 41.51 (2002), pp. 15342–15349. DOI: 10.1021/bi026495h.

[200] E. Sedlák, M. Sprinzl, N. Grillenbeck, and M. Antalík. "Microcalorimetric study of elongation factor Tu from *Thermus thermophilus* in nucleotide-free, GDP and GTP forms and in the presence of elongation factor Ts". *Biochim. Biophys. Acta.* 1596 (2002), pp. 357–365. DOI: 10.1016/S0167-4838(02)00225-X.

[201] V. Granata, G. Graziano, A. Ruggiero, G. Raimo, M. Masullo, P. Arcari, L. Vitagliano, and A. Zagari. "Stability against temperature of *Sulfolobus solfataricus* elongation factor 1α, a multi-domain protein". *Biochim. Biophys. Acta.* 1784 (2008), pp. 573–581. DOI: 10.1016/j.bbapap.2007.12.018.

[202] B. Keller, X. Daura, and W. van Gunsteren. "Comparing geometric and kinetic cluster algorithms for molecular simulation data". *J. Chem. Phys.* 132 (2010), p. 074110. DOI: 10.1063/1.3301140.

[203] H. Yu, A. N. Gupta, X. Liu, K. Neupane, A. M. Brigley., I. Sosova, and M. T. Woodside. "Energy landscape analysis of native folding of the prion protein yields the diffusion constant, transition path time and rates". *Proc. Natl. Acad. Sci. U.S.A.* 109 (2012), p. 14452. DOI: 10.1073/pnas.1206190109.

[204] A. Möglich, K. Joder, and T. K. Thomas. "End-to-end distance distributions and intrachain diffusion constants in unfolded polypeptide chains indicate intramolecular hydrogen bond formation". *Proc. Natl. Acad. Sci. U.S.A.* 103 (2006), pp. 12394–12399. DOI: 10.1073/pnas.0604748103.

[205] J. S. Nicolis. *Chaos and information processing : A heuristic outline*. Singapore, Teaneck, NJ: World Scientific, 1991.

[206] H. Nymeyer, A. Garcia, and J. Onuchic. "Folding funnels and frustration in off-lattice minimalist protein landscapes". *Proc. Natl. Acad. Sci. U.S.A.* 95 (1998), pp. 5921–5928. DOI: 10.1073/pnas.95.11.5921.

[207] R. Best and G. Hummer. "Coordinate-dependent diffusion in protein folding". *Proc. Natl. Acad. Sci. U.S.A.* 1088-1093 (2010). DOI: 10.1073/pnas.0910390107.

[208] K. Abel, M. D. Yoder, R. Hilgenfeld, and F. Jurnak. "An α to β conformational switch in EF-Tu". *Structure* 4 (1996), pp. 1153–1159. DOI: 10.1016/S0969-2126(96)00123-2.

[209] G. Polekhina, S. Thirup, M. Kjeldgaard, P. Nissen, C. Lippmann, and J. Nyborg. "Helix unwinding in the effector region of elongation factor EF-Tu–GDP". *Structure* 4 (1996), pp. 1141–1151. DOI: 10.1016/S0969-2126(96)00122-0.

[210] E. Villa, J. Sengupta, L. G. Trabuco, J. LeBarron, W. T. Baxter, T. R. Shaikh, R. A. Grassucci, P. Nissen, M. Ehrenberg, K. Schulten, and J. Frank. "Ribosome-induced changes in elongation factor Tu conformation control GTP hydrolysis". *Proc. Natl. Acad. Sci. U.S.A.* 106.4 (2009), pp. 1063–1068. DOI: 10.1073/pnas.0811370106.

[211] H. Berchtold, L. Reshetnikova, C. Relser, N. Schirmer, M. Sprinzl, and R. Hilgenfeld. "Crystal structure of active elongation factor Tu reveals major domain rearrangements". *Nature* 365.6442 (1993), pp. 126–132. DOI: 10.1038/365126a0.

[212] M. Kjeldgaard, P. Nissen, S. Thirup, and J. Nyborg. "The crystal structure of elongation factor EF-Tu from *Thermus aquaticus* in the GTP conformation". *Structure* 1.1 (1993), pp. 35–50. DOI: 10.1016/0969-2126(93)90007-4.

[213] T. Kawashima, C. Berthet-Colominas, M. Wulff, S. Cusack, and R. Leberman. "The structure of the *Escherichia coli* EF-Tu·EF-Ts complex at 2.5 Å resolution". *Nature* 379.6565 (1996), pp. 511–518. DOI: 10.1038/379511a0.

[214] L. Vogeley, G. Palm, J. Mesters, and R. Hilgenfeld. "Conformational change of elongation factor Tu (EF-Tu) induced by antibiotic binding. Crystal structure of the complex between EF-Tu·GDP and aurodox". *J. Biol. Chem.* 276.20 (2001), pp. 17149–17155. DOI: 10.1074/jbc.M100017200.

[215] Y. Wang, Y. Jiang, M. Meyering-Voss, M. Sprinzl, and P. Sigler. "Crystal structure of the EF-Tu – EF-Ts complex from *Thermus thermophilus*". *Nature Struct. Biol.* 4.8 (1997), pp. 650–656. DOI: 10.1038/nsb0897-650.

[216] G. Andersen, L. Pedersen, L. Valente, I. Chatterjee, T. Kinzy, M. Kjeldgaard, and J. Nyborg. "Structural basis for nucleotide exchange and competition with trna in the yeast elongation factor complex eEF1A:eEF1Bα". *Molecular Cell* 6.5 (2000), pp. 1261–1266. DOI: 10.1016/S1097-2765(00)00122-2.

[217] G. Andersen, L. Valente, L. Pedersen, T. Kinzy, and J. Nyborg. "Crystal structures of nucleotide exchange intermediates in the eEF1A-eEF1Bα complex". *Nature Struct. Biol.* 8.6 (2001), pp. 531–534. DOI: 10.1038/88598.

[218] D. L. Miller and H. Weissbach. *Molecular mechanisms of protein biosynthesis*. Ed. by H. Weissbach. Academic Press, 1977, pp. 323–373.

[219] A. Adamczyk and A. Warshel. "Converting structural information into an allosteric-energy-based picture for elongation factor Tu activation by the ribosome". *Proc. Natl. Acad. Sci. U.S.A.* 108.24 (2011), pp. 9827–9832. DOI: 10.1073/pnas.1105714108.

[220] J. Chodera, W. Swope, J. Pitera, C. Seok, and K. Dill. "Use of the weighted histogram analysis method for the analysis of simulated and parallel tempering simulations". *J. Chem. Theory Comput.* 3.1 (2007), pp. 26–41. DOI: 10.1021/ct0502864.

[221] F. Forcellino and P. Derreumaux. "Computer simulations aimed at structure prediction of supersecondary motifs in proteins". *Proteins: Struct., Funct., Genet.* 45 (2001), pp. 159–166. DOI: 10.1002/prot.1135.

[222] A. García, R. Blumenfeld, G. Hummer, and J. Krumhansl. "Multi-basin dynamics of a protein in a crystal environment". *Physica D* 107 (1997), pp. 225–239. DOI: 10.1016/S0167-2789(97)00090-0.

[223] D. Wales. "Energy landscapes: Some new horizons". *Curr. Opin. Struct. Biol.* 20 (2010), pp. 3–10. DOI: 10.1016/j.sbi.2009.12.011.

[224] M. Hinczewski, Y. von Hansen, J. Dzubiella, and R. Netz. "How the diffusivity profile reduces the arbitrariness of protein folding free energies". *J. Chem. Phys.* 132.24 (2010), p. 245103. DOI: 10.1063/1.3442716.

[225] P. Derreumaux. "A diffusion process-controlled Monte Carlo method for finding the global energy minimum of a polypeptide chain. I. Formulation and test on a hexadecapeptide". *J. Chem. Phys.* 106.12 (1997), pp. 5260–5270. DOI: 10.1063/1.473525.

[226] W. Song, G. Wei, N. Mousseau, and P. Derreumaux. "Self-assembly of the beta 2-microglobulin NHVTLSQ peptide using a coarse-grained protein model reveals beta-barrel species". *J. Phys. Chem. B* 112 (2008), pp. 4410–4418. DOI: 10.1021/jp710592v.

[227] Y. Chebaro and P. Derreumaux. "Targeting the early steps of Aβ16-22 protofibril disassembly by N-methylated inhibitors: A numerical study". *Proteins* 75 (2009), pp. 442–452. DOI: 10.1002/prot.22254.

[228] A. Barducci, M. Bonomi, and P. Derreumaux. "Assessing the quality of the OPEP coarse-grained force field". *J. Chem. Theory Comput.* 7.6 (2011), pp. 1928–1934. DOI: 10.1021/ct100646f.

[229] R. Zhou, B. Berne, and R. Germain. "The free energy landscape for β hairpin folding in explicit water". *Proc. Natl. Acad. Sci. U.S.A.* 98 (2001), pp. 14931–14936. DOI: 10.1073/pnas.201543998.

[230] H.-X. Zhou and F. Dong. "Electrostatic contributions to the stability of a thermophilic cold shock protein". *Biophys. J.* 84.4 (2003), pp. 2216–22. DOI: 10.1016/S0006-3495(03)75027-9.

[231] A. D. M. Chapman, A. Cortés, T. R. Dafforn, A. R. Clarke, and R. L. Brady. "Structural basis of substrate specificity in malate dehydrogenases: Crystal structure of a ternary complex of porcine cytoplasmic malate dehydrogenase, α-Ketomalonate and TetrahydoNAD". *J. Mol. Biol.* 285.2 (1999), pp. 703–712. DOI: 10.1006/jmbi.1998.2357.

[232] N. Coquelle, E. Fioravanti, M. Weik, F. Vellieux, and D. Madern. "Activity, stability and structural studies of lactate dehydrogenases adapted to extreme thermal environments". *J. Mol. Biol.* 374.2 (2007), pp. 547–562. DOI: 10.1016/j.jmb.2007.09.049.

[233] G. Somero. "Adaptation of enzymes to temperature: Searching for basic "strategies"". *Comp. Biochem. Phys. B* 139 (2004), pp. 321–333. DOI: 10.1016/j.cbpc.2004.05.003.

[234] P. Minárik, N. Tomaásková, M. Kollárová, and M. Antalík. "Malate dehydrogenases – structure and function". *Gen. Physiol. and Biophys.* 21.3 (2002), pp. 257–265.

[235] B. Dalhus, M. Saarinen, U. Sauer, P. Eklund, K. Johansson, A. Karlsson, S. Ramaswamy, A. Bjørk, B. Synstad, K. Naterstad, R. Sirevåg, and H. Eklund. "Structural basis for thermophilic protein stability: Structures of thermophilic and mesophilic malate dehydrogenases". *J. Mol. Biol.* 318.3 (3, 2002), pp. 707–721. DOI: 10.1016/s0022-2836(02)00050-5.

[236] R. Talon, N. Coquelle, D. Madern, and E. Girard. "An experimental point of view on hydration/solvation in halophilic proteins". *Front. Microbiol.* 5.FEB (2014). DOI: 10.3389/fmicb.2014.00066.

[237] C.-H. Hung, T.-S. Hwang, Y.-Y. Chang, H.-R. Luo, S.-P. Wu, and C.-H. Hsu. "Crystal structures and molecular dynamics simulations of thermophilic malate dehydrogenase reveal critical loop motion for co-substrate binding". *PLoS One* 8.12 (2013), e83091. DOI: 10.1371/journal.pone.0083091.

[238] Y.-Y. Chang, C.-H. Hung, T.-S. Hwang, and C.-H. Hsu. "Cloning, overexpression, purification and crystallization of malate dehydrogenase from *Thermus thermophilus*". *Acta Crystallogr. Sect. F* 69.11 (2013), pp. 1249–1251. DOI: 10.1107/S174430911302472X.

[239] B. Hess, C. Kutzner, D. Van Der Spoel, and E. Lindahl. "GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation". *J. Chem. Theory Comput.* 4.3 (2008), pp. 435–447. DOI: 10.1021/ct700301q.

[240] G. Torrie and J. Valleau. "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling". *J. Comput. Phys.* 23.2 (1977), pp. 187–199. DOI: 10.1016/0021-9991(77)90121-8.

[241] D. Madern. "Molecular evolution within the L-malate and L-lactate dehydrogenase super-family". *J. Mol. Evol.* 54 (2002), pp. 825–840. DOI: 10.1007/s00239-001-0088-8.

[242] A. Bjørk, D. Mantzilas, R. Sirevåg, and V. Eijsink. "Electrostatic interactions across the dimer-dimer interface contribute to the pH-dependent stability of a tetrameric malate dehydrogenase". *FEBS Lett.* 553 (2003), pp. 423–426. DOI: 10.1016/S0014-5793(03)01076-7.

[243] A. Bjørk, B. Dalhus, D. Mantzilas, R. Sirevåg, and V. Eijsink. "Large improvement in the thermal stability of a tetrameric malate dehydrogenase by single point mutations at the dimer–dimer Interface". *J. Mol. Biol.* 341 (2004), pp. 1215–1226. DOI: 10.1016/j.jmb.2004.06.079.

[244] J. Lee and S.-H. Kim. "Water polygons in high-resolution protein crystal structures". *Prot. Sci.* 18.7 (2009), pp. 1370–1376. DOI: 10.1002/pro.162.

[245] P. Mark and L. Nilsson. "Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K". *J. Phys. Chem. A* 105.43 (2001), pp. 9954–9960. DOI: 10.1021/jp003020w.

[246] W. Jorgensen and C. Jenson. "Temperature dependence of TIP3P, SPC, and TIP4P water from NPT Monte Carlo simulations: Seeking temperatures of maximum density". *J. Comput. Chem.* 19.10 (1998), pp. 1179–1186. DOI: 10.1002/(SICI)1096-987X(19980730)19:10<1179::AID-JCC6>3.0.CO;2-J.

[247] T. Sun, F.-H. Lin, R. L. Campbell, J. S. Allingham, and P. L. Davies. "An antifreeze protein folds with an interior network of more than 400 semi-clathrate waters". *Science* 343.6172 (2014), pp. 795–798. DOI: 10.1126/science.1247407.

[248] M. M. Teeter. "Water structure of a hydrophobic protein at atomic resolution: Pentagon rings of water molecules in crystals of crambin". *Proc. Natl. Acad. Sci. U.S.A.* 81 (1984), pp. 6014–6018. DOI: 10.1073/pnas.81.19.6014.

[249] M. I. Klotz. "Protein hydration and behavior; many aspects of protein behavior can be interpreted in terms of frozen water of hydration". *Science* 128.3328 (1958), pp. 815–822. DOI: 10.1126/science.128.3328.815.