IN VITRO AND *IN VIVO* PROTEIN FOLDING UNDER STRESS

BY

KAPIL DAVE

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Biophysics and Computational Biology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Professor Martin Gruebele, Chair
Professor Robert B. Gennis
Professor Deborah E. Leckband
Assistant Professor Thomas E. Kuhlman

ABSTRACT

Proteins are subject to a variety of stresses in biological organisms, including pressure and temperature, which are the easiest stresses to simulate by molecular dynamics simulations. The thesis will focus on discussing the effect of pressure and thermal stress on proteins including some of the fast-folding model proteins, who's *in vitro* folding can be fully simulated on computers and compared directly with experiments. Pressure and temperature are prototypical perturbations that illustrate how close many proteins are to instability, a property that cells can exploit to control protein function. I will conclude with some recent in-cell experiments, and progress being made in measuring protein stability and function inside live cells under high pressure conditions.

In chapters 2 and 3, fast-folding WW domains were studied (best-characterized systems for comparing experiments with simulations) by T-jump relaxation in conjunction with protein engineering. Chapter 1 is a comprehensive data set of mutational Φ-values (ΦM) as indicators for folding transition-state structure of 65 side chain, 7 backbone hydrogen bond, and 6 deletion and /or insertion mutants within loop 1 of the 34-residue hPin1 WW domain. We probed the robustness of the two hydrophobic clusters in the folding transition state, and discussed how local backbone disorder in the native-state can lead to non-classical ΦM-values (ΦM > 1) in the rate-determining loop 1 substructure, and conclusively identify mutations and positions along the sequence that perturb the folding mechanism from loop 1-limited toward loop 2-limited folding. In chapter 2 we mutated the FBP 28 WW domain (formin-binding protein; Leu26 by Asp26 or Trp26) to alter the folding scenario from three-state folding toward two-state or downhill folding at temperatures below the melting point of the protein. The investigation was conducted using a combination of simulations over a broad temperature range with experimental temperature-jump data. Chapter 4 is focused on how attaching fluorescent protein tags to a host protein in vitro has a large non-additive effect on its folding free energy. We compared an unlabeled, three singly-labeled, and a doubly-labeled enzyme PGK (phosphoglycerate kinase). Two mechanisms for non-additivity were proposed. In the "quinary interaction" mechanism, two tags interact transiently with one another, relieving the host protein from unfavorable tag–protein interactions. In the "crowding" mechanism, adding two tags provides the minimal crowding necessary to overcome destabilizing interactions of individual tags with the host protein. Both of these mechanisms affect protein stability in cells; they must

also be considered for tagged proteins used for reference in vitro. In Chapter 5 we showed that the protein unfolding/refolding reaction can be driven by a periodic thermal excitation above the reaction threshold. We were also able to speed up the reaction from an undetectable to a detectable rate by the addition of artificial thermal noise. A maximum in the recovered signal as a function of thermal noise was seen, a stochastic resonance. The study alluded that correlated noise is a physically and chemically plausible mechanism by which cells could modulate biomolecular dynamics during threshold processes such as signaling. Chapter 6 explores folding competing with misfolding or aggregation on the μs time scale using tethered WW domains. Tethered protein construct was engineered by linking two or more copies of the fast folding Fip35 WW domain with a flexible linker. We observed that adding more monomer units led to thermodynamic destabilization and slower folding rates, along with an abrupt onset of protein-protein interaction. Kinetics were determined by performing ultrafast laser temperature jump experiments at different temperatures and denaturant concentration. A simple multimeric network model is also proposed for globally fitting the thermodynamics and kinetics data. In the final chapter 7 of this thesis folding of an enzyme phosphoglycerate kinase (PGK) was studied under high pressure stress in different bacterial cytoplasm. The motivation was to understand how cell is capable of modulating the stability of its proteome when subjected to external stress especially high hydrostatic pressure. The thermodynamic stability of PGK was measured in two different strains Wildtype MG1655 and known pressure resistant J1 strain. These results were compared to in vitro experiments to reveal that cellular environment has an overall stabilizing effect on the protein thermodynamic stability but different cellular cytoplasm doesn't affect the stability of PGK significantly.

To my parents, for the hardships they have dealt with to support me and my education

ACKNOWLEDGMENTS

"*University is the place where knowledge thrives and inspiration is floating,*

I seek it and make it my own, whether its source be living or non-living"

Kapil Dave

First and foremost I would like to thank my advisor Prof. Martin Gruebele. It is my immense pleasure to have interacted and work closely with Martin for the past four years. He is an awesome scientist full of enthusiasm and new ideas. The one incidence that happened during my struggle with the laser system in the lab which I can't forget is Martin came and saw me working on the laser and left suggesting to change a mirror and few hours later I did the same and Boom! The laser started lasing. I was amazed with the way he saw the problem and addressed it. I am grateful to have the freedom that Martin offers to his graduate students, and it's his trust and expectations from me that kept me going during my PhD. I like the winning spirit in Martin not only in science but also in the races and challenges he overcomes. He serves as an inspiration for me. I have learnt from him in order to get what you desire you need to work hard and plan.

Second, I want to thank all the members in the Gruebele group for their company and support: Drishti Guin, Shahar Sukenik, Tanya Perlova, Ruopei Fang, Anne Jean Wirth, Shu-Han Chao, Timothy Chen, Mayank M. Boob, Aniket Ravan, Meredith Rickard, Caitlin Davis, and Lydia Kisley. I want to express my special thanks to Dr. Max Platkov for providing me with valuable suggestions and advice on both professional and personal fronts during his brief but productive summer at UIUC. We had a great time working on stochastic resonance experiments and sharing ideas over late night coffee visits.

I am greatly indebted to Cindy Dodds for keep a track on my progress all these years and helping me with any all administrative obstacles. I would love to thank Karen Watson, Beth Myler, Theresa Struss and the IMP office for all the praise and sweet treat alerts.

I would like to thank my committee Prof. Deborah Leckband, Prof. Robert Gennis and Assistant Professor Tom Kuhlman. I am extremely fortunate to have Tom help me with my

bacterial imaging experiments. He was always there to provide me with any support I needed to finish my experiments.

The University of Illinois at Urbana-Champaign is a beautiful place to live and work as a student. The sports and recreational facilities CRCE and ARC are one of the best in the country. I have found friends here at UIUC with whom I have had fun and wonderful time all these years. I would like to thank all of them but specially Heena Gajjar, Saloni Chawla, Nitesh Shashikanth, Punit Singhvi, Mamata Guragain and Soham Mujumdar. I am fortunate enough to have meet John and Daine who served as parents and guardian which made the place away from home to feel more like home. My friends back in India Mehreen Khaleel, Ginny Karir, and D. Jeiyendira Pradeep have constantly believed in me and even though we were thousands of mile apart they never complained and stayed in contact.

Last but not least, I would like to thank my parents and my sister Ronak Dave for their love and support. It is because of my parent's hardships and love that I am able to achieve this feat in my life - without them I would never would have been possible.

TABLE OF CONTENTS

APPENDIXES

LIST OF ABBREVIATIONS

NMR	Nuclear Magnetic Resonance
FReI	Fluorescence Relaxation Imaging
FRET	Forster Resonance Energy Transfer
PGK	Phosphoglycerate Kinase
D/A	Donor-acceptor Fluorescence Ratio
VlsE	Variable-major Like Sequence, Expressed
CD	Circular Dichroism
MD	Molecular Dynamics
ER	Endoplasmic Reticulum
ATP	Adenosine Triphosphate
IDP	Intrinsically Disordered Proteins
GFP	Green Fluorescent Protein
PMT	Photomultiplier Tube
SR	Stochastic Resonance
FPLC	Fast Pressure Liquid Chromatography
GST	Glutathione S-transferase

CHAPTER 1

Introduction

Protein folding produces much of the cell's signaling, structural and catalytic machinery. It happens first upon ribosomal synthesis [1], often with membrane insertion via the translocon [2] [3], but also later on in the cell: One of the most important things learned from *in vitro* folding experiments is that even cytosolic globular proteins have fairly small folding equilibrium constants. Therefore proteins will unfold and refold many times during their life cycle [4]. With the exception of a few extraordinarily stable proteins [5], relatively low stability goes hand in hand with the flexibility required for protein function. Some proteins even remain unstructured after translation and fold upon binding to specific targets [6] [7].

In vitro studies also taught us that folding is a very fast chemical reaction (microseconds to hours at room temperature). Its free energy barriers $\Delta G^{\dagger}$ must be quite small, in some cases on the order of the thermal energy $k_B T$ [8] [9]. Thanks to small folding free energies and small activation barriers, one might expect that the complex solvation environment of the cell can control protein stability and kinetics, and indeed it can [10]. In addition, a network of chaperones can hold misfolded proteins, direct them towards degradation pathways, or unfold them, giving proteins inside cells another chance to fold autonomously, as most proteins do *in vitro* at low concentration when aggregation is unlikely [11] [12].

The plausibility of *in vivo* effects on folding is apparent from *in vitro* studies: slight temperature changes, addition of small molecules, or crowding by large molecules can shift protein equilibria between unfolded and folded ensembles [13]. Such shifts are often "cooperative," by which we mean that they occur over a narrow range of conditions [14]. While the cell modulates the folding free energy landscape, it does not appear to fundamentally alter the way proteins are observed to fold *in vitro* [15].

This chapter is partially adapted from Gruebele, M.; Dave,K.; Sukenik, S. Globular protein folding in vitro and in vivo. Annual Review of Biophysics. Annual Reviews, **2016**

True understanding requires that one should be able to put something back together again after taking it apart. Protein scientists have gone through this process in a variety of ways. Although *de novo* design of active proteins is still not routine, much progress has been made in that field [16] [17]. Likewise model building has gone well. On the "energetic" side, the energy landscape model has explained many of the general [9] [18] and specific [19] features of folding. On the "structural" side, models have advanced from beads on lattices [20] to all-atom simulations based on empirical force fields [21] [22] [23]. The last 10 years have seen a remarkable confluence of protein (un)folding experiments, protein design, protein landscape models, and simulations of folding. The state-of-the-art is proteins of ≈100 residues, folding faster than a few milliseconds if a direct comparison of simulations and experiments is to be made [24]. Most domains of larger proteins are <150 amino acids long, and such domains usually fold relatively independently from one another [13]. We are thus not far off from the holy grail where folds can be reliably computed, just as the structure of small organic molecules can be computed readily with quantum chemistry packages [25].

Many interesting problems remain to be solved. While computation can predict the fold of some small proteins, it is not yet clear how accurate the predicted mechanisms are. This is partly the fault of experiments, which have difficulty providing structural information on the time scale of the actual reaction (barrier crossing) events. An important question is "How detailed do we really need to be to have useful predictions?" While folding reactions can be described adequately by simple mechanisms along one or two reaction coordinates [26], considerable complexity lurks below this apparent simplicity (Fig. 1) [27] [28]. In particular, the unfolded ensemble has more structure and interesting dynamics than it is often given credit for [29] [30]. And of course there is the question of how cells productively fine-tune the energy landscape of their proteins to enhance survival [31]. Finally, other interesting problem such as the effect of applied force on energy landscapes [32], or misfolding and amyloids [33] will only be discussed briefly.

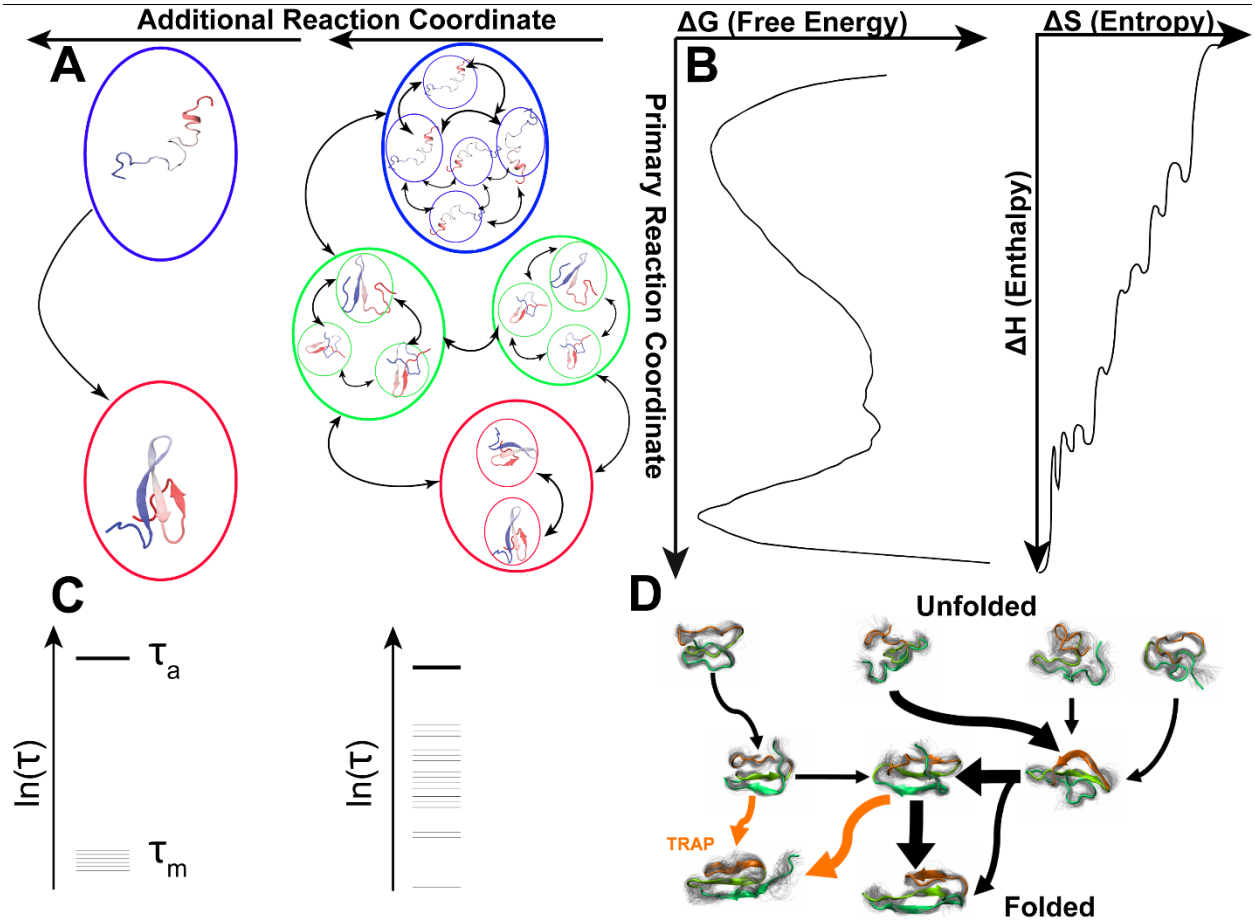

Fig. 1.1: Folding simplicity and complexity: the structural and energetic view. (A) In the most coarse-grained structural picture, only two macrostates ("unfolded," top, "folded," bottom) play a role. The black arrow indicates interconversion through a transition state (whose probabilities of folding and unfolding are pF=pU=0.5). More detailed analysis reveals sub-states of folded [34],unfolded, and transitional ensembles, each containing many microstates. A rich kinetic network occasionally includes parallel paths (if their free energies are within a few kT, so one is not favored over the other). (B) In the most coarse-grained energetic picture, the free energy $\Delta G(x)$ has unfolded and folded minima along just one coordinate x, and the folding enthalpy of a protein is well-funneled as a function of the polypeptide configurational entropy. The lower energy native state has lower configurational entropy and lower enthalpy. (C) For slow folders, the "molecular time" $\tau_m \approx 0.1\text{-}1$ μs during which transition between states occurs is well-separated from the "dwell time" within states, τ_a. For fast folders, or proteins with many folding intermediates covering a wide range of barriers, the time scales overlap. (D) Simplified WW domain kinetic network, showing an actual calculated example of the less coarse-grained picture in (b) [28].

1.1 *In Vitro* protein folding

1.1.1 Structural and energetic models for folding

Perhaps the most basic model used to understand protein folding is the hydrophobic-hydrophilic (HP) residue model [35]. The HP model accounts for hydrophobicity as the major driving force to make compact, de-solvated structures [36] [35] [37], while also allowing local secondary structure formation. Hydrophobicity has been reviewed extensively [38] [39], and while not purely an entropic effect, water molecules avoiding ordered interaction with sidechains buried in the protein's core plays an important role. Several key ideas emerged from such simple models: Certain sequences are more likely to avoid kinetic traps and produce robust folded states that are rapidly accessible from unfolded conformations. Even though no solvent is included explicitly in HP models, hydrophobicity highlights that "the solvent folds, too" when a protein folds. Although two-amino acid alphabets do not fold proteins in practice, alphabets with as few as five residues have been successful [40] and disordered proteins also have reduced alphabets [41]. Of course, a larger alphabet of 20 different amino acid residues still leads to better-packed structures that are more fine-tuned by evolution for function [42]. For a more detailed discussion of theoretical protein models, see the review in this volume by Schuler.

Ideas such as local secondary structure formation or hydrophobicity involve a successive reduction of the search space as the search for the native state goes on. For example, hydrophobicity partitions residues into "more likely inside" and "more likely outside." Go realized that proteins are evolved to have consistent interactions [43], while Frauenfelder proposed hierarchical energy landscapes of native proteins [34]. Such concepts led to a quantitative energy landscape theory of folding [9] [35]. In energy landscape theory, the Levinthal paradox [44] is overcome because enthalpy loss ΔH and entropy loss ΔS_C are correlated as a protein folds, and such enthalpy-entropy compensation [45] overcomes unavoidable enthalpic "noise." The correlation (funnel shape of ΔH as a function of ΔS_C in Fig. 1b) explains why proteins fold over low free energy barriers [8]. The noise in the enthalpy funnel explains traps and intermediates when folding is not perfectly streamlined, or frustrated, in analogy to terminology used in dynamics of glasses. The funneled function $\Delta H\ (\Delta S_C)$ should not be confused with the free energy $\Delta G(x) = \Delta H - T\Delta S$ as a function of reaction progress coordinate x (illustrated for several cases in Fig. 2). Although the funnel is downhill in

enthalpy, the free energy is not necessarily downhill because ΔH and $-T\Delta S$ may not compensate for all values of x.

As computational power has grown, increasingly realistic computer models of folding have become possible [22] [24]. Even downhill folders spend about a microsecond to get to the native state [47] [48], so the major hurdle in computational modeling is the time needed to sample the conformational space before interesting events happen. Coarse graining is a powerful approach that dramatically decreases the computational demands of protein simulations [49] [50]. In a coarse grained model, clusters of atoms are modeled as a unit, interacting via an appropriately averaged force field. In parallel, implicit solvent models greatly reduced the number of atoms tracked in classical molecular dynamics simulations, and yielded interesting folding behavior such as a dominant but parallel pathway [51]. The development of parallel simulation methods greatly improved sampling. Many parallel calculations can be sampled in search of a few successful folding events for comparison with experiment [52]. Independent calculations can be stitched together (Markov state models) to reveal short-lived or long-lived microstates [53]. Replicas can exchange between simulations to provide rapid thermodynamic sampling [54] [55]. For example, replica exchange has computationally revealed "downhill" free energy surfaces for folding [56]. Recently advances in computational power have made possible all-atom single-trajectory protein folding simulations, in which a single protein unfolds and refolds many times in equilibrium [24]. As with experiments, the greatest challenge of simulations is to find the most informative reaction coordinates [57] [58] [59].

1.1.2 Fast folding proteins unite experiment and computation

Some small protein domains fold/unfold in microseconds between just two macrostates (illustrated in Fig. 1.2 A), or even downhill (Fig. 1.2 B). Of course disulfide bridges [60] [61], proline isomers [62], many types of intermediates [63], and domain interactions [64] can complicate the picture in general. Yet small, fast folders reveal the minimal requirements for folding, and currently form the best link between experiment, theory and simulation [59]. Fig. 1.2 A illustrates the free-energy landscape two-state folding, with all highly populated conformations belonging to either the folded or the unfolded ensemble. These ensembles are in local free energy landscape minima, separated by a barrier that needs to be crossed to transition between them. Experimentally, one hallmark of two-state folding is obtaining the

same melting temperature (T_m,) via (sometimes different) spectroscopic measurement techniques that probe different parts of the energy landscape [65] [66].

Such behavior breaks down when intermediates are populated during the course of the folding [67] [68] [48], or in the scenario of downhill folding Fig. 1.2 B [69]. Downhill folding was predicted by energy landscape theory in the special case where decreasing enthalpy and entropy compensated throughout the whole reaction [70] [71]. In addition to the thermodynamic observation of downhill folding [72], the gradual breakdown of timescale separation as downhill folding is approached (Fig. 1.1 D) has also been seen kinetically [73] [48]. Depending on initial conditions and protein stability modified by mutations, proteins can switch from downhill folding at low temperatures to two-state folding at temperatures close to T_m [74] [75], or from downhill folding to folding via intermediates [76]. Two-state and downhill mechanisms are not common in large or multi-domain globular proteins, where the "noise" in the funnel (Fig. 1.1 C, and 1.2 D) is larger, and traps or intermediates occur in the free energy landscape [77].

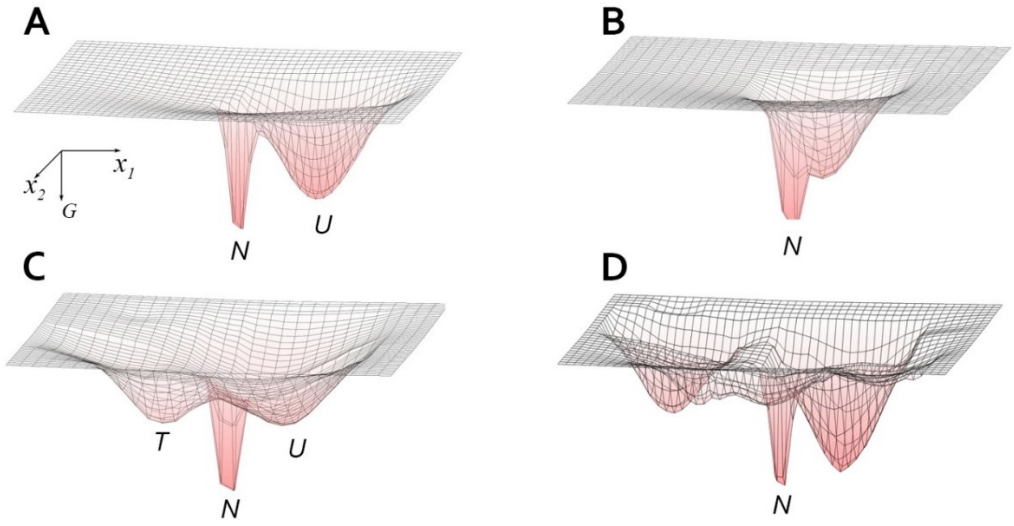

Fig. 1.2: Free-energy landscapes of protein folding highlight several scenarios. A) Scenario of two state folding with well-defined native (labelled *N*) and unfolded (labelled *U*) well separated by a barrier. Axes illustrate the two folding reaction coordinates (x_1 and x_2) and folding free energy (*G*). B) Downhill folding portrayed as native and unfolded well separated by a low lying barrier (~ <3 *kT*). C) The scenario of a free-energy trap (labelled *T*) is added to the two-state scenario. D) The concept of multiple folding pathways available to the protein in case of change in environmental conditions or mutations is shown by the presence of various minima ending up in the native state of the protein. (See also the review in this volume by Barrick describing experimental realization of such landscapes.)

1.1.3 Friction and the speed limit of folding

The scenario of downhill folding represents a case without kinetic barriers, when folding occurs at the "speed limit" [78]. How fast can such proteins actually fold? The lack of a barrier between macrostates means that the rate of folding is limited by polypeptide chain diffusion and solvent friction [79] [80]. Note that a process described as diffusion in a coarse-grained coordinate system can still involve many crossings over small barriers. For example, an individual backbone dihedral angle rotation occurs over a small barrier, but when many Ramachandran angles [81] are coarse grained into a few slower reaction coordinates, the fast motions can be treated as a friction-dependent prefactor in the Arrhenius equation $k_f \sim \nu^{\dagger} e^{-\Delta G^{\dagger}/kT}$ [13]. Here $\nu^{\dagger}$ is the Arrhenius prefactor, $\Delta G^{\dagger}$ is the activation free energy, and k_f is the rate coefficient of the forward reaction. The exact scaling of the prefactor is still under debate [82] [83] [84].

Measurements on cytochrome c chain diffusion [84] [85] estimated a minimum time around ~1 μs for the polypeptide chain to collapse. Secondary structure (helices, beta sheets) can form on a similar time scale [86] [87] [88]. Studies using triplet energy transfer [83] have quantified chain length, location and composition dependence for contact formation, ranging from 10-100 ns. Correspondingly large speed gains have been achieved for small proteins. An example of a protein mutated almost to the speed limit is the GTT variant of a WW domain [89], which was suggested by analysis of a long molecular dynamics trajectory. Another illustration is the three helix bundle $prb_{7\text{-}53}$, in which the wild-type protein is mutated, replacing charged with hydrophobic residues. These computationally designed mutations again pushed folding down to 1 μs [90], close to the theoretical limit.

1.1.4 Trade-off: folding *vs*. function

Globular proteins must attain a well-defined native structure in order to perform the requisite biological functions under specific environmental conditions (pH, temperature, solvent, salts). Even many disordered proteins fold upon binding [91] [92]. The fast folders previously discussed show that evolution for function is an important factor that leads to frustrated folding. For example, when loop 1 of Pin1 WW Domain was truncated to speed up folding, the mutant lost the ability to bind to a phosphorylated target protein that is bound with high affinity [93]. When the beta-bulge of interleukin-1β is replaced by a faster-folding β turn, protein function is again inhibited [94]. Mutations that speed up folding, making shorter loops, more stable helices, or more hydrophobic cores are likely to eliminate charged residues needed for enzymatic function, loops needed for binding, or reduce flexibility needed for docking or substrate diffusion.

The possible explanation is that stabilizing mutations make the native structure too rigid compared to the wild-type, restricting it from sampling other conformational states which facilitate its binding. Low stability can even enhance function: According to the fly-casting mechanism [7], the unfolded form of the protein binds weakly at large distances and folding and binding then go hand-in-hand. Functional proteins are an outcome of co-evolution between the need to fold and the need to perform function [4].

1.1.5 The diversity of folding pathways

There has been a long-standing discussion as to what extent proteins fold through sequential intermediates or parallel pathways; downhill or over obligatory barriers; with or without traps [95] [96] [97] [98] The answer is: all of the above! *In vitro* experiments, theory, and computation all agree that proteins have very shallow free energy landscapes. Depths of valleys and heights of saddle points (barriers) are measured in 10s of kJ/mole, not 100s kJ/mole as for chemical bond-making reactions. On such reaction surfaces, if their dimensionality is low but not necessarily equal to 1, many scenarios are possible. Nonetheless, a given mutant under a given solvent condition will almost always fold via a dominant pathway. In a typical experiment with a signal-to-noise ratio of 50:1, any additional pathways more than $kT\ln(50) \approx 4kT$ up in energy will simply be invisible. In simulations, such events will be rare and also hard

to detect unless many folding/unfolding transitions can be sampled. For this reason, we have coined the phrase "apparent X-folder," where X is the mechanism of choice. Monitoring more reaction coordinates (see review by Englander and Marqusee in this issue describing the monitoring of multiple reaction coordinates using NMR hydrogen exchange), going to higher free energy, or perturbing the system (e.g. temperature, pressure, solution conditions) will always reveal new paths and mechanisms [99] [100] [101] [102] [75].

One case where alternative folding pathways become visible is for multi-repeat proteins. Evidence of parallel folding pathways has been seen by comparing rates for symmetric consensus repeat proteins (CARPs). Folding domains in parallel speeds up overall folding [103]. The increase in folding rates with the chain length of the repeat protein stands in contrast to what is seen for globular proteins, and is clear evidence of parallel folding. Perhaps not surprisingly, gene duplication is a key bootstrap for protein evolution.

Are these alternative paths important in general? This question deserves further investigation. Structure is more robust than mechanism, which is why structure prediction is easier than mechanism prediction [104]. However, the very process of evolution that stabilizes native states *vs.* higher energy states while maximizing function may be the reason for alternative mechanisms and parallel paths. Evolution requires a certain flexibility, and digging too deep a funnel may reduce the evolvability of sequences [4]. Appearance of new function upon mutation must eventually go hand-in-hand with a different folding mechanism and alternative folding pathways.

1.2 Protein folding in-cell

To facilitate proper function, a cell must maintain an internal balance of metabolic, regulation, and transcriptional pathways. In addition, the cell must be able to maintain homeostasis in a changing environment. This is possible thanks to a complex network of regulation, which is carried out primarily by proteins in response to internal and external signals. To this end, cells use a range of strategies to deal with deleterious environmental conditions – from the synthesis of specialized protein machines that ensure proper folding or proteolysis of misfolded proteins [105], to the uptake or synthesis of stabilizing osmolytes, discussed in the previous section [106]. Importantly, many factors in the cell, as well as in the cell's environment, will have dramatic effects on protein folding, as illustrated in Fig. 1.3.

Macromolecules are estimated to take up roughly 0.3-0.5 g per mL of cellular solution [107]. Water content is roughly 70% of total cell mass [108]. Proteins thus take up roughly half of the dry weight of the cell. DNA and lipids take up about a third of the same dry mass, and 10-15% belong to other molecules, mostly low molecular mass species. The small percentage by mass of small molecules is deceptive because these molecules can exist at molar cellular concentrations.

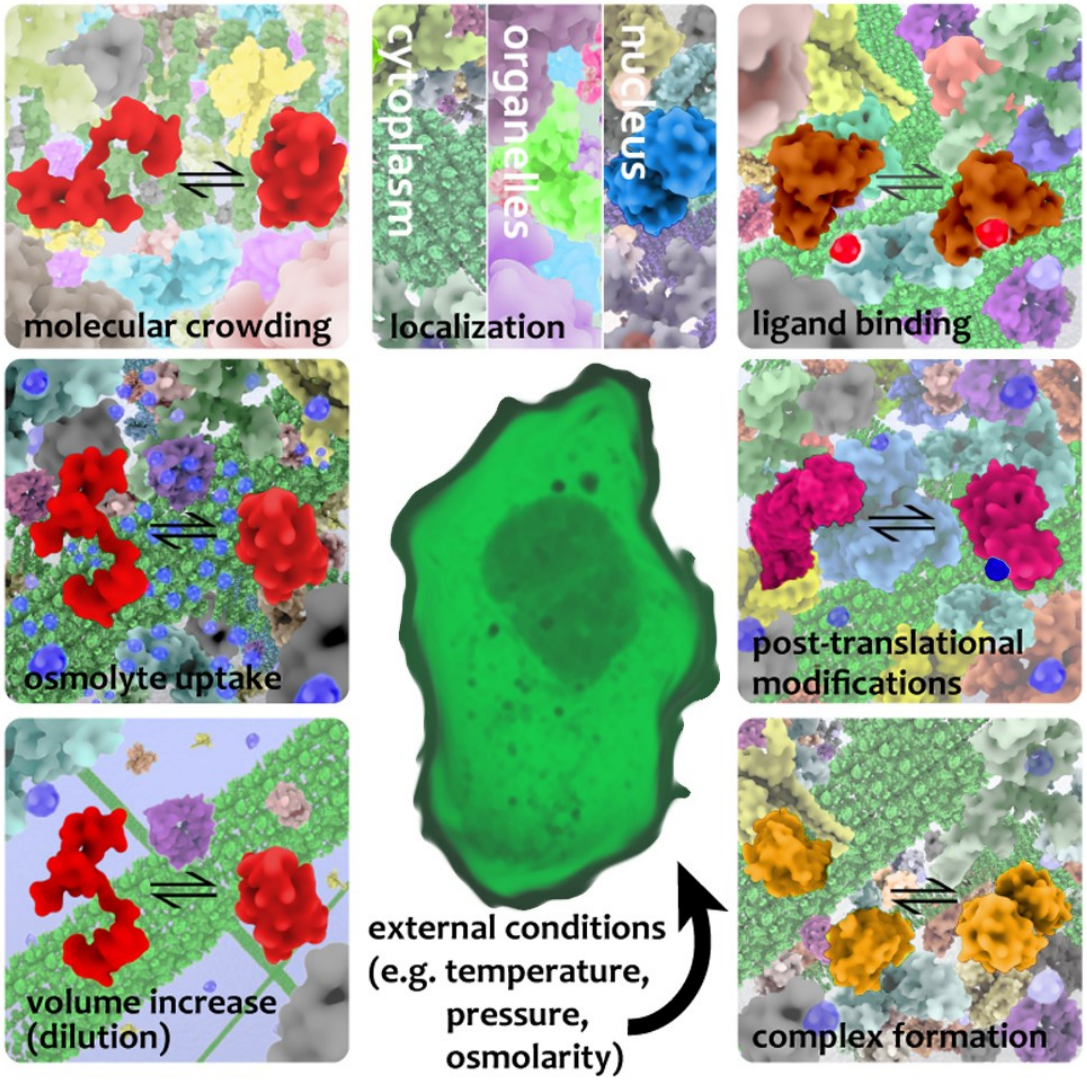

Fig. 1.3: Variations to protein folding in the cell. The different panels illustrate processes and effects that occur within a living cell, and affect protein function as well as folding stability and kinetics. Panels on the left depict cellular processes that affect folding, while panels on the right show protein reactions that occur in the cell, and also affect the protein's folding equilibrium.

As we zoom in to the local environment of a protein, surrounding solution components can vary dramatically due to cytoskeletal and organelle-induced local environments. Taken together, proteins are surrounded by a staggeringly complex cellular environment [109]. In an average mammalian cell, roughly 4 nm of water separate two proteins. This distance also contains other cosolutes, including electrolytes and metabolites. Since a single hydration layer is of the order of the diameter of a water molecule, 4 nm contain ~ 20 such layers – an exceedingly small number compared to many *in vitro* experiments. Water dynamics in those layers is modified by proteins anywhere from 0.2-2 nm from the protein surface, depending on the molecular property being examined [110] [111].

Many processes in the cell occur concomitantly with cellular shape change [112]. These include the obvious cell cycle changes which cause dramatic changes in cell composition and shape [113], but other processes, such as migration, also cause cellular shape change [114]. Such changes trickle down to local solution composition. In terms of kinetics, thermodynamic stability, and protein structure, these changes need not be dramatic to have an effect. As we saw in the previous section, a few kT suffice because many important proteins in the cell (including, IDPs [6] and transcription factors such as p53 [115]) are only marginally stable, and because kinetic barriers for folding are small. Hence, changes to local solvent composition can have a very real effect on cellular function, in both normal and stress conditions.

1.2.1 Cellular effects on protein folding and interactions

How does folding happen in the cell? To a first approximation, as *in vitro*. There are still cooperative folding curves [116], and similar kinetics [117] But to understand how a cell modulates folding, it may be a useful exercise to adopt the "view point" of a protein diffusing in the cytoplasm. Surrounding the protein, are perturbed layers of water, interspersed with abundant dissolved ions, metabolites, sugars, signaling molecules, and short nucleic acids. Potassium, for example, exist in concentrations of ~ 140 μM in the cytoplasm, making these very abundant in the cellular environment. At a distance roughly 10 water layers away, we have larger biomolecules such as other proteins, at a high abundance. In this crowded environment, a protein must remain relatively inert to most solution components. Indeed, bioinformatics studies show a tendency for proteins to use less reactive amino-acids to coat their surface in

the native conformation [118]. Evolution for specific interactions (e.g. signaling) must also evolve against the many non-specific interactions that compete with a protein's interactome.

The stability of the folded state of a protein in a mammalian cell milieu varies widely [119] 120] [121]. At the lower end of this stability range (~8 kJ/mol), over 5% of that protein's population at any given time is unfolded, and subjected to misfolding that occur during folding. This goes on for the lifetime of the protein, estimated to be between half an hour to several days [71], until it is sent to degradation. Cells had to develop complex machinery to monitor initial folding [122] [123], fold unfolded proteins (e.g. hsp70 chaperones [124], and degrade misfolded proteins [125].

The regulatory pathways tied to this machinery, termed collectively the "unfolded protein response" (UPR), are able to detect protein misfolding stress, and act accordingly: slowing down or halting protein synthesis, increasing the specific synthesis of chaperone proteins, uptaking or synthesizing osmolytes. In extreme cases, the UPR can initiate apoptosis, the self-destruction of the cell. Importantly, this machinery is not only initiated at times of duress, but also during protein synthesis, as the nascent chain emerges from the ribosome.[126] [127].

1.2.2 Experimental techniques to monitor protein folding *in-situ*

For decades, protein folding in the cell has been quantified using biochemical methods such as cross-linking and enzymatic digestion, followed by lysis and assaying. This type of methodology is invasive, low in resolution, and cannot observe proteins in their natural environment. Today, new techniques are emerging that enable minimally invasive observation inside cells of protein structure [121], of folding in real-time [128] and with sub-cellular resolution [119], of protein stability [129] and even of single protein molecules [130].

One of the first methods to answer in-cell protein folding questions is the use of live cell NMR [121]. While this technique is technically challenging due to the high concentrations of protein it requires, it has yielded interesting results ranging from gain of structure to decreased stability inside cells [131].

The explosion of fluorescence microscopy techniques has led to the most sensitive probes of protein dynamics in cells. Förster resonance energy transfer (FRET) [132] is utilized extensively today, from single-molecule protein folding experiments [133] [134] [135] to

measurements of protein folding in living cells [117] [129]. Robust and "red" probes that avoid auto fluorescence of the cell have enabled even single-molecule FRET in living cells [130].These studies, together with the development of new fluorescent probes [136] [137], reveal a protein folding environment in the cell that is far from homogeneous [128]. Folding thermodynamics and kinetics are affected not only by spatial localization [119], but also by temporal changes in cell cycle [113]. In addition, protein identity plays a major role in determining whether it is stabilized or destabilized in the cell [138] [120] [10].The picture that emerges from these studies is that of a complex, non-uniform, and dynamic system, where the solvent environment of a protein in the cell can control folding and activity.

1.3 Summary and Outlook

With a firm basis of *in vitro* and computational studies now established, folding science can focus on questions such as the effect of residual structure in unfolded states, and the effect of complex environments, including in the cell, on folding. The marginal stability of most proteins opens up control of folding *in situ* as a new area of study. The increased cross talk between protein science, computation and cell biology will lead to a better understanding of how folding, function and protein evolution are connected.

1.4 References

[1] L.D. Cabrita, C.M. Dobson, J. Christodoulou, Protein folding on the ribosome, Curr. Opin. Struct. Biol. 20 (2010) 33–45.

[2] K.G. Fleming, Energetics of membrane protein folding, Annu. Rev. Biophys. 43 (2014) 233–255.

[3] A. Rath, C.M. Deber, Protein structure in membrane domains, Annu. Rev. Biophys. 41 (2012) 135–155.

[4] M. Gruebele, Downhill protein folding: evolution meets physics, Comptes Rendus Biol. 328 (2005) 701–712.

[5] K. Xia, S. Zhang, B. Bathrick, S. Liu, Y. Garcia, W. Colon, Quantifying the Kinetic Stability of Hyperstable Proteins via Time-Dependent SDS Trapping, Biochemistry. 51 (2012) 100–107. doi:10.1021/bi201362z.

[6] V.N. Uversky, C.J. Oldfield, A.K. Dunker, Intrinsically disordered proteins in human

diseases: Introducing the D(2) concept, Annu. Rev. Biophys. 37 (2008) 215–246. doi:10.1146/annurev.biophys.37.032807.125924.

[7] B.A. Shoemaker, J.J. Portman, P.G. Wolynes, Speeding molecular recognition by using the folding funnel: The fly-casting mechanism, Proc. Natl. Acad. Sci. USA. 97 (2000) 8868–8873.

[8] D. Baker, Metastable states and folding free energy barriers, Nat. Struct. Biol. 5 (1998) 1021–1024. doi:10.1038/4130.

[9] J.D. Bryngelson, J.N. Onuchic, N.D. Socci, P.G. Wolynes, Funnels, Pathways, and the Energy Landscape of Protein-Folding - a Synthesis, Proteins-Structure Funct. Genet. 21 (1995) 167–195.

[10] I. Guzman, H. Gelman, J. Tai, M. Gruebele, The extracellular protein VlsE is destabilized inside cells, J. Mol. Biol. 426 (2014) 11–20.

[11] D. Thirumalai, G.H. Lorimer, Chaperonin-mediated protein folding, Annu. Rev. Biophys. Biomol. Struct. 30 (2001) 245–269. doi:10.1146/annurev.biophys.30.1.245.

[12] Y. Cho, X. Zhang, K.F. Pobre, Y. Liu, D.L. Powers, J.W. Kelly, L.M. Gierasch, E.T. Powers, Individual and collective contributions of chaperoning and degradation to protein homeostasis in E. coli, Cell Rep. 11 (2015) 321–333. doi:10.1016/j.celrep.2015.03.018.

[13] M. Gruebele, THE FAST PROTEIN FOLDING PROBLEM, Annu. Rev. Phys. Chem. 50 (1999) 485–516. doi:doi:10.1146/annurev.physchem.50.1.485.

[14] C.B. Anfinsen, Principles that Govern the Folding of Protein Chains, Science (80-.). 181 (1973) 223–230. doi:10.1126/science.181.4096.223.

[15] M. Guo, Y. Xu, M. Gruebele, Temperature dependence of protein folding kinetics in living cells, Proc. Nat. Acad. Sci. USA. 109 (2012) 17863–17867. doi:10.1073/pnas.1201797109.

[16] W.F. DeGrado, C.M. Summa, V. Pavone, F. Nastri, A. Lombardi, De novo design and structural characterization of proteins and metalloproteins, Annu. Rev. Biochem. 68 (1999) 779–819. doi:10.1146/annurev.biochem.68.1.779.

[17] Z. Li, Y. Yang, J. Zhan, L. Dai, Y. Zhou, Energy functions in de novo protein design: current challenges and future prospects, Annu. Rev. Biophys. 42 (2013).

[18] K.A. Dill, J.L. MacCallum, The Protein-Folding Problem, 50 Years On, Science (80-.). 338 (2012) 1042–1046. doi:10.1126/science.1219021.

[19] J.J. Portman, S. Takada, P.G. Wolynes, Variational Theory for Site Resolved Protein Folding Free Energy Surfaces, Phys. Rev. Lett. 81 (1998) 5237–5240. http://link.aps.org/doi/10.1103/PhysRevLett.81.5237.

[20] M. Karplus, A. Sali, E. Shakhnovich, Kinetics of protein folding, Nature. 373 (1995) 665. http://dx.doi.org/10.1038/373665a0.

[21] B.R. Brooks, C.L. Brooks, A.D. MacKerell, L. Nilsson, R.J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A.R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R.W. Pastor, C.B. Post, J.Z. Pu, M. Schaefer, B. Tidor, R.M. Venable, H.L. Woodcock, X. Wu, W. Yang, D.M. York, M. Karplus, CHARMM: The Biomolecular Simulation Program, J. Comput. Chem. 30 (2009) 1545–1614. doi:10.1002/jcc.21287.

[22] Y. Duan, P.A. Kollman, Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution, Science (80-.). 282 (1998) 740–744. doi:10.1126/science.282.5389.740.

[23] J.E. Shea, C.L. Brooks 3rd, From folding theories to folding proteins: a review and assessment of simulation studies of protein folding and unfolding, Annu. Rev. Phys. Chem. 52 (2001) 499–535. doi:10.1146/annurev.physchem.52.1.499.

[24] K. Lindorff-Larsen, S. Piana, R.O. Dror, D.E. Shaw, How fast-folding proteins fold, Science (80-.). 334 (2011) 517–520. doi:10.1126/science.1208351.

[25] L.A. Curtiss, K. Raghavachari, P.C. Redfern, J.A. Pople, Assessment of Gaussian-2 and density functional theories for the computation of enthalpies of formation, J. Chem. Phys. 106 (1997) 1063–1079.

[26] S.S. Cho, Y. Levy, P.G. Wolynes, P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 586–591. doi:10.1073/pnas.0509768103.

[27] K.A. Beauchamp, R. McGibbon, Y.-S. Lin, V.S. Pande, Simple few-state models reveal hidden complexity in protein folding, Proc. Nat. Acad. Sci. USA. 109 (2012) 17807–17813. doi:10.1073/pnas.1201810109.

[28] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, T.R. Weikl, Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations, Proc. Nat. Acad. Sci. USA. 106 (2009) 19011–19016. doi:10.1073/pnas.0905466106.

[29] H.J. Dyson, P.E. Wright, Insights into the structure and dynamics of unfolded proteins from nuclear magnetic resonance, Adv. Protein Chem. 62 (2002) 311–340.

[30] V.A. Voelz, V.R. Singh, W.J. Wedemeyer, L.J. Lapidus, V.S. Pande, Unfolded state dynamics and structure of protein L characterized by simulation and experiment, J. Am. Chem. Soc. 132 (2010) 4702–4709. doi:10.1021/ja908369h.

[31] E.H. McConkey, Molecular evolution, intracellular organization, and the quinary structure of proteins, Proc. Nat. Acad. Sci. USA. 79 (1982) 3236–3240.

[32] M.T. Woodside, S.M. Block, Reconstructing folding energy landscapes by single-molecule force spectroscopy, Annu. Rev. Biophys. 43 (2014) 19–39.

[33] G. Comellas, C.M. Rienstra, Protein structure determination by magic-angle spinning solid-state NMR, and insights into the formation, structure, and stability of amyloid

fibrils, Annu. Rev. Biophys. 42 (2013) 515–536.

[34] H. Frauenfelder, S.G. Sligar, P.G. Wolynes, The energy landscapes and motions of proteins, Science (80-.). 254 (1991) 1598–1603. doi:10.1126/science.1749933.

[35] K.A. Dill, Theory for the folding and stability of globular proteins, Biochemistry. 24 (1985) 1501–1509. doi:10.1021/bi00327a032.

[36] H.S. Chan, K.A. Dill, Polymer Principles in Protein Structure and Stability, Ann. Rev. Biophys. Biophys. Chem. 20 (1991) 447–490. doi:doi:10.1146/annurev.bb.20.060191.002311.

[37] K.F. Lau, K.A. Dill, A lattice statistical mechanics model of the conformational and sequence spaces of proteins, Macromolecules. 22 (1989) 3986–3997. doi:10.1021/ma00200a030.

[38] D. Chandler, Interfaces and the driving force of hydrophobic assembly, Nature. 437 (2005) 640–647. http://dx.doi.org/10.1038/nature04162.

[39] N.T. Southall, K.A. Dill, A.D.J. Haymet, A View of the Hydrophobic Effect, J. Phys. Chem. B. 106 (2002) 521–533. doi:10.1021/jp015514e.

[40] S. Roy, M.H. Hecht, Cooperative thermal denaturation of proteins designed by binary patterning of polar and nonpolar amino acids, Biochemistry. 39 (2000) 4603–4607.

[41] E.A. Weathers, M.E. Paulaitis, T.B. Woolf, J.H. Hoh, Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein, FEBS Lett. 576 (2004) 348–352. doi:10.1016/j.febslet.2004.09.036.

[42] A. Wagner, Robustness and evolvability in living systems, Princeton University Press, Princeton, 2005.

[43] N. Go, Theoretical studies of protein folding, Annu. Rev. Biophys. Bioeng. 12 (1983) 183–210. doi:10.1146/annurev.bb.12.060183.001151.

[44] C. Levinthal, Are there pathways for protein folding, J. Chim. Phys. Physico-Chim. Biol. 65 (1968) 44.

[45] J.D. Chodera, D.L. Mobley, Entropy-enthalpy compensation: role and ramifications in biomolecular ligand recognition and design, Annu. Rev. Biophys. 42 (2013) 121.

[46] H.A. Scheraga, M. Khalili, A. Liwo, Protein-folding dynamics: overview of molecular simulation techniques, Annu. Rev. Phys. Chem. 58 (2007) 57–83. doi:10.1146/annurev.physchem.58.032806.104614.

[47] H.S. Chung, K. McHale, J.M. Louis, W.A. Eaton, Single-molecule fluorescence experiments determine protein folding transition path times, Science (80-.). 335 (2012) 981–984. doi:10.1126/science.1215768.

[48] W.Y. Yang, M. Gruebele, Folding at the speed limit, Nature. 423 (2003) 193–197.

[49] M.G. Saunders, G.A. Voth, Coarse-graining methods for computational biology,

Annu. Rev. Biophys. 42 (2013) 73–93.

[50] M. Levitt, A. Warshel, Computer simulation of protein folding, Nature. 253 (1975) 694–698. http://dx.doi.org/10.1038/253694a0.

[51] A. Cavalli, U. Haberthür, E. Paci, A. Caflisch, Fast protein folding on downhill energy landscape, Protein Sci. 12 (2003) 1801–1803. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2323966/.

[52] C.D. Snow, H. Nguyen, V.S. Pande, M. Gruebele, Absolute comparison of simulated and experimental protein-folding dynamics, Nature. 420 (2002) 102–106. doi:10.1038/nature01160.

[53] G.R. Bowman, V.A. Voelz, V.S. Pande, Atomistic folding simulations of the five helix bundle protein λ(6–85), J. Am. Chem. Soc. 133 (2011) 664–667. doi:10.1021/ja106936n.

[54] U.H.E. Hansmann, Parallel tempering algorithm for conformational studies of biological molecules, Chem. Phys. Lett. 281 (1997) 140–150.

[55] D.M. Zuckerman, Equilibrium Sampling in Biomolecular Simulation, Annu. Rev. Biophys. 40 (2011) 41–62.

[56] J.W. Pitera, W.C. Swope, F.F. Abraham, Observation of noncooperative folding thermodynamics in simulations of 1BBL, Biophys. J. 94 (2008) 4837–4846.

[57] R.B. Best, G. Hummer, Reaction coordinates and rates from transition paths, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 6732–6737.

[58] R.O. Dror, R.M. Dirks, J.P. Grossman, H. Xu, D.E. Shaw, Biomolecular simulation: a computational microscope for molecular biology, Annu. Rev. Biophys. 41 (2012) 429–452.

[59] M.B. Prigozhin, M. Gruebele, Microsecond folding experiments and simulations: a match is made, Phys. Chem. Chem. Phys. PCCP. 15 (2013) 3372–3388. doi:10.1039/c3cp43992e.

[60] T.E. Creighton, D.P. Goldenberg, Kinetic role of a meta-stable native-like two-disulphide species in the folding transition of bovine pancreatic trypsin inhibitor, J. Mol. Biol. 179 (1984) 497–526.

[61] D. Fass, Disulfide bonding in protein biophysics, Annu. Rev. Biophys. 41 (2012) 63–79.

[62] J.F. Brandts, H.R. Halvorson, M. Brennan, Consideration of the possibility that the slow step in protein denaturation reactions is due to cis-trans isomerism of proline residues, Biochemistry. 14 (1975) 4953–4963.

[63] H.J. Dyson, M. Rance, R.A. Houghten, R.A. Lerner, P.E. Wright, Folding of immunogenic peptide fragments of proteins in water solution: I. Sequence requirements for the formation of a reverse turn, J. Mol. Biol. 201 (1988) 161–200.

[64] E. Freire, K.P. Murphy, J.M. Sanchez-Ruiz, M.L. Galisteo, P.L. Privalov, The molecular basis of cooperativity in protein folding. Thermodynamic dissection of interdomain interactions in phosphoglycerate kinase, Biochemistry. 31 (1992) 250–256.

[65] G.S. Huang, T.G. Oas, Structure and Stability of Monomeric. lambda. Repressor: NMR Evidence for Two-State Folding, Biochemistry. 34 (1995) 3884–3892.

[66] S.E. Jackson, A.R. Fersht, Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition, Biochemistry. 30 (1991) 10428–10435. doi:10.1021/bi00107a010.

[67] S.W. Englander, T.R. Sosnick, L.C. Mayne, M. Shtilerman, P.X. Qi, Y. Bai, Fast and slow folding in cytochrome c, Acc. Chem. Res. 31 (1998) 737–744.

[68] V. Munoz, Folding plasticity, Nat. Struct. Bio. 9 (2002) 792–794. doi:10.1038/nsb1102-792.

[69] J. Liu, L.A. Campos, M. Cerminara, X. Wang, R. Ramanathan, D.S. English, V. Muñoz, Exploring one-state downhill protein folding in single molecules, Proc. Nat. Acad. Sci. USA. 109 (2012) 179–184.

[70] P.G. Wolynes, Z. Luthey-Schulten, J.M. Onuchic, Fast-Folding Experiments and the Topography of Protein Folding Energy Landscapes, Chem. Biol. 3 (1996) 425–432.

[71] H.-C.S. Yen, Q. Xu, D.M. Chou, Z. Zhao, S.J. Elledge, Global protein stability profiling in mammalian cells., Science. 322 (2008) 918–923. doi:10.1126/science.1160489.

[72] M.M. Garcia-Mira, M. Sadqi, N. Fischer, J.M. Sanchez-Ruiz, V. Munoz, Experimental identification of downhill protein folding, Science (80-.). 298 (2002) 2191–2195.

[73] J. Sabelko, J. Ervin, M. Gruebele, Observation of strange kinetics in protein folding, Proc. Nat. Acad. Sci. USA. 96 (1999) 6031–6036.

[74] F. Liu, M. Gruebele, Tuning λ6-85 towards downhill folding at its melting temperature, J. Mol. Biol. 370 (2007) 574–584.

[75] W.Y. Yang, M. Gruebele, Folding λ-repressor at its speed limit, Biophys. J. 87 (2004) 596–608.

[76] A.J. Wirth, Y. Liu, M.B. Prigozhin, K. Schulten, M. Gruebele, Comparing Fast Pressure Jump and Temperature Jump Protein Folding Experiments and Simulations, J. Am. Chem. Soc. (2015).

[77] D. Barrick, What have we learned from the studies of two-state folders, and what are the unanswered questions about two-state protein folding?, Phys. Biol. 6 (2009) 15001. doi:10.1088/1478-3975/6/1/015001.

[78] J. Kubelka, J. Hofrichter, W.A. Eaton, The protein folding “speed limit,” Curr. Opin. Struct. Biol. 14 (2004) 76–88.

[79] R.B. Best, G. Hummer, Coordinate-dependent diffusion in protein folding, Proc. Nat. Acad. Sci. USA. 107 (2010) 1088–1093.

[80] J. Chahine, R.J. Oliveira, V.B.P. Leite, J. Wang, Configuration-dependent diffusion can shift the kinetic transition state and barrier height of protein folding, Proc. Nat. Acad. Sci. USA. 104 (2007) 14646–14651.

[81] C. Ramakrishnan, G.N. Ramachandran, Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units, Biophys. J. 5 (1965) 909–933.

[82] A. Borgia, B.G. Wensley, A. Soranno, D. Nettels, M.B. Borgia, A. Hoffmann, S.H. Pfeil, E.A. Lipman, J. Clarke, B. Schuler, Localizing internal friction along the reaction coordinate of protein folding by combining ensemble and single-molecule fluorescence spectroscopy, Nat Commun. 3 (2012). doi:10.1038/ncomms2204.

[83] B. Fierz, T. Kiefhaber, End-to-end vs interior loop formation kinetics in unfolded polypeptide chains, J. Am. Chem. Soc. 129 (2007) 672–679. doi:10.1021/ja0666396.

[84] S.J. Hagen, Solvent Viscosity and Friction in Protein Folding Dynamics, Curr. Protein Pept. Sci. 11 (2010) 385–395.

[85] A.C.C. Chang, S.S.C. Chuang, M. Gray, Y. Soong, In-situ infrared study of CO2 adsorption on SBA-15 grafted with γ-(aminopropyl) triethoxysilane, Energy & Fuels. 17 (2003) 468–473.

[86] C.M. Davis, S. Xiao, D.P. Raleigh, R.B. Dyer, Raising the speed limit for β-hairpin formation, J. Am. Chem. Soc. 134 (2012) 14476–14482.

[87] D. De Sancho, R.B. Best, What is the time scale for α-helix nucleation?, J. Am. Chem. Soc. 133 (2011) 6809–6816.

[88] D. Poland, H.A. Scheraga, Phase transitions in one dimension and the helix—coil transition in polyamino acids, J. Chem. Phys. 45 (1966) 1456–1463.

[89] S. Piana, K. Sarkar, K. Lindorff-Larsen, M. Guo, M. Gruebele, D.E. Shaw, Computational design and experimental testing of the fastest-folding beta-sheet protein, J. Mol. Biol. 405 (2011) 43–48.

[90] Y. Zhu, X. Fu, T. Wang, A. Tamura, S. Takada, J.G. Saven, F. Gai, Guiding the search for a protein's maximum rate of folding, Chem. Phys. 307 (2004) 99–109.

[91] A.K. Dunker, J.D. Lawson, C.J. Brown, R.M. Williams, P. Romero, J.S. Oh, C.J. Oldfield, A.M. Campen, C.M. Ratliff, K.W. Hipps, Intrinsically disordered protein, J. Mol. Graph. Model. 19 (2001) 26–59.

[92] K. Sugase, H.J. Dyson, P.E. Wright, Mechanism of coupled folding and binding of an intrinsically disordered protein, Nature. 447 (2007) 1021–1025.

[93] M. Jäger, J. Zhang, J. Bieschke, H. Nguyen, G. Dendle, M. Bowman, J. Noel, M. Gruebele, J. Kelly, The structure-function-folding relationship in a WW domain, Proc.

Nat. Acad. Sci. USA. 108 (2006) 10648–10653.

[94] S. Gosavi, P.C. Whitford, P.A. Jennings, J.N. Onuchic, Extracting function from a β-trefoil folding motif, Proc. Nat. Acad. Sci. USA. 105 (2008) 10384–10389.

[95] R.L. Baldwin, The nature of protein folding pathways: the classical versus the new view, J. Biomol. NMR. 5 (1995) 103–109.

[96] R.B. Dyer, Ultrafast and downhill protein folding, Curr. Opin. Struct. Biol. 17 (2007) 38–47.

[97] S.W. Englander, Protein folding intermediates and pathways studied by hydrogen exchange, Annu. Rev. Biophys. Biomol. Struct. 29 (2000) 213–238.

[98] J.B. Udgaonkar, Multiple routes and structural heterogeneity in protein folding, Annu. Rev. Biophys. 37 (2008) 489–510.

[99] S.J. Kim, Y. Matsumura, C. Dumont, H. Kihara, M. Gruebele, Slowing down downhill folding: a three-probe study, Biophys. J. 97 (2009) 295–302.

[100] F. Liu, D. Du, A.A. Fuller, J.E. Davoren, P. Wipf, J.W. Kelly, M. Gruebele, An experimental survey of the transition between two-state and downhill protein folding scenarios, Proc. Nat. Acad. Sci. USA. 105 (2008) 2369–2374.

[101] F. Liu, Y.G. Gao, M. Gruebele, A Survey of λ Repressor Fragments from Two-State to Downhill Folding, J. Mol. Biol. 397 (2010) 789–798.

[102] H. Nguyen, M. Jäger, A. Moretto, M. Gruebele, J.W. Kelly, Tuning the free-energy landscape of a WW domain by temperature, mutation, and truncation, Proc. Nat. Acad. Sci. USA. 100 (2003) 3948–3953.

[103] T. Aksel, D. Barrick, Direct observation of parallel folding pathways revealed using a symmetric repeat protein system, Biophys. J. 107 (2014) 220–232.

[104] D. Baker, A surprising simplicity to protein folding, Nature. 405 (2000) 39–42. doi:10.1038/35011000.

[105] M. Schroder, R.J. Kaufman, The mammalian unfolded protein response, in: Annu. Rev. Biochem., Annual Reviews, Palo Alto, 2005: pp. 739–789. doi:10.1146/annurev.biochem.73.011303.074134.

[106] P.H. Yancey, Water Stress, Osmolytes and Proteins, Integr. Comp. Bio. 41 (2001) 699–709. doi:10.1093/icb/41.4.699.

[107] S.B. Zimmerman, S.O. Trach, Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of Escherichia coli, J. Mol. Biol. 222 (1991) 599–620. doi:10.1016/0022-2836(91)90499-V.

[108] S. Cayley, B.A. Lewis, H.J. Guttman, M.T. Record, Characterization of the cytoplasm of Escherichia coli K-12 as a function of external osmolarity. Implications for protein-DNA interactions in vivo, J. Mol. Biol. 222 (1991) 281–300. doi:10.1016/0022-

2836(91)90212-O.

[109] D.R. Nelson, Biophysical dynamics in disorderly environments, Annu. Rev. Biophys. 41 (2012) 371–402.

[110] S. Ebbinghaus, S.J. Kim, M. Heyden, X. Yu, U. Heugen, M. Gruebele, D.M. Leitner, M. Havenith, An extended dynamical hydration shell around proteins, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 20749–20752. doi:10.1073/pnas.0709207104.

[111] B. Halle, Protein hydration dynamics in solution: a critical survey, Philos. Trans. R. Soc. London Ser. B-Biological Sci. 359 (2004) 1207–1223. doi:10.1098/rstb.2004.1499.

[112] S.F. Pedersen, E.K. Hoffmann, J.W. Mills, The cytoskeleton and cell volume regulation, Comp. Biochem. Physiol.A Mol. Integr. Physiol. 130 (2001) 385–399. doi:10.1016/S1095-6433(01)00429-9.

[113] A.J. Wirth, M. Platkov, M. Gruebele, Temporal Variation of a Protein Folding Energy Landscape in the Cell, J. Am. Chem. Soc. 135 (2013) 19215–19221. doi:10.1021/ja4087165.

[114] E.K. Hoffmann, I.H. Lambert, S.F. Pedersen, Physiology of cell volume regulation in vertebrates., Physiol. Rev. 89 (2009) 193–277. doi:10.1152/physrev.00037.2007.

[115] M.T. Record Jr, C.F. Anderson, T.M. Lohman, Thermodynamic analysis of ion effects on the binding and conformational equilibria of proteins and nucleic acids: the roles of ion association or release, screening, and ion effects on water activity, Q. Rev. Biophys. 11 (1978) 103–178.

[116] Z. Ignatova, L.M. Gierasch, Monitoring protein stability and aggregation in vivo by real-time fluorescent labeling, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 523–528. doi:10.1073/pnas.0304533101.

[117] S. Ebbinghaus, A. Dhar, D. McDonald, M. Gruebele, Protein folding stability and dynamics imaged in a living cell, Nat. Methods. 7 (2010) 319–323. doi:10.1038/nmeth.1435.

[118] E.D. Levy, S. De, S. a Teichmann, Cellular crowding imposes global constraints on the chemistry and evolution of proteomes., Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 20461–20466. doi:10.1073/pnas.1209312109.

[119] A. Dhar, K. Girdhar, D. Singh, H. Gelman, S. Ebbinghaus, M. Gruebele, Protein stability and folding kinetics in the nucleus and endoplasmic reticulum of eucaryotic cells, Biophys. J. 101 (2011) 421–430.

[120] S. Ghaemmaghami, T.G. Oas, Quantitative protein stability measurement in vivo, Nat. Struct. Mol. Bio. 8 (2001) 879–882.

[121] A.E. Smith, Z. Zhang, G.J. Pielak, C. Li, NMR studies of protein folding and binding in cells and cell-like environments, Curr. Opin. Struct. Biol. 30 (2015) 7–16.

doi:10.1016/j.sbi.2014.10.004.

[122] P.L. Clark, Protein folding in the cell: reshaping the folding funnel, Trends Biochem. Sci. 29 (2004) 527–534.

[123] D. V Fedyukina, S. Cavagnero, Protein folding at the exit tunnel, Annu. Rev. Biophys. 40 (2011) 337–359.

[124] F.U. Hartl, M. Hayer-Hartl, Molecular chaperones in the cytosol: from nascent chain to folded protein, Science (80-.). 295 (2002) 1852–1858. doi:10.1126/science.1068408.

[125] M.H. Glickman, A. Ciechanover, The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction, Physiol. Rev. 82 (2002) 373–428. doi:10.1152/physrev.00027.2001.

[126] M.S. Evans, T.F. Clark, P.L. Clark, Conformations of co-translational folding intermediates, Protein Pept. Lett. 12 (2005) 189–195.

[127] E.T. Powers, R.I. Morimoto, A. Dillin, J.W. Kelly, W.E. Balch, Biological and chemical approaches to diseases of proteostasis deficiency., Annu. Rev. Biochem. 78 (2009) 959–991. doi:10.1146/annurev.biochem.052308.114844.

[128] S. Ebbinghaus, M. Gruebele, Protein folding landscapes in the living cell, J. Phys. Chem. Lett. 2 (2011) 314–319.

[129] Z. Ignatova, L.M. Gierasch, Inhibition of protein aggregation in vitro and in vivo by a natural osmoprotectant, Proc. Nat. Acad. Sci. USA. 103 (2006) 13357–13361.

[130] I. König, A. Zarrine-Afsar, M. Aznauryan, A. Soranno, B. Wunderlich, F. Dingfelder, J.C. Stüber, A. Plückthun, D. Nettels, B. Schuler, Single-molecule spectroscopy of protein conformational dynamics in live eukaryotic cells, Nat. Methods. 12 (2015) 773–779.

[131] D.I. Freedberg, P. Selenko, Live cell NMR, Annu. Rev. Biophys. 43 (2014) 171–192.

[132] H. Edelhoch, L. Brand, M. Wilchek, Fluorescence Studies with Tryptophyl Peptides*, Biochemistry. 6 (1967) 547–559.

[133] P.R. Banerjee, A. a Deniz, Shedding light on protein folding landscapes by single-molecule fluorescence., Chem. Soc. Rev. 43 (2014) 1172–1188. doi:10.1039/c3cs60311c.

[134] A. Hoffmann, D. Nettels, J. Clark, A. Borgia, S.E. Radford, J. Clarke, B. Schuler, Quantifying heterogeneity and conformational dynamics from single molecule FRET of diffusing molecules: recurrence analysis of single particles (RASP)., Phys. Chem. Chem. Phys. PCCP. 13 (2011) 1857–1871. doi:10.1039/c0cp01911a.

[135] B. Schuler, H. Hofmann, Single-molecule spectroscopy of protein folding dynamics-expanding scope and timescales, Curr. Opin. Struct. Biol. 23 (2013) 36–47. doi:10.1016/j.sbi.2012.10.008.

[136] A.J. Boersma, I.S. Zuhorn, B. Poolman, A sensor for quantification of macromolecular crowding in living cells, Nat Meth. 12 (2015) 227–229. doi:10.1038/nmeth.3257http://www.nature.com/nmeth/journal/v12/n3/abs/nmeth.3257.html#supplementary-information.

[137] D. Gnutt, M. Gao, O. Brylski, M. Heyden, S. Ebbinghaus, Excluded-Volume Effects in Living Cells, Angew. Chemie Int. Ed. (2014) 2548–2551. doi:10.1002/anie.201409847.

[138] L.A. Benton, A.E. Smith, G.B. Young, G.J. Pielak, Unexpected effects of macromolecular crowding on protein stability, Biochemistry. 51 (2012) 9773–9775.

CHAPTER 2

High-resolution mapping of the folding transition state of a WW domain

WW domains are β sheet modular protein domains of 30-65 residues in length that modulate specific interactions with proline-rich protein ligands. WW domains have proven to be an excellent model for ultrafast folding experiments, for mechanistic experimental studies on the folding of a simple β sheet structure, and for benchmarking computational folding scenarios [1-3]. The best characterized natural WW domains to date are the hPin1 WW domain from human peptidyl-prolyl *cis-trans* isomerase Pin1 [3], and the FBP28 WW domain from formin-binding protein 28 [4], with limited data available for a third WW domain, the hYAP65 WW domain from human Yes-Kinase associated kinase [5]. Mutational Φ_M value analysis suggest that formation of loop 1 in WW domains is mostly rate limiting (Φ_M values > 0.80) [6].

In FBP28 WW and hYap65 WW, the N-terminal loop 1 sequence folds into a 5-residue type-I G-bulge turn, the statistically preferred conformation among WW domains. The longer, intrinsically disordered 6-residue loop 1 in hPin1 WW appears to have been selected for function. Its unusual loop conformation (type II-turn intercalated in a 6-residue loop) may position the side chains of residues S16 and R17 for optimal ligand binding [7]. Replacing the hPin1 loop 1 with the turn of FBP28 WW to make the FiP WW domain increases stability by up to 7 kJ/mole and speeds up folding from ~ 80 μs to ~ 13 μs, but compromises function [7]. A similar frustration of folding by function has also been observed in other cases, such as frataxin [8]. For WW domains with their loop 1 substructure optimized for folding thermodynamics and kinetics, formation of loop 2 becomes competitive as the rate-limiting step for folding. Indeed, further optimization of the loop 2 sequence in FiP (FiP N30G/A31T/Q33T, FiP-GTT hereafter) produced a WW domain with a folding relaxation time of ~ 4 μs, approaching the speed limit for folding [9].

This chapter is adapted from K Dave, M Jäger, H Nguyen, JW Kelly, M Gruebele. High-Resolution Mapping of the Folding Transition State of a WW Domain. Journal of molecular biology, 2016

Here we report an in-depth study of temperature jump kinetics for 78 mutants of the hPin1 WW domain (Table 2.1) that also includes data from two more limited, previous Φ value analyses [6, 7, 10, 11]. 45 mutants were amenable for ΦM value analysis, providing energetic constraints for structural mapping of the folding transition state of hPin1 WW. Multiple side chain substitutions at some key sequence positions (e.g. within the hydrophobic cores or loop 2) allow us to calculate error-weighted average ΦM values that are more likely to be a robust representation of transition state vs. native state free energy changes than single (e.g. Ala) substitutions. We also identify substitutions that are not suitable for ΦM value analysis, and discuss the reasons. This approach has been used by Davidson and co-workers to investigate 'conservatism' of substitutions at several sites of the SH3 domain [12]. Although wild type hPin1 WW and its variants fold more slowly than the redesigned loop 1 variant FiP, their folding rates are still in the microsecond range that is now within the reach of fast folding simulations. As computation of folding in the 50-500 μs range becomes feasible, we believe that the data presented in this study will prove to be a rich resource for detailed comparisons, providing constraints on mechanisms and rate changes deduced from molecular dynamics simulations, which are still debated in the literature [9, 13-15].

2.1 Methods

2.1.1 Nomenclature

Residues of the hPin1 WW domain are abbreviated by a single capital letter, followed by the number of the residue in the sequence (e.g. W11). Amino acids are also abbreviated using the standard three letter code (e.g. Trp for tryptophan). Classical side chain mutants are indicated by single letter code (e.g. W11F), with the first and second letters representing the wild type and replacing residue, respectively, and the number indicates the sequence position. Non-classical backbone hydrogen bond mutations are also designated by single letter code. The first letter represents the mutated residue, and the same letter in small capitals is used for the

replacing residue (e.g. S16s) to distinguish a non-classical amide-to-ester mutation from their classical counterparts.

2.1.2 Protein expression and sample preparation

The wild type hPin1 WW domain and mutants thereof with classical side chain mutations were prepared recombinantly, as described in detail in another publication [10]. hPin1 WW variants with amide-to-ester mutations were synthesized chemically, as described in detail in [16]. Protein identity and purity was ascertained by electrospray mass spectrometry, SDS-PAGE, and reversed-phase HPLC chromatography.

2.1.3 Experimental procedures

Equilibrium unfolding of hPin1 WW was monitored by far-UV spectroscopy at 229 nm as described in detail in [10]. Unfolding transitions were analyzed by using a two-state model, where the folding free energy ΔGf is expressed by a quadratic Taylor series approximation: $\Delta G_f(T)=\Delta G_f^{(1)}(T_m)\cdot(T-T_m)+\Delta G_f^{(2)}(T_m)\cdot(T-T_m)^2$. The two coefficients $\Delta G_f^{(i)}(T_m)$, i=1…2, represent the temperature-dependent free energy of folding, and Tm is the nominal midpoint of thermal denaturation ($\Delta G_f(T_m) = 0$). The inclusion of the quadratic term was necessary to fit the data of most mutants within experimental uncertainty. For selected mutants, the transition was also analyzed by expressing $\Delta G_f(T)$ in terms of a constant heat capacity formula. As shown previously for the hYap65 WW domain, both procedures yield nearly identical results [31].

Laser temperature jumps around the protein's melting temperature were measured for each mutant as described in detail elsewhere [44, 45]. Briefly, a 10 ns Nd:YAG pulse Raman-shifted in H2 heated the sample solution by ~ 5-10 °C, inducing kinetic relaxation of the WW domain to the new thermal equilibrium. 285 nm UV pulses, spaced 1 ns apart from a frequency-tripled, mode-locked titanium:sapphire laser, excited tryptophan fluorescence in the hPin1 WW domain. Fluorescence emission was digitized in 0.5 ns time steps by a miniature photomultiplier tube with a 0.9 ns full-width-half-maximum response time. The sequence of fluorescence decays *f(t)* was fitted within measurement uncertainty by the linear combination $a_1f_1(t)+a_2f_2(t)$ of decays just before and 0.5 ms after the T-jump. The normalized fraction $f(t)=a_1/(a_1+a_2)$ from t≈2 μs to t=0.5 ms was fitted to a single exponential decay $\exp[-k_{obs}t]$

where $k_{obs}=k_f+k_u$. Thus the signal extraction and data analysis are consistently two-state. The observed relaxation rate coefficient was combined with the equilibrium constant K_{eq} to compute the forward reaction rate coefficient $k_f=k_{obs}K_{eq}/(1+K_{eq})$. k_f was measured for several temperatures (typically around 10) below and above T_m, and $\Delta G_f^{\dagger}(T)$ was determined as a function of temperature using the relationship $k_f=A^{\dagger}\cdot\exp(-\Delta G_f^{\dagger}(T)/RT)$ with the quadratic Taylor approximation $\Delta G^{\dagger}_f(T)=\Delta G_f^{\dagger(0)}(T_m)+\Delta G_f^{\dagger(1)}(T_m)(T-T_m)+\Delta G_f^{\dagger(2)}(T_m)(T-T_m)^2$, as well as expansions about the temperature of maximal stability (T_0), or the Gibbs-Helmholtz formula (see SI). The three coefficients $\Delta G_f^{\dagger(i)}$, i=0…2, represent the temperature-dependent activation barrier. The frequency of activation A† was fixed at 500 ns-1, near the lower end of estimates of the folding speed limit [1], and the two coefficients $\Delta G_f^{\dagger(1)}(T_m)$ and $\Delta G_f^{\dagger(2)}(T_m)$ also incorporate some effects of temperature-dependent solvent friction. Because previous ΦM analyses utilized a faster ad hoc frequency of 50 ns-1, the ΦM values of published mutants are shifted by a small constant from the recalculated values of these mutants in this study. Least squares fitting was carried out using IGOR Pro (Wavemetrics). Protein visualization was rendered using Pymol and Weblab viewer software packages (Accelerys, San Diego) [46].

2.2 Results and Discussion

After a brief review of hPin1 WW structure and native state interactions (Fig. 2.1, section 1), we begin our discussion of the results in section 2 with the mutational phi-value (ΦM) analysis, focusing on which mutants are likely to be reliable reporters for transition state structure (Fig. 2.2). Next, a temperature-dependent phi-value (ΦT) analysis is used in section 3 to identify mutations that perturb the folding mechanism and whose perturbing effect escapes detection by inspection of the mutational ΦM values only (Fig. 2.3). The consensus set of 39 non-perturbing mutants with reliable ΦM values is employed in section 4 to analyze the transition state structure of hPin1 WW (Figs 2.4-2.7). Section 5 looks at various loop 1 insertion and deletion variants within the rate-limiting loop 1 substructure (Fig. 2.8). A hypothetical "hybrid" Φ_M map for the ultrafast folding hPin1 WW variant FiP (Fig. 2.9) to benchmark recent molecular dynamics simulations concludes the paper.

2.2.1 Overview of hPin1 WW structure and native state interactions

Two types of interactions help stabilize and specify the three-stranded β sheet structure of the hPin1 WW domain. The first type is mediated by the side chains of conserved hydrophobic residues that form two segregated hydrophobic clusters, one on each side of the β sheet (Fig. 2.1a). The second type of interaction involves a network of 10 backbone-backbone and 4 backbone-side chain hydrogen bonds (Fig. 2.1b). Hydrophobic cluster 1 is formed by the side chains of residues L7, P8, W11, Y24 and P37. The N-terminal Trp (W11 in hPin1 WW) and the C-terminal Pro (P37 in hPin1 WW) are absolutely conserved in WW domains. Mutation of residues W11, Y24 and P37 to Ala or Leu in hPin1 WW results in partially unfolded, or fully unfolded protein, even at low temperature (4° C) (Fig. 2.1c and [10]). As hydrophobic cluster 1 does not contribute to ligand binding, these medium-long range side chain interactions appear to have evolved to maximize thermodynamic stability of hPin1 WW, rather than its biological function. Hydrophobic core 2 lies on the ligand-binding face of the three-stranded β sheet, and is formed by the side chains of residues R14, Y23 and F25 (Fig. 2.1a). These residues are only moderately conserved in WW domains, presumably because hydrophobic core 2 contributes to ligand binding. Ala mutations of residues 14, 23 and 25 in hPin1 WW, although severely destabilizing the native state ($\Delta\Delta Gf \sim 9$ kJ/mole) (Fig. 2.1c), allow folding into the native state structure under the most favorable folding conditions (4 °C). Using amide-to-ester mutagenesis, we showed that the degree of destabilization of the native state upon eliminating a backbone hydrogen bond is strongly context-dependent [16]. Hydrogen bonds near the two loop substructures are less influential than hydrogen bonds that are protected within a hydrophobic core. The side chain amino group of N26 (β strand 2) forms a hydrogen bond with the backbone carbonyl group of P9 and to the indole ring of W11, thus linking β strands 1 and 2 of the three-stranded β sheet. Like the hydrophobic core 1 residues (W11, Y24 and P37 in hPin1 WW), the Asn in strand 2 (N26 in hPin1 WW) is highly conserved among WW domains and N26A or N26L mutations unfold hPin1 WW (Fig. 2.1c) [10].

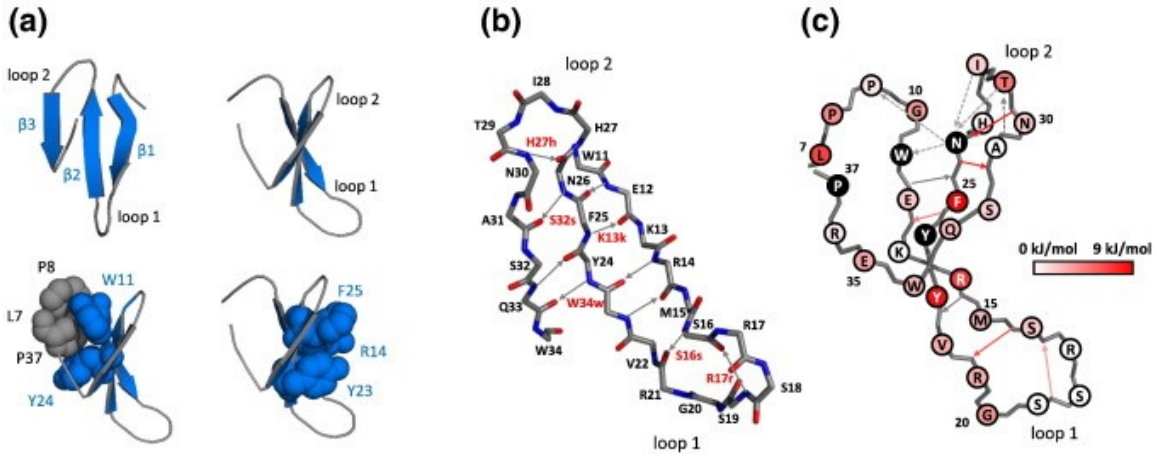

Fig. 2.1: Overview of hPin1 WW structure and native-state interactions. (a) Structural cartoon of the hPin1 WW fold, highlighting the two hydrophobic clusters (cores) that protrude from either side of the three-stranded β-sheet. The individual β-sheet are color coded blue, while the loop segments and the N- and C-terminal extensions are shown in gray. Side-chain contacts that constitute the hydrophobic clusters are shown as van der Waals surfaces. (b) Backbone representation of the three-stranded β-sheet region (residues W11–W34), highlighting the 10 backbone hydrogen bonds that connect the three β-strands and stabilize the three-stranded β-sheet topology. Hydrogen bonds that were perturbed by amine-to-ester mutations for ΦM analysis are labeled in red. Residues are labeled in single letter code and are numbered. (c) Quantitative analysis of a complete Ala scan, replacing each of the 33 non-Alanine residues individually with Ala. Destabilizations calculated at 55 °C range from near zero to ~ 9 kJ/mol and are mapped onto the backbone structure of the folded protein. Four Ala mutants (labeled black) were either completely or significantly unfolded, even at low temperature (4 °C). For these four mutants, ΔΔG must exceed 9 kJ/mol, but no accurate thermodynamic data can be derived in aqueous buffer without invoking stabilizing co-solvents.

2.3 ΦM-value analysis

The mutational ΦM value = ΔΔGf†/ ΔΔGf quantifies changes in the free energy of activation (ΔΔGf†) relative to the ground state free energy of folding (ΔΔGf) between wild type and mutant proteins [17, 18] Computational modeling of ΦM values is now possible for WW domains [14, 19], making direct comparisons with experiments achievable. To obtain accurate ΦM values that truly represent transition state energetics, one must design non-

disruptive mutants that differ sufficiently in thermodynamic stability from the wild type reference protein [20-23], but are not so different that the folding landscape is substantially altered. A generally accepted strategy for ΦM value analysis is to use conservative hydrophobic deletion mutations (e.g. Ile/Leu → Val → Ala ; Thr → Ser; Phe → Leu → Ala). This strategy avoids mutants that increase side chain size or introduce new functional groups (i.e. Ser → Thr, Phe → Trp), as well as mutation of solvent-exposed charged residues with long-range electrostatic interactions and/or protein-solvent interactions (e.g. Glu → Ala, Tyr → Phe). Several of the mutations that we employed in our previous side chain ΦM analysis of hPin1 WW [6] do not meet these requirements. This has been discussed in detail in the literature [22].One in four mutants studied here has a thermodynamic stability very close to wild type hPin1 WW (ΔΔGf < 1 kJ/mole, ΔTm < 2.5 °C, with a typical error in Tm of 0.5 – 1 °C). These mutants were excluded from the ΦM analysis discussed herein. Their thermodynamic and kinetic data (Table 2.1) should nonetheless provide a valuable resource for benchmarking upcoming molecular dynamics simulations because most of these mutants fold on the microsecond to millisecond time scale, accessible to all atom explicit [24], implicit [14] and coarse grained simulations [25]. We calculated ΦM values at three representative temperatures (50 °C, 55 °C and 60 °C) (Table 2.1), where experimental data was available for almost all mutants without the need for error-prone extrapolation. For some of the more stable loop 1 deletion variants, we only report ΦM values at 55 and/or 60 °C.

2.3.1 Outliers in the analysis

At 55 °C, the ΦM values of the mutants that potentially qualify for ΦM analysis (ΔΔGf < 1 kJ/mole and ΔTm < 2.5 °C) range from -0.20 (L7I) to 2.56 (S16A) (Fig. 2.2a, Fig. 2.2b, Table 2.1). With the exception of some loop 1 mutants that only slightly destabilize the domain, there is no correlation between the magnitude of a ΦM value and the extent of destabilization (ΔΔGf in Fig. 2.2a and Fig. 2.2b). Except for mutants E12Q, I28A, and Y23F, the estimated error in Φm was less than 10 %. A surprisingly high fraction of mutants yield ΦM values that lie outside the classical range of ΦM values (in particular ΦM > 1). Almost all mutants with non-classical ΦM values map to the hydrophobic core 1 and loop 1 substructures in native hPin1 WW, pointing to the importance of these substructures for transition state energetics. Mutant L7I yields the only negative ΦM value, which is, however, not supported by the L7A and L7V mutations (Fig. 2.2a). Also the large ΦM value of V22A (β strand 2) can neither be cross-

validated by ΦM values of immediate sequence neighbors (R21A/H, Y23L/A) nor by its cross-strand neighbor (M15A, β strand 1). Finally, the ΦM value of Y23F is almost twice as high as the ΦM values of Y23L and Y23A that target the same residue (Fig. 2.2a). Y23F deletes a solvent-exposed hydroxyl-group that should not affect the side chain packing of hydrophobic core 1. Its unusual ΦM value most likely reports on changes in solvation, rather than packing of the core. Mutants L7I, V22A and Y23F were thus excluded from further analysis.

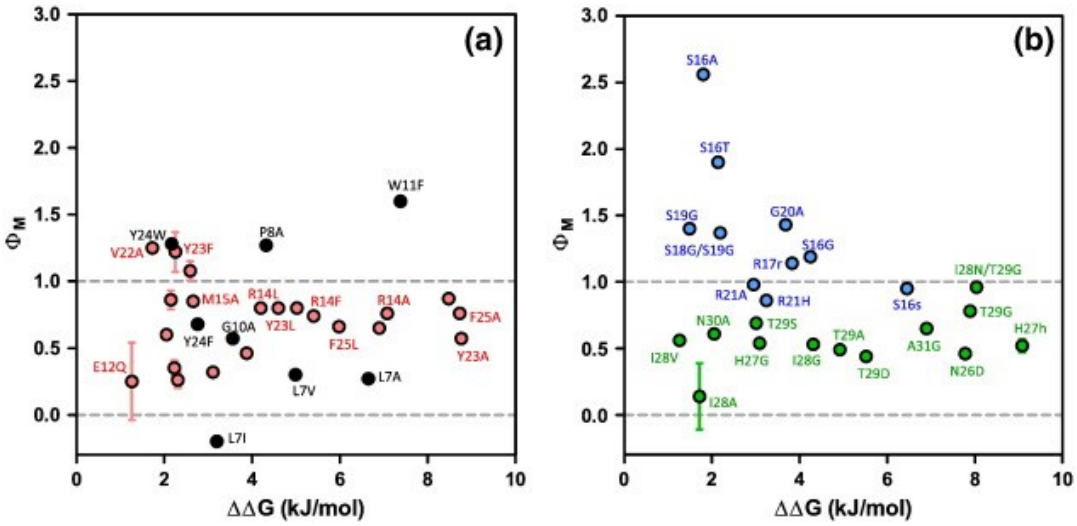

Fig. 2.2: ΦM-Value analysis at 55 °C. (a) Plot of the ΦM-value versus the difference in free energy between wild type and mutant (ΔΔG, in kilojoules per mole) for β strand (filled red circles) and hydrophobic cluster 1 mutants (filled black circles). (b) Plot of the ΦM-value versus the difference in free energy between wild type and mutant (ΔΔG, in kilojoules per mole) of loop 1 (filled blue circles) and loop 2 mutants (filled green circles). Errors in Φ_M that exceed the symbol size are shown explicitly. For clarity, individual ΦM-values are labeled with single letter code.

Table 2.1:

Table 2.1: Thermodynamic and kinetic data for wild type hPin1 WW and mutants thereof

Variant	T_m (°C)	$\Delta G_f^{(1)}$	$\Delta G_f^{(2)}$	$\Delta G^{\ddagger(0)}$	$\Delta G^{\ddagger(1)}$	$\Delta G^{\ddagger(2)}$	Φ_M (50 °C)[1]	Φ_M (55 °C)[1]	Φ_M (60 °C)[1]	Ref.
1. Wildtype and single-site mutants										
wt hPin1	58.6	0.403	0.00272	14.92	0.206	0.00472	-	-	-	[10]
K6A	59.4	0.400	0.00153	11.16	0.166	0.00173	-	-	-	[10]
K6M	58.1	0.414	0.00180	11.76	0.215	0.00162	-	-	-	N [3]
L7A	37.8	0.301	0.00022	13.16	0.136	0.00192	0.23 (0.02)	0.27 (0.02)	0.31 (0.03)	[6, 10]
L7I	49.3	0.318	0.00050	12.66	0.157	0.00141	-0.21 (0.04)	-0.20 (0.04)	-0.26 (0.04)	[10]
L7V	44.0	0.321	0.00041	13.56	0.176	0.00218	0.23 (0.02)	0.30 (0.02)	0.37 (0.02)	[10]
P8A	47.4	0.361	0.00293	18.56	0.139	0.00237	1.29 (0.01)	1.27 (0.01)	1.23 (0.01)	[10]
P9A	56.0	0.397	0.00229	19.10	0.214	0.00272	-	-	-	[10]
G10A	49.0	0.348	0.00151	15.23	0.153	0.00341	0.52 (0.02)	0.57 (0.02)	0.61 (0.02)	[10]
W11F	3.05	0.308	-0.00050	21.62	0.134	0.00399	1.42 (0.01)	1.58 (0.01)	1.79(0.01)	[10]
E12A	52.6	0.373	0.00104	14.33	0.201	0.00396	0.15 (0.12)	0.26 (0.06)	0.36 (0.05)	[10]
E12Q	55.4	0.385	0.00308	14.62	0.179	0.00421	0.22 (0.35)	0.25 (0.30)	0.25 (0.29)	[6, 10]
K13A	59.6	0.385	0.00285	16.11	0.187	0.00139	-	-	-	[10]
K13V	62.8	0.401	0.00322	15.85	0.215	0.00213	-	-	-	N
K13Y	51.7	0.349	0.00237	16.63	0.125	0.00120	1.09 (0.07)	1.09 (0.07)	1.01 (0.08)	N
R14A	39.2	0.347	0.00074	17.21	0.081	0.00464	0.72 (0.01)	0.76 (0.01)	0.82 (0.01)	[10]
R14F	45.2	0.388	0.00195	16.87	0.087	0.00517	0.76 (0.01)	0.74 (0.01)	0.73 (0.01)	[6]
R14L	47.8	0.367	0.00234	16.31	0.145	0.00482	0.77 (0.01)	0.80 (0.01)	0.84 (0.01)	N
M15A	51.8	0.380	0.00289	15.88	0.168	0.00434	0.81 (0.02)	0.84 (0.02)	0.85 (0.02)	[6, 10]
S16A	54.0	0.380	0.00313	18.63	0.205	0.00372	2.44 (0.03)	2.56 (0.02)	2.62 (0.02)	[10]
S16G	47.6	0.369	0.00194	17.75	0.174	0.00452	1.13 (0.01)	1.19 (0.01)	1.25 (0.01)	[10]
S16T	53.2	0.398	0.00325	18.01	0.161	0.00401	1.99 (0.02)	1.90 (0.02)	1.78 (0.01)	[6]
R17A	58.8	0.391	0.00232	19.23	0.221	0.00276	-	-	-	[10]

Cont'd

R17G	57.3	0.374	0.00277	18.76	0.241	0.00301	-	-	-	[10]
S18A	58.4	0.398	0.00185	22.34	0.238	0.00614	-	-	-	[10]
S18G	56.5	0.440	0.00227	16.49	0.231	0.00670	-	-	-	[10]
S19G	54.8	0.384	0.00248	16.29	0.176	0.00432	1.38 (0.04)	1.40 (0.01)	1.41 (0.04)	[6, 10]
G20A	48.9	0.355	0.00270	18.11	0.217	0.00216	1.33 (0.01)	1.43 (0.01)	1.50 (0.01)	[10]
R21A	50.9	0.369	0.00144	16.54	0.138	0.00181	1.00 (0.02)	0.98 (0.02)	0.94 (0.02)	[10]
R21H	50.0	0.359	0.00130	16.31	0.138	0.00127	0.86 (0.02)	0.86 (0.02)	0.83 (0.02)	N
R21L[4]	55.9	0.521	-0.00010	15.63	0.217	0.00111	-	-	-	N
V22A	54.2	0.403	0.00116	16.29	0.155	0.00146	1.36 (0.05)	1.25 (0.04)	1.12 (0.04)	[6, 10]
Y23A	33.9	0.328	0.00098	15.99	0.114	0.00193	0.55 (0.01)	0.57 (0.01)	0.58 (0.01)	[10]
Y23F	52.8	0.376	0.00254	16.54	0.208	0.00141	1.11 (0.02)	1.23 (0.02)	1.27 (0.02)	[10]
Y23L	45.3	0.313	0.00153	16.24	0.155	0.00159	0.74 (0.01)	0.80 (0.01)	0.84 (0.01)	[6, 10]
Y24F	51.4	0.363	0.00279	15.49	0.163	0.00392	0.64 (0.02)	0.68 (0.02)	0.71 (0.02)	[10]
Y24W	52.9	0.357	0.00230	16.72	0.139	0.00436	1.27 (0.02)	1.28 (0.02)	1.30 (0.02)	[10]
F25A	32.5	0.316	0.00042	16.92	0.155	0.00098	0.72 (0.01)	0.76 (0.02)	0.79 (0.02)	[10]
F25L	42.5	0.340	0.00202	15.85	0.156	0.00239	0.62 (0.01)	0.66 (0.01)	0.68 (0.01)	[6, 10]
N26D	36.0	0.327	0.00044	14.56	0.133	0.00211	0.42 (0.01)	0.46 (0.02)	0.50 (0.03)	[6, 10]
H27A	57.7	0.388	0.00262	14.76	0.207	0.00245	-	-	-	[10]
H27G	50.5	0.367	0.00130	15.20	0.148	0.00197	0.53 (0.02)	0.54 (0.02)	0.52 (0.02)	[10]
I28A	54.2	0.379	0.00165	14.35	0.150	0.00404	0.17 (0.22)	0.14 (0.25)	0.08 (0.44)	[6, 10]
I28G	47.2	0.363	0.00105	14.93	0.181	0.00326	0.46 (0.01)	0.53 (0.01)	0.60 (0.01)	[10]
I28V	55.4	0.382	0.00328	15.01	0.164	0.00413	0.58 (0.12)	0.56 (0.10)	0.50 (0.12)	[10]
T29A	44.3	0.317	0.00100	14.80	0.152	0.00205	0.44 (0.01)	0.49 (0.01)	0.53 (0.01)	[10]
T29D	42.9	0.338	0.00009	14.38	0.159	0.00262	0.38 (0.01)	0.44 (0.01)	0.51 (0.01)	[6]
T29G	34.4	0.316	0.00001	15.32	0.200	0.00243	0.68 (0.01)	0.79 (0.01)	0.91 (0.02)	[10]

Cont'd

T29S	50.8	0.373	0.00159	15.57	0.170	0.00278	0.65 (0.03)	0.70 (0.03)	0.72 (0.04)	[10]
N30A	53.3	0.372	0.00208	15.02	0.278	0.00302	0.31 0.07)	0.61 (0.03)	0.89 (0.03)	[10]
A31G	40.9	0.359	0.00197	15.45	0.186	0.00311	0.58 (0.01)	0.65 (0.01)	0.70 (0.01)	[6, 10]
A31S	57.7	0.381	0.00283	15.76	0.133	0.00373	-	-	-	[10]
S32G	50.1	0.335	0.00200	14.46	0.145	0.00198	0.29 (0.03)	0.32 (0.03)	0.30 (0.04)	[10]
S32T	61.7	0.398	0.00356	14.70	0.100	0.00240	-	-	-	[6]
Q33A	53.1	0.332	0.00103	15.13	0.171	0.00326	0.50 (0.04)	0.60 (0.04)	0.70 (0.04)	N
W34A	52.9	0.386	0.00067	14.75	0.118	0.00295	0.43 (0.06)	0.35 (0.06)	0.24 (0.10)	[6, 10]
W34F	58.0	0.399	0.00326	15.81	0.251	0.00212	-	-	-	[10]
E35Q	53.1	0.380	0.00280	15.67	0.221	0.00265	0.72 (0.09)	0.87 (0.06)	0.96 (0.06)	[10]
E35A	50.3	0.369	0.00283	16.13	0.154	0.00203	0.82 (0.07)	0.83 (0.07)	0.79 (0.06)	[10]
R36A	56.7	0.357	0.00225	16.44	0.117	0.00231	-	-	-	[10]
S38A	59.1	0.393	0.00204	17.13	0.174	0.00327	-	-	-	[10]
S38G	58.2	0.411	0.00382	18.43	0.245	0.00295	-	-	-	[10]
S38T	58.2	0.390	0.00327	18.22	0.232	0.00337	-	-	-	N
2. Double-site mutants										
S18G/S19G	53.0	0.382	0.00163	16.88	0.169	0.00246	1.36 (0.02)	1.37 (0.02)	1.36 (0.02)	N
S19G/G20S	56.7	0.393	0.00288	16.88	0.169	0.00246	-	-	-	N
I28N/T29G	36.4	0.352	0.00024	15.25	0.287	0.00387	0.79 (0.01)	0.96 (0.01)	1.14 (0.01)	N
3. Loop1 insertion and deletion mutants [2]										
var1 (FiP)	77.5	0.428	0.00327	10.65	0.2052	0.00532	-	-	0.92 (0.01)	[7]
var2	69.2	0.425	0.00191	13.01	0.2305	0.00457	-	0.84 (0.01)	0.91 (0.01)	[7]
var3	68.1	0.422	0.00220	12.07	0.2126	0.00498	-	-	1.18 (0.01)	[7]
var4	62.0	0.393	0.00228	13.92	0.1931	0.00216	-	1.28 (0.07)	1.24 (0.04)	[7]
	47.7	0.396	0.00139	18.73	0.1310	0.00256	1.34 (0.01)		1.32 (0.01)	N
var5 (+1G)	50.9	0.366	0.00347	16.47	0.2360	0.00281	0.94 (0.01)	1.32 (0.01)	1.09 (0.01)	N
var6 (+2G)								1.09 (0.01)		

Cont'd

4. Backbone hydrogen bond amide-to-ester mutants

K13k	46.4	0.410	0.0010	16.52	0.21	0.00100	0.79 (0.01)	0.80 (0.01)	0.77 (0.01)	[16]
S16s	42.2	0.400	-0.0005	17.37	0.25	0.00120	0.91 (0.01)	0.95 (0.01)	0.97 (0.01)	[16]
R17r	49.1	0.400	0.0016	17.20	0.22	0.00300	1.08 (0.03)	1.14 (0.03)	1.19 (0.03)	[16]
V22v	56.7	0.420	0.0034	16.64	0.33	0.00340	-	-	-	[16]
H27h	38.7	0.420	0.0031	14.83	0.16	0.00560	0.46 (0.01)	0.52 (0.01)	0.57 (0.01)	[16]
S32s	41.5	0.510	0.0010	14.70	0.50	0.00090	0.72 (0.01)	0.87 (0.01)	0.98 (0.01)	[16]
W34w	49.5	0.430	0.0032	14.74	0.19	0.00840	0.39 (0.03)	0.46 (0.02)	0.57 (0.01)	[16]

[1] Mutants that differ < 1 kJ/mole in stability from wild type hPin1 WW resulted in large errors in Φ_M, so no Φ_M-values are listed. Φ_M-value were also not calculated at 50 and/or 55 °C for the more stable loop 1 deletion mutants with thermodynamically optimized loop 1 substructures, to avoid errors in Φ_M due to extrapolation of the data. Rounded errors in Φ_M of all other mutants are given in brackets.

[2] Var1: Type-I G-bulge turn, sequence: SADGR. Var2: Type-I G-bulge turn, sequence: SSSGR. Var3: Type-I' turn, sequence: SNGR. Var4: Type-I' turn, sequence: SSGR. Var5: Single Gly insertion, sequence: SRSSGGR. Var6: Double Gly insertion, sequence: SRSSGGGR.

[3] N= new mutant.

[4] Mutant R21L forms a dimer at protein concentrations employed for T-jump relaxation (10-30 μM) and was thus excluded from Φ_M analysis.

2.3.2 Probing key residues for stability by multiple mutations

Several residues critical for thermodynamic stability, i.e. R14, Y23 and F25 that constitute hydrophobic core 2 (Fig. 2.1a), and T29 in loop 2 of hPin1 WW (Fig. 2.1b), were probed by multiple mutations (vertical ΦM analysis). We find excellent agreement between the ΦM value of the non-conservative mutants R14F/L and the classical R14A mutant, and the ΦM values of the Leu and Ala mutants of F25 differ by 0.10 units (Fig. 2.2a, Table 2.1). This is clear evidence that hydrophobic cluster 2, although moderately conserved among WW domains, is rather robust towards perturbation by single side chain modifications. Loop 2 of hPin1 WW is formed by residues H27-N30, and adopts a αR-αR-αR-αL, or παL-conformation, with the first three residues being in a right-handed helical conformation, and N30 being in a left-handed helical conformation. The παL-conformation is very common among four residue loops and is also found in the homologous hYap65 and FBP28 WW domains. We probed the contribution of T29 to transition state structure and energetics by the three classical mutations T29S/A/G. The non-conservative T29D mutation was also included in the analysis, as T29D is found in the homologous hYap65 WW domain, and T29D was utilized in our first ΦM analysis study of hPin1 WW [6]. The ΦM value of T29A (0.49 ± 0.01) is closest to the error-weighted average

ΦM value (0.53), with T29D yielding a slightly lower value (ΦM = 0.44 ± 0.01) while T29S (ΦM = 0.69 ± 0.02) and T29G (ΦM = 0.79 ± 0.01) yielded higher values. Of all these, only the glycine mutant lies more than a standard deviation from the average. We also studied a double-mutant, I28N/T29G, which replaces the base of the helical παL-turn with a sequence (Asn-Gly) that has a high propensity to form a tight 4-residue type-I' turn, a common loop type seen in hairpin structures. I28N/T29G is one of the most destabilized loop 2 mutants (ΔΔGf = 8 kJ/mol) and has a large ΦM value (0.96 ± 0.01). The larger ΦM value shows that loop 2 can become rate limiting when destabilized, moving the transition state towards the native state. As shown in the next section (ΦT analysis), mutants T29G and I28N/T29G are perturbing mutants in that they shift the folding transition state with respect to wild type hPin1 WW, so both mutants are not reliable reporters of the unperturbed wild type transition state structure.

2.3.3 Perturbation of hydrophobic cluster 1 disrupts the folding transition state

Molecular dynamics simulations of the fast-folding FiP variant of hPin1 WW suggest that hydrophobic cluster 1 is only weakly formed in the transition state. The simulated ΦM values for hydrophobic core 1 residues (L7: -0.30 ± 0.50, P8: -0.3 ± 0.1, W11: ~ 0.4, Y24: 0.32 ± 0.1, P37: ~ 0) suggest that the native W11-Y24 side chain interaction is partially developed in the folding transition state, while other hydrophobic core contacts (e.g. P37 sandwiched between W11 and Y24 (Fig. A.1)) must develop after crossing the folding barrier [17, 26, 27].Because of its importance for stability (Fig. 2.1c), hydrophobic cluster 1 proves to be difficult to map experimentally by ΦM analysis. Even though the negative ΦM value of L7I (within error) agrees with the value from simulations, its ΦM value is not supported by L7A and L7V mutations. Mutating residues W11, Y24 and P37 to either Ala or Leu resulted in unfolded proteins. Mutants P8A, W11F and Y24W, although (severely) destabilized, unfold cooperatively upon heating but yield non-classical ΦM values significantly higher than the ΦM values of other hydrophobic core 1 mutations (L7I/A/V, G10A, Y24F). As the W11F mutant of hPin1 WW folds into a native-like structure with a rigid core (Fig. A.2), and because the conservative W11F mutation is unlikely to perturb unfolded state structure significantly, the high ΦM value of W11F most likely results from a perturbation of transition state energetics, rather than ground state effects. The Y24W mutation replaces the phenol-moiety of Y24 with the indole ring of Trp. The larger side chain enables "gain-of-interactions" in the denatured and transition state ensembles, as well as steric clashes in the native state that are not present in the

wild type protein. The ΦM values of mutants G10A (0.57 ± 0.02) and Y24F (0.68 ± 0.02) agree reasonably well with simulation, but we observed that neither mutation is ideal for transition state mapping. Surface-exposed G10 acts as a hinge residue in hydrophobic core 1 formation, so it does not contribute to the side chain packing of the hydrophobic core per se, and Y24F removes a solvent-exposed OH-group without perturbing the side chain packing of the core (Fig. A.1). Like Y23F in hydrophobic core 2, its ΦM value may primarily report on changes in protein solvation energetics, rather than genuine hydrophobic core contacts. Unlike the disruptive mutations P8A, W11F and Y24W, mutants G10A and Y24F were included in further analysis. In summary, the large number of disruptive hydrophobic core 1 mutants, the strong effect of the W11F mutation on the hPin1 WW folding kinetics, and the intermediate ΦM values of the non-disruptive mutants L7A/V/I, G10A and Y24F, suggest that while hydrophobic cluster 1 is only partially structured in the transition state, it is very important for protein stability.

2.3.4 Non-classical Φ_M values in loop 1

The intrinsically dynamic loop 1 substructure of hPin1 WW (Fig. A.3) was probed by both side chain and backbone hydrogen bond mutagenesis. Mutation S16s deletes the backbone hydrogen bond between residues S16 and R21, while mutation R17r weakens, but does not eliminate, the backbone hydrogen bond between residue S16 and S19 (Fig. 2.1b). Mutants S16G, S19G, S18G/S19G and G20A perturb the native state by changing the backbone entropy. Supporting our previous hypothesis that loop 1 formation is rate-limiting for hPin1 WW folding, all ten loop 1 mutants exhibit high ΦM values close to or larger than 1 (Fig. 2.2b). The highest ΦM values were calculated for mutants S16A (2.56 ± 0.02) and S16T (1.78 ± 0.02). The ΦM value of S16A is about twice as high as that of all other loop 1 mutants, and is a clear outlier. From the structure of the folded hPin1 WW domain it is not immediately obvious why S16A would perturb transition state energetics and slow down folding so much, but similar observations have been made with the fynSH3 domain [28], where a T47A substitution produces a ΦM value twice as high as that of T47S and T47G.Mutants S16G, R17r, S19G, S18G/S19G and G20A all share ΦM values > 1 (ΦM = 1.14-1.43). Mutants S16G, R17r and G20A are significantly less stable than S19G and S18G/S19G, so at least their non-classical ΦM values cannot be attributed to artifacts due to small differences in the stability between wild type and mutant proteins (ΔΔGf). ΦM values close to 1 are obtained for side chain mutants

R21A/H (loop 1/β strand 2 interface) and for mutant S16s that eliminates the backbone hydrogen bond between residues S16 and R21 that closes the 6-residue loop conformation. Except for S16A and S16T, all these mutants are used for further analysis.

2.4 Φ_T-value analysis

In folding studies that employ chemical denaturants (urea, guanidine hydrochloride) as the perturbation, transition state locations can be calculated from an analysis of the V-shaped folding relaxation rate vs. denaturant concentration plot, also known as "chevron plot." The Tanford βT value from this analysis is an indicator of the relative compactness of the folding transition on the reaction coordinate in terms of solvent accessible surface area [29]. Using temperature as perturbant by analogy [6, 30, 31], a mutant's ΦT value ($\Phi T = \frac{\partial \Delta G^{\dagger}/\partial T}{\partial \Delta G/\partial T} = \frac{\Delta S^{\dagger}}{\Delta S}$) can be used as a quantitative, entropic reaction coordinate that describes how much the transition state shifts along the reaction coordinate because of the mutation. It is worth emphasizing that the ΦT value reports on the overall changes in entropy (i.e. it also includes changes in protein solvation), not just protein conformational entropy. Because the ΦT value is calculated from two derivatives, it is also sensitive to the quality of the raw data with the best results obtained at temperatures close to the midpoint of unfolding (Tm). We first calculated ΦT values directly by taking the derivatives of the second order Taylor series in Table 2.1. Some of the quadratic coefficients have larger errors than others, and this results in unphysical values of ΦT (Fig. A.4A), of the temperature of maximal stability T_0 (where ΔG is at a minimum), and of heat capacities. We therefore also analyzed the data by Taylor series expanding the free energy around the temperature of maximal stability using $\Delta G = \Delta G_0 + \Delta G^{(2)} (T\text{-}T_0)^2$. This "ΦT T0-fit" yields essentially the same ΦM values as the Taylor expansion about Tm in Table 2.1 (Fig. A.4B), and ΦT values with more realistic T0 for all proteins, so we opt to discuss the "ΦT Tm-fit" throughout this paper. For completeness, we summarize the connection between the Taylor expansion and the common Gibbs-Helmholtz expansion (in terms of the more physical parameters ΔH0, ΔS0 and ΔCP) in the SI, and provide a table of heat capacities (Table A.4).Mutations N30A, T29G, I28N/T29G, S32s and W11F had ΦT values > 0.7 (Fig. 2.3, dotted horizontal line), which we chose as a reasonable cut-off for distinguishing between conservative and perturbing mutants because the ΦM values of mutants W11F, T29G and I28N/T29G either stand-out as clear outliers or are not cross-validated by

other mutants (Fig. 2.2b). In these mutants, the transition state shifts closer to the native state such that their ΦM values are no longer reliable indicators of the unperturbed "wild type" transition state ensemble, and thus must be excluded from consensus ΦM analysis. Excluding the abovementioned 5 outliers, the remaining mutants fall within a 25 % interval around the average ΦT value of 0.50 (Fig. 2.3, horizontal dashed line). Loop 2 mutants in general tend to have higher ΦT values, indicative that loop 2 can compete with loop 1 for becoming rate-limiting at higher temperatures. The ±0.2 spread in the transition state locations as quantified by ΦT is similar to that reported for the FBP28 WW domain, analyzed using Tanford's βT value [32]. Even though the individual ΦT values were measured with high precision (error in ΦT ~ 0.02), the systematic error in ΦT may be substantially larger. This is best seen when we compare the ΦT values of multiple mutations for one residue. Mutants R21A and R21H have very similar ΦM values (0.95 and 0.89) and essentially identical ΦT values (0.44 and 0.45), while mutants R14A, R14L and R14F also have similar ΦM values, but their ΦT values that span 25 %.The most dramatic shift in ΦT is found for the I28N/T29G mutant, whose large ΦM value (0.96 ± 0.02) also poorly agrees with other loop 2 mutants (Fig. 2.2b, Table 2.1). The double mutation I28N/R29G replaces the central two residues of loop 2 with a sequence that has a strong propensity to fold into a tight type-I' turn, suggesting that loop 2 is particularly prone to mutations that introduce residues that have a low propensity to adopt the helical αR-αR-αR-αL backbone conformation that is required to form loop 2. Indeed, the statistically preferred residues at position 29 are Ser and Thr, and at position 30, Arg, Lys, Gly or Asn. glycine (position 29) and alanine (position 30) are rare, or not found at all among WW domains. For mutant W11F, the shift in ΦT is accompanied by a very large ΦM value that clearly stands out as a outlier from the mutant pool (Fig. 2.2a), while the perturbing effect (shift in ΦT) seen for loop 2 mutants T29G, I28N/T29G, N30A and S32s results in more subtle abnormalities in ΦM that are more difficult to identify by merely looking at the context-dependent ΦM values alone (Fig. A.5). A third class of mutants (e.g. P8A, S16A, V22A and Y24W) shows clear outlier ΦM values, but normal ΦT values.

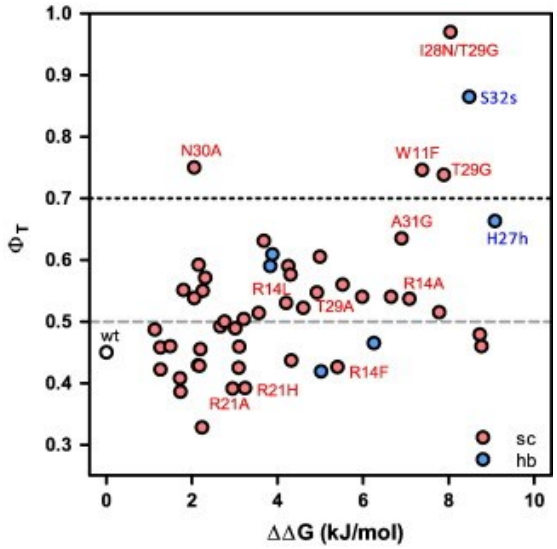

Fig. 2.3: ΦT Analysis at 55 °C: Plot of the ΦT-value for wild-type hPin1 WW and mutants thereof versus the change in free energy (ΔΔG, in kilojoules per mole) between wild type and mutant. ΦT-Values are calculated using the T0-fitting procedure (for details, see Appendix A). ΦT-values of side-chain and backbone hydrogen bond mutants are color coded red and blue, respectively. Except for the obvious five outliers (mutants W11F, T29G, I28N/T29G, N30A, S32s), the ΦT-values are within a ± 25% error margin of the average ΦT (0.50, dashed gray horizontal line). The outlier ΦT-values (N0.70, dotted gray line) are indicative of perturbing mutations that shift the transition-state ensemble along the reaction coordinate closer to the native-state. Mutational ΦM-values calculated from these mutants are no longer reliable indicators of the unperturbed "wild-type" transition-state ensemble, and must be excluded from the consensus ΦM analysis of hPin1 WW transition-state structure

2.5 High-resolution mapping of the folding transition state of hPin1 WW

2.5.1 General features of the transition state

Our approach for mapping the folding transition state of hPin1 WW was to pick the most conservative mutant set with ΦM values that were not outliers, based on cross-validation by multiple mutations, sequence neighbors, and backbone hydrogen bond neighbors, and whose ΦT values indicate no excessive shift of the transition state. Thirty-nine mutants (34 side chain and 5 backbone hydrogen bond variants) fulfill these criteria and form a consensus set for transition state analysis (Fig. 2.4a, Table 2.2). Except for S19G and I28V, all mutants had $\Delta\Delta G_f$

> 2 kJ/mol, close to or above the empirical cutoff (> 2.50 kJ/mol) for reliable ΦM analysis [33], and except for mutants I28A and E35Q/A, statistical errors in ΦM were small. Several residues (L7, E12, R14, R21, Y23, F25, I28, T29) in hPin1 WW were probed by more than one side chain mutation. For these residues, we can calculate more robust (and more representative) error-weighted average ΦM values from the side chain ΦM values of individual mutations (Table 2.2). Mapping the (error-weighted average) side chain ΦM values onto the Cα-backbone of the folded protein reveals that loop 1 (S16-R21) is substantially more structured in the transition state than loop 2 (H27-N30) and hydrophobic cluster 1 (Fig. 2.4b). The (error weighted) average side chain ΦM plot is a smooth function of sequence (Fig. 2.5a, solid red line), indicating that the formation of transition state structure is governed mainly by local interactions. Even without the outlier mutants S16A/T, a peak at loop 1 is obvious (see Fig. A.5 for an extended plot, including outliers). While hydrophobic cluster 1 contacts (probed by L7V/A, G10 and Y24F) are essential for hPin1 WW stability, their contribution to the folding rate is small, and folding of hPin1 WW is rate-controlled by the loop 1 substructure that contributes only slightly to thermodynamic stability. The high side chain ΦM value of the C-terminal E35, although corroborated by two mutants (E35A/Q), may not truly report on transition state structure. E35 is a charged residue and solvent-exposed in the folded protein. Except for mutant S16A, we find good agreement between the ΦM values of individual Ala mutants and the consensus average ΦM value (Fig. A.5).

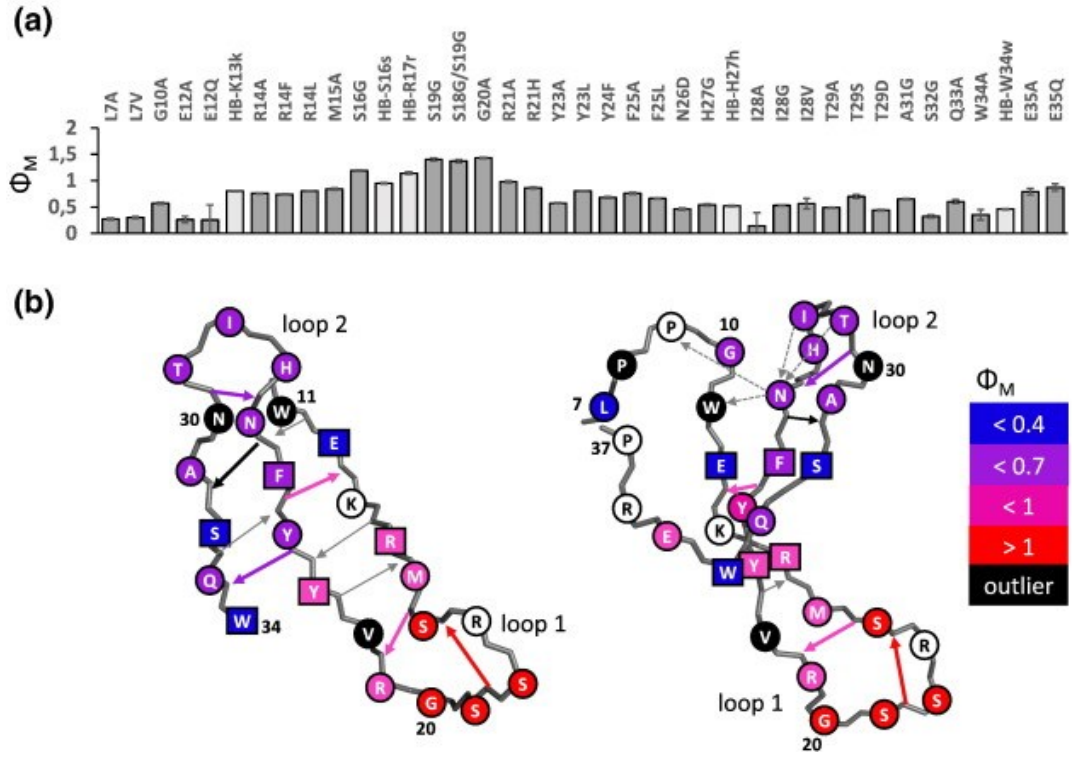

Fig. 2.4: Analysis of the folding transition state of the hPin1 WW domain. (a) ΦM-Values of the 34 single and double mutants (dark gray) and the 5 amide-to-ester backbone hydrogen bonds mutants (light gray) that qualify for ΦM analysis, and that were used for consensus ΦM mapping of the folding transition state. (b) ΦM Map of the folding transition state, with ΦM-values for 25 of the 34 residues (single letter representation) mapped onto the backbone structure of the N-terminally truncated folded protein (residues 6–39). Left panel: residues W11–W34 that define the three-stranded β-sheet. Right panel: residues L7–P37 that includes hydrophobic cluster 1 and the N- and C-terminal extensions. For clarity, ΦM-values were grouped and color coded (0 b ΦM b 0.40, blue; 0.4 b ΦM b 0.7, purple, ΦM b 1, pink; ΦM N 1, red). Residues for which classical hydrophobic deletion mutagenesis yields very high, or negative, ΦM-values that are not supported by other mutations or structural context are color coded black. Residues for which no mutant is suitable for ΦM analysis are color coded white. Backbone hydrogen bonds that were studied by amide-to-ester mutagenesis are indicated by arrows (same color code as for side chains). Data used to render the figure are provided in Tables 2.1 and 2.2.

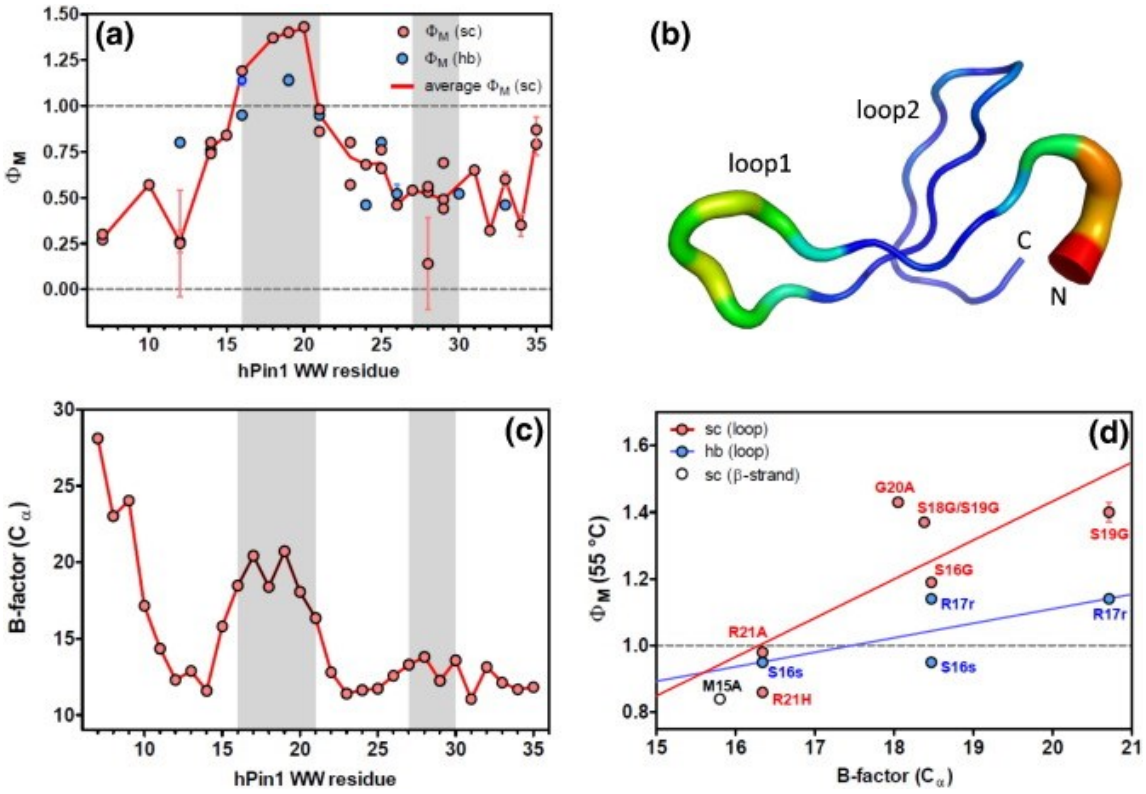

Fig. 2.5: ΦM versus sequence map and ΦM versus backbone disorder correlation. (a) Plot of ΦM-values versus the hPin1 WW sequence used for transition-state analysis. Individual side-chain ΦM-values are color coded red, while those calculated from backbone hydrogen bond mutants are color coded blue. The solid red line represents the error-weighted average trend of the side-chain ΦM (see Table 2.2 for data). The gray bars indicate the regions of loop 1 and loop 2. (b) Tube plot showing the distribution of thermal B factors from the X-ray crystal structure [17] along the backbone of hPin1 WW domain. (c) Plot of thermal B factors versus the hPin1 WW sequence, showing a pronounced maximum in loop 1, and a smaller maximum in loop 2. (d) Correlation between ΦM-values and thermal B factors for residues M15–R21 with increased local backbone disorder at 55 °C. Side chain (sc) loop 1 mutants are color coded red and backbone hydrogen bond mutants (hb) are color coded blue. The solid lines represent best fits of the experimental data.

Table 2.2:

Table 2.2: Summary of Φ_M values of consensus mutants used for transition state mapping at 55 °C

Residue	Mutation	Type [1]	ΔΔG (kJ/mol)	Φ_M (55 °C)	Average Φ_M (sc)	Average Φ_M (hb)
L7	L7A	sc	6.65	0.27 (0.02)	0.28	-
	L7V	sc	5.00	0.30 (0.02)		
G10	G10A	sc	3.56	0.57 (0.02)	0.57	-
E12	E12A	sc	2.31	0.26 (0.06)	0.26	0.80
	E12Q	sc	1.26	0.25 (0.29)		
	K13k	hb	5.01	0.80 (0.01)		
R14	R14A	sc	7.08	0.76 (0.01)	0.77	-
	R14F	sc	5.41	0.74 (0.01)		
	R14L	sc	4.18	0.80 (0.01)		
M15	M15A	sc	2.66	0.84 (0.02)	0.84	-
S16	S16G	sc	4.25	1.19 (0.01)	1.19	1.01
	S16s	hb	6.45	0.95 (0.01)		
	R17r	hb	3.38	1.14 (0.02)		
S18 [3]	S18G/S19G	sc	2.19	1.37 (0.02)	1.37	-
S19	S19G	sc	1.49	1.40 (0.03)	1.40	1.19
G20	G20A	sc	3.68	1.43 (0.01)	1.42	-
R21	R21A	sc	2.95	0.98 (0.02)	0.92	0.95
	R21H	sc	3.24	0.86 (0.02)		
Y23	Y23A	sc	8.77	0.57 (0.01)	0.72	-
	Y23L	sc	4.60	0.80 (0.01)		
Y24	Y24F	sc	2.76	0.68 (0.02)	0.68	0.46
F25	F25A	sc	8.73	0.76 (0.02)	0.69	0.80
	F25L	sc	5.98	0.66 (0.01)		
N26	N26D	sc	7.79	0.46 (0.02)	0.46	0.52
	H27h	hb	9.08	0.52 (0.01)		
H27	H27G	sc	3.09	0.54 (0.02)	0.54	-
I28	I28A	sc	1.72	0.14 (0.25)	0.52	-
	I28V	sc	1.26	0.56 (0.10)		
	I28G	sc	4.31	0.53 (0.01)		
T29	T29A	sc	4.92	0.49 (0.01)	0.49	-
	T29S	sc	3.01	0.70 (0.04)		
	T29D	sc	5.52	0.44 (0.01)		
N30	H27h	hb	9.08	0.52 (0.01)	-	0.52
A31	A31G	sc	6.87	0.65 (0.01)	0.65	-
S32	S32G	sc	3.10	0.32 (0.03)	0.32	-
Q33	Q33A	sc	2.05	0.60 (0.04)	0.60	0.46
	W34w	hb	3.87	0.46 (0.01)		
W34	W34A	sc	2.23	0.35 (0.10)	0.35	-
E35	E35A	sc	3.27	0.83 (0.06)	0.85	-
	E35Q	sc	2.14	0.87 (0.07)		

[1] Type of mutation: side chain (sc), backbone hydrogen bond (hb). [2] Error weighted average Φ_M-value for residues probed my multiple mutations. [3] Φ_M-value of the S18G/S19G was assigned to S18.

2.5.2 Correlation between native-state disorder and non-classical Φ_M-values in loop 1

Here we propose the hypothesis that φM values >1 in loop 1 (see section 2) are due to native-state backbone dynamics. An NMR-solution structure of the apo-form of the isolated WW domain implies that loop 1 is intrinsically dynamic [34] (Fig. A.3), and this dynamic

nature appears to be preserved in the high-resolution X-ray structure (1.35 Å) of hPin1 WW in the context of the full-length hPin1 rotamase (Fig. 2.5b). Except for M15A in β strand 1, all mutations that yield non-classical ΦM values > 1 mutate residues that map onto the intrinsically more disordered loop 1 region, and the concordance between the average consensus ΦM values (Fig. 2.5a) and the thermal B factors (a convenient measure for native-state conformational disorder) (Fig. 2.5c) is striking. The reasonable correlation between the local disorder of a loop 1 residue and the magnitude of its ΦM value (Fig. 2.5d) suggests that the ΦM values in loop 1 are shifted upward further, from values near 1 that are indicative of the importance of loop 1 in the transition state, to even larger values indicative of native state disorder. A more disordered loop 1 may better accommodate mutations that change backbone and sidechain entropy or perturb backbone hydrogen bonds, and thus yields a lower ΔΔGf (and a higher ΦM value), if at the same time the transition state is more sensitive to such mutations because other robust structure (e.g. hydrophobic core 1) have not yet formed.

2.5.3 Correlation between side chain and backbone hydrogen bond ΦM values

Hydrophobic cluster 2 (R14-Y23-F25) that stabilizes the N-terminal β-hairpin is loosely formed in the transition state, making an average of 73 % of its native contacts in the transition state (R14 = 77 %, Y23 = 72 %, F25 = 69 %, each calculated from the error-weighted average ΦM, Table 2.2). The ΦM value of mutant K13k that weakens the E12-F25 backbone hydrogen bond (0.80 ± 0.02) agrees well with the side chain Φm values of hydrophobic core 2 that protects the hydrogen bond from solvent in native hPin1 WW, suggesting that the E12-F25 backbone hydrogen bond and hydrophobic cluster 2 form cooperatively in the folding transition state. To test whether this correlation between backbone hydrogen bond and side chain ΦM values generally holds for hPin1 WW, it is helpful to compare the backbone and side chain ΦM values at the level of individual residues. We thus assign the ΦM value of a perturbed backbone hydrogen bond to the two residues that form such a bond, not the residue that is mutated to perturb the hydrogen bond (as done in a previous study [16]). For example, mutation S16s eliminates the S16-R21 backbone hydrogen bond by replacing the amide moiety of the M15-S16 backbone peptide bond that acts as a hydrogen bond donor to form the backbone hydrogen bond with the carbonyl moiety of residue R21 with an ester moiety that cannot engage in backbone hydrogen bond formation (Fig. 2.1b). Here, we assign the ΦM of the S16s mutant to both residue S16 and R21. Likewise, mutation K13k perturbs, but does not eliminate, the

backbone hydrogen bond between residues E12 and F25, by weakening the hydrogen bond acceptor (backbone carbonyl) of E12 (Fig. 2.1b). Here, however, it would be more correct to assign the ΦM of K13k not to residue K13 but to residues E12 and F25 that form the backbone H, even though formally, the amide-moiety of residue K13 is mutated.

Overall, we find good agreement between the "residue-assigned" backbone ΦM values (Fig. 2.5a, filled blue circles) and the ΦM values from classical side chain mutation (Fig. 2.5a, filled red circles), in particular within the hairpin 2 region (Table 2.2). As the strength of a hydrogen bond is strongly dependent on the distance between the hydrogen bond donor (backbone amide) and hydrogen bond acceptor (backbone carbonyl), even fractional backbone hydrogen bond ΦM values of ~0.5 imply that loop 2 is highly compact or that the measured fractional ΦM values within hairpin 2 represent ensemble averages with about 50 % of the molecules having hairpin 2 fully formed in the transition state ensemble (ΦM ~1), while in the other half of molecules hairpin 2 is disordered (ΦM ~ 0). Such a scenario has been predicted in less extreme form from Markov-State-modeling of hPin1 WW folding [35-37]. The poor agreement between the side chain and backbone ΦM values calculated for residue E12 probably stem from the removal of a solvent-exposed charged residue by mutations E12A/Q. Long-range electrostatic effects may play a role instead of just local contacts.

2.5.4 Variation of transition state structure with temperature

Probing the folding kinetics not just at a single temperature, but over a wider range of temperatures (here, 50, 55 and 60 °C), reveals the robustness of the transition state ensemble against thermodynamic stress. Folding studies at various temperatures also identify 'borderline' mutations that perturb the folding mechanism under increased thermal stress, but whose disruptive nature might escape detection under more favorable folding conditions. On average, the ΦM values increase by 0.07 units (Fig. 2.6a) and the ΦT value increases by 0.15 units (Fig. 2.6b) upon raising the temperature from 50 to 60 °C (for data, see Table A.1, A.2). This suggests that the folding transition state becomes more structured and native-like at higher temperature, and the transition state ensemble shifts along the reaction coordinate closer to the native state, in agreement with Hammond's postulate [38]. A plot of ΦM (60°C)/ΦM(50°C) vs. sequence in Fig.2. 6c reveals that structure within hairpin 1 (residues 12-25) at best changes only weakly with temperature. In contrast the loop 2 region (residues 27-30), the third β strand (residues 31-34) and hydrophobic core 1 (probed by L7A and L7V) increase by a larger margin

and beyond experimental uncertainty. The absolute changes in ΦM are, however, rather small such that hairpin 1 still dominates transition state structure at higher temperatures. The Ala mutant W34A may show unusual temperature tuning (although it has a large error bar in Fig. 2.6c), and we speculate on a possible origin in the SI.

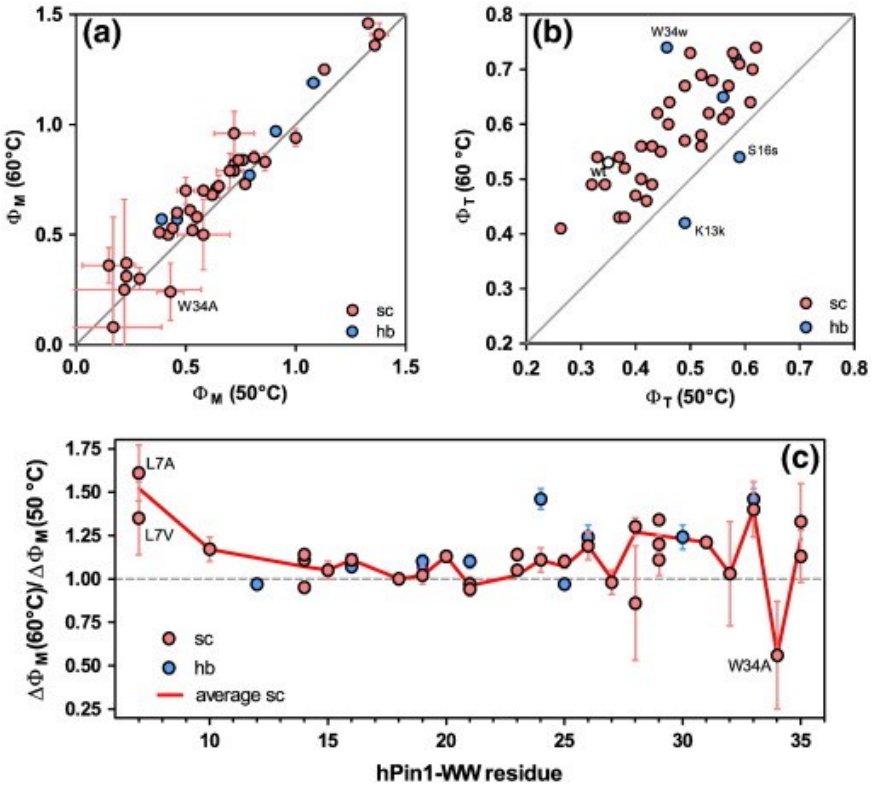

Fig. 2.6: Variation of transition-state structure with temperature. (a) Plot of ΦM (60 °C) versus ΦM (50 °C). On average, ΦM-values increase by 0.07 units when raising the temperature from 50 °C to 60 °C, suggesting that the transition-state overall gains native structure upon heating. (b) Plot of ΦT (60 °C) versus ΦT (50 °C). On average, ΦT-values increase by 0.15 units when raising the temperature from 50 °C to 60 °C, suggesting that the transition state becomes more native-like at elevated temperature, consistent with Hammond's postulate. (c) Plot of the ΦM (60 °C)/ΦM (50 °C) ratio versus the residue number of the hPin1 WW sequence. Data from individual side-chain mutants are color coded red. Data from individual backbone hydrogen bond mutants are color coded blue. The solid red line represents the error-weighted average side-chain trend. For clarity, the side-chain data of E12 (large errors, see Table 2.2) are not shown.

2.5.5 Average fraction of native contacts and its temperature dependence

For the set of consensus mutants depicted in Fig. 2.4a, we calculate an average ΦM value of 0.68 ± 0.04 at 55 °C, which is higher than the overall average ΦT value (0.50 at 55 °C, excluding the 5 outliers discussed in sections 3 and 4). Mutants with a higher slope of ΔG vs. T (folding cooperativity) have a higher melting temperature (Tm) (Fig. 2.7a, where ΔG=0 at T=Tm for all mutants). The average slope is +0.0017 kJ/mole/K, indicative of a negative folding entropy $\Delta S=-(\partial \Delta G/\partial T)$, and increases by about 0.1 kJ/mole/K over the 35-60 °C range of melting temperatures. The size-dependence of ΔS for folding has been discussed in the literature [39, 40]. From the temperature dependence of the folding barrier on protein stability (Fig. 2.7b), we calculate a slope $(\partial \Delta G^{\dagger}/\partial T) \approx 0.0024$ kJ/mole/K (0.0028 for all mutants listed in Table 2.1). The ratio of the two slopes (activated/ground) is ~ 0.70 (0.63 for all mutants listed in Table 2.1). This value is also higher than the average ΦT value of 0.50, and suggests that there is a significant unfolding cooperativity effect in the folding transition state, although not as high as the unfolding cooperativity seen in the native protein. The ΦT value thus seems to slightly overestimate the distance of the transition state to the native state.

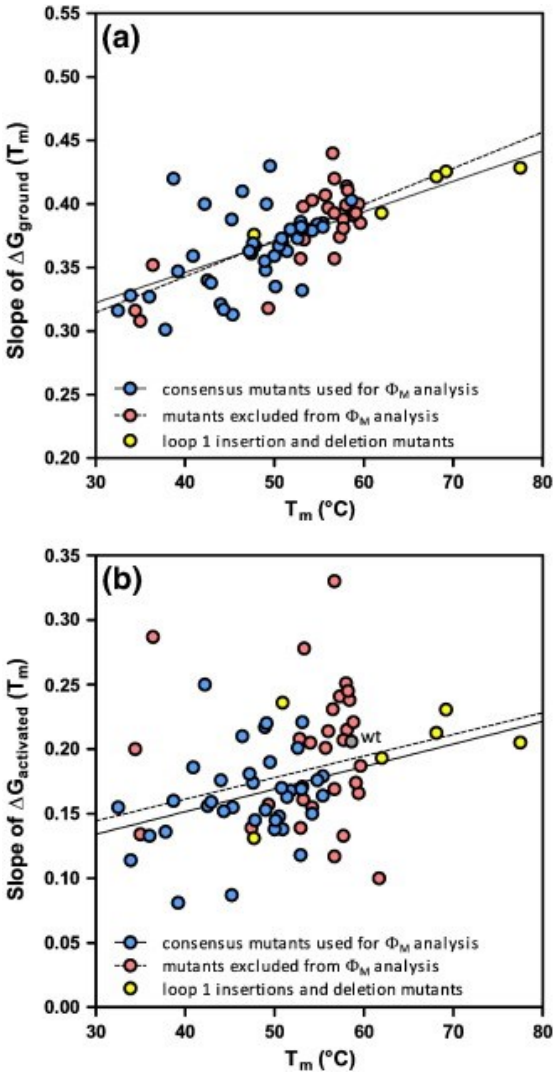

Fig. 2.7: Average number of native contacts in the folding transition state. (a) Slope of the ground-state free energy ($\partial\Delta G_{ground}(T)/\partial T$) of the 39 consensus mutants used for ΦM analysis (filled red circles, solid black line) or the entire set of single and double mutants (excluding the 6 loop 1 insertion and deletion variants) (filled gray circles, dashed black line) at the midpoint of unfolding (T = Tm, with $\Delta G_{ground}(Tm= 0)$. (b) Corresponding plot as in panel A showing the slope of the free energy of activation ($\partial\Delta G_{activated}(T)/\partial T$) at the midpoint of unfolding (T = Tm). The ratio of the two slopes (activated/ ground) of ~0.70 for the 39 consensus mutants (0.63 for the entire mutant set) suggests that about 70% of the native contacts are developed in the folding transition state, a value that agrees well with the average calculated from the ΦM data (Table 2.2), but that is higher than the average ΦT-value (0.50). The loop 1 insertion and deletion variants that change local changes in backbone topology (filled yellow circles) were excluded from the fit, but their values agree well with the extrapolated fits of the mutants with the 6-residue wild-type hPin1 WW loop 1.

2.6 ΦM analysis of loop 1 insertion and deletion mutants

2.6.1 Mutant design and structural analysis

We recently designed and biophysically characterized several hPin 1 WW variants in which the wild type loop 1 sequence is replaced by either a 5-residue type-I G-bulge turn (the preferred loop type in WW domains) or tighter, 4-residue type-I' turns that are not found among WW domains [7] (Fig. 2.8a). The X-ray structures of the most stable type-I G-bulge variant (var1, or FiP, loop sequence: SADGR) and the most stable type-I' turn variant (var3, loop sequence: SNGR) have been solved at 1.90 and 1.50 Å resolution, respectively. Both variants essentially superimpose with the wild type structure (1.35 Å resolution), except for the redesigned loop 1 region (Fig. 2.8b). The thermal B factors of the FiP variant are consistently lower than that of wild type hPin 1 WW, while those of var3 are higher (Fig. A.6). While the difference in the absolute values of the thermal B factors may result from different crystal packings, we note that turn 1 in the X-ray structure of FiP appears to be conformationally rigid, consistent with NMR-solution data of the same turn in its natural FBP28 WW context (APPENDIX AFig.3). The 4-residue type-I' turn of variant 3 shows a relative maximum in the B factor similar that of loop 1 in wild type hPin1 WW, suggesting that the type-I' turn, although stabilizing and hastening hPin1 WW folding, is conformationally flexible in the folded protein.

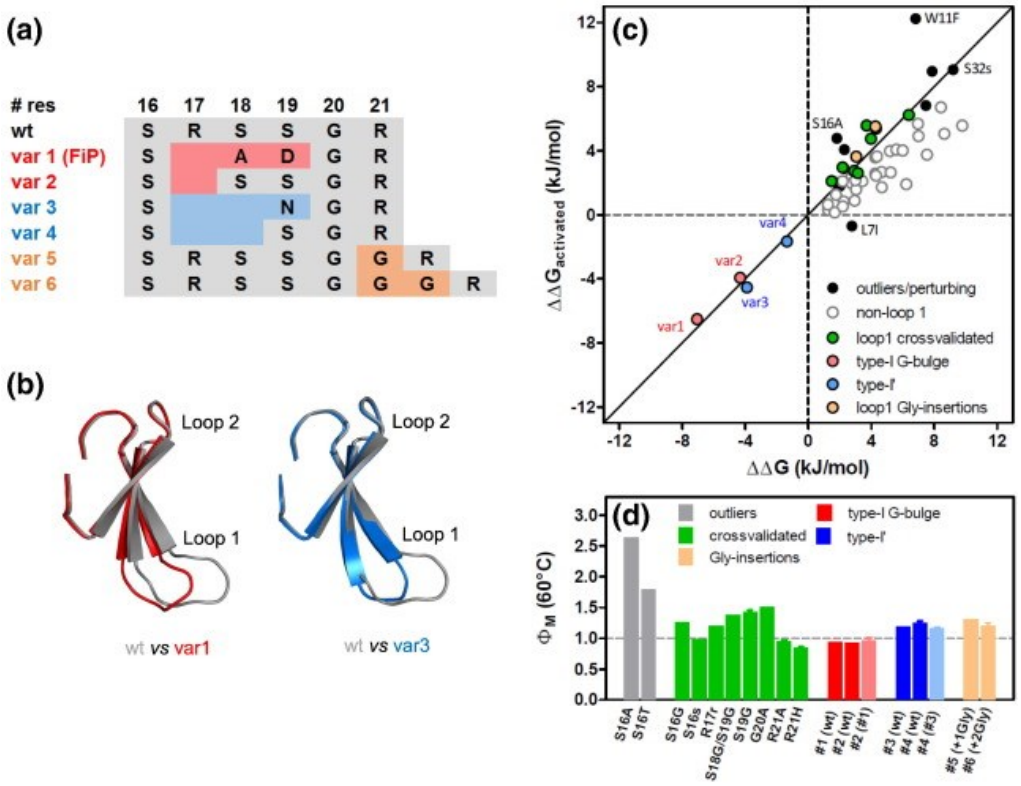

Fig. 2.8: ΦM Analysis of hPin1 WW variants with loop 1 deletions or insertions mutations. (a) Loop 1 sequences of the hPin1 WW loop 1 deletions or insertions variants. Wild-type residues are numbered and color coded gray. Mutated or deleted residues in the loop deletion variants are color coded red (type I G-bulge turn) and blue (type I′ turn), while the inserted Gly residues in the loop 1 insertion mutants are highlighted in orange. (b) Superposition of the high-resolution X-ray structures of type I G-bulge variant FiP (1.90 Å resolution, color coded red, left) and the type I′ variant 3 (1.50-Å resolution, color coded blue, right) with wild-type hPin1 WW structure (1.35 Å resolution, color coded gray). (c) Brønsted plot for folding of the loop 1 variants of hPin1 WW at 60 °C, rendered from the data provided in Table A.2. Filled red circles: 5-residue type I G-bulge turn mutants (var1, var2). Filled blue circles: 4-residue type I' turn variants (var3, var4). Filled green circles: cross-validated loop 1 side-chain and backbone hydrogen bond mutants (6-residue wild-type loop 1 context). Filled orange circles: Gly insertion variants (var5, var6). Filled black circles: outlier/perturbing mutants. Open light gray circles: non-loop 1 consensus mutants. The solid black line is the line predicted for ΦM = 1. (d) Bar plot of ΦM-values for selected mutants shown in panel C. ΦM-Values calculated for the redesigned loop 1 variants using wild-type hPin1 WW as reference are color coded red (5-residue type I G-bulge variants) and blue (4-residue type I' variants). ΦM-Values calculated for variants 2 and 4 in the type I G-bulge (var1, FiP) and type I' context (var3) are shown in light red and light blue, respectively.

2.6.2 Group ΦM analysis and ΦM vs. ΔΔGf correlation

At 60 °C, and using wild type hPin1 WW as the reference protein, we calculate ΦM values of 0.92 ± 0.01 for FiP and 0.91 ± 0.01 for the related variant 2. Both ΦM values are cross-validated by the ΦM value of variant 2 calculated with FiP as "pseudo wild type" reference (0.94 ± 0.05) (Fig. 2.8d), demonstrating that ΦM analysis is surprisingly robust towards more severe sequence manipulations that simultaneously alter sequence and local chain topology. The ΦM values of FiP and related variant 2 also agree well with the ΦM values of mutants R21A, R21H and S16s (ΦM = 0.83-0.97) measured in the wild type loop context (Fig. 2.8 c,d). This correlation is remarkable in that the mutants differ by up to 15 kJ/mole in stability. It further implies that in the strictly sequential folding model (loop 1 first, then loop 2) proposed for FiP by Shaw et al., the energy barrier of the second transition (loop 2 nucleation) must be sufficiently small for FiP-variant 2 to yield a ΦM value = 0.94 ± 0.05 (Fig. A.7A). The GTT variant of FiP with an optimized loop 2 structure, however, significantly accelerates FiP folding (by a factor of three), suggesting that loop 2 formation in FiP is associated with a non-negligible barrier and rate-limiting for folding (Fig. A.7B). Both observations are contradictory and difficult to reconcile in the framework of a sequential model, but perfectly compatible with a simple two-state mechanism, as in the latter case, stabilizing loop 1 and loop 2 mutations may additively lower the (single) transition barrier (Fig. A.7C). Type-I' turn variants also hasten wild type hPin1 WW folding, but by a smaller margin than in FiP. In contrast, the two Gly insertion variants 6 and 7 (both less stable than wild type) slow down folding, presumably because of an increased entropic penalty to form the longer 7- or 8-residue loop 1 substructure. All four variants yield ΦM values greater than 1, similar in magnitude to the ΦM values of wild type mutants S16G, S18G, S18G/S19G and G20A (Fig. 2.8d). As for wild type hPin1 WW (Fig. 2.5), increased local backbone dynamics around the type-I' turn may cause the already high ΦM values to fall outside the classical range.

2.6.3 Hypothetical hybrid ΦM-map of FiP and comparison with MD-simulations

ΦM values are determined experimentally as a ratio of logarithms of rates to logarithms of equilibrium constants. This can be simulated directly by computation (using long trajectories or multiple shorter trajectories with Markov analysis to obtain rate and equilibrium constants),

or it can be done by examining structure near the transition state (which has a Pfold ≈1/2 folding probability) and comparing with native structure (based on native contacts). In principle, the kinetic/energetic method is the more direct comparison, but structural information may have smaller error bars than energy information, so there is a tradeoff between the two approaches. Extensive data sets such as those in the present paper should become amenable to both approaches in the next few years, to test the merits of the structural vs. energetic approach to simulated ΦM values in detail. Here we present a brief comparison of our results, adapted to the FiP modification (see loop mutants in Table 2.1 for example) of WW domain, and comparing with ref. [14], which presents both structure-based (native side chain contacts) and energy based (long trajectory kinetics) ΦM values. In the case of [14], the difference between experiment and the two computational approaches still exceeds the difference between the computations, so it appears that force field errors currently still dominate over errors caused by the structural approximation. We assume that replacing the wild type hPin1 WW loop with the FiP loop 1 sequence only affects the local loop 1 energetics. This assumption is justified by the smooth dependence of ΦM on sequence, and by the nearly superimposable loop 2 and hydrophobic core 1 substructures of FiP and wild type hPin1 WW (Fig. 2.8b). A hypothetical "hybrid" ΦM-map can be rendered for the ultrafast-folding FiP variant by combining the loop 1 ΦM value of FiP variant 2 (0.94 ± 0.05, measured with FiP as the "pseudo wild type" reference) with the non-loop 1 ΦM values obtained with wild type hPin1 WW (Fig. 2.9, red symbols and solid red line). For loop 1 and its immediate sequence neighbors, our putative "hybrid" ΦM map (60 °C) agrees well with the simulated ΦM map calculated at slightly higher temperature (75 °C) [14]. This reinforces our hypothesis (previous paragraph) that replacing loop 1 in wild type hPin1 WW with more stable sequences hastens folding without changing the folding mechanism - either loop type is substantially (or fully) formed in the folding transition state. The ΦM values within the loop 2 region, however, do not agree very well. Here, the experimental ΦM values clearly suggest more structure within hairpin 2 than the MD-simulation [14]. As loop 2 slightly gains structure with temperature this discrepancy should be even more pronounced at 75 °C (the temperature used for MD-simulations).Shaw et al. argue that the folding mechanism of FiP is a direct consequence of the difference in the thermal stability of the N- and C-terminal hairpins. Although the isolated hairpins fold about one order of magnitude faster than full-length FiP and at similar rates in simulations, hairpin 1 with the optimized loop 1 sequence is significantly more stable (25 % folded hairpin at equilibrium) than hairpin 2 (4 % folded hairpin at equilibrium), such that loop 1 nucleation is expected to kinetically outperform loop 2 nucleation. Although plausible, this model does not take into

account the aforementioned significant (approximately 3-fold) increase in the folding rate that is seen experimentally with the GTT-FiP variant. In hPin1 WW with the unstable and intrinsically flexible 6-residue loop 1 sequence, isolated hairpin 1 is expected to be much less stable, perhaps even less stable than isolated hairpin 2. This would open up three possible folding scenarios: With both hairpins being similarly unstable, folding could occur through parallel pathways nucleated by either loop substructure (scenario 1), as predicted from Markov-state-modeling of hPin1 WW folding. In this case, the experimentally measured ΦM values for the loop 1 and loop 2 regions would directly describe the relative flux along either pathway. In the simplest, and most extreme case, the hairpin whose loop segment nucleates folding is fully formed in the transition state (ΦM ~ 1) while the other hairpin is completely unstructured (ΦM ~ 0). For loop 2, we find average ΦM values of ~ 0.60 at 60 °C. Therefore, if that extreme model applied, one would expect ΦM values of only ~ 0.40 for loop 1, which is clearly not what we observe experimentally (average ΦM > 0.9 at 60 °C). Alternatively, both loop substructures may fluctuate between an open and a closed state, although not necessarily a native-like state, however a native-like N-terminal hairpin is mandatory for barrier-limited folding into the native state (scenario 2). In this model, loop 1 residues will by necessity yield the highest ΦM values, while the loop 2 ΦM values will be reporters about the equilibrium ratio of the open and closed hairpin 2 conformations before their interaction with the structured N-terminal hairpin occurs. As loop 2 formation could either occur before or after loop 1/hairpin 1 formation, hairpin 1 would "catalyze" the final transition of hairpin 2 from the closed to the native state. This folding model is unlikely for wild type hPin1 WW domain because an increase in temperature should shift the loop 2 equilibrium towards the open (less structured) conformation, so the loop 2 ΦM should decrease with temperature, rather than (slightly) increase. It may, however, become a dominant mechanism in fast-folding WW domains such as FiP. The most likely folding model for hPin1 WW thus remains a two-state folding mechanism, in which folding and docking of the hairpins occurs in a concerted fashion. The measured ΦM values would then imply that the N-terminal hairpin is mainly formed in the transition state, while the second hairpin and the hydrophobic core are in the process of being formed in the transition state. Two-state folding of not only wild type hPin1 WW, but also the FiP variant, would also better explain why certain FiP variants such as FiP-GTT with stabilizing mutations within loop 2 and β strand 3 speed up its folding despite high ΦM values near unity in the hairpin 1 turn region.

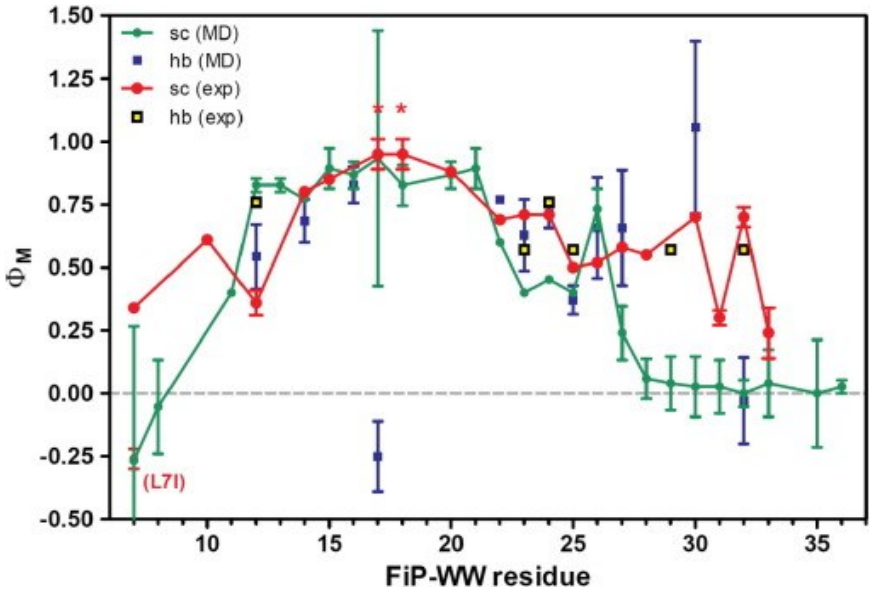

Fig. 2.9: Hypothetical "hybrid" ΦM-map for the fast-folding FiP variant of hPin1 WW. Hypothetical side-chain ΦM-map (red circles and solid red line) for the fast-folding FiP variant of hPin1 WW, rendered with side-chain ΦM-values of non-loop 1 mutants measured with wild-type hPin1 WW as reference (see Fig. 2.3, Table A.2 for details) and the side-chain ΦM-value for loop 1 FiP WW variant 2 (loop 1 sequence: SSSGR) measured with FiP as "pseudo wild-type" reference (loop 1 sequence: SADGR). As two residues were replaced simultaneously in FiP variant 2 (A18S, D19S; see Fig. 2.8a), the ΦM-value calculated for variant 2 (ΦM = 0.94 ± 0.05) was assigned to either mutated residue (labeled by asterisks) in FiP. For residues that are probed by multiple side-chain mutations, the error-weighted average ΦM-value is shown (see Table A.2 for details). Experimentally measured backbone hydrogen bond ΦM-values (filled yellow squares) are those measured for wild-type hPin1 WW and are assigned to the two residues that engage in the perturbed hydrogen bond (see Table A.2 for details). The simulated side-chain and backbone hydrogen bond ΦM-values and associated errors are shown in green and blue, respectively and were rendered from Fig. 2E in Ref. [15]. Residue numbers correspond to the 33-residue FiP sequence and thus account for the shorter loop 1 substructure (deletion of Arg17 of wild-type hPin1 WW).

2.7 Conclusions

ΦM-value analysis can provide valuable information about the structure of folding transition states by correlating changes in mutationally induced stability and folding kinetics. In its simplest manifestation, ΦM-value analysis can be affected by probe perturbation of the

folding mechanism, and by a trickle-down effect of mutations that lowers the structural resolution. Such trickle down effects can arise for instance from native state flexibility, or from solvent interactions that do not report on genuine structure per se. Here we present a comprehensive ΦM-value analysis with horizontal (sequence), vertical (multiple mutations at a single site) and chemical depth (side chain and "residue-assigned" backbone hydrogen bond mutations) to identify reliable mutations that can act as probes of the folding mechanism. The "conservatism" of mutations with respect to the folding mechanism is ascertained by multiple side chain substitutions at the same site (L7, E12, R14, S16, Y23, Y24, F25, I28 and T29), verification of individual ΦM values by cross-β strand neighbors (M15 vs. V22, E12 vs. F25), residue assigned ΦM values from backbone hydrogen bond mutagenesis (e.g. S16A/G/T vs. S16s, N26D vs. H27h) or immediate sequence neighbors (R21-V22-Y23 series), and temperature tuning (outliers in ΦT).For some residues (R14, T29), ΦM values calculated from non-conservative mutations agree well with ΦM values calculated from more conservative and structurally less perturbative mutations, while other mutations yield ΦM values that primarily report on the energetics of polar or charged residues with solvent (e.g. Y23F, E12A/Q, E35A/Q). Another subclass of mutations that target the flexible loop 1 substructure of hPin1 WW (S16G, R17r, S19G, S18G/S19G, G20A) yield ΦM values that lie clearly outside the classical range (ΦM > 1). Based on the correlation with X-ray B factors, their high ΦM values result at least in part from increased local backbone dynamics in the native state. Although Ala mutations overall appear to be reliable reporters of transition state structure, as often assumed in the literature, we also identify clear outliers (P8A, S16A and V22A). Another Ala-mutant (W34A) shows an unusual dependence on temperature tuning. Its ΦM value decreases with temperature, suggesting that the smaller Ala residue perturbs non-native interactions that are stable at low temperature, yet nevertheless speed up folding. Aside from obvious mutant outliers that can be easily identified by cross-validating their ΦM values with different mutants at the same sequence location, another subset of mutants perturb transition state structure and shift the transition state ensemble to a more native-like ensemble state, as evidenced by large ΦT values for such mutations. Four of the five mutants that shift the transition state position in Fig. 2.5 map to the loop 2 region or immediately flanking residues. Although not dominating transition state structure, the wild type sequence of loop 2 can be perturbed sufficiently to affect folding rates. The ease with which the folding mechanism of the hPin1 WW domain can be changed by what appears to be subtle sequence modifications or perturbations of intermolecular forces (e.g. weakening a single, partially solvent-exposed backbone hydrogen bond as in amide-to-ester mutant S32s) argues against two-state folding with a well-defined,

robust and narrow transition state and suggests a more complex, multidimensional energy surface with additional local extrema waiting to become rate limiting for folding, as shown experimentally and computationally for the FBP28 WW domain [4, 41]. The hPin1 WW domain is thus an apparent two-state folder, but not by a wide margin. Using a more expanded set of consensus mutants, a detailed map of the folding transition state was generated that now covers 76 % of the hPin1 sequence (previous coverage: 50 %). Many of our earlier findings are supported in the present study, but some interpretations need to be modified or revisited. Loop 2 and β strand 3, which define the C-terminal hairpin in folded hPin 1 WW, appear to be more structured in the transition state than thought previously, and the discrepancy in the backbone and side chain ΦM values within the loop 1 substructure can now be attributed to local backbone disorder in the folded protein, rather than a genuine variation in backbone and side chain structure. In fact, by assigning backbone hydrogen bond to the two residues that constitute the bond, we found good agreement between the ΦM values measured by side chain and backbone hydrogen bond perturbation for most positions. The mutants with a thermodynamically and kinetically optimized loop 1 substructure agree well with the native-like ΦM values of the highly destabilized loop 1 variants R21A/H and S16s mutants that perturb the 6-residue wild type hPin1 WW loop. Clearly, in both wild type hPin1 and the redesigned variants, the tip of the loop/turn is fully developed in the transition state. These observations and the fact that stabilizing loop 2 in the already fast folding FiP domain further speeds up folding by a factor of 3 are difficult to reconcile in a truly sequential (framework) model for folding, making a simple two-state folding mechanism more likely. Alternatively, as suggested by some simulations [35, 42] and experiments [43] of fast-folding WW domains, loop 2 could actually form before or after loop 1, or fluctuate between folded and unfolded conformations before loop 1 forms, while loop 1 remains rate-limiting due to its larger activation barrier. Additional experiments with mutations targeting loop 2 in FiP are needed to further discriminate between these alternatives.

2.8 References

[1] J. Kubelka, J. Hofrichter, W.A. Eaton, The protein folding 'speed limit', Curr. Opin. Struct. Biol. 14 (2004) 76-88.

[2] R.D. Schaeffer, A. Fersht, V. Daggett, Combining experiment and simulation in protein folding: closing the gap for small model systems, Curr. Opin. Struct. Biol. 18 (2008) 4-9.

[3] F. Cecconi, C. Guardiani, R. Livi, Testing simplified proteins models of the hPin1 WW domain, Biophys. J. 91 (2006) 694-704.

[4] H. Nguyen, M. Jäger, A. Moretto, M. Gruebele, J.W. Kelly, Tuning the free-energy landscape of a WW domain by temperature, mutation, and truncation, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 3948-3953.

[5] H.I. Chen, M. Sudol, The WW domain of Yes-associated protein binds a proline-rich ligand that differs from the consensus established for Src homology 3-binding modules, Proc. Natl. Acad. Sci. U. S. A. 92 (1995) 7819-7823.

[6] M. Jager, H. Nguyen, J.C. Crane, J.W. Kelly, M. Gruebele, The folding mechanism of a beta-sheet: The WW domain, J. Mol. Biol. 311 (2001) 373-393.

[7] M. Jäger, Y. Zhang, J. Bieschke, H. Nguyen, M. Dendle, M.E. Bowman, et al., Structure–function–folding relationship in a WW domain, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 10648-10653.

[8] S. Gianni, C. Camilloni, R. Giri, A. Toto, D. Bonetti, A. Morrone, et al., Understanding the frustration arising from the competition between function, misfolding, and aggregation in a globular protein, Proc. Natl. Acad. Sci. U. S. A. 111 (2014) 14141-14146.

[9] S. Piana, K. Sarkar, K. Lindorff-Larsen, M. Guo, M. Gruebele, D.E. Shaw, Computational Design and Experimental Testing of the Fastest-Folding β-Sheet Protein, J. Mol. Biol. 405 (2011) 43-48.

[10] M. Jäger, M. Dendle, J.W. Kelly, Sequence determinants of thermodynamic stability in a WW domain—An all-β-sheet protein, Protein Science : A Publication of the Protein Society 18 (2009) 1806-1813.

[11] S. Deechongkit, H. Nguyen, E.T. Powers, P.E. Dawson, M. Gruebele, J.W. Kelly, Nature 430 (2004) 101.

[12] J.G.B. Northey, K.L. Maxwell, A.R. Davidson, Protein Folding Kinetics Beyond the Φ Value: Using Multiple Amino Acid Substitutions to Investigate the Structure of the SH3 Domain Folding Transition State, J. Mol. Biol. 320 (2002) 389-402.

[13] K. Lindorff-Larsen, S. Piana, R.O. Dror, D.E. Shaw, How Fast-Folding Proteins Fold, Science 334 (2011) 517-520.

[14] D.E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R.O. Dror, M.P. Eastwood, et al., Atomic-Level Characterization of the Structural Dynamics of Proteins, Science 330 (2010) 341-346.

[15] S.V. Krivov, The Free Energy Landscape Analysis of Protein (FIP35) Folding Dynamics, J. Phys. Chem. B 115 (2011) 12315-12324.

[16] S. Deechongkit, P.E. Dawson, J.W. Kelly, Toward Assessing the Position-Dependent Contributions of Backbone Hydrogen Bonding to β-Sheet Folding Thermodynamics Employing Amide-to-Ester Perturbations, J. Am. Chem. Soc. 126 (2004) 16762-16771.

[17] M. Petrovich, A.L. Jonsson, N. Ferguson, V. Daggett, A.R. Fersht, phi-Analysis at the experimental limits: Mechanism of beta-hairpin formation, J. Mol. Biol. 360 (2006) 865-881.

[18] N.R. Guydosh, A.R. Fersht, A Guide to Measuring and Interpreting phi-values, Protein Folding Handbook (2005) 445-453.

[19] T.R. Weikl, Transition States in Protein Folding Kinetics: Modeling Φ-Values of Small β-Sheet Proteins, Biophys. J. 94 (2008) 929-937.

[20] M.A. De Los Rios, B.K. Muralidhara, D. Wildes, T.R. Sosnick, S. Marqusee, P. Wittung-Stafshede, et al., On the precision of experimentally determined protein folding rates and φ-values, Protein Science : A Publication of the Protein Society 15 (2006) 553-563.

[21] I. Ruczinski, K.W. Plaxco, Some recommendations for the practitioner to improve the precision of experimentally determined protein folding rates and Φ values, Proteins 74 (2009) 461-474.

[22] I. Ruczinski, T.R. Sosnick, K.W. Plaxco, Methods for the accurate estimation of confidence intervals on protein folding φ-values, Protein Sci. 15 (2006) 2257-2264.

[23] A.N. Naganathan, V. Muñoz, Insights into protein folding mechanisms from large scale analysis of mutational effects, Proceedings of the National Academy of Sciences 107 (2010) 8611-8616.

[24] P. Ferrara, A. Caflisch, Folding simulations of a three-stranded antiparallel beta-sheet peptide., Proc. Natl. Acad. Sci. 97 (2000) 10780-10785.

[25] R. Zhou, G.G. Maisuradze, D. Suñol, T. Todorovski, M.J. Macias, Y. Xiao, et al., Folding kinetics of WW domains with the united residue force field for bridging microscopic motions and experimental measurements, Proceedings of the National Academy of Sciences 111 (2014) 18243-18248.

[26] A.R. Fersht, Φ value versus ψ analysis, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 17327-17328.

[27] T. Sharpe, A.L. Jonsson, T.J. Rutherford, V. Daggett, A.R. Fersht, The role of the turn in beta-hairpin formation during WW domain folding, Protein Sci. 16 (2007) 2233-2239.

[28] S.L. Lin, A. Zarrine-Afsar, A.R. Davidson, The osmolyte trimethylamine-N-oxide stabilizes the Fyn SH3 domain without altering the structure of its folding transition state, Protein Sci. 18 (2009) 526-536.

[29] J.L. Howland. Structure and Mechanism in Protein Science. A guide to Enzyme Catalysis and Protein Folding: Alan Fersht, W.H. Freeman and Company, New York, 1999, 631 pp, ISBN 0-7167-3268-8, $53.002001.

[30] J. Ervin, M. Gruebele, Quantifying Protein Folding Transition States with Φ(T), J. Biol. Phys. 28 (2002) 115-128.

[31] J.C. Crane, E.K. Koepf, J.W. Kelly, M. Gruebele, Mapping the transition state of the WW domain β-sheet1, J. Mol. Biol. 298 (2000) 283-292.

[32] C. Tanford, Protein denaturation, Adv. Protein Chem 24 (1970) 95.

[33] A.R. Fersht, S. Sato, Φ-Value analysis and the nature of protein-folding transition states, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 7976-7981.

[34] J.A. Kowalski, K. Liu, J.W. Kelly, NMR solution structure of the isolated Apo Pin1 WW domain: Comparison to the x-ray crystal structures of Pin1, Biopolymers 63 (2002) 111-121.

[35] D.L. Ensign, V.S. Pande, The Fip35 WW Domain Folds with Structural and Mechanistic Heterogeneity in Molecular Dynamics Simulations, Biophys. J. 96 (2009) L53-L55.

[36] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, T.R. Weikl, Constructing the Equilibrium Ensemble of Folding Pathways from Short Off-Equilibrium Simulations, Proc. Nat. Acad. Sci. USA 106 (2009) 19011-19016.

[37] T.J. Lane, G.R. Bowman, K. Beauchamp, V.A. Voelz, V.S. Pande, Markov state model reveals folding and functional dynamics in ultra-long MD trajectories, J. Am. Chem. Soc. 133 (2011) 18413-18419.

[38] G.S. Hammond, A correlation of reaction rates, J. Am. Chem. Soc. 77 (1955) 334-338.

[39] G.P. Brady, K.A. Sharp, Entropy in protein folding and in protein—protein interactions, Curr. Opin. Struct. Biol. 7 (1997) 215-221.

[40] M.S. Li, D.K. Klimov, D. Thirumalai, Finite Size Effects on Thermal Denaturation of Globular Proteins, Phys. Rev. Lett. 93 (2004) 268107.

[41] G.G. Maisuradze, R. Zhou, A. Liwo, Y. Xiao, H.A. Scheraga, Effects of Mutation, Truncation, and Temperature on the Folding Kinetics of a WW Domain, J. Mol. Biol. 420 (2012) 350-365.

[42] S. a Beccara, T. Škrbić, R. Covino, P. Faccioli, Dominant folding pathways of a WW domain, Proceedings of the National Academy of Sciences 109 (2012) 2330-2335.

[43] A.J. Wirth, Y. Liu, M.B. Prigozhin, K. Schulten, M. Gruebele, Comparing Fast Pressure Jump and Temperature Jump Protein Folding Experiments and Simulations, J. Am. Chem. Soc. 137 (2015) 7152-7159.

[44] R.M. Ballew, J. Sabelko, C. Reiner, M. Gruebele, Rev. Sci. Instrum. 67 (1996) 3694.

[45] J. Ervin, J. Sabelko, M. Gruebele, J. Photochem. Photobiol., B 54 (2000) 1.

[46] W.L. DeLano. The PyMOL Molecular Graphics System

[47] M. Jager, Y. Zhang, M.E. Bowman, J.P. Noel, J.W. Kelly. 2F21: human Pin1 Fip mutant. Worldwide Protein Data Bank; 2006.

CHAPTER 3

Eliminating a protein folding intermediate by tuning a local hydrophobic contact

It is well-known that folding intermediates play an important role in protein folding process. They can be a cause of less efficient folding, and the same time may help to describe the subdomain architecture of a protein, or assist experimentalists to identify fundamental mechanistic details in protein folding by providing additional snapshots of the folding reaction. Moreover, folding intermediates on or off the main folding pathway are a common route to the formation of oligomers and amyloid fibrils, which are linked to a variety of fatal neuro-degenerative protein diseases [1–5]. Preventing the population of such intermediates, whether they lie on or off the dominant folding pathway, offers one solution to the protein related diseases.

The triple- β -stranded WW domain from the formin-binding protein 28 (FBP28) (PDB ID: 1E0L) [6] is a useful model system for studying protein re-design to eliminate intermediates. Folding of the FBP28 WW domain has been studied extensively by both experiments [3,7–14] and simulations[4,15–26]. The mechanism by which this protein folds to the native structure is sensitive to both its sequence and its solvation environment. Near its melting temperature [9], or in a denaturant [7], FBP28 (wild type) is an apparent two-state folder. Its turn 1 sequence has been used to engineer other WW domains into fast apparent two-state folders [27]. Closer to its physiological melting temperature and in the absence of a denaturant, experiments using tryptophan-fluorescence detection revealed slow concentration-independent biphasic kinetics attributed to a folding intermediate [9]. That assignment was also supported by simulations[16,17]. FBP28 readily forms fibrils under similar experimental conditions; hence, the biphasic kinetics has been attributed by Ferguson et al.[3] to an off-pathway intermediate that is a gateway for oligomer formation.

This chapter is adapted from K Dave, K Kachlishvili, M Gruebele, HA Scheraga, GG Maisuradze. Eliminating a Protein Folding Intermediate by Tuning a Local Hydrophobic Contact. Journal of Physical Chemistry B, 2016

Whether the intermediate is on- or off-pathway, truncation of the FBP28 sequence at the C terminus restored apparent two-state folding, showing how sensitive the folding mechanism of FBP28 is to amino acid sequence [9].The experimental results make it unlikely that the strand-crossing hydrophobic cluster of residues Tyr11, Tyr19, and Trp30 is associated with the intermediate [9]. Instead, two other causes have been implicated by simulations: [16,17] slower formation of turn 2 contacts relative to turn 1 (also seen in closely related FiP35 WW domains [28], and a surface-exposed local hydrophobic contact between Leu26 and Tyr21 that assists the correct registry of hairpin 2. A very general scenario consistent with all the data has been described in ref. 26: the wild type and many mutants fold through an intermediate with just one turn formed. This intermediate can become short-lived and invisible to experiment if one of the barriers separating it from the folded or unfolded state is much larger than the other [25]. Tuning solvent conditions or mutating the sequence can alter the barrier heights to reveal the intermediate or suppress it.

In recent computational work [26], two of six FBP28 mutants [Leu26Asp (PDB ID: 2n4r) and Leu26Trp (PDB ID: 2n4t)] (Fig. 3.1) folded through downhill and two-state folding scenarios in ~ 15% of folding molecular dynamics (MD) trajectories. Both hairpins in these trajectories fold through the mechanism proposed by Matheson and Scheraga [29],which is based on transient hydrophobic interactions, and considers the nucleation process as an initial aspect of folding. Thus, these mutations may restore more rapid folding mechanisms over multi-state folding. The Leu26Asp/Trp mutations alter the local Tyr21-Leu26 hydrophobic side-chain interaction and packing at the site implicated in registry of strands 2 and 3 [16,17].Here, we combine new simulations of protein backbone fluctuations over a wide temperature range with temperature jump experiments of the two mutants. We show that Leu26Asp and Leu26Trp both reduce formation of a folding intermediate at low temperature. In addition, Leu26Trp significantly speeds up folding at all temperatures, moving the system closer to downhill folding. We explain these findings in terms of hydrophobic interactions [30–32].

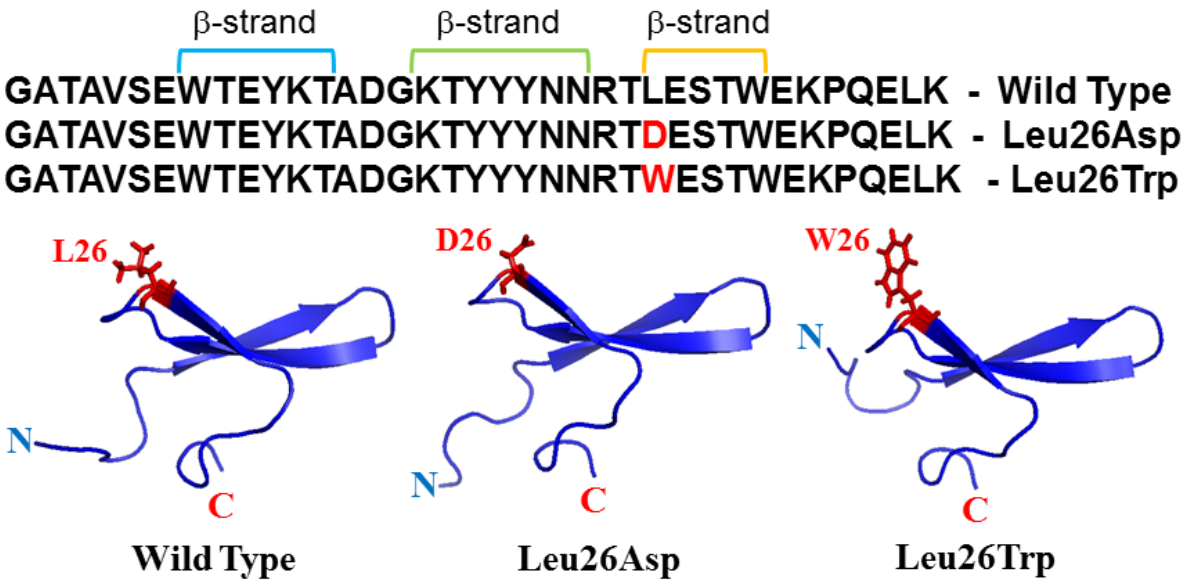

Fig. 3.1: The sequences and cartoon representation of experimental NMR structures of the wild type, Leu26Asp and Leu26Trp mutants of the FBP28 WW domain. Mutated residues are highlighted in red in the sequences, and are highlighted in red and represented in sticks in experimental NMR structures.

3.1 Methods

3.1.1 MD simulations

We performed 100-ns all-atom MD simulations of the mutants Leu26Asp, Leu26Trp and the wild type FBP28 WW domain at 250 K, 275 K, 300 K, 325 K, 350 K, 375 K, 400 K, and 425 K in explicit water [Simple Point Charge (SPC) water model] with the GROMACS package [33] using the all-atom OPLS force field [34]. The structures of Leu26Asp, Leu26Trp and the wild type were taken from the NMR model 1 of refs. 26 and 6, respectively. The coordinates were saved every 1 ps. Periodic boundary conditions were applied. The distances of 1 nm (250 – 325 K) and 1.3 nm (350 – 425 K) were assigned between the protein and the sides of the unit supercell in order to avoid any interaction between the proteins of the neighboring supercells. The temperature of the MD simulations was kept at 250K, 275K, 300K, 325K, 350K, 375K, 400K, and 425K, respectively, with a v-rescale thermostat [35], and the pressure (Parrinello-Rahman barostat)[36] was kept at 1 bar. The steepest descent algorithm

with tolerance of 100 kJ/mol/nm and maximum step size 0.01 nm was used for energy minimization. The particle-mesh-Ewald method [37,38] was used for calculating long-range electrostatic interactions, and a distance of 1.0 nm was used for the van der Waals cutoff. After the desired temperature was reached, an equilibration of 0.3 ns duration was performed with random initial conditions generated by using a random seed for the initial velocities.

3.1.2 Thermodynamic Characterization

Leu26Asp and Leu26Trp were both custom-synthesized (Genscript corp., NJ) to > 98% purity. The peptides were then dissolved in sodium phosphate buffer (pH=7.0) to a required concentration. Thermal unfolding for both Leu26Asp and Leu26Trp was measured by tryptophan fluorescence and circular dichroism. The Leu26Trp mutant contains an extra tryptophan residue compared to wild type, and fluoresces more strongly than the Leu26Asp mutant. Both types of measurements were carried out with 10 μM protein dissolved in 10 mM sodium phosphate buffer (pH=7.0). Fluorescence spectroscopy was carried out using a Cary Eclipse fluorescence spectrophotometer equipped with programmable temperature control (Varian) with excitation and emission slit widths kept at 5 nm. Tryptophan was excited at 280 nm, and emission was collected from 290-450 nm. Circular dichroism was measured using a JASCO spectrophotometer with Peltier temperature control (JASCO Inc, Easton MD). All spectra were recorded from 200-250 nm at a scan rate of 50 nm/min with 1 nm resolution and are an average of 5-10 accumulations. Measurements were conducted in a 2 mm path length quartz cuvette. Thermodynamic denaturation signals $S(T)$, obtained by fluorescence and by far-UV CD spectroscopy, were fitted to a two-state model in equation (1a,b) to obtain the denaturation midpoints with respect to temperature (T_m). S_U and S_F are unfolded and folded baseline and $\Delta G(T)$ is the free energy change.

$$S(T) = S_U + S_F e^{-\Delta G(T)/RT} / (1 + e^{-\Delta G(T)/RT}) \quad (1a)$$

$$\Delta G(T) = g_X (T - T_m) \quad (1b)$$

3.1.3 Kinetics Experiments

Laser temperature jumps were carried out using a Surelite Q-switched Nd:YAG laser (Continuum Inc., Santa Clara, CA), with details of the instrument mentioned elsewhere

[39,40].The jump size was 7-8 °C. The exact size of the jump was calibrated by comparing the fluorescence decays *f* of tryptophan (300 μM solution) after the jump with the corresponding decay at an equilibrium temperature several degrees higher. Fluorescence decays were excited at 280 nm by a tripled, mode-locked Ti:sapphire laser every 12.5 ns for a total of 1 ms. The temperature jump was set to occur 153.75 μs after the oscilloscope was triggered to start data collection. The sampling frequency was 10 Giga-samples per second. Thus each fluorescence decay was sampled at 100 picosecond intervals, or 125 times before the next decay was excited. The signal was usually 50-60 mV. Sample concentrations were 100 μM for both mutants as measured by the absorption signal at 280 nm.

3.1.4 Kinetics Data Analysis

Kinetics data were analyzed using MATLAB (Mathworks Inc., Natick, MA) and IGOR Pro (Wavemetrics Inc., Lake Oswego, OR). A fluorescence decay f(t) was collected every 12.5 ns. 100 of these were binned into intervals of 1.25 μs. Thus the protein kinetics could be followed with 1.25 μs time resolution. The decays f(t) were fitted to a linear combination of the decay f1 averaged between 153.75 and 28.75 μs before the T-jump, and the decay f2 averaged over the final 125 μs of data collection, where the protein had equilibrated. The relative lifetime shift as a function of time, χ(t), was then obtained (see Results for definition of χ). The χ(t) traces were fitted to a double or single exponential function starting at t=0, where the T-jump occurred (see Results).

3.2 Results

3.2.1 Structural fluctuations of Leu26Asp and Leu26 Trp mutants vs. temperature

Concerted or sequential formation of two hairpins determines the folding mechanism. If formation of hairpin 1 is assisted by global hydrophobic collapse (e.g. by the core at Tyr 11/Tyr19/Trp30), whereas formation of hairpin 2 is delayed by comparison, this results in an intermediate state, making the protein fold through a three-state scenario [26]. If formation of hairpin 2 is assisted by the Matheson-Scheraga mechanism, then temperature may have a strong effect on the relative propensities and kinetics of hairpins 1 and 2 forming. In the wild type, two phases appear very pronounced in the relaxation kinetics at low temperature, but not at high temperature [9].Therefore, it is of interest to know whether the change of temperature

plays a significant role in the formation of the intermediate state of the Leu26Asp and Leu26Trp mutants of the FBP28 WW domain.

To answer this question, we investigated the backbone fluctuations of native FBP28 at eight temperatures from 250 to 425 K, by performing all-atom MD simulations for Leu26Asp, Leu26Trp, and the wild type. A detailed analysis of the two-dimensional free-energy landscapes (FELs) along the dihedral angles ϕ and ψ of each residue, and of the contributions of the principal modes to the mean-square-fluctuations (MSF) along the angles ϕ and ψ was conducted.

Inspection of the MD trajectories showed that structural fluctuations of all proteins increase with temperature; however, all systems remain mainly in their native states except for very high temperatures (400, 425 K). It has been shown that the dynamics in the native state are controlled by the same energy landscape that guides the entire folding process [41]. Hence, it is of interest to investigate how the dynamics of the backbone change with an increase of temperature in the native state, and whether these changes determine the folding scenario of the system.

We exclude the < 275 K and > 375 K temperatures from discussion in the main text because they cannot be reached in our experimental analysis [see Fig. B.1 and Fig. B.2 in the Appendix B for full simulations and additional results].

3.2.2 Reducing backbone motion to a few collective modes

The structural mean-square-fluctuations of the dihedral angles ϕ and ψ can be decomposed into collective (principal) modes by using dihedral principal component analysis (dPCA) [42–44]. The dPCA facilitates a projection of the dihedral-angle coordinates of a protein onto a few relevant coordinates along which the FELs and the collective modes of the protein can be analyzed. These modes have "frequencies" and directions corresponding to the eigenvalues and eigenvectors of the dPCA covariance matrix [22,44,45]. The projection of the trajectory on the eigenvector is named the principal component. The modes with the largest eigenvalues λ_k (named slow modes) contribute the most to the structural fluctuations of the protein. The contribution of the i^{th} dihedral angles ϕ and ψ to a mode k is the so-called influence ν_i^k, and the mean-square-fluctuation at residue i is given by [22,44,45]

$$MSF_i = \sum_k \lambda_k \nu_i^k. \quad (2)$$

Fig. 3.2 illustrates the percentages of the total fluctuations captured by the principal modes derived from dPCA for Leu26Asp (panel A), Leu26Trp (panel B), and the wild type (panel C) at five different temperatures (see Fig. B.1 for the full temperature range). We list only the first several modes, the sum of which captures ~ 50% of the total fluctuations. It is well established that, if the principal modes are able to capture ≥ 40% of the total fluctuations, the FELs constructed along those principal components can describe the folding dynamics correctly [20]. The percentage of the total fluctuations captured by principal modes changes with temperature. For example, in the MD trajectories of Leu26Asp, ~ 40% of the total fluctuations can be captured by the first two modes at 275 K, but the first seven modes are required at 375 K. We have also calculated the contributions of the first k principal modes (k is the number of modes capturing at least 40% of the total fluctuations) to the MSF_i along the angles ϕ and ψ at five different temperatures [Fig. 3.2, right side (see Fig. B.1 for the full temperature range)]. Based on our earlier results, [20,26] Fig. 3.2 enabled us to determine the folding scenario of each system at any particular temperature. Most of the residues in the MD trajectories of Leu26Asp at low temperatures from 275 K to 325 K move in a concerted fashion. Contributions to the fluctuations in that temperature range are almost identical: in addition to the termini, they are localized at the second turn – the main factor in the emergence of the intermediate state [4][16–18,20,22][24–26].This localization indicates that Leu26Asp is a three-state folder at lower temperature, with turn 2 unraveling first. The result is different for Leu26Asp at high temperatures. At 350 and 375 K, mainly the termini contribute to fluctuations. Since there is never a dominant contribution from the second turn alone, Leu26Asp can fold through either downhill or two-state folding scenarios at these temperatures (Fig 3.2).

In contrast, Leu26Trp exhibits downhill or two-state folding scenarios in the MD trajectories even from 275 K to 350 K (Fig. 3.2). At very high temperatures (> 375 K), contributions to the fluctuations in the trajectory come from not only the N- and C-termini, but also from the first and second turns, and eventually the first and third β-strands, which indicates onset of multi-state folding outside the experimental temperature range (Fig. B.1). Similar multi-state folding scenario is observed for Leu26Asp at very high temperatures, 400 and 425 K. It should be noted that the wild type folds through either downhill or two-state folding scenario at very low temperatures (250 - 275 K), and changes to three-state folding at the lowest experimentally reachable temperatures, which is in agreement with our earlier experiments [9].

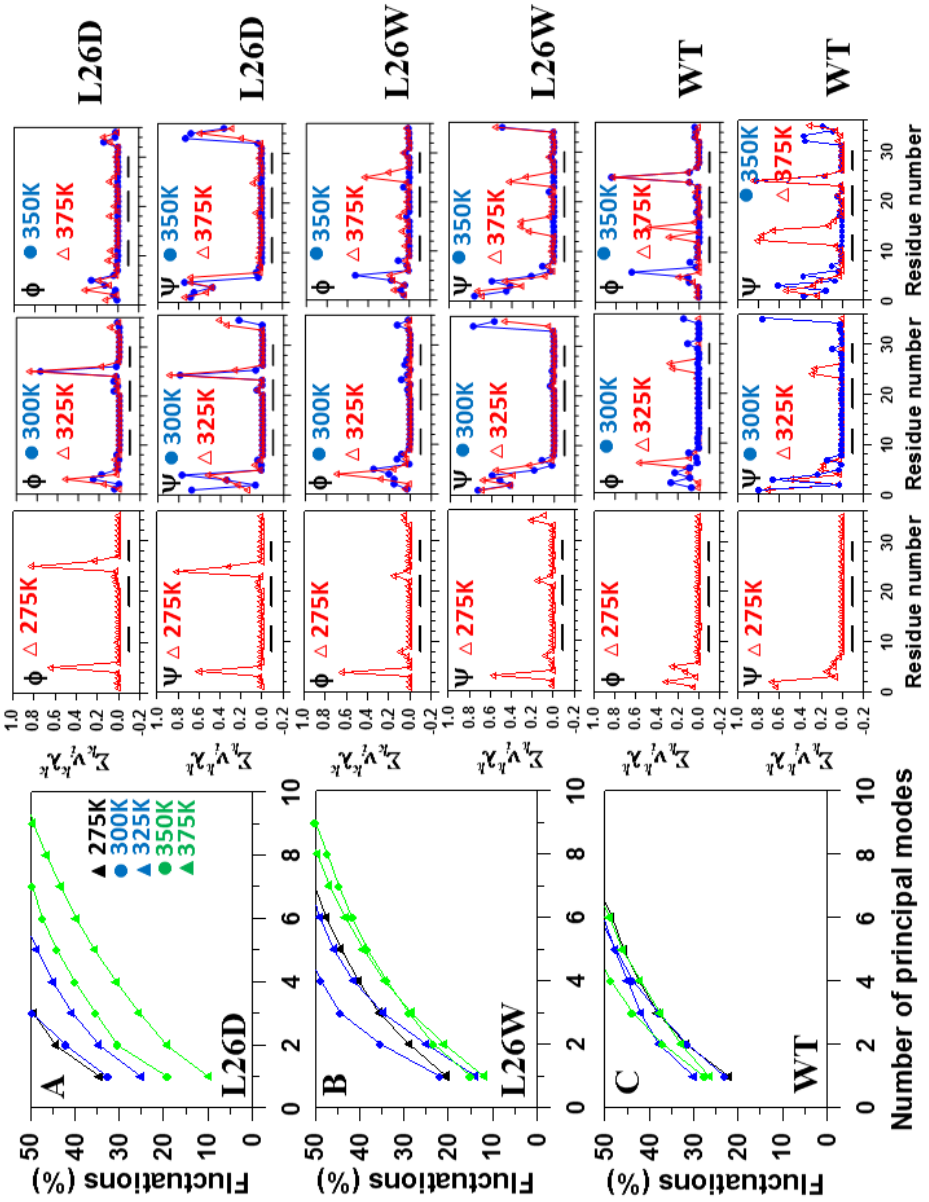

Fig. 3.2: Percentages of the total fluctuations captured by the principal components for Leu26Asp (panel A), Leu26Trp (panel B) and the wild type (panel C) at five different temperatures. The panels on the right represent contributions of the first *k* collective modes (*k* is the number of modes capturing at least 40% of the total fluctuations) to the MSF along the angles ϕ and ψ at five different temperatures (275 K, 300 K, 325 K, 350 K, and 375 K) for Leu26Asp, Leu26Trp and the wild type. The black bars above each x-axis label the β-strand locations.

3.2.3 Pinpointing the origin of the change in the folding mechanism

In order to support the three-state folding of Leu26Asp at low temperature, and the two-state or downhill folding of Leu26Trp at all experimentally reachable temperatures, we examined the backbone fluctuations of each system by building two-dimensional FELs along the dihedral angles ϕ_i and ψ_i of each residue (Fig. B.2).

The change in backbone dynamics with increasing temperature is similar for all mutants and wild type, except for a few residues (discussed below) that play a vital role in determining the folding scenario. As expected, in all proteins the amplitudes of the fluctuations of the ends (ϕ_i, ψ_i, i = 1 – 7, 31 – 36) gradually become larger with increase of temperature, which is manifested in the larger regions explored by these residues (the larger amplitudes indicating cold denaturation at 250 K being the exception). These results show that the N-terminal region is even less stable than the C-terminal region. The turn 1 residues (ϕ_i, ψ_i, i = 14 - 16) do not respond significantly to the increase of temperature up to 375 K, whereas ϕ_{25}, ψ_{25} in turn 2 are influenced by the temperature change. The fluctuations of the threonine 25 residue are strongly correlated with those of its neighboring 26th residue, which belongs to the third β-strand. All β-strands retain stability almost entirely until $T \geq 425$ K. Only some edges between turns and β-strands (ϕ_i, ψ_i, i = 12, 13, 17, 30) exhibit instability at higher experimentally unreachable temperatures. The exception is residue 26, which is the most "sensitive" residue to the temperature change among the residues pertaining to the β strands; however, it reacts differently to the temperature change in each protein. In particular, the FEL along the angles ϕ_{26} and ψ_{26} of the Leu26Asp mutant exhibits multiple minima starting from 275 K, but not close to T_m (350 K), where the number of minima reduces to one main deep minimum and one shallow minimum (Fig. B.2A). Thus, Leu26Asp locally recapitulates a multi-state to two-state transition as temperature is increased.

In contrast, the FELs along angles φ_{26} and ψ_{26} of the Leu26Trp mutant exhibit one deep minimum from 250 K to 350 K (close to T_m). Leu26Trp becomes unstable at higher (≥ 375 K) temperatures (Fig. B.2B). Thus, Leu26Trp is even closer to two-state or downhill folding than Leu26Asp.

Since none of the other residues pertaining to the β-strands or turns (except for threonine 25, which is correlated with aspartate 26 in Leu26Asp) are affected by the increase of temperature up to 350 K, we can conclude that residues threonine 25 and aspartate 26 are the "key players" that determine the folding scenario. We thus predict that the Leu26Asp mutant

shows some three-state behavior at low temperature, but rapidly switches to two-state folding at higher temperature, whereas the Leu26Trp mutant is a two-state or downhill folder over the experimentally reachable temperature range. The wild type is more of a three-state folder or multi-state folder than either of these mutants [except for the very low temperature region (250 – 300 K), in which it can fold through either a downhill or two-state folding scenario (Fig. B.2C)]. I next consider experimental data to test this prediction.

3.2.4 Experimental thermal melts

Differential scanning calorimetry was previously measured in ref. 26 to obtain information about the changes in heat capacity during the unfolding process. Here, thermal melts at different denaturant concentrations were collected for the Leu26Asp and Leu26Trp mutants of the FBP WW domain. Two different probes were used see Fig. 3.3. The tryptophan fluorescence emission was monitored by exciting the protein at 280 nm, providing information about the local environment around the fluorophore. We report the average wavelength <λ>, where I is intensity, λ is wavelength, and j indexes the wavelengths in the range of 300 – 450 nm,

$$\langle\lambda\rangle = (\sum_j \lambda_j I_j) \,/\, (\sum_j I_j). \tag{3}$$

Circular dichroism at 227 nm was used as a global probe to monitor secondary structure changes during protein denaturation. Temperature *vs.* wavelength measurements at 0 to 3 M GuHCl concentrations were performed to obtain more accurate melting temperatures. The full data set of thermal/GuHCl denaturation data was fitted globally for each mutant (Fig. B.3).

Leu26Trp is consistently less stable than Leu26Asp by all probes (Table 3.1). Different probes reveal different melting temperatures for the same protein, suggesting that it is not an ideal two state folder (Table 3.1). Different probes overlap only at high temperature (Fig. B.3). This observation is consistent with downhill folding[46,47] or an intermediate state below the melting transition, and with two-state folding above the melting transition. The thermal melts are reversible for both Leu26Asp and Leu26TrP plot shown in Appendix B (Fig. B.4).

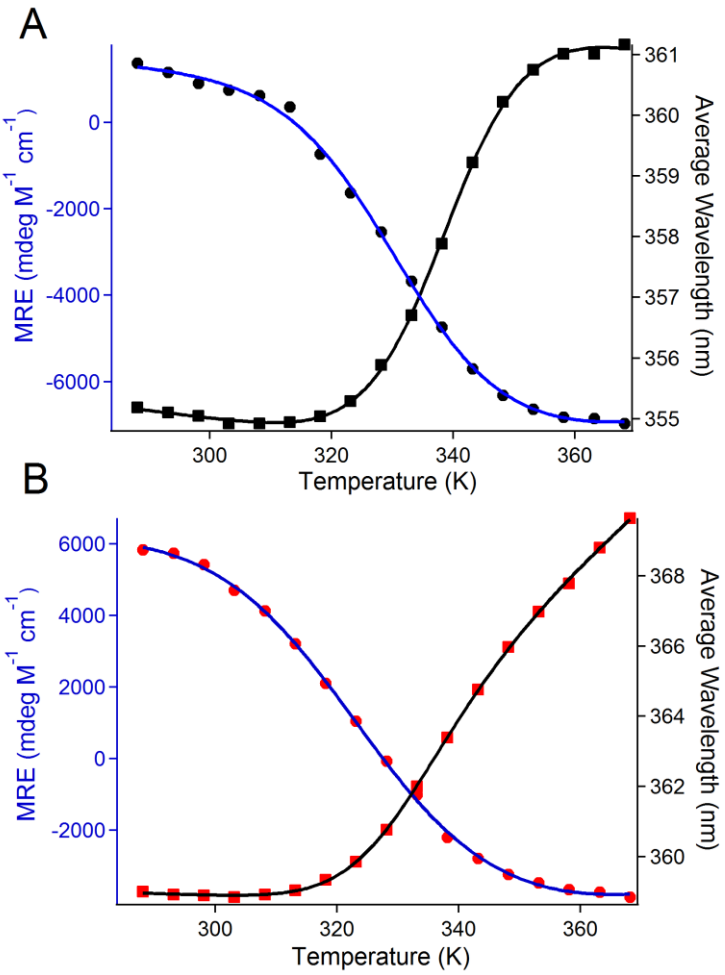

Fig. 3.3: Thermal denaturation of Leu26Asp (A) and Leu26Trp (B) monitored by using circular dichroism at 227 nm (blue left axis) and by the average wavelength of tryptophan fluorescence emission (right black axis). The thermodynamic fits are shown as smooth curves, with T_m in Table 3.1 and plots of calculated fraction folded *vs.* temperature for both mutants in the Appendix B (Fig. B.5).

Table 3.1: Two state thermodynamic fitting results for Leu26Asp and Leu26Trp using different probes. Data are shown in Fig. 3.3 and Fig. 3.4 for Leu26Asp and for Leu26Trp. One standard deviation uncertainties are shown in parentheses when available

Protein	T_m, K (Fluorescence wavelength shift)	T_m, K (Circular Dichroism)	T_m, K (DSC, ref. 26)	T_m Standard Deviation, K
Leu26Asp	339(1)	329(1)	334	5
Leu26Trp	331(2)	324(1)	328	3.5

3.2.5 Experimental kinetics data

Temperature jump experiments were performed at three different final temperatures at or below the T_m to obtain the relaxation kinetics (313 to 333 K for Leu26Asp, 313 to 325 K for less stable Leu26Trp). We measured a tryptophan fluorescence decay profile $f(t)$ every 1.25 microseconds. Upon T-jump, the profile $f(t)$ changes shape as the protein equilibrates towards more unfolded state. We fitted $f(t)$ to a linear combination of the fluorescence decay before the T-jump (f_1) and after equilibration (f_2), or $f(t)=\chi(t)f_1+[1-\chi(t)]f_2$ [39,40]. The fitted value of $\chi(t)$ tracks the change in fluorescence decay lifetime as the protein equilibrates after the T-jump. The results and least-squares fits to single or double exponential functions

$$\chi(t) = A_0 + A_1 e^{t/\tau_1} + A_2 e^{t/\tau_2} \quad (4)$$

are shown in Fig. 3.4 and Table 3.2. About 40% of the fluorescence lifetime change ($A_1+A_2\approx0.4$) is resolved; the rest occurs in <1 μs and is attributed to the intrinsic dependence of the tryptophan lifetime on temperature [39].

At low temperature, the Leu26Asp mutant has a small but significant slow phase ($A_2/(A_1+A_2)\approx17\%$) of 130 μs, in addition to a fast phase of 20 μs (Fig. 3.4 and Table 3.2). Both of these are considerably slower than the measured speed-limit of the WW domain ≈2 μs [48–50].Thus, we attribute the kinetics at low temperature to three- (or multi-state) folding, although the slow phase is not as large as was measured for the wild type [9]. At higher temperature, the slow phase vanishes, and the faster phase speeds up to ≈8 μs (Fig. 3.4 and Table 3.2), not far from the measured folding speed limit for the FiP35 WW domain. Thus, we attribute the folding of Leu26Asp to fast two state folding, approaching downhill folding.

The Leu26Trp mutant has a smaller slow phase even at low temperature ($A_2/(A_1+A_2)$ < 10%) (Table 3.2). The slow phase also disappears at high temperature, where this mutant folds in ≈7 μs (Fig. 3.4 and Table 3.2). Thus Leu26Trp is closer to two-state folding or downhill folding than Leu26Asp. Both mutants fold at least twice as fast as the wild type, whose fastest phase does not drop below 14 μs at its melting temperature [9].

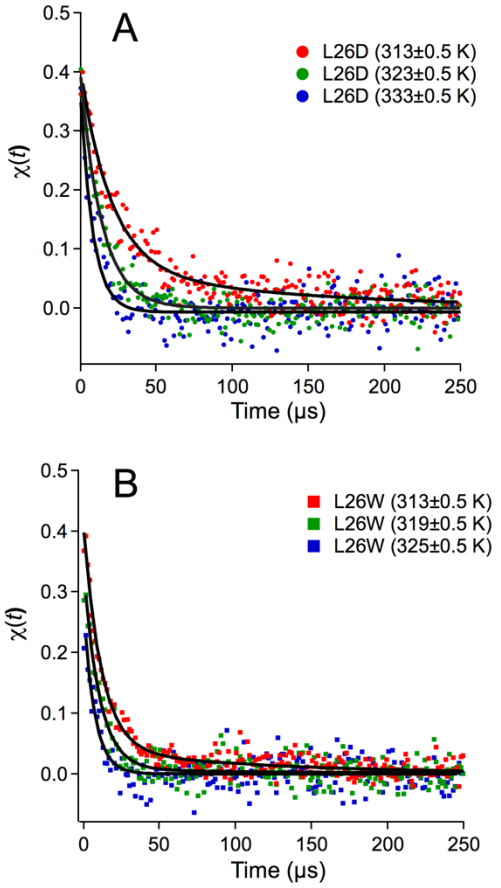

Fig. 3.4: (A) Relaxation kinetics of the Leu26Asp mutant of FBP28 at three different temperatures shown in the Figure legend. The black traces correspond to the single or double exponential fits of the data as shown in Table 3.2. The relative amplitude of the slower phase is negligible at > 313 K, and much smaller than for the wild type in ref. 9 at 313 K. (B) Analogous data for the Leu26Trp mutant (at lower temperatures due to its reduced stability). This mutant folds faster than wild type or Leu26Asp and has an even smaller slow phase in Table 3.2.

Table 3.2: Single or double exponential fits to the data in Fig. 3.4. One standard deviation uncertainties are given in parentheses.

Protein, T	A_1	τ_1 (μs)	A_2	τ_2 (μs)
Leu26Asp, 313 K	0.34 (0.02)	20 (2)	0.07 (0.01)	130 (37)
Leu26Asp, 323 K	0.40 (0.01)	13.9 (0.5)	-	-
Leu26Asp, 333 K	0.39 (0.03)	7.6 (0.9)	-	-
Leu26Trp, 313 K	0.37 (0.01)	11.7 (0.6)	0.04 (0.01)	113 (17)
Leu26Trp, 319 K	0.34 (0.02)	11.4 (0.7)	-	-
Leu26Trp, 325 K	0.30 (0.01)	7.3 (0.8)	-	-

3.3 Discussion and conclusion

Based on our simulations, we make two general predictions: (i) The Leu26 and Thr25 positions are critical to the folding mechanism of the FBP28 WW domain because they alter a surface-exposed local hydrophobic contact that forms hairpin 2, as predicted in refs. 16, 26. (ii) The Leu26Asp and Leu26Trp mutants affect this interaction differently and differ from the wild type. The Leu26Asp mutant is a three-state folder at low temperature because of the slow correct formation of turn 2, but becomes two-state or downhill folder at higher temperature. The Leu26Trp mutant is a two-state or downhill folder almost over the whole experimentally reachable temperature range. Both are closer to two-state folding than the wild type.

Our experimental data largely validates this prediction: The Leu26Asp mutant has the largest standard deviation of T_m by three different probes (Table 3.1), and the largest slow phase amplitude at low temperature, but only a single exponential phase at high temperature. This observation is consistent with a transition from three-state folding to two-state or downhill folding as the temperature is raised. The Leu26Trp mutant has a smaller standard deviation of T_m, a smaller slow phase at low temperature, and folds with a faster single exponential phase than Leu26Asp at high temperature. This observation is consistent with Leu26Trp being closer to the two-state/downhill limit than Leu26Asp.

Thus, we propose that both mutants undergo a transition from three-state towards fast two-state folding (approaching downhill folding) at higher temperature, but the Leu26Trp mutant is almost two-state already even at low temperature. This is exactly the trend predicted from simulation and dPCA analysis, although evidence of three-state folding of Leu26Trp at low temperature cannot be discerned in the simulations. The downhill folding time of ≈3 μs/0.5 μs calculated for FBP28 Leu26Asp/Trp in ref. 26 (after adjustment for coarse-graining in UNRES [51,52] MD simulations) differs from the fast experimental phase of 7 μs observed here. This is not surprising especially for coarse-grained force fields, in which averaging out the fast motions of the secondary degrees of freedom, at the coarse-grained level, makes the free-energy barriers lower than those at the atomic level. Thus, it appears that the simulations overestimate how close Leu26Trp already is to fast two-state/downhill folding, but correctly predict the change in mechanism going from Leu26Asp to Leu26Trp with increasing temperature. Such agreement shows that well-calibrated modern force fields such as Optimized Potentials for Liquid Simulations (OPLS) [34] can give insight into mechanistic details of folding, not just whether a protein folds to a certain native structure or not. This was also shown for FiP35 and its GTT triple mutant using the CHARMM22* force field [53] (FiP35 is a 35 residue, engineered WW domain that combines human Pin1 WW domain with the shorter loop 1 of FBP WW domain, so FiP = "FBP in Pin." The FiP35 mutant "GTT" contains mutations N26G, A27T, and S28T, including position 26, which is important based on the simulations presented here.)

Finally, ref. 22 shows that three-state folding can be partly 'hidden' when the barriers connecting the intermediate to other states are asymmetrical, resulting in a single experimental activated time scale. It is worth noting that this effect may contribute to the rapid smoothing of folding kinetics to a single time scale (Table 3.2) when the temperature is raised. We tested this possibility further with a quantitative investigation of the Leu26Asp mutant residence times in the intermediate state from MD simulation. We find that the time, spent in the intermediate state by Leu26Asp, oscillates within a 23 – 34 ns range at low temperatures (except for 250 K), but then drops down to ~ 5 ns, and then starts slowly increasing again with increasing temperature (Fig. 3.5). The simulations thus are consistent with higher free energy landscape roughness (intermediates) at low temperature, less roughness (downhill folding) at higher temperature, and again increased roughness at the highest temperature, where an intermediate may be hidden due to a large activation energy differences leading from either the

native or unfolded states to the intermediate. (This, of course, assumes a native-intermediate-unfolded topology of the landscape, which the present experiments cannot prove or disprove.)

The mutants and wild type have similar patterns of dynamics in terms of backbone fluctuations. The main difference was that two key residues lost their resistance to fluctuations at a temperature much below T_m in the wild type, whereas they retained stability almost over the whole experimentally reachable temperature range in Leu26Trp. (The key residues of Leu26Asp also lose their resistance to fluctuations at a temperature << T_m; however, they regain the stability close to T_m.) In the particular case of the Leu26Asp mutation, removing a local hydrophobic interaction with tyrosine 21 [16] seems to be the key. The key residues forming the intermediate state were identified as 25 and 26 by MD simulations of folded state fluctuations in the 275 to 375 K range (see Figures B.1 and B.2 for the full temperature range).

These findings can be corroborated by the results obtained from NMR experiments:[26] In Leu26Asp, the aspartic acid 26 side chain is consistently oriented toward the tyrosine 21 hydroxyl, which suggests the presence of a water-mediated hydrogen bond that stabilizes that specific orientation, which may allow some "flexibility" during the correct formation of turn 2. In other words, it may either speed up (two-state or downhill folding) or slow down (three-state folding) the correct formation of turn 2 in contrast to the wild type, in which surface-exposed hydrophobic contact enforces the slow correct formation of turn 2.

Our results indicate that the speed of correct formation of turn 2 depends on temperature. In particular, a water-mediated hydrogen bond is strong at low temperatures and plays an important role in slowing down the formation of turn 2. It weakens with the increase of temperature, while hydrophobic interactions between Tyr20 and Pro33, and Tyr19 and Trp30 become stronger,[30] and enforce the fast correct formation of turn 2. For the other mutant, Leu26Trp, the interaction between Trp26 and Tyr21 seems to play a crucial role in fast formation of turn 2. The point is that hydrophobic interactions between aromatic residues contribute substantially to protein stability [31].Aromatic-aromatic interactions are stronger than those between aliphatic and aromatic residues at all temperatures (until the protein starts unfolding),[30–32] hence, enforcing the fast formation of hairpin 2 almost over the whole experimentally reachable temperature range.

In the end, the flexibility of leucine 26,[54] compared to tryptophan, may be the main reason for slow correct formation of turn 2 in wild type, explaining the three-state folding scenario of wild type at low temperatures, observed here and previously.

In this study, by performing T-jump experiments, we have experimentally validated theoretical findings (this work and ref. 26) that a mutant of the FBP28 WW domain, Leu26Asp,

can reduce the intermediate state population at lower temperature relative to the wild type, and eliminate it entirely at high temperature. Another mutant, Leu26Trp, reduces the intermediate population even more at low temperature. Protein folding intermediates are associated with formation of amyloid fibrils, which are responsible for a number of degenerative protein-related disorders. Based on our results, it is possible to re-design proteins with very few mutations (even just a single mutation) to avoid folding intermediates. The extensive truncation of the N- and C- termini done in ref. 9 to reduce three-state folding is not necessary to approach two-state folding. A single carefully chosen residue can have a similar effect. However, the possibility of kinetically hidden intermediates should always be kept in mind when a mechanism apparently changes from three- to two-state folding. However, it should be noted that, the recent studies on other domains showed the similar results. In particular, investigations of folding mechanisms of a fluorescent variant of PDZ2 from PTP-BL [55] and the measles virus X domain [56] revealed that folding can be tuned from a three-state to a two-state under stabilizing conditions (e.g. in the presence of sodium sulfate) and by mutation, respectively.

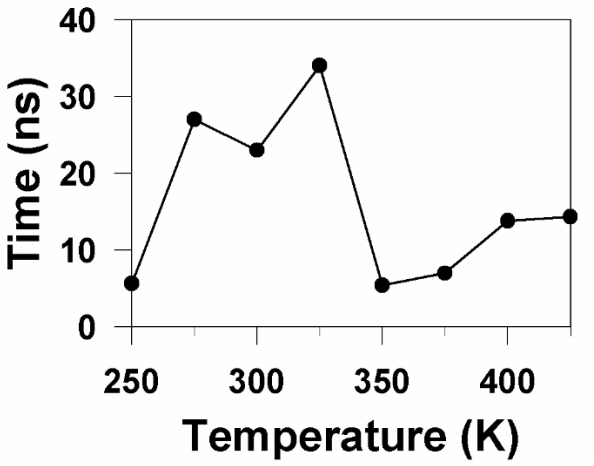

Fig. 3.5: Time spent in the intermediate state, *vs.* temperature, for Leu26Asp.

3.4 References

[1] J. Guijarro, M. Sunde, J. Jones, Amyloid fibril formation by an SH3 domain, Natl. Acad. …. (1998). http://www.pnas.org/content/95/8/4224.short (accessed May 4, 2017).

[2] M. Ramírez-Alvarado, J. Merkel, A systematic exploration of the influence of the protein stability on amyloid fibril formation in vitro, Proc. (2000). http://www.pnas.org/content/97/16/8979.short (accessed May 4, 2017).

[3] N. Ferguson, J. Berriman, M. Petrovich, Rapid amyloid fiber formation from the fast-folding WW domain FBP28, Proc. (2003). http://www.pnas.org/content/100/17/9814.short (accessed May 4, 2017).

[4] Y. Mu, L. Nordenskiöld, J. Tam, Folding, misfolding, and amyloid protofibril formation of WW domain FBP28, Biophys. J. (2006). http://www.sciencedirect.com/science/article/pii/S0006349506725802 (accessed May 4, 2017).

[5] P. Neudecker, P. Robustelli, A. Cavalli, P. Walsh, Structure of an intermediate state in protein folding and aggregation, (2012). http://science.sciencemag.org/content/336/6079/362.short (accessed May 4, 2017).

[6] M. Macias, V. Gervais, C. Civera, Structural analysis of WW domains and design of a WW prototype, Nat. Struct. (2000). http://www.nature.com/nsmb/journal/v7/n5/abs/nsb0500_375.html (accessed May 4, 2017).

[7] M. Jäger, H. Nguyen, J.C. Crane, J.W. Kelly, M. Gruebele, The folding mechanism of a beta-sheet: the WW domain., J. Mol. Biol. 311 (2001) 373–393. doi:10.1006/jmbi.2001.4873.

[8] N. Ferguson, C. Johnson, M. Macias, Ultrafast folding of WW domains without structured aromatic clusters in the denatured state, Proc. (2001). http://www.pnas.org/content/98/23/13002.short (accessed May 4, 2017).

[9] H. Nguyen, M. Jäger, A. Moretto, M. Gruebele, J.W. Kelly, Tuning the free-energy landscape of a WW domain by temperature, mutation, and truncation, Proc. Nat. Acad. Sci. USA. 100 (2003) 3948–3953.

[10] N. Ferguson, J. Becker, H. Tidow, General structural motifs of amyloid protofilaments, Proc. (2006). http://www.pnas.org/content/103/44/16248.short (accessed May 4, 2017).

[11] F. Liu, D. Du, A.A. Fuller, J.E. Davoren, P. Wipf, J.W. Kelly, M. Gruebele, An experimental survey of the transition between two-state and downhill protein folding scenarios, Proc. Nat. Acad. Sci. USA. 105 (2008) 2369–2374.

[12] M. Jager, S. Deechongkit, E. Koepf, H. Nguyen, Understanding the mechanism of β-sheet folding from a chemical and biological perspective, Peptide. (2008). http://onlinelibrary.wiley.com/doi/10.1002/bip.21101/full (accessed May 4, 2017).

[13] C. Davis, R. Dyer, WW domain folding complexity revealed by infrared spectroscopy, Biochemistry. (2014). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4151701/ (accessed May 4, 2017).

[14] M. Petrovich, A. Jonsson, N. Ferguson, V. Daggett, Φ-analysis at the experimental limits: mechanism of β-hairpin formation, J. Mol. (2006). http://www.sciencedirect.com/science/article/pii/S0022283606006516 (accessed May 4, 2017).

[15] N. Ferguson, J. Pires, F. Toepert, Using flexible loop mimetics to extend Φ-value analysis to secondary structure interactions, Proc. (2001). http://www.pnas.org/content/98/23/13008.short (accessed May 4, 2017).

[16] J. Karanicolas, C. Brooks, The structural basis for biphasic kinetics in the folding of the WW domain from a formin-binding protein: Lessons for protein design?, Proc. Natl. (2003). http://www.pnas.org/content/100/7/3954.short (accessed May 4, 2017).

[17] J. Karanicolas, C. Brooks, Integrating folding kinetics and protein function: Biphasic kinetics and dual binding specificity in a WW domain, Sci. United States …. (2004). http://www.pnas.org/content/101/10/3432.short (accessed May 4, 2017).

[18] G. Maisuradze, A. Liwo, H. Scheraga, Principal component analysis for protein folding dynamics, J. Mol. Biol. (2009). http://www.sciencedirect.com/science/article/pii/S0022283608012886 (accessed May 4, 2017).

[19] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, T.R. Weikl, Constructing the

equilibrium ensemble of folding pathways from short off-equilibrium simulations, Proc. Nat. Acad. Sci. USA. 106 (2009) 19011–19016. doi:10.1073/pnas.0905466106.

[20] G. Maisuradze, A. Liwo, Relation between free energy landscapes of proteins and dynamics, J. Chem. Theory. (2010). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3633568/ (accessed May 4, 2017).

[21] S. Piana, K. Sarkar, K. Lindorff-Larsen, M. Guo, M. Gruebele, D.E. Shaw, Computational design and experimental testing of the fastest-folding beta-sheet protein, J. Mol. Biol. 405 (2011) 43–48.

[22] G. Maisuradze, R. Zhou, A. Liwo, Y. Xiao, Effects of mutation, truncation, and temperature on the folding kinetics of a WW domain, J. Mol. (2012). http://www.sciencedirect.com/science/article/pii/S0022283612003695 (accessed May 4, 2017).

[23] Dominant folding pathways of a WW domain, Natl. Acad. …. (2012). http://www.pnas.org/content/109/7/2330.short (accessed May 4, 2017).

[24] G. Maisuradze, A. Liwo, P. Senet, Local vs global motions in protein folding, Theory Comput. (2013). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3727290/ (accessed May 4, 2017).

[25] R. Zhou, G. Maisuradze, D. Suñol, Folding kinetics of WW domains with the united residue force field for bridging microscopic motions and experimental measurements, Proc. (2014). http://www.pnas.org/content/111/51/18243.short (accessed May 4, 2017).

[26] G. Maisuradze, J. Medina, Preventing fibril formation of a protein by selective mutation, Proc. (2015). http://www.pnas.org/content/112/44/13549.short (accessed May 4, 2017).

[27] H. Nguyen, M. Jäger, J. Kelly, Engineering a β-sheet protein toward the folding speed limit, J. Phys. (2005). http://pubs.acs.org/doi/abs/10.1021/jp052373y (accessed May 4, 2017).

[28] A.J. Wirth, Y. Liu, M.B. Prigozhin, K. Schulten, M. Gruebele, Comparing Fast Pressure Jump and Temperature Jump Protein Folding Experiments and Simulations, J. Am. Chem. Soc. (2015).

[29] R.M. Jr, H. Scheraga, A method for predicting nucleation sites for protein folding based on hydrophobic contacts, Macromolecules. (1978). http://pubs.acs.org/doi/pdf/10.1021/ma60064a038 (accessed May 4, 2017).

[30] G. Némethy, H. Scheraga, The structure of water and hydrophobic bonding in proteins. iii. The thermodynamic properties of hydrophobic bonds in proteins1, 2, J. Phys. Chem. (1962). http://pubs.acs.org/doi/pdf/10.1021/j100816a004 (accessed May 4, 2017).

[31] S. Burley, G. Petsko, Aromatic-aromatic interaction: a mechanism of protein structure stabilization, Science (80-.). (1985). http://go.galegroup.com/ps/i.do?p=AONE&sw=w&issn=00368075&v=2.1&it=r&id=GALE%7CA3847736&sid=googleScholar&linkaccess=fulltext (accessed May 4, 2017).

[32] A. de Araujo, T. Pochapsky, B. Joughin, Thermodynamics of interactions between amino acid side chains: experimental differentiation of aromatic-aromatic, aromatic-aliphatic, and aliphatic-aliphatic, Biophys. J. (1999). http://www.sciencedirect.com/science/article/pii/S0006349599773893 (accessed May 4, 2017).

[33] S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit, (2013). http://bioinformatics.oxfordjournals.org/content/early/2013/02/21/bioinformatics.btt055.short (accessed May 4, 2017).

[34] W. Jorgensen, D. Maxwell, Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids, J. Am. (1996). https://www.researchgate.net/profile/Julian_Tirado-Rives/publication/220038598_Development_and_Testing_of_the_OPLS_All-Atom_Force_Field_on_Conformational_Energetics_and_Properties_of_Organic_Liquids/links/54c83ba90cf238bb7d0dd3d0/Development-and-Testing-of-the-OPLS-All-Atom-Force-Field-on-Conformational-Energetics-and-Properties-of-Organic-Liquids.pdf (accessed May 4, 2017).

[35] G. Bussi, D. Donadio, M. Parrinello, Canonical sampling through velocity rescaling, J. Chem. Phys. (2007). http://aip.scitation.org/doi/abs/10.1063/1.2408420 (accessed May

4, 2017).

[36] M. Parrinello, A. Rahman, Polymorphic transitions in single crystals: A new molecular dynamics method, J. Appl. Phys. (1981). http://aip.scitation.org/doi/abs/10.1063/1.328693 (accessed May 4, 2017).

[37] U. Essmann, L. Perera, M.L. Berkowitz, T. Darden, H. Lee, L.G. Pedersen, A smooth particle mesh Ewald method, J Chem Phys. 103 (1995) 8577–8593. doi:10.1063/1.470117.

[38] T. Darden, D. York, L. Pedersen, Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems, J. Chem. Phys. (1993). http://aip.scitation.org/doi/abs/10.1063/1.464397 (accessed May 4, 2017).

[39] R. Ballew, J. Sabelko, C. Reiner, A single-sweep, nanosecond time resolution laser temperature-jump apparatus, Rev. Sci. (1996). http://aip.scitation.org/doi/abs/10.1063/1.1147137 (accessed May 4, 2017).

[40] J. Ervin, J. Sabelko, M. Gruebele, Submicrosecond real-time fluorescence sampling: application to protein folding, J. Photochem. (2000). http://www.sciencedirect.com/science/article/pii/S1011134400000026 (accessed May 4, 2017).

[41] H. Lammert, J. Noel, E. Haglund, A. Schug, Constructing a folding model for protein S6 guided by native fluctuations deduced from NMR structures, J. (2015). http://aip.scitation.org/doi/abs/10.1063/1.4936881 (accessed May 4, 2017).

[42] Y. Mu, P. Nguyen, G. Stock, Energy landscape of a small peptide revealed by dihedral angle principal component analysis, Proteins Struct. Funct. (2005). http://onlinelibrary.wiley.com/doi/10.1002/prot.20310/full (accessed May 4, 2017).

[43] G. Maisuradze, D. Leitner, Free energy landscape of a biomolecule in dihedral principal component space: sampling convergence and correspondence between structures and minima, Proteins Struct. Funct. (2007). http://onlinelibrary.wiley.com/doi/10.1002/prot.21344/full (accessed May 4, 2017).

[44] A. Altis, P. Nguyen, R. Hegger, Dihedral angle principal component analysis of molecular dynamics simulations, J. Chem. (2007). http://aip.scitation.org/doi/abs/10.1063/1.2746330 (accessed May 4, 2017).

[45] Y. Cote, P. Senet, P. Delarue, Anomalous diffusion and dynamical correlation between the side chains and the main chain of proteins in their native state, Proc. (2012). http://www.pnas.org/content/109/26/10346.short (accessed May 4, 2017).

[46] R. Zwanzig, Simple model of protein folding kinetics, Proc. Natl. Acad. (1995). http://www.pnas.org/content/92/21/9801.short (accessed May 4, 2017).

[47] M. Sadqi, D. Fushman, V. Muñoz, Atom-by-atom analysis of global downhill protein folding, Nature. (2006). http://www.nature.com/nature/journal/v442/n7100/abs/nature04859.html (accessed May 4, 2017).

[48] H. Ma, M. Gruebele, Low barrier kinetics: dependence on observables and free energy surface, J. Comput. Chem. (2006). http://onlinelibrary.wiley.com/doi/10.1002/jcc.20311/full (accessed May 4, 2017).

[49] T. Cellmer, E. Henry, J. Hofrichter, Measuring internal friction of an ultrafast-folding protein, Proc. (2008). http://www.pnas.org/content/105/47/18320.short (accessed May 4, 2017).

[50] F. Liu, M. Nakaema, M. Gruebele, The transition state transit time of WW domain folding is controlled by energy landscape roughness, J. Chem. Phys. (2009). http://aip.scitation.org/doi/abs/10.1063/1.3262489 (accessed May 4, 2017).

[51] A. Liwo, C. Czaplewski, J. Pillardy, Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field, J. Chem. (2001). http://aip.scitation.org/doi/abs/10.1063/1.1383989 (accessed May 4, 2017).

[52] G. Maisuradze, P. Senet, C. Czaplewski, Investigation of protein folding by coarse-grained molecular dynamics with the UNRES force field, J. (2010). https://www.ncbi.nlm.nih.gov/pmc/articles/pmc2849147/ (accessed May 4, 2017).

[53] S. Piana, K. Lindorff-Larsen, D. Shaw, How robust are protein folding simulations with respect to force field parameterization?, Biophys. J. (2011). http://www.sciencedirect.com/science/article/pii/S0006349511004097 (accessed May 4, 2017).

[54] E. Azarya-Sprinzak, D. Naor, H. Wolfson, Interchanges of spatially neighbouring

residues in structurally conserved environments., Protein. (1997). http://peds.oxfordjournals.org/content/10/10/1109.short (accessed May 4, 2017).

[55] S. Gianni, N. Calosci, J. Aelen, Kinetic folding mechanism of PDZ2 from PTP-BL, Protein Eng. (2005). http://peds.oxfordjournals.org/content/18/8/389.short (accessed May 4, 2017).

[56] D. Bonetti, C. Camilloni, L. Visconti, S. Longhi, Identification and Structural Characterization of an Intermediate in the Folding of the Measles Virus X domain, J. Biol. (2016). http://www.jbc.org/content/291/20/10886.short (accessed May 4, 2017).

CHAPTER 4

The effect of fluorescent protein tags on phosphoglycerate kinase stability is non-additive

Fluorescent proteins tags have become ubiquitous labels to track diffusion, folding, or binding of a host protein.[1-6] Tags such as AcGFP1 are very stable,[7, 8] making them convenient probes. They are particularly convenient in-cell or *in vivo* because cell or tissue auto-fluorescence can hinder detection of a host protein's intrinsic tryptophan fluorescence,[9] and because tags can be co-expressed with the host protein without dye-labeling and injection.

Part of the reason for the great success of fluorescent protein tags is simply that they are just proteins without specific binding partners. Especially in a cell, they mimic the presence of other cellular proteins, albeit connected to the target by a short linker. The tags will interact with and crowd the host protein, but other proteins in the cell do the same.

Nonetheless, fluorescent protein tags are rather large (27 kDa), and despite being engineered to be monomeric, they are prone to interactions.[10] This raises the question of how much fluorescent protein tags interfere with folding kinetics, stability, or function of the host protein by crowding it or interacting with it. The effect of tags on host proteins is clearly not negligible. For example, it has been shown recently that substituting a small ReAsH tag for a 27 kDa mCherry tag speeds up protein folding *in cells* by a factor of two.[11] Thus reduced chain diffusion due to the tag may contribute to the slower folding kinetics observed in cells.[12]

As experiments with fluorescently tagged proteins evolve towards drawing quantitative conclusions about the target protein, it becomes more important to understand both the magnitude and the mechanism of label effects.[13] Extrapolating from observations of the tagged proteins often requires that effect of the tag be treated as a small perturbation, and often assumes that multiple tags (to monitor either intra- or inter- molecular interactions[14]) will perturb the system in a predictable, additive fashion.

This chapter is adapted from K Dave, H Gelman, CTH Thu, D Guin, M Gruebele. The effect of fluorescent protein tags on phosphoglycerate kinase stability is nonadditive. Journal of Physical Chemistry B, 2016

The appropriateness of this assumption places limits on the accuracy of conclusions drawn from the study of tagged proteins. Here I address both the mechanism and the additivity of fluorescent tag effects *in vitro* to provide a reference for *in-cell* experiments. Our results falsify the assumption of additivity for the host protein I study.

Fluorescently-labeled phosphoglycerate kinase (PGK) is used extensively for studies of folding in the cellular milieu and *in vitro*.[12, 15-22] It has been used in both singly-labeled[14] and doubly labeled[12, 23] versions. I compare five constructs of PGK to investigate how this host protein interacts with its fluorescent tags *in vitro*. I denature PGK with heat and pressure to compare the thermodynamic stability of unlabeled PGK, PGK labeled with either AcGFP1 or mCherry, and PGK labeled with both fluorescent tags (Fig. 4.1). When stability of all five constructs is measured by intrinsic tryptophan fluorescence or circular dichroism, I observe that the addition of either individual tag is destabilizing. Thus, destabilizing interactions between PGK and the tag must outweigh any stabilizing effect of crowding by the tag. In contrast, the addition of a second tag doesn't further destabilize PGK. The two tags either sufficiently crowd PGK to overcome the individual destabilizing interactions, or they interact with one another to divert some of the destabilizing interactions away from PGK, or both.

To our surprise, I also found that fluorescence spectroscopy of AcGFP1 alone, but not of mCherry, can be used to detect unfolding of the attached PGK in the singly labeled construct. The green fluorescent protein emission wavelength is sensitive to the conformation of PGK and its spectral shift can be used as a probe of PGK conformation throughout the unfolding transition. I propose that this occurs through differential interaction of the folded and unfolded PGK with hydrophobic regions on the GFP surface, modulating its structural fluctuations with a noticeable effect on the chromophore.[24, 25] This sensitivity is disrupted in the doubly labeled construct, suggesting that the second fluorescent protein disrupts these interactions. This observation could prove useful in cases where a protein with two large tags does not express well in cells.

Fluorescence is very sensitive and non-destructive, but the non-additive effect of fluorescent protein tags on protein stability shows that competition between destabilizing tag-host interaction, tag-tag-interaction, and host crowding already occurs in the *in vitro* model protein, let alone in cells. The destabilizing effect of tags on the host protein may be lessened in cells, where other biomolecules compete to interact with both the tags and the host protein. This may explain some of the stabilization of proteins observed in cells. Although the "apples-to-apples" comparison of tagged protein *in vitro* and in-cell is a valid one, one must keep in mind that it only highlights the cell's effect on the labeled protein. Endogenously expressed

label-free proteins may react differently to the cellular milieu. Despite the complications introduced by the use of fluorescent tags for quantitative measurements, they still fill an essential role in in-cell studies. NMR, infrared absorption and mass spectrometry are label-free,[26-28] but they require either high protein concentrations in the cell (NMR, IR), or they can be destructive to the cell (MS). Comparing stability of the same protein in-cell *vs. in vivo* by a range of methods will be the best solution to assess the different challenges posed by different techniques.

4.1 Methods

4.1.1 Protein sample preparation

Yeast phosphoglycerate kinase (PGK) mutant Y122W/W308F/W333F, with a melting temperature of *ca.* 40 °C *in vitro*, was the basis for all tagged constructs. I expressed the untagged PGK (P), and three fluorescent constructs: PGK labeled with either AcGFP1 (GP) or mCherry (CP) at the N-terminus, mCherry (PC) at the C-terminus, and the doubly labeled FRET construct with the donor AcGFP1 at the N-terminus and the acceptor mCherry at the C-terminus (GPC) (Fig. 4.1). For simplicity, I refer to them as GP, CP, PC, GPC, and P for "bare" PGK. All five proteins were expressed in *E. coli* BL21 cells (DE3 CodonPlus(RIPL), Agilent), and purified as described elsewhere.[21] The purified proteins were dialyzed in 10 mM phosphate buffer at pH 6.8. Pressure thermodynamics were conducted under the same buffer conditions. Temperature thermodynamics were measured in UK buffer (25 mM Tris-HCl, pH 7.5, 5 mM $MgCl_2$, 10 mM KCl, 1 mM EDTA). Protein concentration varied between experiments; I did not observe any effect of concentration on observed stability (Fig. C.1 and C.2). The addition of DTT to either buffer did not affect the observed stabilities, so cysteine-mediated interactions are not significant (1 Cys on the surface of PGK, 2 in AcGFP1). No difference was observed in PGK or fluorescent protein stability over the range of buffer conditions used here.

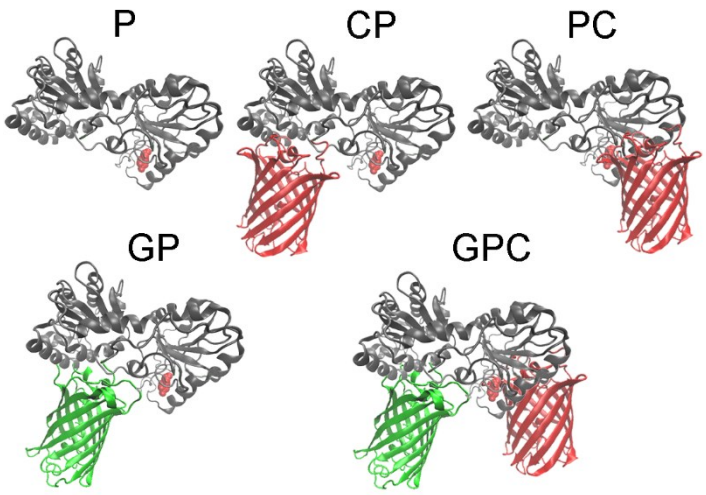

Fig. 4.1: Schematic ribbon structures of PGK (P, showing the tryptophan reporter), PGK labeled with AcGFP1 (GP), PGK labeled with mCherry at either terminus (CP and PC) or doubly-labeled PGK with both tags (GPC).

4.1.2 Pressure and temperature unfolding thermodynamics

Temperature denaturation of all constructs was measured by tryptophan fluorescence and circular dichroism. Tagged proteins were also studied by direct excitation of the fluorescent protein tags (GP, CP, PC, GPC), or FRET (GPC).

Pressure denaturation was measured by tryptophan fluorescence in an ISS cell as described in [29] and by direct excitation of and FRET between the fluorescent protein tags, as for temperature denaturation. A rectangular quartz cuvette with a path length of 4 mm holds the sample in the pressure cell. Measurements are taken every 100 bar from 1 to 1800 bar with a wait time of approximately 8 minutes at each pressure to allow equilibration. Spectrophotometric grade ethyl alcohol (95.0%, A.C.S. reagent; Acros Organics) was used as pressurization fluid.

Fluorescence spectroscopy was carried out using a Cary Eclipse fluorescence spectrophotometer equipped with programmable temperature control (Varian) with excitation and emission slit widths kept at 5 nm. Tryptophan excitation was 280 nm and emission was collected from 290 – 450 nm. AcGFP1 was excited at 475 nm and emission collected from 485 – 560 nm. mCherry was excited at 585 nm and emission collected from 595 – 750 nm. For each fluorescence emission spectrum, the average wavelength $\langle\lambda\rangle$ was calculated by equation (1) where I is intensity and λ wavelength :[10]

$$\langle\lambda\rangle = (\sum_j \lambda_j I_j) / (\sum_j I_j) \quad (1)$$

The same wavelength range was used in all cases to obtain consistent results. I confirm that the starting conformations for the pressure and temperature denaturation experiments are the same by showing that the initial tryptophan emission spectra for both experiments (which both start at ~ 25 °C, 1 bar) are same (Fig. C.3).

FRET measurements of GPC stability were conducted by excitation at 475 nm and collecting emission from 485 – 700 nm. The reported donor/acceptor (D/A) ratio is calculated by dividing the integrated intensity from 485 - 560 nm (D) by the integrated intensity from 585 – 700 nm (A).

Circular dichroism was measured using a JASCO spectrophotometer with Peltier temperature control (JASCO Inc, Easton MD). All spectra were recorded from 250 – 200 nm at a scan rate of 50 nm/min with 1 nm resolution and are an average of 5-10 accumulations. Measurements were conducted in a 2 mm path length quartz cuvette and, unless otherwise noted, at a protein concentration of 2 to 5 μM.

All thermodynamic denaturation signals $S(X)$, where X is temperature or pressure, were fitted to a two-state model separately for temperature and pressure denaturation

$$S(X) = S_U + S_F e^{-\Delta G(X)/RT} / (1 + e^{-\Delta G(X)/RT}) \quad (2a)$$

$$\Delta G(X) = g_X (X - X_m) \quad (2b)$$

to obtain the denaturation midpoints with respect to temperature (T_m) and pressure (P_m). In the main paper, I focus on P_m and T_m, but values of the cooperativity parameters g and signal linear baselines $S_{U,F}$ were also obtained (see Appendix C). Note that PGK is at least a three-state folder, but I focus here on the earliest transition. The higher transition observed by temperature unfolding shows the same ordering of melting temperatures as the lowest transition (see Appendix C, Fig. C.1).I confirm that both temperature and pressure denaturation are reversible by titrating to the start of the unfolded baseline (45 °C and 900 bar, respectively) and then returning to the starting condition (Fig. C.4). I also report the fraction folded ([F]/([F]+[U]) given by setting S_U=0 and S_F=1 in eq. 2b.

4.2 Results

4.2.1 AcGFP1 and mCherry do not show evidence of denaturation

GFP has been shown to be very stable to thermal and pressure denaturation.[30] Here I characterize the fluorescent tags using different perturbations and a variety of probes. When temperature melts for AcGFP1 and mCherry are monitored by exciting tryptophan at 280 nm and detecting integrated fluorescence from 290 to 450 nm, there is no change in the average wavelength over the temperature range from 20 to 65 °C. The average tryptophan emission wavelength of the tag proteins is rather long (see Fig. C.5).

A cooperative transition also was not observed in the 20 to 65 °C temperature range when the fluorescent proteins were directly excited at 475 nm (AcGFP1) or 585 nm (mCherry) (see Fig. C.6). In order to monitor secondary structure of these tags, I also measured mean residue ellipticity (MRE) by circular dichroism (CD) over a similar temperature range (10 – 70 °C). The structure of these tags remains almost unchanged based on CD, bolstering again the claim of stability over a wide temperature range (Fig. C.7).

I also observe no cooperative pressure unfolding transitions for AcGFP1 or mCherry between 1 and 1800 bar as monitored by tryptophan fluorescence or by direct fluorophore excitation (Fig. C.8, C.9). From all the above measurements it is evident that, within our experimental temperature and pressure range, these fluorescent protein labels are stable.

4.2.2 PGK is destabilized by single fluorescent tags

The triple mutant Y122W/W308F/W333F of wildtype yeast PGK[12, 22, 23, 31] was tagged at the N-terminus with green fluorescent protein (GP), or mCherry (CP) to see what effect these tags have on the thermodynamic stability of unlabeled protein (P). The first cooperative transition for unfolding shown in Figures 4.2 and 4.3 was quantified by fitting it to a two-state model.[32] The melting temperatures T_m and pressures P_m are summarized in Table 4.1, and the cooperativity parameters g_T and g_P are shown in Appendix C Tables C.1 and C.2.

The midpoint of the unfolding transition for GP and CP is decreased by several °C relative to P when thermal denaturation is detected by tryptophan fluorescence (Fig. 4.2) or circular dichroism (Fig. 4.3 and Fig. C.10). PC is slightly stabilized relative to P when probed by tryptophan fluorescence, but is destabilized when detected by CD. All CD-detected melting points are higher than the fluorescence-detected melting points, evidence for the known multi-state unfolding for PGK.[33]

The tryptophan fluorescence of PGK undergoes a further transition at temperatures >45 °C, beyond the in-cell range (Fig. C.1).I performed temperature melts at different concentrations,

but the transition was still observed (Fig. C.2). It appears to be a genuine three-state transition with a hyperfluorescent intermediate, which has been observed in the literature.[33, 34] The label-dependence of this transition matches the first unfolding transition shown in Fig.4.2, so the conclusions are independent of which transition is discussed.

Tryptophan fluorescence was also used as a probe to observe unfolding under pressure (Fig. 4.4). Pressure denaturation of all the singly-labeled constructs occurs at lower pressure than for the unlabeled protein, consistent with destabilization of PGK by single tags upon thermal denaturation. Fig. 4.4C shows the calculated folded populations from the fit in Table 4.1. The cooperativity parameters "*g*" (Tables C.1 and C.2) did not show any strong trends.

The unfolding transitions monitored by fluorescence appear smaller for the labeled proteins. The effect is caused by background from the tryptophans in AcGFP1 (1 tryptophan) and/or mCherry (3 tryptophans). Since the fluorescent protein tags have relatively red fluorescence and do not undergo any transition (Fig. C.5 to C.9), their contribution reduces the fraction of tryptophan fluorescence from PGK and shifts the native baseline to longer wavelength. Although good signal-to-noise ratio still allowed reliable extraction of T_m and P_m for tagged constructs, I also performed singular value decomposition (SVD) analysis of the tryptophan emission spectra (see Appendix C). 95-98% of the signal change is accounted for by the first two SVD components. The second SVD component undergoes a transition very near the reported melting temperature or pressure for each analyzed variant (Fig. C.11-C.12). Error analysis also shows that the fitted transition midpoints are accurate (e.g. Appendix C Fig. C.13).

It is therefore evident that placing a single tag on the protein mostly decreases its thermodynamic stability irrespective of denaturation method or probe method, as long as the same probes are compared. Only the PC construct deviates from this general pattern upon thermal denaturation for one probe (tryptophan fluorescence wavelength).

Table 4.1: Stability of protein constructs with respect to pressure and temperature as monitored by CD and tryptophan fluorescence.

Protein	Temperature Denaturation Midpoint (°C)		Pressure Denaturation Midpoint (Bar) (280 nm excitation)
	Measured via Fluorimeter (280 nm excitation)	Measured via Circular Dichroism (CD)	
P	40 (±1)	52 (±1)	1100 (±10)
GP	38 (±1)	43 (±1)	760 (±20)
CP	38 (±1)	42 (±1)	815 (±30)
PC	43(±1)	45 (±1)	880(±10)
GPC	44(±1)	45 (±1)	770 (±30)

4.2.3 Destabilization by two tags is not an additive effect

The doubly labeled construct of PGK was employed to study the effect of adding an additional tag to the protein. The AcGFP1 fluorophore was attached at the N-terminus and mCherry was attached at the C-terminus (GPC). The GPC construct is typical of those used in published FRET folding studies.[22]

Thermal denaturation of GPC probed by tryptophan fluorescence (Fig. 4.2) or circular dichroism (Fig. 4.3) shows that the additional tag does not destabilize PGK further than either individual label. In fact, PGK recovers some or all of its unlabeled stability (Table 4.1). Pressure denaturation detected via tryptophan fluorescence is also highly non-additive, although the doubly labeled PGK is not significantly more stable with respect to pressure than the singly labeled constructs (Fig. 4.4, Table 4.1).

Thus the effect of the two tags on PGK is non-additive by thermal and pressure denaturation, whether tryptophan fluorescence or secondary structure is detected. By all probes and all denaturation methods, the doubly labeled construct was more stable than expected for the sum of the singly labeled effects, even if GP and PC (not CP) were used as reference.

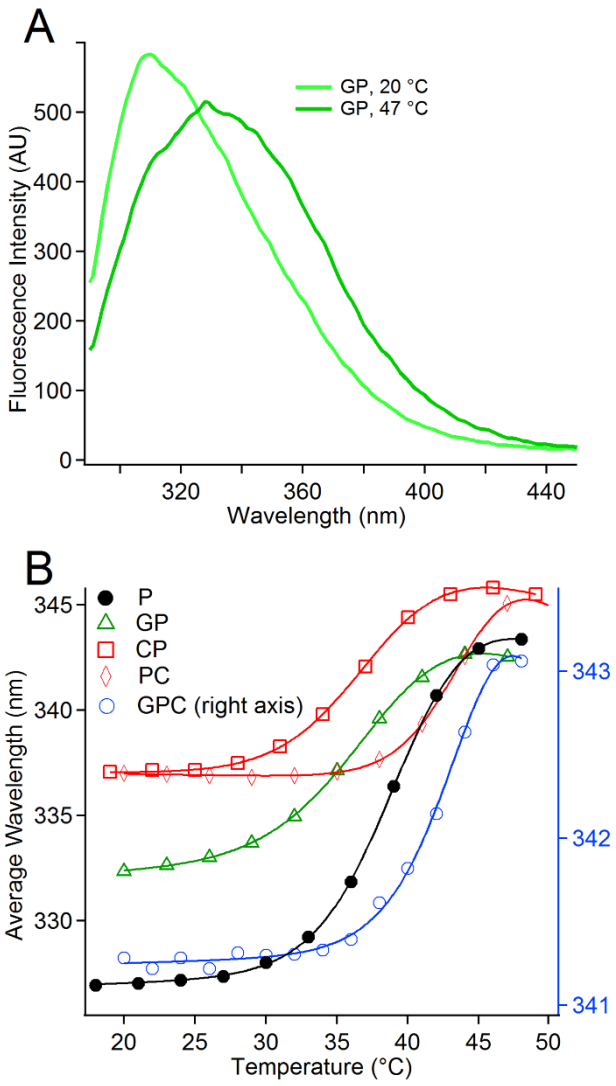

Fig. 4.2: Temperature denaturation of the protein constructs as monitored by tryptophan. **A**. Emission spectrum of GP at 20 °C and 47 °C. **B.** Average wavelength to monitor the unfolding midpoint for the first unfolding transition of P (black), GP (green), CP (red squares), PC (red diamonds) and GPC (blue). The smaller wavelength shift of GPC (right axis) is caused by signal contribution from the two stable labels (1 tryptophan in AcGFP1, 1 in PGK, and 3 in mCherry).

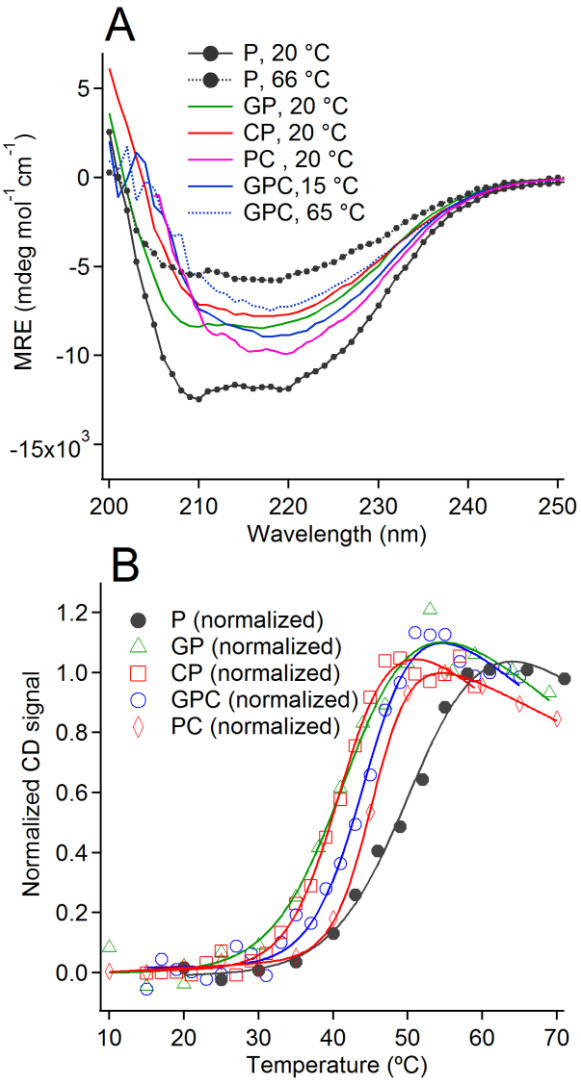

Fig. 4.3: A. Comparison of secondary structure of all protein constructs represented by absolute mean residue ellipticity (MRE) from 200 to 250 nm; P (black), GP (green), CP (red squares), CP (red diamonds) and GPC (blue). Dashed curves show representative spectra at high temperatures of GP (green) and GPC (blue) showing significant loss of secondary structure. **B.** Scaled MRE *vs.* temperature for all the protein constructs. The tags are thermally stable (see Appendix C), so the melting curve monitors PGK denaturation. Absolute MRE is shown in Fig. C.5.

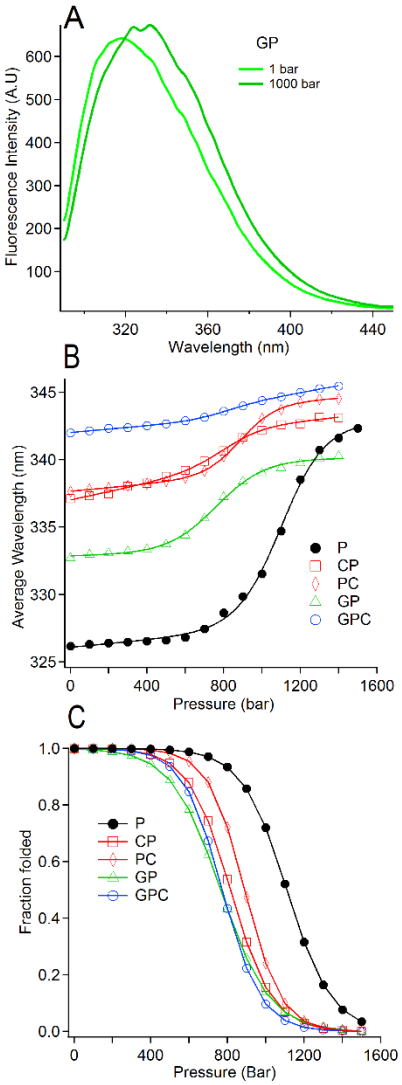

Fig. 4.4: Pressure denaturation for all of the protein constructs monitored by tryptophan fluorescence **A.** Representative emission spectrum of GP at 1 bar (light shade), 1000 bar (medium shade) and 2000 bar (dark shade). **B.** Comparison of average wavelength for GP (green), CP (red), CP (red diamonds) and GPC (blue) and P (black). As for the temperature denaturation curves, the change in average wavelength of the tagged constructs isn't as dramatic as observed for P due to the contribution of additional tryptophan in the stable AcGFP1 and mCherry to the overall signal. **C**. Plot of fraction folded for all the protein constructs *vs*. pressure. The order of stabilities is more obvious than in **B**., where background fluorescence from tag tryptophans reduces the apparent wavelength shift observed for tagged proteins.

4.2.4 FRET-detected unfolding of PGK in GPC

Unlike tryptophan fluorescence and circular dichroism, Förster Resonant Energy Transfer (FRET) cannot serve as a universal comparison between constructs because it can be measured only for GPC. I measured FRET because of its relevance for in-cell experiments. While thermal denaturation of PGK has been studied by FRET,[12, 21] pressure denaturation has not yet been reported by FRET.

I report the ratio of donor (AcGFP1) to acceptor (mCherry) fluorescence signal D/A excited at 475 nm in Fig. 4.5, and the melting temperature obtained in Table 4.2. A higher donor to acceptor ratio indicates an increase in the proportion of the protein population that is unfolded. The decrease of the mCherry peak as the PGK in GPC pressure-unfolds is easily seen in the inset of Fig. 4.5A. FRET between the tags of GPC clearly reports on both temperature- and pressure-denaturation, with midpoints consistent with the tryptophan-detected transitions within fitting error.

Table 4.2: Pressure and temperature midpoints for constructs tagged with GFP or AcGFP1 and mCherry and monitored by fluorescence excited at 475 nm (GP), or by FRET Donor/Acceptor ratio (GPC).

Protein	**Temperature Denaturation Midpoint (^{o}C) 475 nm excitation**	**Pressure Denaturation Midpoint (bar) 475 nm excitation**
GP	46 (±1) average wavelength	680 (±10) average wavelength
GPC	42 (±2) D/A	770 (±10) D/A

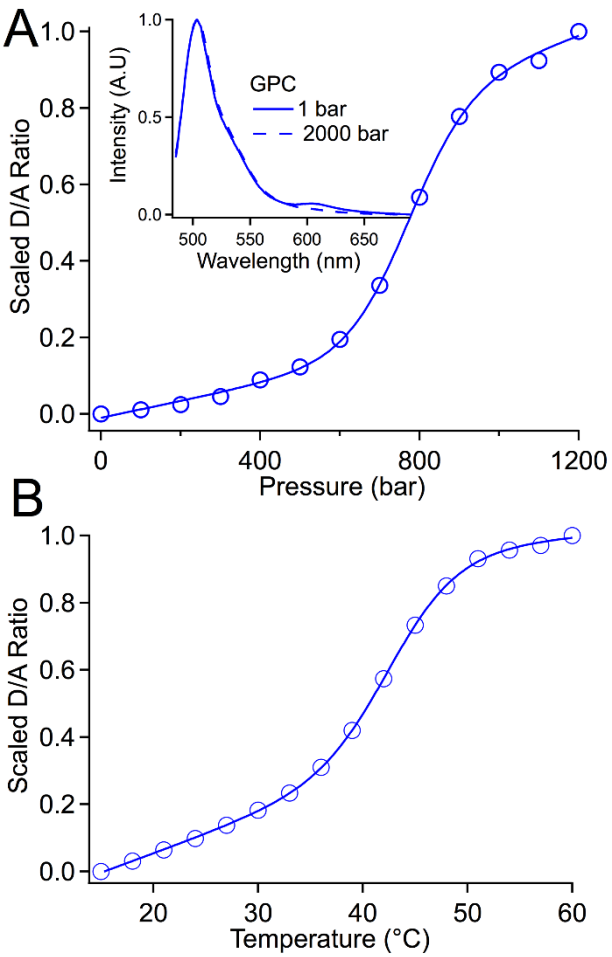

Fig. 4.5: FRET detection of GPC unfolding by pressure and temperature **A.** The normalized donor to acceptor ratio *vs.* pressure shows a cooperative unfolding transition with respect to pressure. *Inset***:** Raw fluorescence intensity *vs.* wavelength plot showing energy transfer at lower pressure (Solid blue trace) as evidenced by a second significant emission maximum and reduction in FRET as the protein unfolds at higher pressure (dashed blue trace). **B**. Normalized donor to acceptor ratio with respect to temperature shows a cooperative unfolding transition.

4.2.5 Unfolding of PGK can be monitored by shift in the GFP emission spectrum alone

I decided to study also the average emission wavelength of AcGFP1 for GP and GPC excited at 475 nm, and of mCherry for CP (as the most direct comparison with GP) excited at 585 nm, in analogy to the tryptophan emission experiments. No significant wavelength shift was observed for CP or GPC at pressures up to 2000 bar (Fig. 4.6A). To our surprise, the GP construct showed a small but highly cooperative wavelength shift under increasing temperature

and pressure (Fig. 4.6B, Fig. C.14, Table 4.2). The thermal denaturation midpoint measured by AcGFP1 wavelength shift is 46 °C, consistent with the CD-detected T_m of GP (Table 4.1). The pressure midpoint is 680 bar, lower than with any other probes.

The directly excited GP fluorescence emission is notably red-shifted relative to both AcGFP1 and GPC under both native and denaturing conditions (Fig. 4.6A), indicating a perturbation of the chromophore in the presence of both folded and unfolded PGK. Thus it appears that labeling only with AcGFP1 enables detection of the unfolding transition of PGK by wavelength shift of the fluorescent tag emission alone.

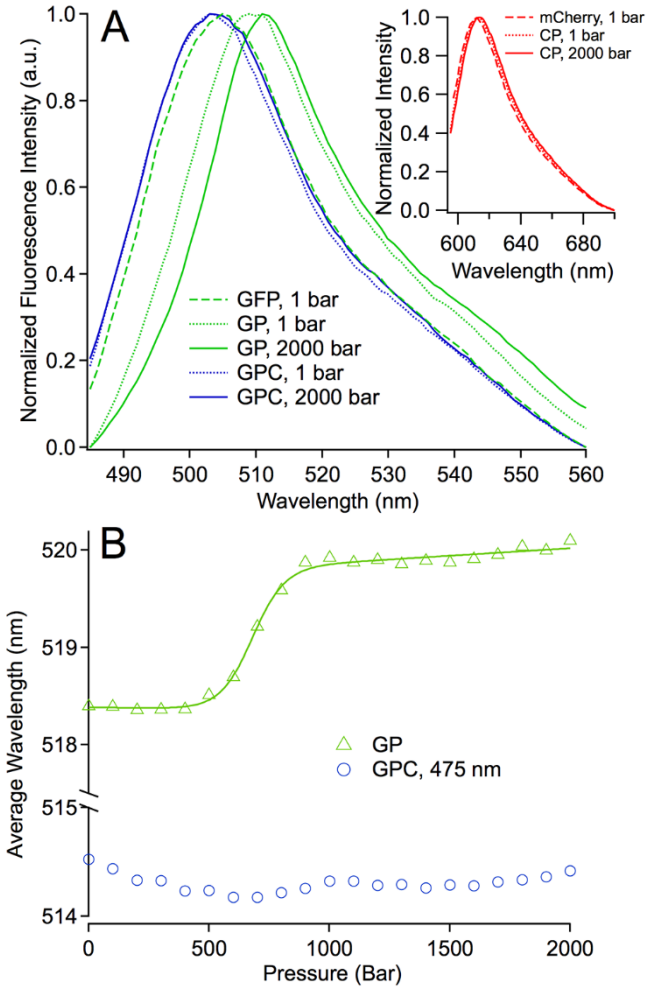

Fig. 4.6: Shift in AcGFP1 emission spectrum **A.** Normalized fluorescence intensity *vs.* wavelength plot for GP at 1 bar (dashed green line), GP (dotted green line: 1 bar; solid green line: 2000 bar) and GPC (dotted blue line: 1 bar; solid blue line: 2000 bar). *Inset:* Normalized fluorescence intensity *vs.* wavelength for mCherry and CP showing no significant emission shift in the same pressure range **B.** Average GFP emission wavelength of GP (green) is sensitive to GP unfolding while emission from GPC (blue) shows no cooperative transition.

4.3 Discussion and conclusion

When a protein is tagged or surrounded by proteins in the cell, there are two major influences on its stability. The mere presence of other proteins excludes volume near the host protein. Such crowding generally destabilizes unfolded states by lowering their conformational entropy. In addition, protein-protein interactions can stabilize or destabilize the host. For example, the unfolded state may be stabilized by interacting with hydrophobic surface patches on other proteins, thus opposing the crowding effect.

I propose two non-mutually exclusive mechanisms to account for the non-additivity I observe for singly- *vs.* doubly-tagged PGK: 1) The "crowding mechanism:"[35, 36] volume exclusion by two labels overcomes the destabilizing interactions of the host protein with individual labels. 2) The "quinary interaction mechanism:"[27, 37, 38] electrostatically or hydrophobically mediated contacts of the labels with one another reduces the destabilizing interaction of the labels with the host protein.

It was previously shown that the melting temperature of label-free PGK linearly increases when the simple crowder Ficoll is added.[23] PGK on its own responds as expected to crowding. Our results show that the addition of a single large fluorescent protein tag (with the exception of one probe for one denaturation method) destabilizes PGK. Therefore, a model where the fluorescent protein tags act as inert crowders cannot explain their effect on PGK stability. I assign this destabilization to interaction of the host protein with the fluorescent tag. Our observation of the AcGFP1 spectral shift under native conditions (Fig. 4.6) suggests that the tag interacts with both folded and unfolded PGK. A similar effect was measured by Sokolovski *et al.* for EnHD attached to eGFP, independent of linker length.[39] Based on MD simulations, they attributed the effect to entropic stabilization of the unfolded EnHD via surface interactions with eGFP.

Comparison of the CP and PC constructs suggests that fluorescent tags can specifically interact with their host. Labeling with mCherry has a much smaller effect on stability at the C-terminus than at the N-terminus. Therefore, a minimal model of the tag-protein interaction must take into account tag location and possibly the chemical properties of the protein regions most accessible to the tag. Specific interaction may explain why some proteins lose activity when labeled in one configuration, but are unaffected by labeling in another configuration.[40]

There is additional evidence for tag-protein interaction. AcGFP1 fluorescence depends significantly on whether AcGFP1 is isolated, in GP, or in GPC. The fluorescence wavelength shift of isolated AcGFP1 compared to GP is large (Fig. 4.6A), indicating an interaction between

the two protein surfaces that modulates the electronic properties of the AcGFP1 fluorophore. In contrast, mCherry does not show any shift in the equivalent CP construct. When mCherry is attached to GP to make GPC, the AcGFP1 fluorescence almost reverts to the AcGFP1 monomer fluorescence. This is a strong indication that AcGFP1 interacts with the PGK surface (reducing PGK stability and shifting AcGFP1 fluorescence), and that this interaction is disrupted by the presence of mCherry.

The sensitivity of the AcGFP1 fluorescence emission wavelength to the folding of PGK is further evidence of a tag-protein interaction. It can be used to monitor PGK unfolding in the GP construct without FRET (Fig. 4.6B), as an alternative to tryptophan fluorescence, circular dichroism, or FRET labeling. Monitoring PGK unfolding by AcGFP1 wavelength shift produces midpoint temperatures closer to the CD result than to the tryptophan result, indicating that the AcGFP1 fluorescence is a more global reporter of unfolding than tryptophan fluorescence.

I tested the additivity of the folding free energy with the dual-labeled GPC. Depending on the probe monitored, PGK with two tags either recovers some stability compared to singly-tagged PGK (e.g. circular dichroism-detected thermal denaturation), or is of similar stability as singly-tagged PGK (e.g. pressure denaturation), or is even more stable than unlabeled PGK (e.g. tryptophan fluorescence-detected thermal denaturation). Two tags do not destabilize the protein by the sum of the individual tag effects, as would be expected from a simple additive model. This is particularly clear for the CD-detected thermal denaturation and pressure denaturation, where the uncertainties are smaller relative to the shift than for fluorescence-detected thermal denaturation. The observation can be explained by more effective crowding in the presence of two tags, and/or tag-tag interaction competing with the unfavorable tag-protein interaction. The back-shift of Ac1GFP wavelength when the mCherry tag is added favors the latter scenario.

The differences observed between different probes and denaturation methods are not unexpected for a multi-state folder like PGK. Such differences may inform the mechanism through which the tags and protein interact. Conventional chemical denaturants and temperature are, in many cases, treated as causing small, additive perturbations to protein stability.[41] This assumption justifies the extrapolation of phenomena observed under denaturing conditions (e.g. unfolded baselines) to the native conditions of theoretical interest. The non-additive effect of fluorescent tags on protein stability[39] and folding kinetics,[11] clearly indicates that such tags are a different class of "perturbation" than the well-understood solvent manipulations (temperature, pressure, denaturants).

Our observations have implications for comparing folding *in vitro* with folding in-cell, where crowding and protein-protein interactions both play a role. Initial in-cell studies have shown both protein stabilization and de-stabilization inside cells.[22, 27, 28, 42-44] Putting tags on PGK already introduces a combination of protein-tag interaction (destabilizing for most single tags), crowding of a protein by two tags (stabilizing), and tag-tag interaction (reduces protein destabilization by the individual tags). Furthermore, fluorescent protein fusions are prone to proteolysis in the cell, raising the possibility that a population of proteins assumed to be homogeneous may actually include proteins with different numbers of intact tags.[45] In-cell tagged protein experiments clearly show that the cell affects protein stability, but the effect may be different on unlabeled endogenous proteins than on tagged proteins. For this reason, a comparison of different assays (NMR, fluorescence mass spectrometry) on the same target protein will be an important next step for in-cell studies.

Ironically, the effect of tags is probably smaller in cells than *in vitro* because other biomolecules in the cell compete to interact with both the tags and the host protein. Nonetheless, it will be important to compare results from different labeling schemes to ensure that tags have a minimal effect on the behavior of a host protein inside the cell.

4.4 References

[1] Tsien RY. The green fluorescent protein. Annu Rev Biochem. 1998;67:509-44.

[2] Chudakov DM, Lukyanov S, Lukyanov KA. Fluorescent proteins as a toolkit for in vivo imaging. Trends Biotechnol. 2005;23:605-13.

[3] Zhang J, Campbell RE, Ting AY, Tsien RY. Creating new fluorescent probes for cell biology. Nat Rev Mol Cell Biol. 2002;3:906-18.

[4] Jones SA, Shim SH, He J, Zhuang X. Fast, three-dimensional super-resolution imaging of live cells. Nat Methods. 2011;8:499-508.

[5] Crivat G, Taraska JW. Imaging proteins inside cells with fluorescent tags. Trends Biotechnol. 2012;30:8-16.

[6] Toomre D, Bewersdorf J. A new wave of cellular imaging. Annu Rev Cell Dev Biol. 2010;26:285-314.

[7] Gurskaya N, Fradkov A, Pounkova N, Staroverov D, Bulina M, Yanushevich Y, et al. A colourless green fluorescent protein homologue from the non-fluorescent hydromedusa aequorea coerulescens and its fluorescent mutants. Biochem J. 2003;373:403-8.

[8] Matz MV, Fradkov AF, Labas YA, Savitsky AP, Zaraisky AG, Markelov ML, et al. Fluorescent proteins from nonbioluminescent anthozoa species. Nat Biotechnol. 1999;17:969-73.

[9] Stephens DJ, Allan VJ. Light microscopy techniques for live cell imaging. Science. 2003;300:82-6.

[10] Rouget J-B, Schroer MA, Jeworrek C, Pühse M, Saldana J-L, Bessin Y, et al. Unique features of the folding landscape of a repeat protein revealed by pressure perturbation. Biophys J. 2010;98:2712-21.

[11] Gelman H, Wirth AJ, Gruebele M. Developing ReAsH as a quantitative probe of in-cell protein dynamics. Biochemistry. 2015;under review.

[12] Dhar A, Ebbinghaus S, Shen Z, Mishra T, Gruebele M. The diffusion coefficient for PGK folding in eukaryotic cells. Biophys J. 2010;99:L69-L71.

[13] Chien P, Gierasch L. Challenges and dreams: physics of weak interactions essential to life. Mol Biol Cell. 2014;25:3474-7.

[14] Guo M, Gelman H, Gruebele M. Coupled protein diffusion and folding in the cell. PLoS ONE. 2014;9:e113040.

[15] Beechem JM, Sherman MA, Mas MT. Sequential domain unfolding in phosphoglycerate kinase: Fluorescence intensity and anisotropy stopped-flow kinetics of several tryptophan mutants. Biochemistry. 1995;34:13943-8.

[16] Watson HC, Walker NPC, Shaw PJ, Bryant TN, Wendell PL, Fothergill LA, et al. Sequence and structure of yeast phosphoglycerate kinase. EMBO J. 1982;1:1635-40.

[17] Griko YV, Venyaminov SY, Privalov PL. Heat and cold denaturation of phosphoglycerate kinase (Interaction of domains). FEBS Lett. 1989;244:276-8.

[18] Haran G, Haas E, Szpikowska BK, Mas MT. Domain motions in phosphoglycerate kinase: determination of interdomain distance distributions by site-specific labeling and time-resolved fluorescence energy transfer. Proc Natl Acad Sci USA. 1992;89:11764-8.

[19] Gast K, Damaschun G, Desmadril M, Minard P, Müller-Frohne M, Pfeil W, et al. Cold denaturation of yeast phophoglycerate kinase: which domain is more stable? FEBS Lett. 1995;358:247-50.

[20] Lillo MP, Beechem JM, Szpikowska BK, Sherman MA, Mas MT. Design and characterization of a multisite fluorescence energy-transfer system for protein folding studies: a steady-state and time-resolved study of yeast phosphoglycerate kinase. Biochemistry. 1997;36:11261-72.

[21] Ebbinghaus S, Dhar A, McDonald JD, Gruebele M. Protein folding stability and dynamics imaged in a living cell. Nat Methods. 2010;7:319-23.

[22] Dhar A, Girdhar K, Singh D, Gelman H, Ebbinghaus S, Gruebele M. Protein stability and folding kinetics in the nucleus and endoplasmic reticulum of eucaryotic cells. Biophys J. 2011;101:421-30.

[23] Dhar A, Samiotakis A, Ebbinghaus S, Nienhaus L, Homouz D, Gruebele M, et al. Structure, function, and folding of phosphoglycerate kinase are strongly perturbed by macromolecular crowding. Proc Natl Acad Sci USA. 2010;107:17586-91.

[24] Konold P, Regmi CK, Chapagain PP, Gerstman BS, Jimenez R. Hydrogen bond flexibility correlates with stokes shift in mPlum variants. J Phys Chem B. 2014;118:2940-8.

[25] Xu MY, George DK, Jimenez R, Markelz AG. Protein resilience and fluorescent protein resistance to photobleaching. Biophys J. 2014;106:459A-A.

[26] Charlton LM, Pielak GJ. Peeking into living eukaryotic cells with high-resolution NMR. Proc Natl Acad Sci USA. 2006;103:11817-8.

[27] Monteith WB, Cohen RD, Smith AE, Guzman-Cisneros E, Pielak GJ. Quinary structure modulates protein stability in cells. Proc Nat Acad Sci USA. 2015;112:1739-42.

[28] Ghaemmaghami S, Oas TG. Quantitative protein stability measurement in vivo. Nat Struct Biol. 2001;8:879-82.

[29] Prigozhin MB, Liu Y, Wirth AJ, Kapoor S, Winter R, Schulten K, et al. Misplaced helix slows down ultrafast pressure-jump protein folding. Proc Nat Acad Sci USA. 2013;110:8087-92.

[30] Herberhold H, Marchal S, Lange R, Scheyhing C, Vogel R, Winter R. Characterization of the pressure-induced intermediate and unfolded state of red-shifted green fluorescent protein—a static and kinetic FTIR, UV/VIS and fluorescence spectroscopy study. J Mol Biol. 2003;330:1153-64.

[31] Osváth S, Sabelko JJ, Gruebele M. Tuning the heterogeneous early folding dynamics of phosphoglycerate kinase. J Mol Biol. 2003;333:187-99.

[32] Wirth AJ, Liu Y, Prigozhin MB, Schulten K, Gruebele M. Comparing fast pressure jump and temperature jump protein folding experiments and simulations. J Am Chem Soc. 2015;137:7152-9.

[33] Missiakas D, Betton J, Minard P, Yon JM. Unfolding-refolding of the domains in yeast phosphoglycerate kinase: Comparison with the isolated engineered domains. Biochemistry. 1990;29:8683-9.

[34] Ervin J, Larios E, Osvath S, Schulten K, Gruebele M. What causes hyperfluorescence: folding intermediates or conformationally flexible native states? Biophys J. 2002;83:473-83.

[35] Minton AP. Excluded volume as a determinant of macromolecular structure and reactivity. Biopolymers. 1981;20:2093-120.

[36] Stagg L, Zhang S-Q, Cheung MS, Wittung-Stafshede P. Molecular crowding enhances native structure and stability of α/β protein flavodoxin. Proc Natl Acad Sci USA. 2007;104:18976-81.

[37] McConkey EH. Molecular evolution, intracellular organization, and the quinary structure of proteins. Proc Natl Acad Sci U S A. 1982;79:3236-40.

[38] Wirth AJ, Gruebele M. Quinary protein structure and the consequences of crowding in living cells: Leaving the test-tube behind. BioEssays. 2013;35:984-93.

[39] Sokolovski M, Bhattacherjee A, Kessler N, Levy Y, Horovitz A. Thermodynamic protein destabilization by GFP tagging: A case of interdomain allostery. Biophys J. 2015.

[40] Swulius MT, Jensen GJ. The helical MreB cytoskeleton in *Escherichia coli* MC1000/pLE7 is an artifact of the N-terminal yellow fluorescent protein tag. J Bacteriol 2012;194:6382-6.

[41] Gelman H, Perlova T, Gruebele M. Dodine as a protein denaturant: The best of two worlds? J Phys Chem B. 2013;117:13090-7.

[42] Ignatova Z, Gierasch LM. Monitoring protein stability and aggregation in vivo by real-time fluorescent labeling. Proc Nat Acad Sci USA. 2004;101:523-8.

[43] Ignatova Z, Krishnan B, Bombardier JP, Marcelino AMC, Hong J, Gierasch LM. From the test tube to the cell: Exploring the folding and aggregation of a beta-clam protein. Biopolymers. 2007;88:157-63.

[44] Guzman I, Gelman H, Tai J, Gruebele M. The extracellular protein VlsE is destabilized inside cells. J Mol Biol. 2014;426:11-20.

[45] Huang L, Pike D, Sleat DE, Nanda V, Lobel P. Potential pitfalls and solutions fot the use of fluorescent fusion proteins to study the lysosome. PloS One. 2014;9:e88893.

CHAPTER 5

Environmental fluctuations and stochastic resonance in protein folding

Weak biological signals below the detection or reaction threshold can be amplified by the addition of noise. The recovered signal is maximized at a certain noise level, resulting in a stochastic resonance.[1] Biological examples range from predating fish generating weak periodic sound waves that are detected by their crayfish prey only when random environmental noise is added,[2] to amplification of electrical membrane signals due to membrane channel voltage fluctuations and many others on different size-scales.[3] The process is illustrated in Fig. 5.1 and requires a sub-threshold signal, a detection threshold, and noise that modulates the sub-threshold signal by just the right amount: too little noise, and the signal remains below the threshold; too much noise, and the signal is swamped by the noise.

Many biomolecular reactions exhibit thresholds, and are thus candidates for stochastic resonance at the molecular level. For example, protein folding is a cooperative process with a sharp transition between folded and unfolded state (as a function of pH, denaturant, temperature, crowding, etc.)[4]. Likewise, protein-RNA binding curves have a sigmoid concentration dependence.[5] Such systems, when poised just below the cooperative threshold, are sensitive to environmental fluctuations. Biomolecular binding and stability inside a cell could be modulated by thermal fluctuations near mitochondria, fluctuations of hydrophobic patches in contact with a protein, or fluctuations in excluded volume as macromolecules jam and unjam inside the cell.[6] Whether such modulation has adaptive consequences for the cell remains unknown.

This chapter is adapted from K Dave, A Davtyan, GA Papoian, M Gruebele, M Platkov. Environmental Fluctuations and Stochastic Resonance in Protein Folding.ChemPhysChem, 2016

Recently, it has been proposed that cooperative kinetics could be driven by a periodic perturbation, and that parameters such as rate coefficients or equilibrium constants could be extracted from such data[7, 8] Indeed, DNA hairpin folding,[9] DNA hybridization in live cells,[10] and protein folding[11] all have been analyzed by driving the reactions with periodic temperature modulation. From such experiments it is a small step to add artificial noise to the periodic perturbation, or to use colored noise with a frequency cutoff to drive the system. Such "artificial thermal noise" is not limited to the kT level, but acts in analogy to thermal noise driving single molecule reactions.

Here I present modulated folding kinetics of the FRET-labeled protein VlsE, a genetically highly variable extracellular membrane protein used by the Lyme disease agent *B. burgdorferii* during host invasion.[12] I drive the folding reaction experimentally with a periodic temperature perturbation, scanning the frequency of the perturbation.[11] A two-state kinetic model[8, 11] is shown to fit the FRET data that monitors the periodic folding/unfolding of VlsE. Reaction parameters such as the activation barrier are extracted from the data. The question then arises whether noise modulation could accelerate the reaction when modulated below the reaction threshold, i.e. whether folding is subject to stochastic resonance. The problem is tractable computationally with a coarse-grained native structure-based model,[13, 14] and interesting properties emerge: for instance, the mean first passage time for folding decreases the most when a protein is driven by noise with a spectrum peaked just above the folding rate k_{obs}. I then follow up on the simulations with analogous experiments, by driving VlsE with a sub-threshold sine wave of frequency $\nu \approx k_{obs}$. No reaction is seen, but adding noise indeed induces the folding/unfolding reaction, peaked at specific noise amplitude in the experiments. Thus, it is at least physically and chemically possible, although it remains biologically unproven, that environmental noise in cells can modulate cooperative biomolecular reactions poised near the reaction threshold, and that such modulation could have an adaptive advantage.

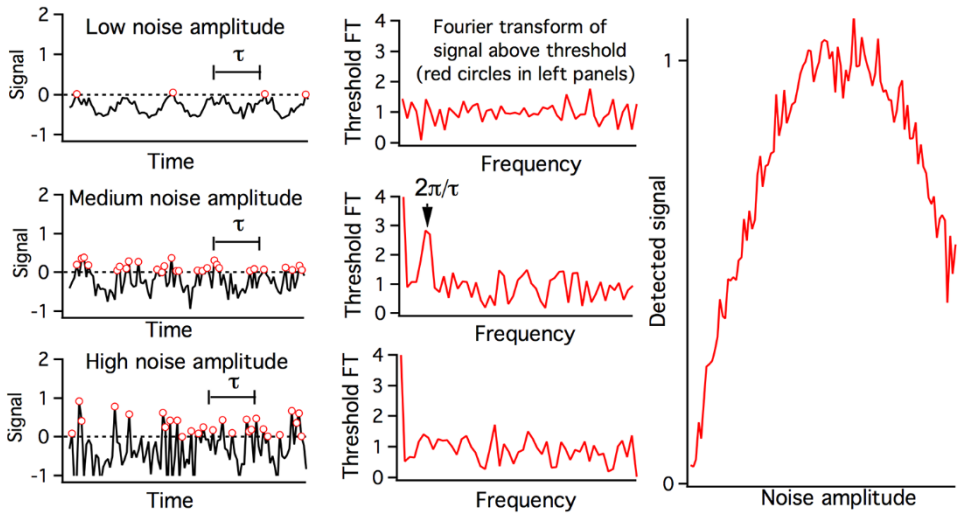

Fig. 5.1: A cartoon of the stochastic resonance mechanism. The left column shows a periodic signal below the detection threshold (dotted line referenced at 0), with increasing noise added. The middle column shows the Fourier transform of the signal detected above threshold, scaled so the baseline noise in each FT is equal. At low noise amplitude no signal peak is detected within the background noise, at high noise amplitude only a noise spectrum is detected. At medium noise amplitude, the noise is modulated at the signal frequency, and a signal can be detected. Thus there is a stochastic resonance in the detected signal as a function of noise amplitude (right panel).

5.1 Methods

5.1.1 Sample

Protein expression was reported previously,[12] so I describe mainly slight differences here. An Ac1GFP-VlsE-mCherry plasmid obtained from Genscript was transformed into *E. Coli* P-lysis cells. The bacterial colonies were later grown into lysogeny media (LB) containing chloramphenicol antibiotics to an OD of around 0.6. At this OD isopropyl thiogalactopyranoside (1 mM IPTG) was added to induce protein expression. Cells were left to grow overnight (≈ 12 hours) at room temperature. Later the cells were collected by centrifugation and sonicated to get cell lysate. Cell lysate was applied to a nickel-nitrilotriacetic acid (Ni-NTA) column which has high affinity towards the histidine tag, protein was purified according the Qiagen protocol.[12]

5.1.2 Apparatus and Measurement procedure

The experimental setup was developed In-house on our live-cell instrument, [15] as described previously.[11] Briefly, a blue LED (470nm, 400 mW) excites the GFP donor; an inverted epifluorescent microscope with a 40x objective illuminates the protein sample, and collects the donor and acceptor fluorescence separately after splitting by a dichroic filter. A frame-rate of 110 Hz was used and data was collected for 11 sec in order to probe and compare the dynamic range of VlsE folding/unfolding kinetics.

The sample chamber was made using double-sided tape of approximately 120 μm height (Grace, Secureseal 654006) on a glass slide and coverslip. The experiments were conducted using VlsE protein concentration of up to 10 μM, with no signs of aggregation over the entire average temperature range (T_0=25-39 °C).

The temperature modulation was performed above the reaction threshold of the protein for the sine wave-driven experiment in Fig. 5.4, and below the reaction threshold (about 38 °C in Fig. 5.2) for the stochastic resonance experiment, where added temperature noise makes the harmonic modulation detectable. Periodic and random temperature modulation (see Fig. 5.6) was achieved by heating the sample with an infrared laser (m2K Lasers, λ=2200 nm, up to 700 nm) which is attached to a computer-controlled power supply (LDC340). The sample base-temperature was set by using two PID-controlled heating-resistors and a Peltier chip to within 0.1° of a user-selected setpoint temperature target in a range of 12 °C up to 50 °C. These resistors, the Peltier chip and heat-conducting copper ribbons as well as fan-cooled heat dissipation fins were attached to the sample chamber through a layer of heat-conduction compound, and the assembly was mounted on the microscope stage.

It is known that the protein VlsE-FRET folds with a folding time $\tau_{obs}=k_{obs}^{-1}\approx 0.7\ s^{-1}$ at 38.3 °C.[12] Based on that and the calculations, I chose to induce stochastic resonance by modulating the temperature on the VlsE-FRET protein with a sine-wave whose period is slightly faster than the folding rate (2 Hz) and below the folding rate (1 Hz), to see if the stochastic resonance weakens or shifts with driving frequency.

The green and red fluorescence coming from the protein were imaged onto a charge-coupled device camera. The fluorescence recorded by the camera exhibited a photobleaching and quantum yield temperature dependence of the donor and acceptor fluorophores.

Photobleaching resulted in a linear decrease of signal over the 11 s time scale of the experiment, and was taken into account by a linear scaling, after which the output could be fitted to phase shifted sine waves. The same correction was used for analogous noise experiments taken under the same conditions. The quantum yield of Ac1GFP and mCherry depends linearly on temperature over the small temperature range used here (20-40 °C). As a result intensity modulation occurs together with temperature modulation, and was taken into account by the fitting model.

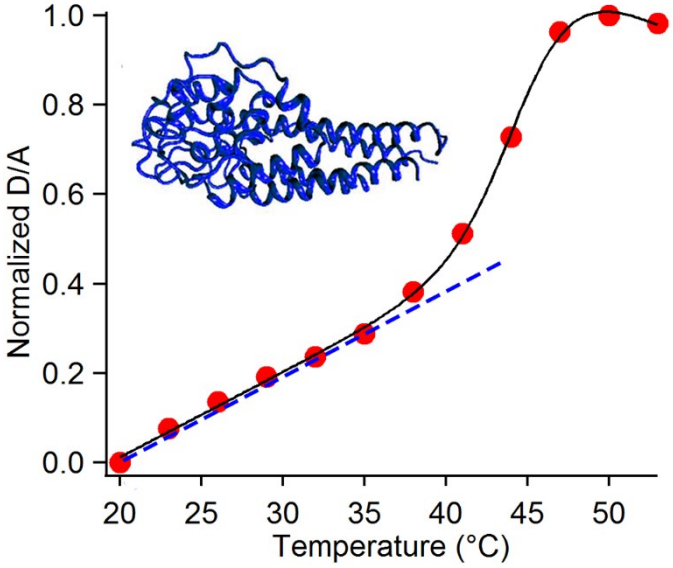

Fig. 5.2: Thermal melt of VlsE- FRET (*in vitro*). Red markers: experimental data; black curve: two-state model fit; blue dashed line: folded state D/A baseline. The reaction threshold for protein unfolding lies at *ca.* 38 °C, and the equilibrium constant $K_{eq} \approx 1$ at the melting temperature of 42±2 °C.

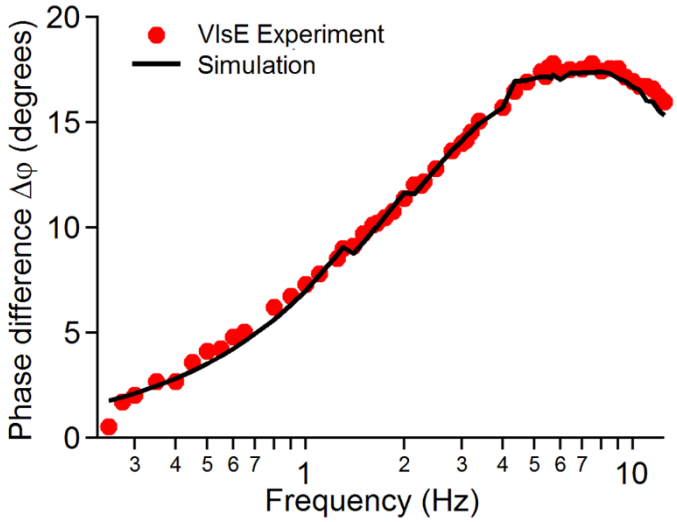

Fig. 5.3: Experimental phase shift between the donor and acceptor fluorescence (red circles) of the VLSE protein. The least-squares fit to a two-state model including temperature-dependent donor and acceptor quantum yield is shown as a black curve. Small glitches in the black fit are numerical errors due to sampling the phase at 0.5° increments in the simulation (see Appendix D).

5.1.3 Data analysis

For the analysis of periodic modulation data, I used the same algorithm presented previously for the analysis of PGK.[11] This is similar to algorithms proposed by Lemarchand and coworkers,[8] and verified by Brownian dynamics in the accompanying paper in this issue.[16] The new addition to the model is the capability to optimize model parameters by a least-squares algorithm, and the code used in this paper is available in Appendix D. Briefly, the donor and acceptor signals at each driving frequency $\omega=2\pi\nu$ were least square fitted to a sine wave

$$S(t)=A\sin(2\pi\nu t+\varphi). \tag{4}$$

Then the phase difference $\Delta\varphi(\nu)$ was calculated, as plotted in Fig. 5.4. This phase difference was simulated as follows:[11]

$$\Delta G(t) = \delta g_1 (T(t) - T_m) \tag{5}$$

$$\Delta G^{\dagger}(t) = \delta g_0^{\dagger} + \delta g_1^{\dagger} (T(t) - T_m) \tag{6}$$

$$\mathrm{QY}_i(t) = 1 + QY_{1i} (T(t) - T_m) \tag{7}$$

were assumed for the folding free energy, activation barrier, and relative quantum yield of "i"=donor or acceptor. δg_1 and T_m were obtained from a thermodynamic fit to Fig. 5.2. $\delta g_0^{\dagger}$, $\delta g_1^{\dagger}$ are kinetic fitting parameters, and $QY_{1D} = -0.011$ and $QY_{1D} = -0.010$ were fixed at the known relative quantum yield slopes of Ac1GFP and mCherry.[17] Rate coefficients for the forward/backward reactions were calculated as $k_m exp[-(\Delta G^{\dagger}(t) \pm {}^1/_2\, \Delta G(t))/RT$. The two-state kinetic master equation $[\dot{F}] = -k_U(t)$[F]+ $k_f[U]$) was then solved, where [U]=C-[F] is the unfolded protein concentration and C is the total protein concentration. From the folded and unfolded concentrations, the observed donor and acceptor fluorescence signals were computed as $D(t) = QY_D \cdot (D_F[F](t) + D_U[U](t))$ and $A(t) = QY_A \cdot (A_F[F](t) + A_U[U](t))$. Here Di and Ai are four constants between 0 and 1 to account for the relative donor and acceptor fluorescence in the folded and unfolded states. The signals A(t) and D(t) were fitted to sine waves just like the experimental data (eq. 4), and the phase difference of the resulting sine waves was evaluated. Adjustable parameters were then optimized by least-squares fitting.

For the sub-threshold modulation + noise experiments (Fig. 5.6), the fast Fourier Transform amplitude of the donor and acceptor signals was calculated using Matlab (Mathworks). This results in a baseline from quantum yield modulation, but any stochastic resonance of comparable magnitude can be seen easily when the FT is plotted as a function of noise amplitude (Fig. 5.7). The baseline is due to fast (<0.1 s) response of the fluorophores to temperature, whereas the ≥0.5 s response of the reaction is delayed. Thus one could improve the stochastic resonance signal further by zeroing out the in-phase component of the FT. However, for unknown sub-threshold modulation waveforms to be detected, this phase is not known, and thus I did not make use of this information in Fig. 5.7, unlike Fig. 5.4.

5.1.4 Native structure-based model potential and dynamics with periodic and random environmental modulation

More details on the Native structure-based model and molecular dynamics simulation and its comparison with kinetic master equation models can be found in the companion computational theory paper.[16] Here I focus on the Native structure-based model and folding rate simulations for 1SRL. Similar results were found for the larger protein PGK,[16] supporting the idea that the observed resonance effect is universal.

For this study I used a Gō-like model developed by Onuchic and coworkers.[14] According to this model, the energy of a specific conformation of a protein is given by a sum of bond

distance/angle/contact potential terms shown in the accompanying paper.[16]

Molecular dynamics simulations of PDB model protein 1SRL (an SH3 domain from tyrosine kinase) were carried out with the Go-like potential described above, but with ε in the native contact energy term modulated as follows about its average value ε_0: A sinusoidal wave with amplitude $\delta\varepsilon$ and period τ, or Gaussian noise with amplitude $\delta\varepsilon$ and correlation time τ was added at each time step, to act like the modulation and artificial thermal noise in the experiments. Each folding time (average mean first passage time) was computed from 1000 trajectories as a function of $\delta\varepsilon$ and τ. All simulations were started in random conformations with low native contact order Q. A fraction of native contacts above Q=0.8 was considered folded and yielded one mean first passage time for each trajectory.

For 1SRL I used the following model parameter: ε_0=2.3 kJ/mole. This results in 2.5 k_BT free energy barrier for folding vs. about 9 k_BT barrier for VISE used in the experiments (see Table 5.1) and ensures that the folding reaction can be seen over the course of computationally feasible simulations. Thus the absolute time scales of the simulation in Fig. 5.5 and experiment in Fig. 5.7 cannot be compared. Additionally, the reduction in the number of degrees of freedom in the coarse grained native structure-based model results in smoother free energy landscape and thus even faster dynamics for 1SRL protein. Consequently, the times in Fig. 5.5 that are on the sub-picosecond time scale cannot be directly related to the experimental times, and only the trends as a function of $\delta\varepsilon$ and τ should be considered.

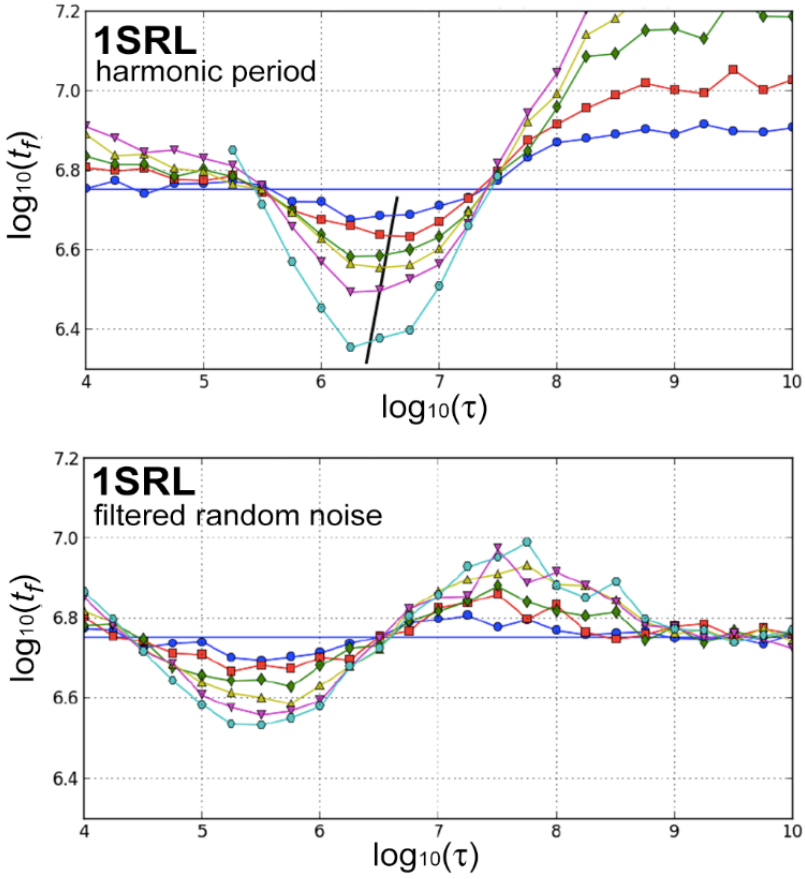

Fig. 5.4: Computational prediction of periodic-driven response and noise-driven response of a model protein. Top: For reference, a native structure-based (Gō-like) model protein (PDB code: 1SRL, SH3 domain) is subject to harmonic temperature modulation$\varepsilon(t) = \varepsilon + \delta\varepsilon\sqrt{2}sin(2\pi t/\tau + \phi)$. The average first-passage time *vs.* period τ of the driving waveform is plotted. Several modulation amplitudes are shown. $\sqrt{\delta\varepsilon^2/\varepsilon^2}$ covers the range 0.03 (dark blue), 0.04, 0.05, 0.06, 0.07, 0.1 (light blue). The black line shows that optimal driving frequency and amplitude are correlated. Bottom: Same model, but driven by Gaussian noise with a correlation time τ, obtained by solving the Langevin equation (2) in the text. When τ is equal or faster than the natural rate k_{obs} of the unperturbed system, the first passage time of the driven system decreases (reaction rate k_f increases) as the amplitude increases. t_f and τ on those plots are in the units of femtoseconds.

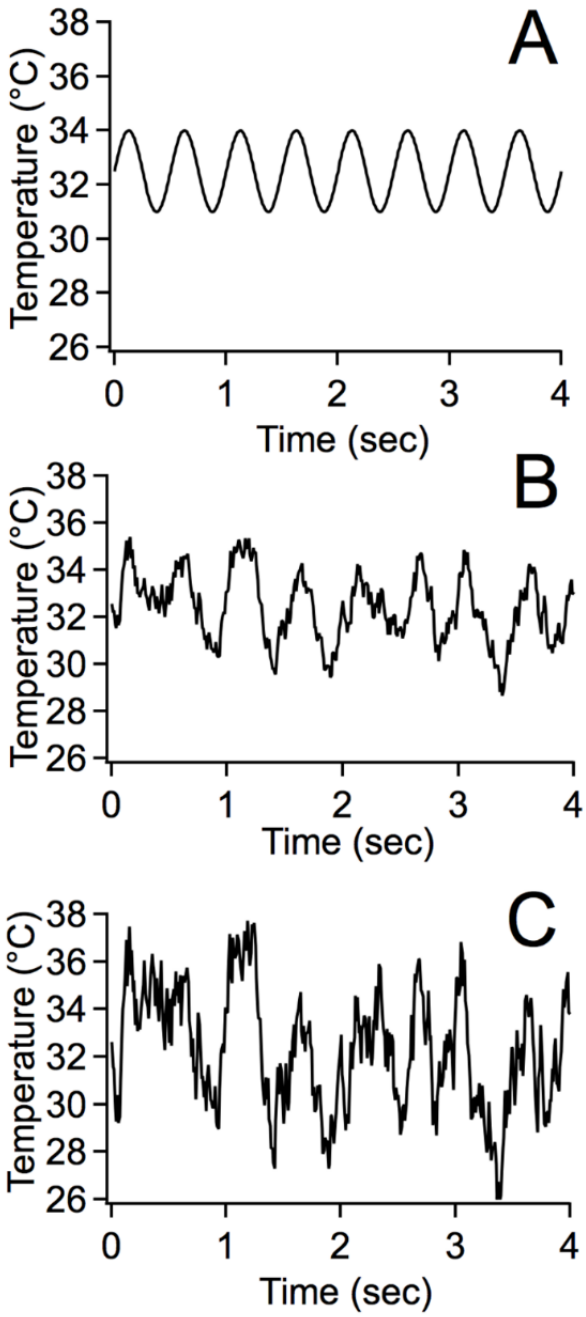

Fig. 5.5: The 2 Hz sine wave + noise signal used to drive VlsE into stochastic resonance as shown in Fig. 5.7, (A) with 0 °C RMS temperature noise, (B) with 1 °C RMS temperature noise (near stochastic resonance when T_0=32-33 °C), and (C) with 2.25 °C RMS temperature noise (above stochastic resonance maximum).

5.2 Results

5.2.1 Overview

The extracellular protein VlsE is a large and relatively slow-folding protein ($k_{obs} \approx (0.7\ s)^{-1}$ at 38.3 °C).[12] VlsE is the largest known two-state folder,[18] so it should obey simple unimolecular kinetics $\Delta c(t) \sim e^{-k_{obs}t}$. Here k_{obs} is the measured rate, which for a two-state system equals the sum of the rates of folding and unfolding, or $k_{obs} = k_f + k_u$. This simple behavior is in contrast to the enzyme PGK, a multi-state folder whose modulation kinetics was studied previously.[11] In order to detect reversible modulation of the protein population between the folded and unfolded states, VlsE was FRET-labeled with Ac1GFP at the N-terminal, and with mCherry at the C-terminal along with a His tag for purification. The protein was then subjected to thermal modulation: first with variable-frequency sine waves to corroborate its activation barrier and folding rate by the modulation approach; then, based on encouraging molecular dynamics and Brownian dynamics simulation results, with a sinusoidal signal + variable thermal noise amplitude, to detect stochastic resonance.

5.2.2 Fluorescence-detected thermal unfolding of VlsE

In order to locate the optimal temperature range for modulation experiments, fluorimeter temperature melts were detected by the FRET Donor/Acceptor (D/A) ratio using ≈2 μM protein solution. Protein concentrations up to 10 μM showed no signs of aggregation over the temperature range of subsequent thermal modulation experiments (20-38 °C).

Thermal unfolding of VlsE is a nonlinear threshold process. The temperature unfolding data in Fig. 5.2 was fitted to a sigmoidal two-state model (see Methods). The midpoint of the thermal unfolding transition of VlsE-FRET was obtained to be T_m=42±2 °C, in agreement with previous work.[12, 18]The onset of the unfolding reaction occurs at approximately 38 °C, where the D/A ratio begins to differ substantially for the almost linear native state baseline (dashed blue line in Fig. 5.2). The baseline is due to temperature-dependent quantum yields of AcGFP1 and mCherry donor and acceptor labels.[12, 15]

5.2.3 Periodic thermal modulation

In our earlier study of the enzyme PGK I showed that thermal modulation can be used to study protein folding reaction kinetics.[11] VlsE folding can be driven by a periodic waveform, and kinetic parameters such as the activation free energy $\Delta G^{\dagger}$ can be obtained, in analogy to measuring fluorescence lifetimes by periodic modulation instead of a fast excitation pulse.[19, 20]

The experiment is illustrated in Fig. 5.3 (A and B). Thermal modulation was performed with a waveform-controlled 2200 nm infrared laser about an average temperature of T0=38 °C, in order to maximize the signal without inducing protein aggregation. The ≈10 μM protein solution was subjected to a periodic temperature waveform

$$T(t) = T_0 + \delta T(t) = T_0 + \frac{1}{2}\Delta T sin(2\pi\nu t + \phi) \quad (1)$$

at the sample slit. As discussed in detail previously,[11] the green donor and red acceptor fluorescence signals collected at the CCD camera are affected by two processes. 1) The quantum yield of the fluorescence labels decreases linearly with temperature.[15] This process causes each of the green and red signals to be 180° out of phase with T(t). In that case, the relative phase $\Delta\varphi$ between green and red is 0° (Fig. 5.3A). 2) The folding reaction causes green and red FRET signals to be 180° out of phase relative to each other (unfolding = more green/less red, refolding = less green/more red). Moreover, the red signal is in phase with T(t) for slow modulation frequency $\nu << k_{obs}$, but up to 90° out of phase with T(t) for fast modulation frequency $\nu >> k_{obs}$ as the protein folding reaction cannot track rapid variations in temperature, which is further elaborated elsewhere. [16]

The resulting phase shift $\Delta\varphi(\nu)$ between the red and green output signals (Fig. 5.3B) can be used to extract the folding/unfolding kinetics from the data. I use a kinetic two-state model with time-dependent free energy and quantum yields.[11] From the model parameters, a two-state kinetic master equation is solved, the time-dependent rate coefficients can be calculated, and donor and acceptor fluorescence signals are calculated. Finally the simulated phase shift $\Delta\varphi(\nu)$ is calculated from the simulated fluorescence signals and compared with the obtained experimental phase shift between the green and red acquired signals (see Methods). The suitability of such simple kinetic master equation models[8] has been tested by comparison with Brownian dynamics simulations.[16]

Fig. 5.4 shows the experimental data (red) and the computed phase shift (black), after the model parameters have been optimized by least squares fitting. Table 5.1 shows the optimized model parameters and 1 σ uncertainty for the model fit in Fig. 5.4 (see Methods, Data Analysis). As expected, the phase difference $\Delta\varphi$ between green and red is 0° at low driving frequency. $\Delta\varphi$ increases as ν approaches the unimolecular reaction rate k_{obs}, Eventually $\Delta\varphi$ decreases again: If the reaction is driven at $\nu > k_{obs}$, the reaction amplitude decreases, so the quantum yield modulation, which has $\Delta\varphi = 0$ dominates the signal. The activation barrier determined for this reaction is $\Delta G^{\dagger}$ = 22.3±0.1 kJ/mole, assuming a prefactor of $k_m \approx (5\ \mu s)^{-1}$ in the equation $k_i = k_m \exp(-\Delta G^{\dagger}/RT)$ for the rate coefficients. The value of k_m is chosen close to the "speed limit" of protein folding,[21, 22] with an upwards adjustment because VlsE is much larger than the mini proteins for which the speed limit has been estimated.[23, 24] At 38 °C, the fitted kinetic and thermodynamic parameters yield a reaction rate $k_{obs} \approx (0.7\ s)^{-1}$, in agreement with conventional T-jump measurements.[12]

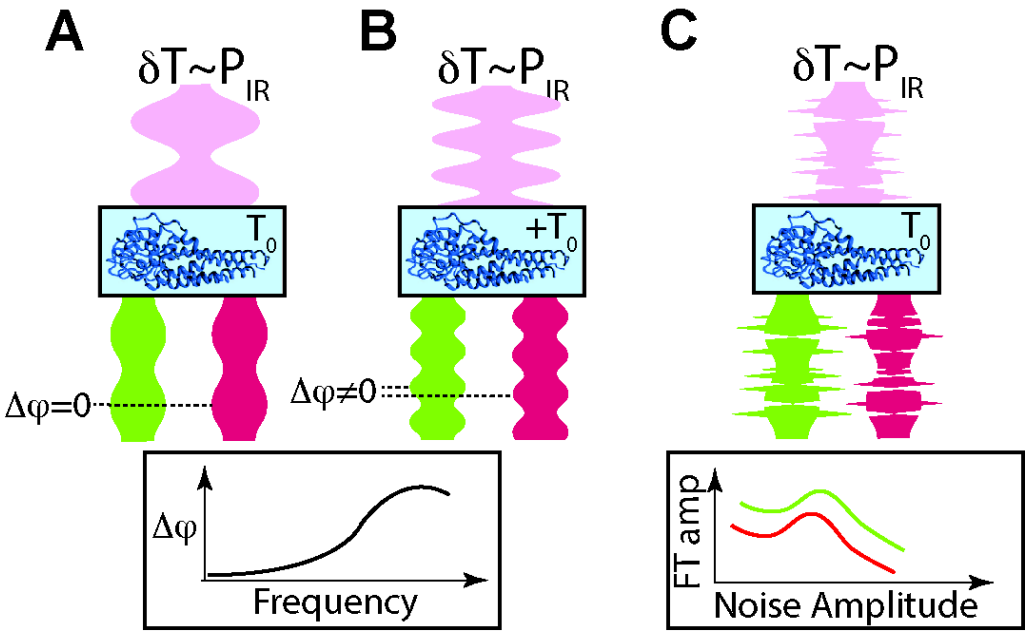

Fig. 5.6: Schematic of the experiment. A 2200 nm IR pulse (pink) periodically modulates the sample above and below the average temperature T_0. The green (donor) and red (acceptor) fluorescence of the labeled VlsE (sample in blue box) oscillates as a result. (A) When the average temperature T_0 is too low to allow significant reaction, donor and acceptor fluorescence signals are modulated in phase because both have a quantum yield that decreases with temperature.[33] (B) When the average temperature $+T_0$ is sufficiently high for reaction to occur and the modulation frequency comparable to or faster than the reaction rate k_{obs}, the green and red fluorescence contain components shifted by 180° relative to one another, and (red) up to 90° with respect to the temperature modulation. This is a 'low pass filter' effect of the reaction when driven too fast. The plot of phase shift *vs*. applied modulation frequency at constant modulation amplitude can be used to fit k_{obs} and determine the activation barrier. (C) If the IR modulation is below the reaction threshold ~ 38 °C, but an increasing amount of temperature noise is added, a stochastic resonance can be detected at the driving frequency above the background signal due to quantum yield modulation.

Table 5.1: Experimental folding parameters and FRET parameters for VlsE, obtained by least squares fitting of the measured denaturation curve in Fig. 5.2 and the measured phase curve in Fig. 5.4.

Fitting parameter	Value and fitting uncertainty (1 standard deviation)
T_m	42 ± 2 °C
δg_1	1380 ± 180 kJ/mole/K
k_m	5 μs^{-1} (fixed)
$\Delta G^{\dagger}$	22.3 ± 0.1 kJ/mole
A_F	0.4 ±0.1
D_F	0.6 ±0.1
A_U	0.17 ±0.05
D_u	0.83 ±0.05

5.2.4 Computational prediction of stochastic resonance in a folding reaction

So far, our results are analogous to what was found for PGK.[11] To see if folding/unfolding can be accelerated by application of artificial thermal noise, I performed molecular dynamics simulations on a small model protein with a temperature-dependent native structure-based model potential [13, 14]. The goal was not to simulate VlsE, which folds far too slowly for realistic simulation with an all-atom force field, but to obtain in general the effect of noise amplitude and correlation time on reaction rate (see Methods and accompanying theory paper[16]).

In our model, the strength ε of the native contact terms in the protein interaction potential was modulated either periodically, or by correlated random noise, where the deviation $\delta\varepsilon(t)$ of ε from its average value is determined by solving the following the Langevin equation

$$\frac{\partial \delta\varepsilon(t)}{\partial t} = -\frac{\delta\varepsilon(t)}{\tau} + \frac{<\delta\varepsilon^2>}{\tau} G(t), \qquad (2)$$

where G(t) is Gaussian white noise. The resulting $\delta\varepsilon(t)$ simulates artificial thermal noise with amplitude $\sqrt{<\delta\varepsilon^2>}$ and correlation time τ. Frequency components higher than $1/\tau$ rapidly diminish in such noise. Fig. 5.6 shows the similar results obtained when the protein is driven periodically at period τ, or by artificial thermal noise with frequency content up to $1/\tau$. The first passage time ($\tau_{MFPT} = 1/k_f$ in our experiments discussed above) decreases (i.e. the reaction speeds up) when the reaction is driven at frequencies comparable to the reaction rate. Additionally, as the noise level $\sqrt{<\delta\varepsilon^2>}$ increases, at driving frequencies just above

resonance, the rate acceleration is continuously enhanced, indicating that the protein approaches stochastic resonance driven by the noise. Therefore it is possible that such a noise-driven rate increase can be observed experimentally for protein folding.

5.2.5 Experimental addition of artificial thermal noise

Next I tested the idea that noise can be utilized to amplify a sub-threshold folding reaction to increase its rate to the detectability limit. I thermally modulated VlsE again as described by eq. (1) slightly above and below the folding rate (2Hz and 1Hz, respectively), but this time kept the average temperature and periodic modulation by itself well below the reaction threshold of *ca.* 38 °C (Fig. 5.6A). I then added increasing amounts of noise to the thermal waveform to drive the system towards the reaction threshold as follows:

$$T(t) = T_0 + \frac{1}{2}\Delta T sin(2\pi\nu t + \phi) + \frac{1}{2}\Delta T_{rand} G_{lp}(t). \quad (3)$$

The random component $G_{lp}(t)$ was obtained by computing Gaussian-distributed pseudo-random numbers $G(t)$ and passing $G(t)$ through a 6dB/octave low-pass filter with a cut-off frequency of 20 Hz. The values of ΔT (3 °C) and T_0 (28 to 33 °C) were chosen so T would remain below the reaction threshold of ~38 °C at all times unless assisted by noise. The Gaussian random noise amplitude was tuned so the root-mean-squared temperature fluctuations (RMS temperature noise in Fig. 5.7) ranged from 0 to 2.25 °C. Fig. 5.6 shows a sample of the periodic+noise waveforms driving the VlsE folding reaction for T_0=32.5 °C and RMS thermal noise of 0, 1, and 2.25 °C.

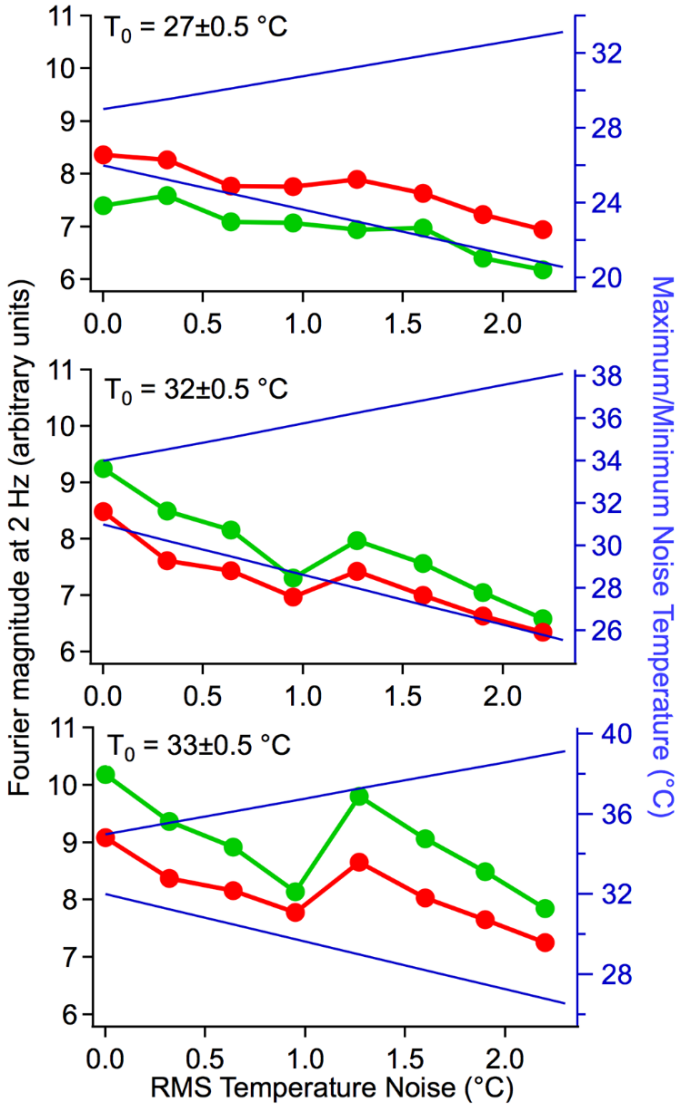

Fig. 5.7: Stochastic resonance, detected by Fourier transform magnitude of the donor (green) and acceptor (red) signals grows in when a root-mean-squared (RMS) temperature noise of ca. 1.2 °C is superimposed on the sub-threshold sine wave modulation in Fig. 5.6A. (A) T_0 = 28 °C. (B) T_0=32 °C. (C) T_0=33 °C. Stochastic resonance grows in at $\sqrt{\delta(T-T_0)^2} \approx 1.25$ °C as the average temperature T_0 is increased. The baseline (*ca*. 6 units of the FT y-axis) is due to modulation of the quantum yield of donor and acceptor. The blue decreasing line is the minimal noise temperature applied on the protein at each RMS temperature, and the blue increasing line is the maximal noise temperature thereof.

5.2.6 Observation of stochastic resonance

I measured the resulting donor and acceptor FRET amplitudes $D(t)$ and $A(t)$, and computed their Fourier transform amplitude at the driving frequency of 2 Hz (see Appendix D Fig. D.1), slightly faster than the natural relaxation rate $k_{obs} \approx 0.7\ s^{-1}$ measured in the previous section and in ref. [12] at 38.3 °C. The sub-threshold periodic modulation alone produces only a slope as a function of RMS temperature noise, due to the modulation of the quantum yield of donor and acceptor, which depends linearly on temperature (Fig. 5.7 top) – just as I observe in our kinetic model simulations. However, as the average temperature is increased from 11 °C below reaction threshold to 5 °C below reaction threshold, a peak can be seen in the signal at 2 Hz at an artificial thermal noise of ≈1.3 °C RMS. At higher noise amplitude, the signal disappears again and returns to the baseline (Fig. 5.7 middle and bottom). A much weaker effect is seen at 1 Hz ($\nu < k_{obs}$) sub-threshold modulation with the same added noise levels (Fig. D.2).

I assign the peak in the 2 Hz signal *vs.* noise amplitude to a stochastic resonance of the folding/unfolding reaction of VlsE, driven by a sub-threshold periodic modulation that produces no detectable reaction on its own, but induces a reaction rate above our detection threshold when noise is added. Adding too much noise (> 2 °C RMS) still produces reaction, but swamps the periodic sub-threshold modulation so the Fourier transform no longer peaks at 2 Hz (see Fig. D.1 for examples of the full Fourier spectra).

5.3 Discussion and conclusion

Stochastic resonance has been observed in a variety of natural phenomena. Macroscopic phenomena include mechanoreceptors in rats [25] and electroreceptors in paddlefish that receive signals more sensitively due to added environmental noise [26]. It also plays a role in biological signal processing, from visual enhancement [27] to neuronal signaling[28].

Stochastic resonance can also affect chemical reactions, which have intrinsically nonlinear rate and equilibrium behavior. This effect is generally observed near unstable points of the reaction's state space as a function of perturbation parameters[29]. Examples include pulsing Belousov-Zhabotinsky reactions,[30] as well as electron transfer reactions[31]. In particular, stochastic resonance plays a role in biochemical reactions, such as cell signaling, where noise due to a small number of signaling molecules can control gene silencing[32].

Here I add protein folding to the list of chemical reactions that exhibit stochastic resonance. The analysis of a small protein using native structure-based coarse-grained simulations shows that when a folding reaction is driven either by periodic perturbation with period τ, or by colored noise with a characteristic cutoff time $\tau \sim 1/k$, a significant decrease in the first passage time (or increase in the forward rate) can be observed. Thus protein folding has sufficiently nonlinear equilibrium curves (Fig. 5.2) to exhibit stochastic resonance. At first, our modulation experiment confirmed that VlsE behaves as a two-state folder under periodic modulation above threshold. I then subjected VlsE to a combination of a sub-threshold temperature perturbation with period τ plus artificial thermal noise. When monitored by Fourier transform at the frequency $\nu=1/\tau$, the fluorescence output signal peaks as a function of noise amplitude, but only when the periodic modulation is close to threshold (average temperatures of 32 or 33 °C *vs.* 27 °C in Fig. 5.7). I also measured the response at v=1 Hz, a factor of 2 below the reaction rate k_{obs}, (Fig. D.2), but there the response is not as evident, in keeping with a slowdown of the reaction when the noise correlation time is slower than the reaction rate (Fig. 5.5). As discussed in the accompanying theory paper,[16] stochastic resonance can also be seen via the phase shift $\Delta\varphi$ of the red and green signal. The predicted phase shifts with realistic FRET input parameters are very small (see accompanying theory paper), and I was not able to use phase shift to identify the noise level or driving frequency that maximize stochastic resonance.

It is not known at present whether cells use stochastic resonance to modulate biomolecule function outside the cases of signaling or visual signal enhancement that have been studied.[27, 32] However, our results show that it is a physically plausible process. There is growing evidence that many proteins in the cell, for example certain intrinsically disordered protein (IDPs), can switch conformation based on small thermal, or other perturbations. Lymphotactin is an example of such a protein whose structure and function are modulated by a small environmental perturbation.[33] It is possible that environmental fluctuations are accelerating protein folding and potentially even protein association reactions, and thus contribute to the cellular control of structure and function of such proteins.

5.4 References

[1] Hänggi P. Stochastic resonance in biology how noise can enhance detection of weak signals and help improve biological information processing. ChemPhysChem. 2002;3:285-90.

[2] Douglass JK, Wilkens L, Pantazelou E, Moss F. Noise enhancement of information-transfer in crayfish mechnoreceptors by stochastic resonance. Nature. 1993;365:337-40.

[3] Wiesenfeld K, Pierson D, Pantazelou E, Dames C, Moss F. Stochastic Resonance on a circle. Phys Rev Lett. 1994;72:2125-9.

[4] Schellman JA. The thermodynamic stability of proteins. Annu Rev Biophys Chem. 1987;16:115-37.

[5] Levy Y, Onuchic JN. Water mediation in protein folding and molecular recognition. Annu Rev Biophys Biomol Struct2006. p. 389-415.

[6] Wirth AJ, Gruebele M. Quinary protein structure and the consequences of crowding in living cells: Leaving the test-tube behind. Bioessays. 2013;35:984-93.

[7] Lemarchand A, Berthoumieux H, Jullien L, Gosse C. Chemical Mechanism Identification from Frequency Response to Small Temperature Modulation. J Phys Chem A. 2012;116:8455-63.

[8] Closa F, Gosse C, Jullien L, Lemarchand A. Identification of two-step chemical mechanisms and determination of thermokinetic parameters using frequency responses to small temperature oscillations. J Chem Phys. 2013;138:244109.

[9] Braun D, Libchaber A. Lock-in by molecular multiplication. Appl Phys Lett. 2003;83:5554-6.

[10] Schoen I, Krammer H, Braun D. Hybridization kinetics is different inside cells. Proc Nat Acad Sci USA. 2009;106:21649-54.

[11] Platkov M, Gruebele M. Periodic and stochastic thermal modulation of protein folding kinetics. J Chem Phys. 2014;141:035103.

[12] Guzman I, Gelman H, Tai J, Gruebele M. The Extracellular Protein VlsE Is Destabilized Inside Cells. J Mol Biol. 2014;426:11-20.

[13] Go N, Noguti T, Nishikawa T. Dynamics of a small globular protein in terms of low-frequency vibrational modes. Proc Nat Acad Sci USA. 1983;80:3696-700.

[14] Clementi C, Nymeyer H, Onuchic JN. Topological and energetic factors: what determines the structural details of the transition state ensemble and “en-route” intermediates for protein folding? an investigation for small globular proteins1. J Mol Biol. 2000;298:937-53.

[15] Ebbinghaus S, Dhar A, McDonald JD, Gruebele M. Protein folding stability and dynamics imaged in a living cell. Nat Meth. 2010;7:319-23.

[16] A. Davtyan MP, M. Gruebele, G. Papoian. The Role of Micro-environmental Fluctuations in Protein Folding. ChemPhysChem. 2015.

[17] Dhar A, Girdhar K, Singh D, Gelman H, Ebbinghaus S, Gruebele M. Protein Stability and Folding Kinetics in the Nucleus and Endoplasmic Reticulum of Eucaryotic Cells. Biophys J. 2011;101:421-30.

[18] Samiotakis A, Wittung-Stafshede P, Cheung MS. Folding, Stability and Shape of Proteins in Crowded Environments: Experimental and Computational Approaches. Int J Mol Sci. 2009;10:572-88.

[19] Alcala JR, Gratton E, Prendergast FG. Resolvability of fluorescence lifetime distributions using phase fluorometry. Biophys J. 1987;51:587-96.

[20] Lakowicz JR, Laczko G, Cherek H, Gratton E, Limkeman M. Analysis of fluorescence decay kinetics from variable-frequency phase shift and modulation data. Biophys J. 1984;46:463-77.

[21] Kubelka J, Hofrichter J, Eaton WA. The protein folding 'speed limit'. Curr Opin Struct Biol. 2004;14:76-88.

[22] Yang WY, Gruebele M. Folding at the speed limit. Nature. 2003;423:193-7.

[23] Liu F, Nakaema M, Gruebele M. The transition state transit time of WW domain folding is controlled by energy landscape roughness. J Chem Phys. 2009;131:195101.

[24] Munoz V, Thompson PA, Hofrichter J, Eaton WA. Folding dynamics and mechanism of β-hairpin formation. Nature. 1997;390:196-9.

[25] Collins JJ, Imhoff TT, Grigg P. Noise-enhanced information transmission in rat SA1 cutaneous mechanoreceptors via aperiodic stochastic resonance. J Neurophysiol. 1996;76:642-5.

[26] Russell DF, Wilkens LA, Moss F. Use of behavioural stochastic resonance by paddle fish for feeding. Nature. 1999;402:291-4.

[27] Simonotto E, Riani M, Seife C, Roberts M, Twitty J, Moss F. Visual perception of stochastic resonance. Phys Rev Lett. 1997;78:1186.

[28] Kato T, Fujita K, Kashimori Y. A neural mechanism of phase-locked responses to sinusoidally amplitude-modulated signals in the inferior colliculus. Biosystems. 2015;134:24-36.

[29] Lemarchand A, Gorecki J, Gorecki A, Nowakowski B. Temperature-driven coherence resonance and stochastic resonance in a thermochemical system. Phys Rev E. 2014;89.

[30] Guderian A, Dechert G, Zeyer KP, Schneider FW. Stochastic resonance in chemistry .1. The Belousov-Zhabotinsky reaction. J Phys Chem. 1996;100:4437-41.

[31] Hromadova M, Valasek M, Fanelli N, Randriamahazaka HN, Pospisil L. Stochastic Resonance in Electron Transfer Oscillations of Extended Viologen. J Phys Chem C. 2014;118:9066-72.

[32] Chattoraj S, Saha S, Jana SS, Bhattacharyya K. Dynamics of gene silencing in a live cell: stochastic resonance. J Phys Chem Lett. 2014;5:1012-6.

[33] Tuinstra RL, Peterson FC, Kutlesa S, Elgin ES, Kron MA, Volkman BF. Interconversion between two unrelated protein folds in the lymphotactin native state. Proceedings of the National Academy of Sciences of the United States of America. 2008;105:5057-62.

CHAPTER 6

Tethered WW domains from monomer to tetramer: folding competing with aggregation

WW domains are a family of fast-folding protein modules with three anti-parallel beta sheet structure (Fig. 6.1). The name came along due to the presence of two highly conserved tryptophan amino acids in these small 30-40 residues domains. WW is a binding module involved in apoptosis, among other functions [1]. WW domain's binding to a target protein is mediated by recognition of a proline rich region, which latches onto its loop 1 and hydrophobic pocket to facilitate the binding process. These versatile domains are also involved in transcriptional regulation [2].

WW domains have proven to be an excellent model for ultrafast folding experiments, for mechanistic experimental studies on the folding of a simple β sheet structure, and for benchmarking computational folding scenarios [3–5]. For WW domains with their loop 1 substructure optimized for folding thermodynamics and kinetics, formation of loop 2 becomes competitive as the rate-limiting step for folding. Indeed, optimization of the loop 2 sequence in FiP (FiP N30G/A31T/Q33T, FiP-GTT) produced a WW domain with a folding relaxation time of ~ 4 μs, approaching the speed limit for folding [6]. Another ultra-fast folding domain is the FBP28 2L (loop 2 replaced by β-hairpin, CLN025 with a ~100 ns folding time) which folds on ~ < 5μs [7].It is now also possible to [8,9] refolded these small proteins completely after pressure jumps *in silico*, joining equilibrium [10–12] and temperature jump simulations [6,13]. In this current study I have engineered a tethered construct by linking two or more copies of the fast folding Fip35 WW domain in the quest to understand fast folding competing with misfolding or aggregation.

Misfolding, binding and aggregation have already been studied extensively by experiments and computations [14,15].The problem that lies ahead is connecting experimental kinetics with current MD capabilities: most misfolding, and aggregation phenomena are really slow. They may take several seconds or hours instead of a few μs to ms. The tethered construct is not only an affordable system for conducting atomistic or coarse-grained MD simulations

but it also creates an effective higher concentration of the protein. This effective higher concentration enables competition between folding and aggregation on μs time scale. Using this approach of linking monomer units together, I can determine the nucleation size for aggregation which is usually very tedious to determine from bulk experiments. Similar tethering studies have been conducted on U1A protein [16], but here I aim at studying folding competing with misfolding/aggregation in μs time scale.

It has been demonstrated that repeat proteins also provide rich insights into both energetics and kinetics of folding. Recent work on a repeat protein model by Robert Best and co-workers revealed that a protein's tendency to misfold depends largely on the relative stability of the domains present in the folded or misfolded intermediates rather than size of the barriers [17]. Another set of experiments on consensus Ankyrin repeat proteins (CARPs) provided evidence of parallel folding pathways. Increase in folding rates with the addition of more repeats and size of the CARPs supported the idea of parallel folding pathways [18]. Ising-like model was employed to analyze repeat-protein thermodynamics and relaxation kinetics [19].

Over the years a lot of progress has been made to understand the dynamics of repeat proteins but yet there exists only few studies that makes a direct comparison of experiments with simulations. One such current investigation on tandem repeats of immunoglobin-like domains of titin claimed that it is not evolutionary beneficial to have higher sequence identity within repeat proteins as the tendency to form more stable misfolded states is more when the neighboring repeat domains have high sequence similarity [20].

Tandem repeats of WW domains are utilized by nature to have better control over cellular regulation.Furman and co-workers published a detailed overview on how tandem repeat module facilitate fine-tuning of regulation inside cells, specifically describing the variety of ways in which two or more tandem repeats of WW domains cooperate or interact in binding to their polyproline rich ligands. A few possible ways as mentioned by the authors are 1) Additive binding- repeats domains bind to their own specific targets contributing to an overall increase in binding affinity. 2) Chaperone effect- one domain assist the binding of other 3) Binding induced binding 4) Adjacent WW domain can change the dynamics and stability of the neighboring domain [1]. The vital role in cellular regulation of the family of tandem repeat WW Domains provides an additional motivation for me to investigate this system in mechanistic details via both ultra-fast laser temperature jump experiments and molecular dynamic simulations. With my tethered protein construct experiments I have observed that

adding more monomer units led to thermodynamic destabilization and slower folding rates, along with an abrupt onset of protein-protein interaction for the tetramer. I performed relaxation kinetics using ultrafast laser temperature jump experiments at different temperature and denaturant concentrations. Finally, I proposed a simplified multimeric network model which can globally fit the thermodynamics and kinetics data. As computation of folding in the 50-500 μs range has become feasible, I believe that my data presented in this study will prove to be a rich resource for detailed comparisons, providing constraints on mechanisms and rate changes deduced from molecular dynamics simulations for folding/misfolding of repeat WW domains.

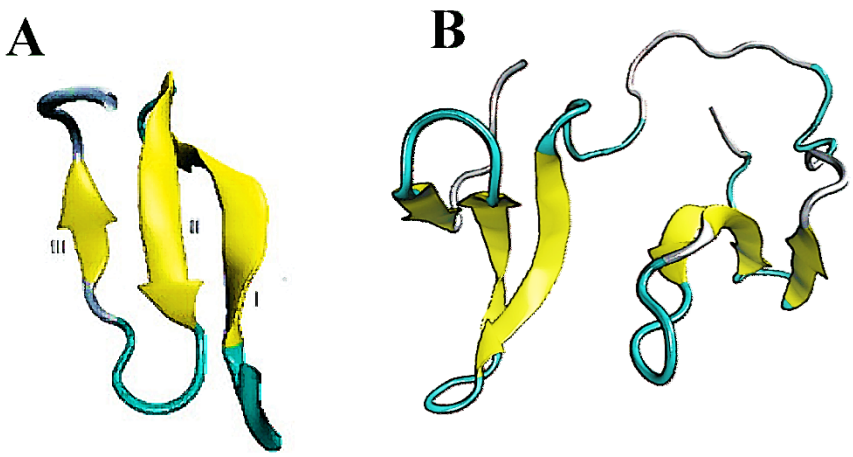

Fig. 6.1: New cartoon representation of Fip35 monomer (pdb code: pin1) (A) and tethered dimer Dfip35 (B). Dfip35 was constructed by connecting two identical Fip35 monomers together via a 10 amino acid flexible linker composed of (GSG) units.

6.1 Methods

6.1.1 Protein sample preparation

For the monomer (Mfip35) and tetramer (Qfip35) constructs a plasmid encoding a fusion protein consisting of Glutathione-S-transferase (GST), a thrombin cleavage site, and protein sequence was cloned into pDream (GenScript) as mentioned in ref.[21]. Briefly, the fusion protein construct was expressed in BL21 (DE3)-RIPL (Agilent) E. coli and captured and purified from the cell extract on an immobilized glutathione resin according to manufacturer's guidelines (GenScript). The protein was eluted by 10 mM glutathione in 50 mM Tris-HCl pH

8.0 and followed by dialysis in 10 mM sodium phosphate buffer. FiP35 was cleaved from the purification tag by overnight incubation with biotinylated thrombin (EMD Millipore). Thrombin was removed by incubation with streptavidin-agarose resin (EMD Millipore) according to manufacturer's protocol. Monomer was purified from cleaved GST via an ultrafiltration cell with 10 kDa cutoff membrane (Millipore) whereas due to comparable size of GST and Qfip35 the separation was performed by passing the cleaved protein solution through a gravity column with immobilized glutathione resin. The procedure for other dimer and trimer construct was similar except both of them were purified using a His-Tag. The presence of single tryptophan on the first β-strand (loop or hairpin 1) of a single WW Domain enabled monitoring of folding via fluorescence.

6.1.2 Temperature unfolding thermodynamics

Temperature denaturation of all constructs was measured by tryptophan fluorescence and circular dichroism. Fluorescence spectroscopy was carried out using a Jasco fluorescence spectrophotometer FP-8300 equipped with programmable temperature control with excitation and emission slit widths kept at 5 nm. Tryptophan excitation was 280 nm and emission was collected from 290 – 450 nm. For each fluorescence emission spectrum, the average wavelength <λ> was calculated by equation (1) where I is intensity and λ wavelength [22]:

$$\langle\lambda\rangle = (\sum_j \lambda_j I_j) / (\sum_j I_j) \quad (1)$$

The same wavelength range was used in all cases to obtain consistent results. Circular dichroism was measured using a JASCO spectrophotometer with Peltier temperature control (JASCO Inc, Easton MD). All spectra were recorded from 250 – 200 nm at a scan rate of 50 nm/min with 1 nm resolution and are an average of 5 accumulations. Measurements were conducted in a 1 mm path length quartz cuvette and, unless otherwise noted, at a protein concentration of 10 μM.

All thermodynamic denaturation signals $S(X)$, where X is temperature, were fitted to a two-state model for temperature denaturation

$$S(X) = S_U + S_F e^{-\Delta G(X)/RT} / (1 + e^{-\Delta G(X)/RT}) \quad (2a)$$

$$\Delta G(X) = g_X (X - X_m) \quad (2b)$$

to obtain the denaturation midpoints with respect to temperature (T_m). All of the protein constructs were reversible in the concentration range used for the experiments.

6.1.3 Temperature jump kinetics

Laser temperature jumps were carried out using a Surelite Q-switched Nd:YAG laser (Continuum Inc., Santa Clara, CA), with details of the instrument mentioned elsewhere [23,24].The jump size was 5-6°C. The exact size of the jump was calibrated by comparing the fluorescence decays *f* of tryptophan (300 μM solution) after the jump with the corresponding decay at an equilibrium temperature several degrees higher. Fluorescence decays were excited at 280 nm by a tripled, mode-locked Ti:sapphire laser every 12.5 ns for a total of 1 ms. The temperature jump was set to occur 153.75 μs after the oscilloscope was triggered to start data collection. The sampling frequency was 10 Giga-samples per second. Thus each fluorescence decay was sampled at 100 picosecond intervals, or 125 times before the next decay was excited. The signal was usually 50-60 mV. Sample concentrations were 40 μM for all of the proteins with the exception of Qfip35 for which only 25 μM was used.

6.1.4 Kinetics data

Kinetics data were analyzed using MATLAB (Mathworks Inc., Natick, MA) and IGOR Pro (Wavemetrics Inc., Lake Oswego, OR). A fluorescence decay f(t) was collected every 12.5 ns. 100 of these were binned into intervals of 1.25 μs. Thus the protein kinetics could be followed with 1.25 μs time resolution. The decays f(t) were fitted to a linear combination of the decay f1 averaged between 153.75 and 28.75 μs before the T-jump, and the decay f2 averaged over the final 125 μs of data collection, where the protein had equilibrated. The relative lifetime shift as a function of time, χ(t). The χ(t) traces were fitted using the model described below.

6.1.5 Multimeric network model

A simplified multimeric network model was built for fitting the experimental data. In this model each monomer units can attain in any of the three forms namely folded (N), misfolded (M) and unfolded (U). For example for the case of monomer there exists only 3 total states whereas for dimer the total possible states will be 3^2 and for general system containing n monomer repeats the total number of states will be calculated as 3^n. States like NU and UN

are distinguishable in my model as in the tethered construct N terminal of one domain is connected to C terminal of the other. The thermodynamics of each of n-mer has been calculated by assuming a Boltzmann's distribution and fitting to experimental fluorescence data. Experimental fluorescence thermal melts do not show two apparent co-operative transitions hence the model considered the misfolded baseline (Sm) to be an average of the folded (Sf) and unfolded baselines (Su) (see Appendix E for details). The model also included pairwise interaction between the folded (NN) and misfolded (MM) units. These nearest neighbor interaction terms were added in the free energy equations in a way that more of these interactions will stabilize the native or misfolded state. This model assumes an off-pathway intermediate meaning that direct N to M transition is forbidden, it has to go to U first and then to M. In order for the model to mimic the T-jump relaxation experiments we first equilibrated the system at the initial temperature to obtain relative concentrations of all the species (dimer NN, NU, UN, NM, MN UM, MU, MM, UU) and later jumped the temperature to the desired experimental temperature to obtain kinetics solving the master equation. Similar types of models have reported earlier for fitting experimental folding data [14,21,22,23].

6.2 Results

6.2.1 Decrease in thermal stability as more monomer units are added

The thermal stability of the tethered n-mer constructs was measured by probing the only tryptophan (present in the hairpin 1 in each monomer WW Domain) over a temperature range of 5-90 °C by both circular dichroism and fluorescence spectroscopy. The thermal melts were performed with varying concentrations of guanidine hydrochloride to obtain the melting temperature (Tm) with better accuracy. It was observed that dimer stability was similar compared to monomer but when more monomer units were tethered the construct became thermally unstable (see Table 6.1 and Appendix E). The expression yield of tetramer construct decreased significantly than the others. The dimer yielded ~ 12 mg for a three liter expression whereas only 3-5 mg of protein was obtained for the tetramer. The tetramer solution also turned turbid as fractions were collected on the FPLC.

Table 6.1: Thermodynamic data for the tethered protein constructs

Fluorimeter			Circular Dichroism	
Protein	**T_m (°C)**	**g_1 (J mol^{-1} K^{-1})**	**T_m (°C)**	**g_1 (J mol^{-1} K^{-1})**
Fip35	**82(1)**	**405(22)**	**78(1)**	**268(10)**
Dfip35	**83(2)**	**290(12)**	**78(1)**	**261(6)**
Tfip35	**79(1)**	**348(8)**	**72(1)**	**312(6)**
Qfip35_GST	**67(3)**	**291(10)**	**64(1)**	**260(2)**
Qfip35_His	**83(2)**	**275(10)**	**67(2)**	**378(16)**

* Qfip35_His is Qfip35 purified using His tag

6.2.2 Effect of purification tag on Qfip35 protein structure

The tetramer (Qfip35) was purified using GST and the tag was later cleaved as mentioned in the method section and ref [28]. In order to confirm the presence of purified protein the sample was run on SDS gel and a clear band at ~ 17 KDa was seen. The purified protein was also characterized using MALDI (see Appendix E) and a clear peak at 17.76 KDa was seen for the cleaved Qfip35. I performed circular dichroism spectroscopy on the sample to my surprise the spectrum didn't show a typical WW domain spectrum (peak at around 227 nm) but instead had a CD looking closer to random coil see (Fig. 6.2). I also conducted thermal melt on GST purified Qfip35 by probing the tryptophan 280 nm and monitoring the spectrum from 290-450 nm (see Fig. 6.3). The fluorimeter traces were noisy but showed co-operative transition. The expression was repeated atleast thrice to get similar results. Interestingly, when Qfip35 purification was conducted using an attached His-tag the CD spectrum now resembled to that of a typical WW domain with a peak at 227 nm (see Fig.6.2).I obtained similar characterization results as before using mass and gel electrophoresis for Qfip35.The thermal denaturation

midpoint for this sample came out to be similar to monomer (Table 6.1). Based on the above results same protein when purified using two different tags resulted in different structural folds.

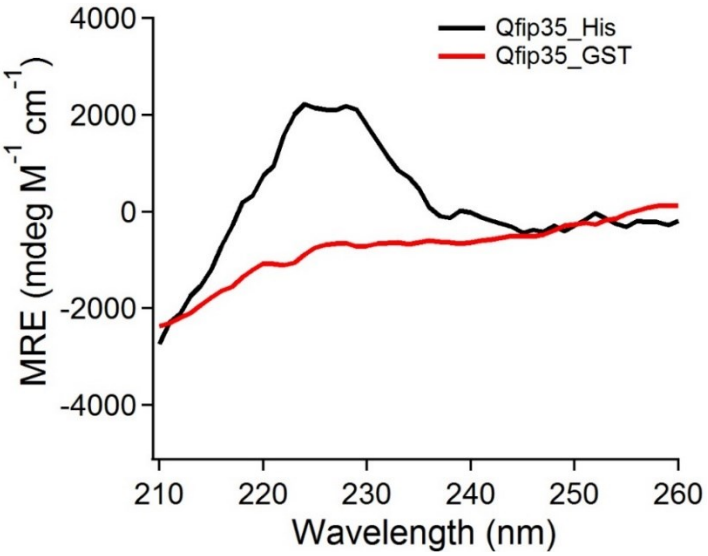

Fig. 6.2: Comparison of Qfip35 (tetramer) expressed and purified using GST and His tag using Circular dichroism at 25 °C. The typical 227 nm peak for the WW domain is present in Qfip35_His protein but not in the spectra obtained for Qfip35_GST

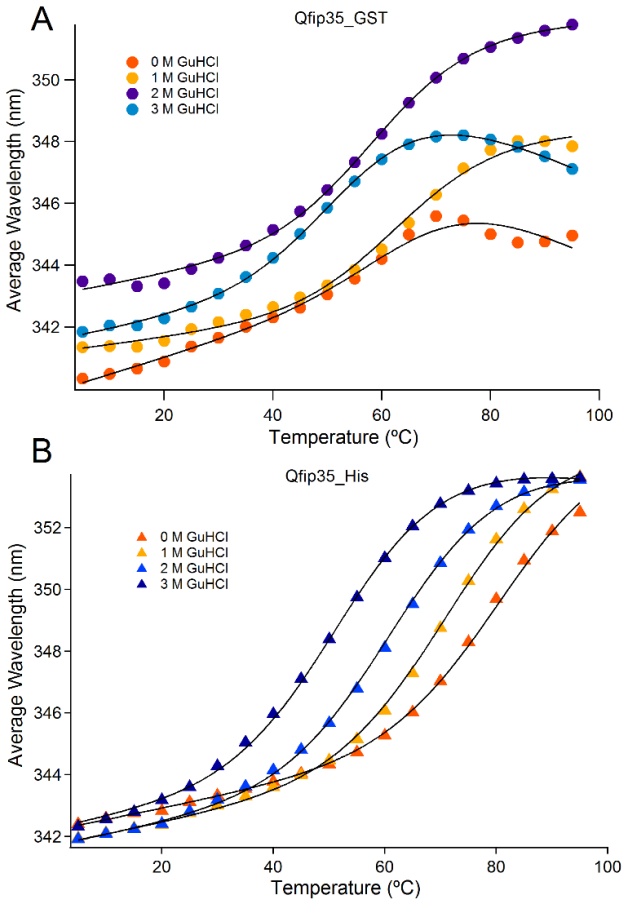

Fig. 6.3: Comparison plot of Average wavelength vs temperature plot for Qfip35_GST and Qfip35_His. Melting temperature is higher and data is less noisy in the case of Histag purification. Addition of GuHCl shifts the folded baseline in Qfip35_GST.

6.2.3 Global fitting of kinetics and thermodynamics using multimeric model

The thermodynamics of monomer to tetramer construct was conducted by probing the tryptophan at 280 nm and monitoring the emission spectrum as function of increasing temperature. The fluorimeter data was collected at varying GuHCl concentration (1,2,3 M) to obtain better unfolding baselines. Circular dichroism spectroscopy was also done on these tethered constructs in order to observe structural changes in the protein when subjected to increase in temperature.

In order to determine the relaxation kinetics I conducted temperature jump relaxation experiments on all the tethered constructs. The jumps were conducted near and below the

melting temperature using our inbuilt ultrafast laser temperature jump setup described above in the method section. The kinetics experiments were done at different temperature and GuHCl concentrations. In order to globally fit the thermodynamics and kinetics data for all of the protein a simplified multimeric model was built descried in method section and Appendix E. Briefly the model consist of 14 parameters overall. The thermodynamics was represented using 12 parameters and the remaining 2 parameters were for barrier heights going from N to U and U to M (see Table 6.2). The unfolded state was used as reference state and the free energy was written as a Taylor series expansion across the Tm (see equation 1 in Appendix E).This simplified model populates the intermediate species (cyan shades see Fig. 6.4) at higher temperature and GuHCl concentration. Data was globally fitted assuming a Boltzmann distribution for a full set of multimeric structures such as for tetramer NNNN or UMMN. The fitted parameter values are shown in Table 6.2. Thermodynamics was fitted with an effective Tm of ~ 83 °C with both unfolded and folded baseline linked across the data set (see Fig. 6.5). The kinetic data was globally fitted with a relatively large (~ 17 KJ/mole) barrier (Fig.6.6).

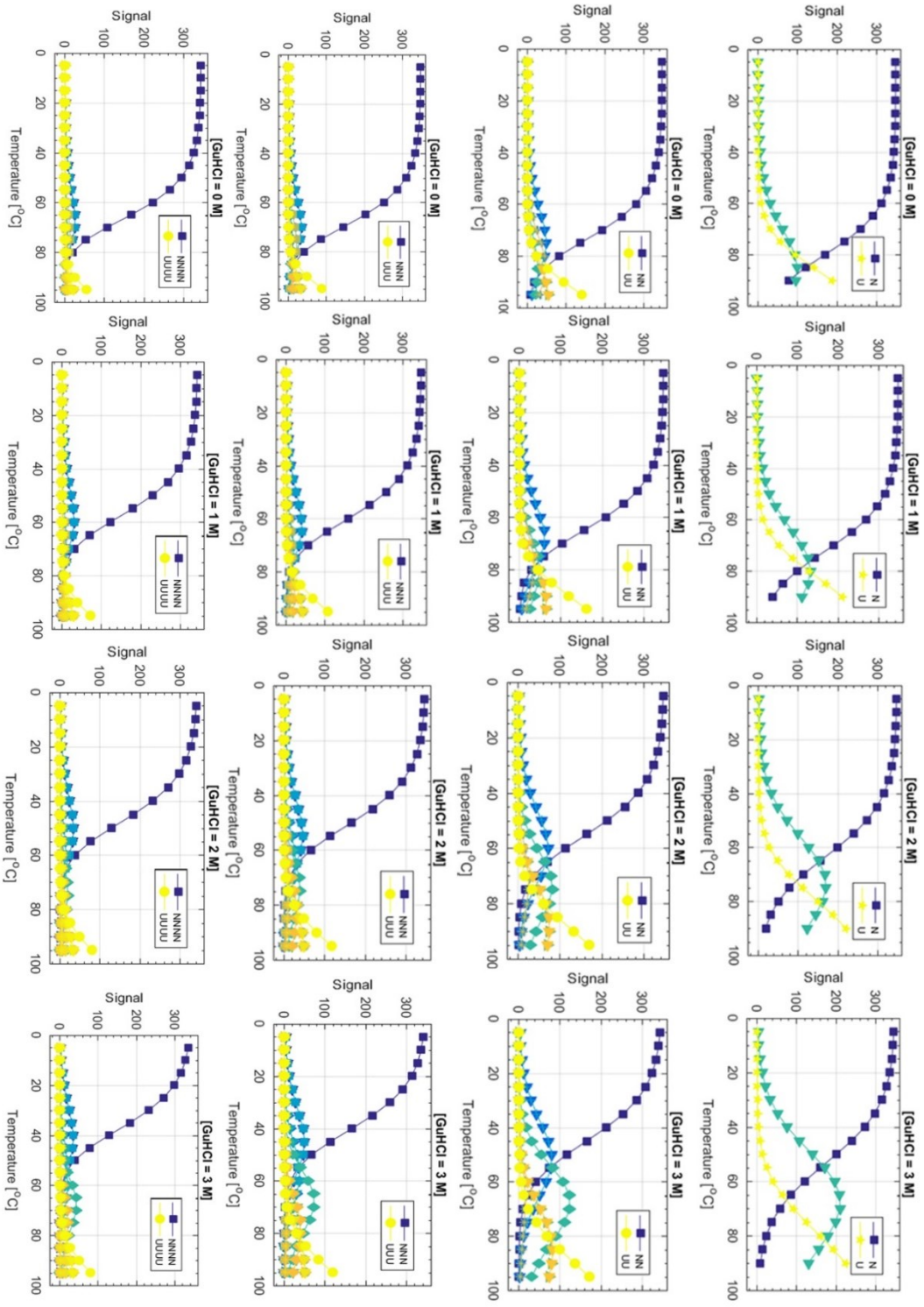

Fig. 6.4: Signal vs Temperature plot Top panel to bottom panel Mfip35 to Qfip35. The change in folded state population is represented as the blue trace similarly the yellow trace represents population change in the unfolded state. Intermediate species are populated at high temperature and denaturant concentration and are plotted in colors other than blue and yellow.

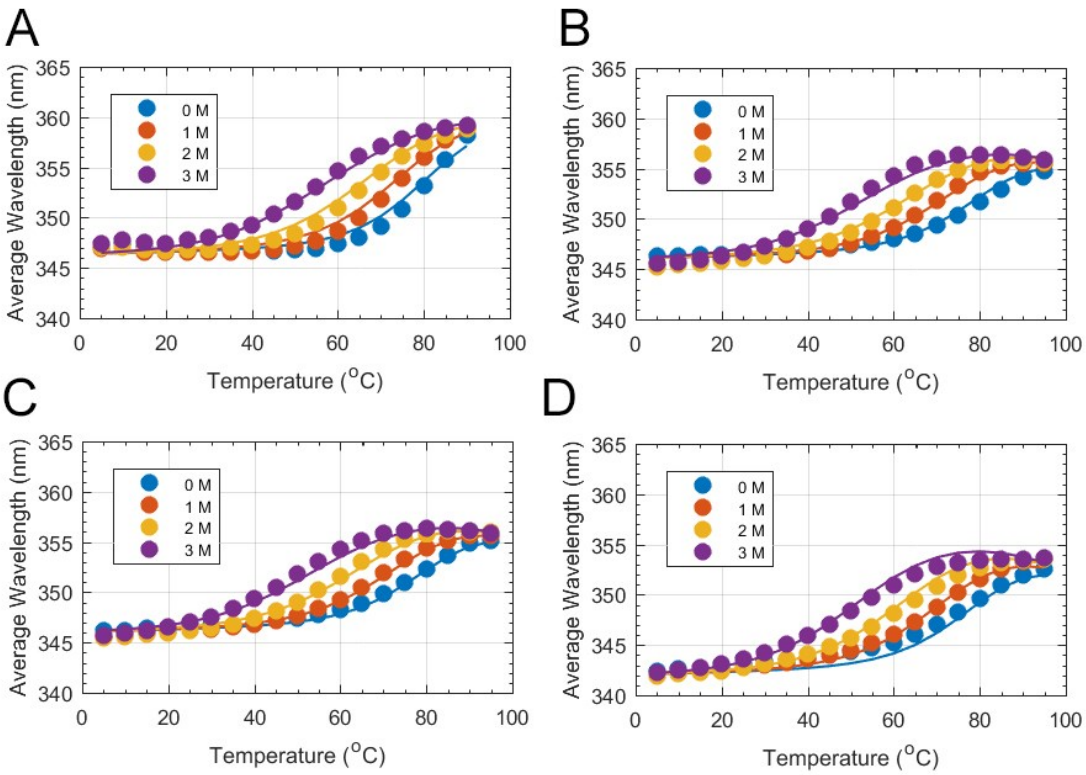

Fig. 6.5: Global Thermodynamic Fitting: Signal (Average wavelength) vs Temperature plot

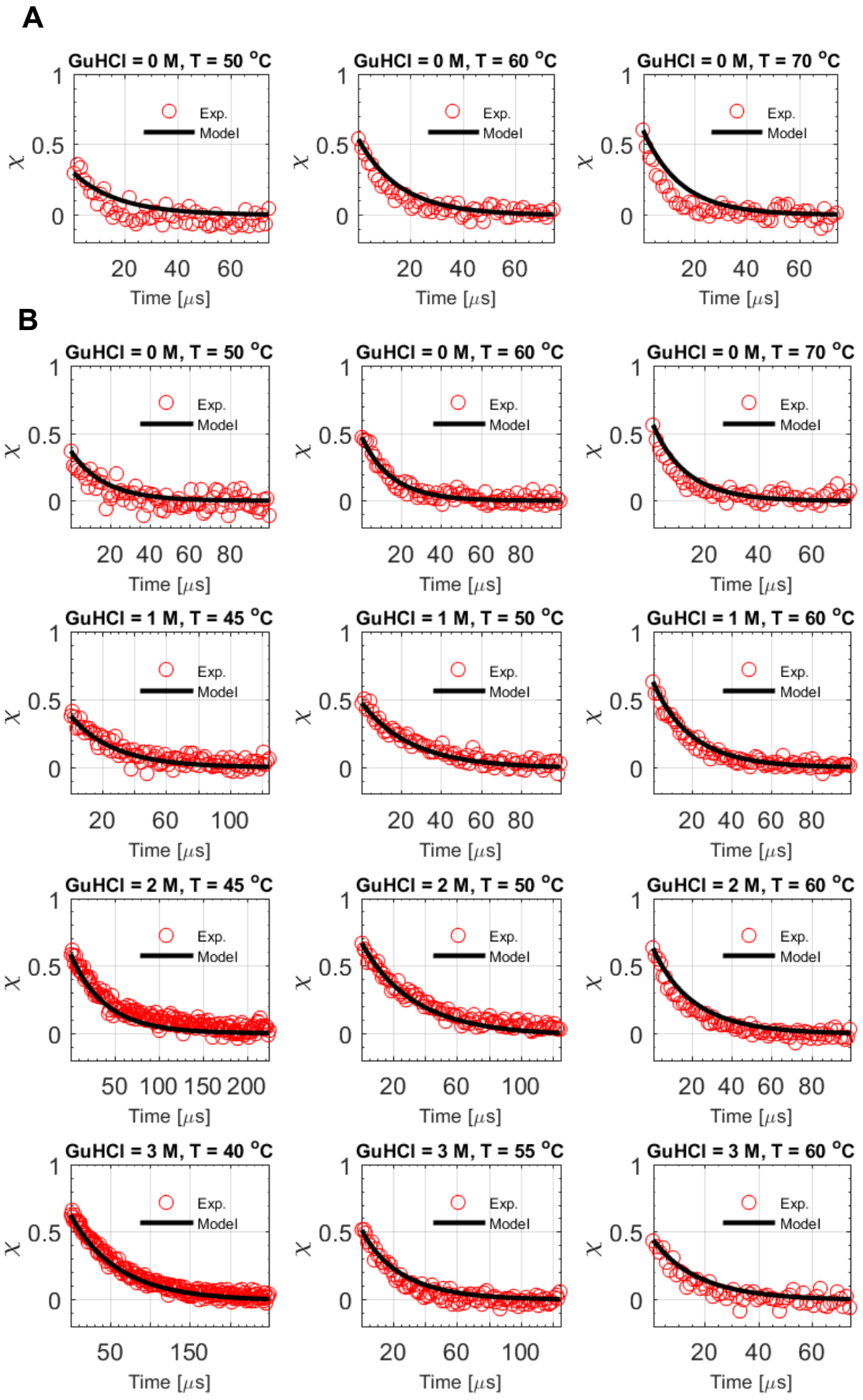

Figure 6.6 (cont.)

Figure 6.6 (cont.)

C

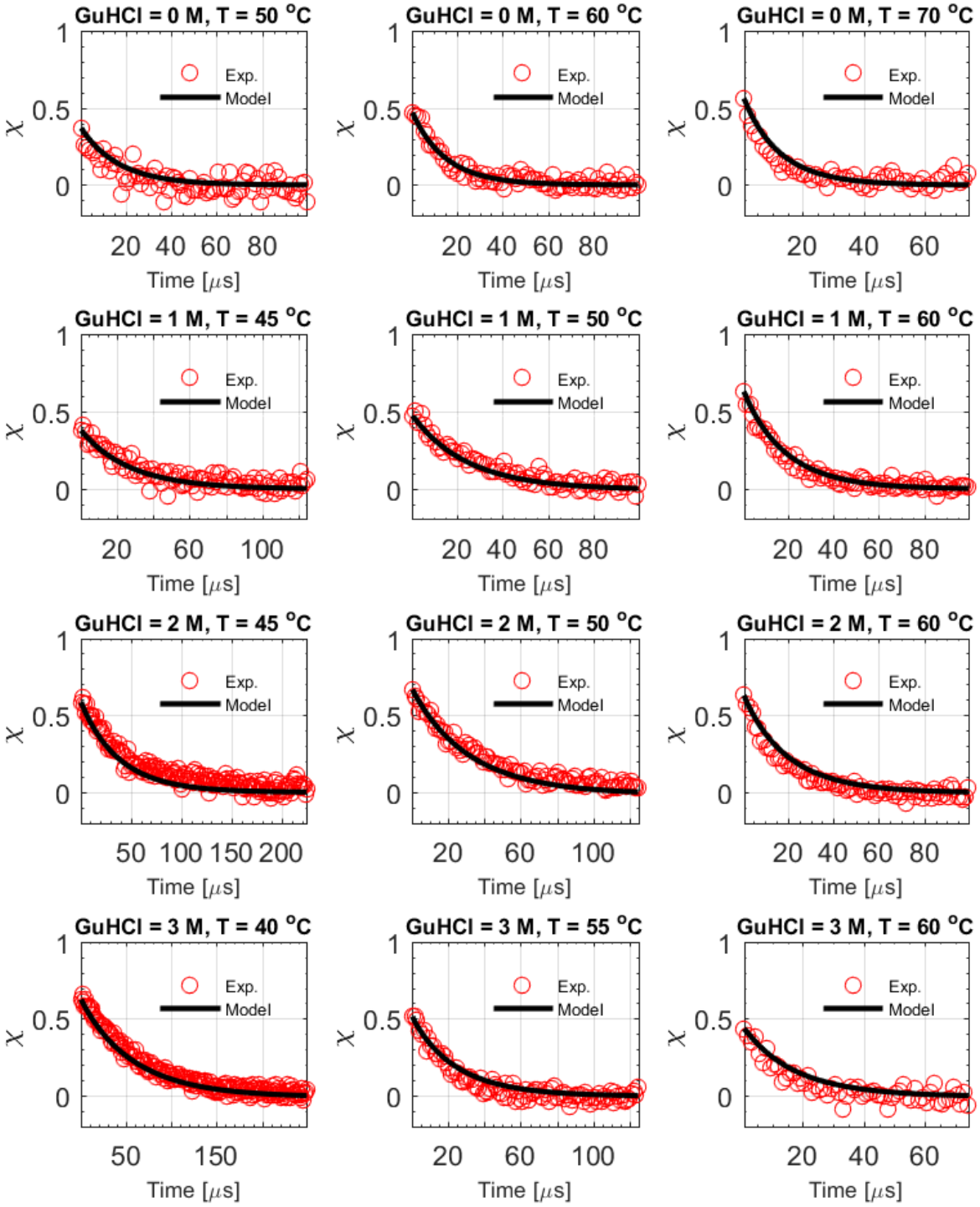

Figure 6.6 (cont.)

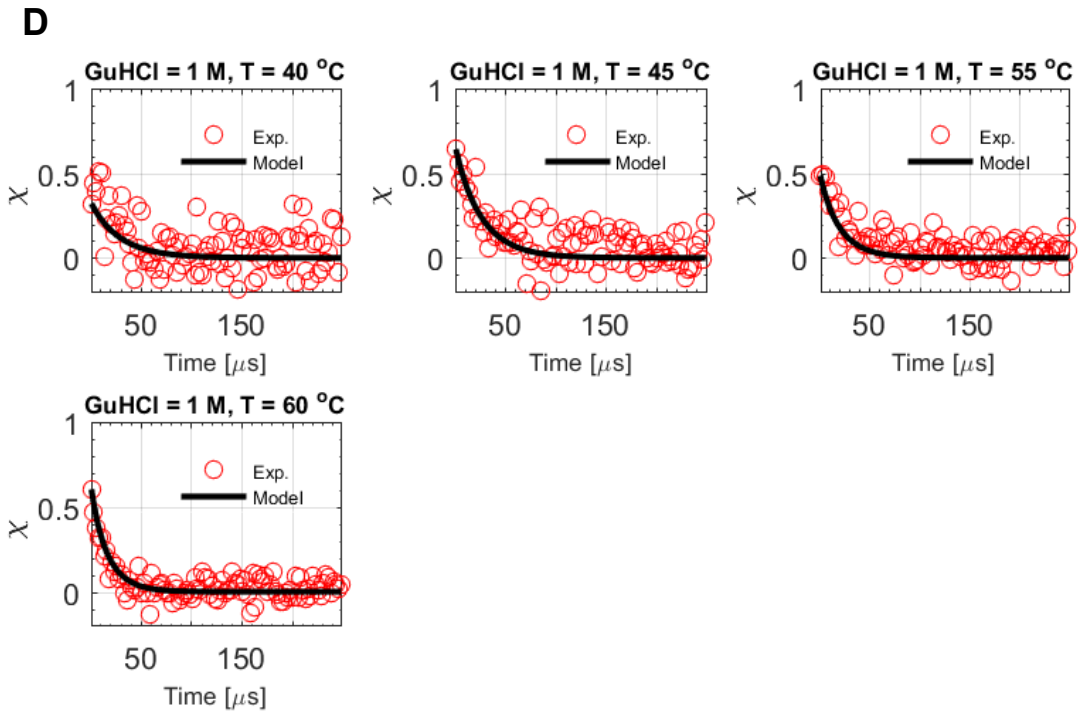

Fig. 6.6: Temperature jump relaxation kinetics χ(t) vs time traces for A) Mfip35 B) Dfip35 C) Tfip35 D) Qfip35_GST

Table 6.2: Fitted model parameters

Parameters	Model fitted values
T_m (°C)	84.3 (0.4)
g31 (J/mol K)	436 (15)
g32 (J/mol K)	201 (13)
go32 (J/mol K)	809 (298)
gg31 (J/mol K)	2481(132)
gg32 (J/mol K)	-129 (115)
gnn (J/mol K)	-2.22E-14
gmm (J/mol K)	-2.22E-14
bu	359.7 (0.2)
mu	-0.19 (0.01)
bf	347.1 (0.2)
mf	0.011 (0.001)
Gǂ13 (J/mol K)	17170 (7972)
Gǂ23 (J/mol K)	2626.5

* N(folded)=1 ; M(Intermediate)=2 ; U(unfolded)=3

*Gǂ13 and Gǂ23 are barrier height going from folded to unfolded or intermediate to unfolded form

6.3 Discussion and conclusion

The idea of systematically designing tethered proteins presented in this chapter to examine the stability and kinetics of different aggregate structures is derived by connecting two interesting repeat protein studies on notch ankyrin repeats [29–31] and Iκβα isoforms [32,33]. These repeat protein studies demonstrated that transient interactions between monomers in a repeat protein can be tweaked by mutations or changing the number of repeats. My approach of tethering creates an effective higher concentration of protein and also speed up aggregation of a fast folding protein. The thermal denaturation experiments on the tethered constructs monitored by circular dichroism spectroscopy revealed that the interaction between the different domains lead to decrease in stability (melting temperature decreases) of the overall construct when more monomer units are added. This trend is consistent regardless of the tetramer being expressed and purified with a GST or histidine tag see Table 6.1. However, when probed by fluorescence a decrease in stability is seen with the exception of Qfip35_His, the stability of this construct is comparable to that of the monomer domain (Mfip35). One possible explanation for this could be that as more units are added to make the tetramer two or three of the domains interact leaving one of WW domain alone which gives rise to an overall increase in stability. It is worth mentioning here that Qfip35_His melting temperature probed by CD and tryptophan fluorescence differ by more than 10 degrees. Experimentally, the hallmark of two-state folding is to obtain the same melting temperature (T_m) via different spectroscopic measurement techniques that each probe different parts of the energy landscape [34,35]. Such behavior breaks down when intermediates are populated during the course of the folding. This evidence highlights that Qfip35_His is not an apparent two state folder. In the literature the effect of GST tag on target protein is not clear and conflicting results exists for the same. In my experiment the tetramer was purified using both the six histidine affinity tag (Histag) and also the GST (26 K Da). The report by Speicher and Harper claimed that using GST as a fusion tag can provide chaperoning which help the target protein to fold properly. The article also reported GST being capable of yielding more soluble protein by avoiding the protein going to the inclusion bodies [36]. On the contrary ref [37] proposed that GST as a fusion tag is a poor solubility and affinity tag as it has four exposed cysteine residues which provide oxidative aggregation. This makes it a bad choice for tagging oligomeric target proteins [38]. In fact I observed in our circular dichroism experiments that QFip35 cd signal varies depending on the tag I used for purification (see Fig. 6.2). The difference seen in the cd structure of the repeat protein may be due to the interference by the GST tag during the

expression and purification steps. This is an interesting finding as now I have a protein which is trapped in some kind of an intermediate state which when subjected to increasing temperature undergoes co-operative unfolding. Kinetic relaxation experiments on the Qfip35_His are yet to be performed and I am working towards that end.

Comparing the thermodynamics of the tethered constructs the average wavelength range for the monomer to trimer is around (345- 358 nm) see Fig. 6.5 but in case of the tetramer folded baseline intercept is shifted more towards the blue (342 nm) indicating that the tryptophan molecules in this protein are more buried (less exposed to water). Fig. 6.7 shows the plot of mean residue ellipticity for all the protein constructs and it is intriguing to see that mre (cd signal normalized for the number of peptide bonds in the protein) for them don't overlap. Dimer and trimer have mre values that are twice as compared to monomer whereas the Qfip35_His has values similar to that of the monomer. One possible explanation for such behavior can be that the typical cd signature (peak at ~227 nm) for family of WW domains arises from tryptophan coupling [39].The difference in mre values between the tethered constructs can come from interaction between the neighboring domains in the dimer and trimer system giving rise to enhanced tryptophan coupling and higher mre values. Whereas for the case Qfip35_His lower mre could be a result of either all four domains being folded independently (there exist no interaction) or two domains interacting while the other two domains are misfolded giving mre values that are similar to that of the monomer.

The simplified model described in the method section was able to fit the relaxation kinetics and thermodynamics globally. The model included the interaction terms gnn and gmm in the free energy of the system but it turns out that their contribution to the overall free energy is not significant. The kinetics fitted to a barrier height going from folded to the unfolded state to be around ~17 KJ/mole. The model provide rate between all of the multimeric states starting from NNNN to UUUU. The model is flexible to include other interaction terms in the free energy equations (see Appendix E for more details).The current study can serve as future benchmark for protein-protein interactions simulations (coarse-grained or all atom) as unfold/ fold on a time scale of few 10's of μs which is not very computationally expensive. These simulations will reveal details about the nature of the misfolded states whether domain swapped structures are formed or the protein from random clumps (hydrophobic interactions).

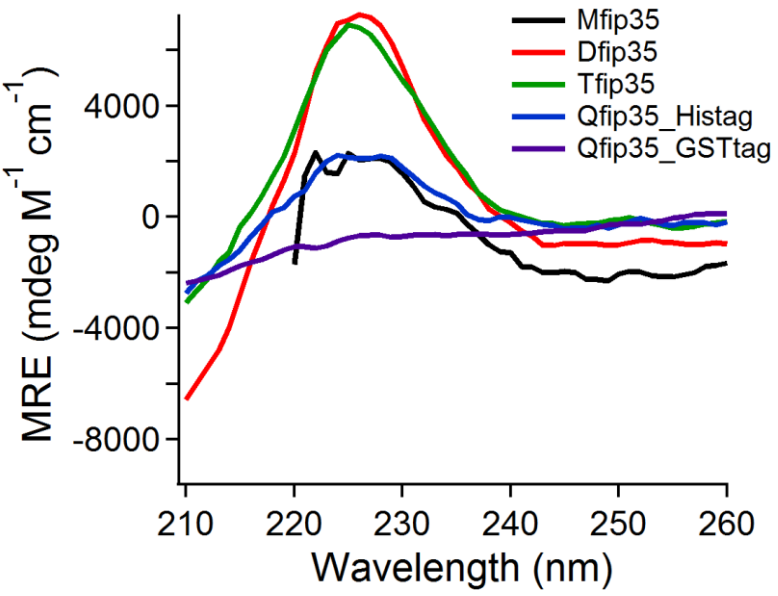

Fig. 6.7: Plot of MRE vs wavelength for all the tethered constructs. Qfip35_GST showed no typical 227 nm peak seen for WW domains.

6.4 References

[1] E.J. Dodson, V. Fishbain-Yoskovitz, S. Rotem-Bamberger, O. Schueler-Furman, Versatile communication strategies among tandem WW domain repeats, Exp. Biol. Med. 240 (2015) 351–360. doi:10.1177/1535370214566558.

[2] R. Yagi, L. Chen, K. Shigesada, A WW domain-containing Yes-associated protein (YAP) is a novel transcriptional co-activator, EMBO. (1999). http://emboj.embopress.org/content/18/9/2551.abstract (accessed May 17, 2017).

[3] J. Kubelka, J. Hofrichter, W.A. Eaton, The protein folding “speed limit,” Curr. Opin. Struct. Biol. 14 (2004) 76–88.

[4] R. Schaeffer, A. Fersht, V. Daggett, Combining experiment and simulation in protein folding: closing the gap for small model systems, Curr. Opin. Struct. Biol. (2008). http://www.sciencedirect.com/science/article/pii/S0959440X0700200X (accessed May 17, 2017).

[5] F. Cecconi, C. Guardiani, R. Livi, Testing simplified proteins models of the hPin1 WW domain, Biophys. J. (2006).

http://www.sciencedirect.com/science/article/pii/S0006349506717684 (accessed May 17, 2017).

[6] S. Piana, K. Sarkar, K. Lindorff-Larsen, M. Guo, Computational design and experimental testing of the fastest-folding β-sheet protein, J. Mol. (2011). http://www.sciencedirect.com/science/article/pii/S0022283610011319 (accessed May 4, 2017).

[7] C.M. Davis, R.B. Dyer, Dynamics of an Ultrafast Folding Subdomain in the Context of a Larger Protein Fold, J. Am. Chem. Soc. 135 (2013) 19260–19267. doi:10.1021/ja409608r.

[8] A. Wirth, Y. Liu, M. Prigozhin, K. Schulten, Comparing fast pressure jump and temperature jump protein folding experiments and simulations, J. (2015). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4794261/ (accessed May 4, 2017).

[9] M. Prigozhin, Y. Liu, A. Wirth, Misplaced helix slows down ultrafast pressure-jump protein folding, Proc. (2013). http://www.pnas.org/content/110/20/8087.short (accessed May 22, 2017).

[10] K. Lindorff-Larsen, S. Piana, R.O. Dror, D.E. Shaw, How fast-folding proteins fold, Science (80-.). 334 (2011) 517–520. doi:10.1126/science.1208351.

[11] D. Shaw, P. Maragakis, K. Lindorff-Larsen, Atomic-level characterization of the structural dynamics of proteins, (2010). http://science.sciencemag.org/content/330/6002/341.short (accessed May 17, 2017).

[12] S. Piana, K. Lindorff-Larsen, Protein folding kinetics and thermodynamics from atomistic simulation, Proc. (2012). http://www.pnas.org/content/109/44/17845.short (accessed May 17, 2017).

[13] M.B. Prigozhin, M. Gruebele, Microsecond folding experiments and simulations: a match is made, Phys. Chem. Chem. Phys. PCCP. 15 (2013) 3372–3388. doi:10.1039/c3cp43992e.

[14] C. Dobson, Principles of protein folding, misfolding and aggregation, Semin. Cell Dev. Biol. (2004). http://www.sciencedirect.com/science/article/pii/S1084952103001137 (accessed May 17, 2017).

[15] R.O. Dror, R.M. Dirks, J.P. Grossman, H. Xu, D.E. Shaw, Biomolecular simulation: a computational microscope for molecular biology, Annu. Rev. Biophys. 41 (2012) 429–452.

[16] W.Y. Yang, M. Gruebele, Binary and ternary aggregation within tethered protein constructs., Biophys. J. 90 (2006) 2930–2937. doi:10.1529/biophysj.105.075846.

[17] P. Tian, R.B. Best, Structural Determinants of Misfolding in Multidomain Proteins, PLoS Comput. Biol. 12 (2016) 1–28. doi:10.1371/journal.pcbi.1004933.

[18] T. Aksel, D. Barrick, Direct observation of parallel folding pathways revealed using a symmetric repeat protein system, Biophys. J. 107 (2014) 220–232.

[19] T. Aksel, D. Barrick, Chapter 4 Analysis of Repeat-Protein Folding Using Nearest-Neighbor Statistical Mechanical Models, 1st ed., Elsevier Inc., 2009. doi:10.1016/S0076-6879(08)04204-3.

[20] A. Borgia, K.R. Kemplen, M.B. Borgia, A. Soranno, S. Shammas, B. Wunderlich, D. Nettels, R.B. Best, J. Clarke, B. Schuler, Transient misfolding dominates multidomain protein folding, Nat. Commun. 6 (2015) 8861. doi:10.1038/ncomms9861.

[21] F. Liu, M. Nakaema, M. Gruebele, The transition state transit time of WW domain folding is controlled by energy landscape roughness, J. Chem. Phys. (2009). http://aip.scitation.org/doi/abs/10.1063/1.3262489 (accessed May 18, 2017).

[22] A.J. Wirth, Y. Liu, M.B. Prigozhin, K. Schulten, M. Gruebele, Comparing Fast Pressure Jump and Temperature Jump Protein Folding Experiments and Simulations, J. Am. Chem. Soc. (2015).

[23] J. Ervin, J. Sabelko, M. Gruebele, Submicrosecond real-time fluorescence sampling: application to protein folding, J. Photochem. (2000). http://www.sciencedirect.com/science/article/pii/S1011134400000026 (accessed May 4, 2017).

[24] R. Ballew, J. Sabelko, C. Reiner, A single-sweep, nanosecond time resolution laser temperature-jump apparatus, Rev. Sci. (1996). http://aip.scitation.org/doi/abs/10.1063/1.1147137 (accessed May 4, 2017).

[25] E. Kloss, N. Courtemanche, D. Barrick, Repeat-protein folding: New insights into origins of cooperativity, stability, and topology, Arch. Biochem. Biophys. 469 (2008)

83–99. doi:10.1016/j.abb.2007.08.034.

[26] V. Muñoz, W.A. Eaton, A simple model for calculating the kinetics of protein folding from three-dimensional structures., Proc. Natl. Acad. Sci. U. S. A. 96 (1999) 11311–6. doi:10.1073/PNAS.96.20.11311.

[27] E. Henry, R. Best, W. Eaton, Comparing a simple theoretical model for protein folding with all-atom molecular dynamics simulations, Proc. (2013). http://www.pnas.org/content/110/44/17880.short (accessed May 18, 2017).

[28] F. Liu, M. Nakaema, M. Gruebele, The transition state transit time of WW domain folding is controlled by energy landscape roughness, J. Chem. Phys. 131 (2009) 0–9. doi:10.1063/1.3262489.

[29] T. Aksel, A. Majumdar, D. Barrick, The contribution of entropy, enthalpy, and hydrophobic desolvation to cooperativity in repeat-protein folding, Structure. (2011). http://www.sciencedirect.com/science/article/pii/S0969212611000281 (accessed May 22, 2017).

[30] E. Vieux, D. Barrick, Deletion of internal structured repeats increases the stability of a leucine-rich repeat protein, YopM, Biophys. Chem. (2011). http://www.sciencedirect.com/science/article/pii/S030146221100192X (accessed May 22, 2017).

[31] T. Aksel, A. Majumdar, D. Barrick, Determinants of Cooperativity in Repeat Protein Folding, Biophys. J. (2011). http://www.cell.com/biophysj/abstract/S0006-3495(10)04523-6 (accessed May 22, 2017).

[32] C.H. Croy, S. Bergqvist, T. Huxford, G. Ghosh, E.A. Komives, Biophysical characterization of the free IκBα ankyrin repeat domain in solution, Protein Sci. 13 (2004) 1767–1777. doi:10.1110/ps.04731004.

[33] D.U. Ferreiro, S.S. Cho, E.A. Komives, P.G. Wolynes, The energy landscape of modular repeat proteins: Topology determines folding mechanism in the ankyrin family, J. Mol. Biol. 354 (2005) 679–692. doi:10.1016/j.jmb.2005.09.078.

[34] G. Huang, T. Oas, Structure and Stability of Monomeric. lambda. Repressor: NMR Evidence for Two-State Folding, Biochemistry. (1995). http://pubs.acs.org/doi/pdf/10.1021/bi00012a003 (accessed May 22, 2017).

[35] S.E. Jackson, A.R. Fersht, Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition, Biochemistry. 30 (1991) 10428–10435. doi:10.1021/bi00107a010.

[36] S. Harper, D.W. Speicher, Purification of proteins fused to glutathione S-transferase., Methods Mol. Biol. 681 (2011) 259–80. doi:10.1007/978-1-60761-913-0_14.

[37] S. Costa, A. Almeida, A. Castro, L. Domingues, Fusion tags for protein solubility, purification and immunogenicity in Escherichia coli: the novel Fh8 system., Front. Microbiol. 5 (2014) 63. doi:10.3389/fmicb.2014.00063.

[38] A. Malhotra, Chapter 16 Tagging for Protein Expression, in: Methods Enzymol., 2009: pp. 239–258. doi:10.1016/S0076-6879(09)63016-0.

[39] R. Woody, Contributions of tryptophan side chains to the far-ultraviolet circular dichroism of proteins, Eur. Biophys. J. 23 (1994) 253–262. doi:10.1007/BF00213575.

CHAPTER 7

Folding under high pressure inside the bacterial cytoplasm

In vitro experiments have painted a rich protein folding landscape over the years. Globular proteins are dynamic in nature with small folding equilibrium constants is one of the most valuable lessons learnt from these experiments. However, *in vitro* experiments lack in providing the complex environment presented by a living cell. The intricate solution environment inside the cell is capable of controlling protein stability and kinetics. Some of the ways in which cells can potentially modulate the function of their proteins are: suppression of transcription[1], interference with mRNA [2], different post-translational modifications [3,4], as well as protein transport, storage and degradation[5–9]. Cell has a highly crowded environment, some of which include tRNA, proteins, osmolytes, complex carbohydrates and other large organelles (Like: golgi bodies, endoplasmic reticulum). These molecules can not only exclude volume but are also known to interact with proteins via "quinary interactions".

The most natural environment for performing folding experiments is inside living cells as they provide conducive surroundings that is highly evolved to modulate the stability of proteins compared to aqueous buffer (*in vitro)*. Using the fast relaxation imaging (FReI) technique it is shown that proteins have higher refolding yields inside living cells compared to *in vitro* experiments a possible result of chaperoning [10]. Chaperones can help proteins refold quickly once they has been unfolded. Recently, it has been shown that studying protein folding kinetics inside mammalian cells subject to stress such as temperature and osmotic changes is feasible. Interesting experiments have been conducted on protein (PGK) stability changes when it is localized in different cellular compartments like cytoplasm, nucleus or endoplasmic reticulum. Protein (PGK) showed different thermal stability when measured in the cytoplasm compared to when it was localized inside the nucleus. As a step forward protein kinetics and stability has also been studied as a function of cell cycle. The results of that investigation showed that the cytoplasmic environment somehow changes when the cell is dividing in order to make the protein several °C more stable, whereas during interphase (the normal metabolic state), PGK is less stable. The possible implication of this study is this cell cycle dependent

folding/unfolding of signaling or cell control proteins can provide a better timing control on their function [11,12]. Other such examples are: RfaH C-terminal domain's secondary structure is modified when crowded by its adjacent domain. Also the protein lymphotactin which is a GPCR-binding chemokine rearranges into a glycosaminoglycan binder when the temperature is tuned across 37 °C [13]. These are demonstrations where small perturbations can completely reshape a protein's structure and therefore function. Cell can exert such effects, ranging from subtle to these two obvious examples, on its proteome.

Recent experiments by Oliverberg and co-workers on beta barrel SOD1 protein inside both mammalian and bacterial cells have shown that stability of SOD1 protein's was lowered in both types of cells compared to *in vitro* experiments. The Tm (melting temperature) as well as Tc (cold denaturation point) shifted to physiological regime inside cells. However, it is worth mentioning here that SOD1 was more stable in the bacterial cells compared to the mammalian cells; intercellular environment of different cell may act on the protein in separate fashion [14]. The emerging picture thus far is that proteins are not just optimized for structure and function but also its interactions (electrostatic and hydrophobic) with the host cell environment plays a vital role. This research embarked questions on physiological indication of marginal stability and constraints on protein behavior across evolutionary diverse organisms.

Finally, as cells are also subject to variations in temperature, pressure (osmotic or hydrostatic), and solute concentrations. All of these effects together can alter protein stability, and could be used by cells to control its proteome's stability and biological function in more subtle ways than just protein synthesis and degradation [15].The focus of this chapter is to investigate how stability of an enzyme PGK is modified in two very different bacterial strains namely J1 strain[16,17] (pressure resistant strain obtained from Dr. Samantha Miller's lab in University of Aberdeen, UK) and MG1655 under high pressure and thermal stress. The motivation behind subjecting bacterial cells to high pressure was to find the reason behind pressure tolerance in 1% bacteria that survive high pressure pasteurization procedure employed by food and juice industry. Whole genome sequencing approach was taken to investigate on any underlying genomic variations that may give rise to pressure tolerance in these treated bacterial cells. The FRET experiments on labelled PGK clearly demonstrate the feasibility of performing high pressure denaturation experiments on proteins inside living bacterial cells.

7.1 Methods

7.1.1 In-cell ReAsH labeling

In-cell ReAsH labeling was carried out according to manufacturer's (Invitrogen) protocols, with some modifications (see Appendix F for more details). In order to perform spectroscopic measurements cells were spun down at 10,000 g for 10 minutes after 12-13 hrs of induction and washed using ice-cold PBS (3X times) and later diluted to 1:8 ratio of concentrated cell stock to PBS buffer pH=7. Undergraduate research student Timothy Chen assisted with the labeling procedure.

7.1.2 *In vitro* ReAsH labeling

In vitro labeling of ACGFP1 tagged PGK tetracysteine (GPGK-tc) containing construct was conducted at 5x ReAsH excess: 10 μM protein and 50 μM ReAsH. Before the protein was labeled it was incubated at room temperature in 1x BAL buffer (250 μM 2,3-dimercapto-1-propanol, Invitrogen) supplemented with 7.5 mM tris-(2-carboxyethyl)phosphine (TCEP) and 2.5 mM EDTA, pH 7. All buffers were degassed using sonication prior to labeling.

Labeling was initiated by the addition of ReAsH to the reaction mixture and was monitored by fluorescence intensity at 610 nm for about 90-120 minutes after initiation. After labeling, excess ReAsH was removed by filtration (Amicon) to a final dilution of >1000. Excess ReAsH was monitored *via* absorbance at 593 nm by UV-Vis spectroscopy and labeling was confirmed by MALDI mass spectrometry.

7.1.3 Pressure and temperature unfolding thermodynamics

Temperature and pressure denaturation of PGK was measured by direct excitation at 475 nm and monitoring the FRET between the fluorescent protein tags AcGFP1 (Donor) and ReAsH (Acceptor). Pressure unfolding measurements were done using rectangular quartz cuvette with a path length of 4 mm holds the sample in the ISS pressure cell. Measurements were done at an interval of 100 bar in the pressure range of 1 to 1700 bar. A wait time of approximately 8 minutes was set at each pressure to allow equilibration. Pressure increment was achieved using an automated pressure generator (HUB 440) by Pressure Biosciences. HUB

440 is capable of generating pressure upto 4000 bar. It can also maintain pressure at a particular set point using an inbuilt PID to an accuracy of around 2-3 bars.

Fluorescence spectroscopy was carried out using JASCO fluorescence spectrophotometer (FP- 8300) equipped with programmable temperature control with excitation and emission slit widths kept at 5 nm. FRET measurements of PGK stability were conducted by excitation at 475 nm and collecting emission from 500 – 700 nm. The reported donor/acceptor (D/A) ratio is calculated by dividing the integrated intensity from 500 - 560 nm (D) by the integrated intensity from 585 – 700 nm (A). The same wavelength range was used in all cases to obtain consistent results.

All thermodynamic denaturation signals $S(X)$, where X is temperature or pressure, were fitted to a two-state model separately for temperature and pressure denaturation

$$S(X) = S_U + S_F e^{-\Delta G(X)/RT} / (1 + e^{-\Delta G(X)/RT}) \quad (1a)$$

$$\Delta G(X) = g_X (X - X_m) \quad (1b)$$

to obtain the denaturation midpoints with respect to temperature (T_m) and pressure (P_m). It was observed that both temperature and pressure denaturation inside the bacterial cells is irreversible as the protein aggregated at high temperature or pressure.

7.2 Results

7.2.1 Higher Labeling efficiency and signal intensity of J1 strain compared to MG1655

The enzyme PGK with Ac-GFP1 and tetra-cysteine (tc) tag was expressed *in-situ* in both J1 and MG1655 strains followed by ReAsH labeling. ReAsH labeling procedure (see Appendix F) and growth conditions were kept uniform for both of the strains. FRET was monitored by exciting the Ac-GFP1 at 475 nm and collecting the fluorescence from 500-700 nm. The advantages of using ReAsH dye is that it is smaller in size compared to fluorescent protein mCherry. ReAsH is also not fluorescent unless it is bound to the tc tag this helps in reducing the background red fluorescence. Successful labeling with ReAsH and energy transfer yield a peak at 610 nm. We observed that J1 cells labeled every single time the labeling reaction was

conducted whereas MG1655 had a success rate of 66 %. We also noticed that J1 strain also had almost 4 fold more signal intensity compared to wildtype MG1655 strain (see Fig.7.1).

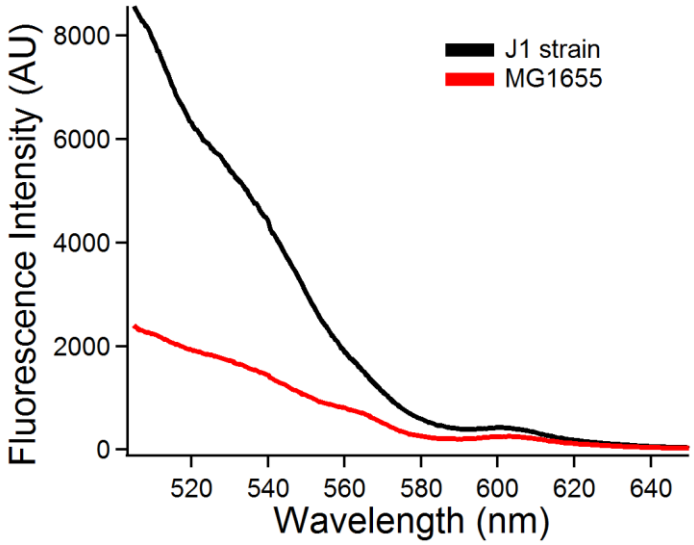

Fig.7.1: Raw fluorescence intensity vs wavelength plot for J1 and MG1655 strain. The cells were excited at 475 nm (GFP) with a PMT=700V and FRET was measured between AcGFP1 and ReAsH dye.

7.2.2 Pressure stabilization of PGK inside living bacterial cells

I measured the thermodynamic stability of FRET-Labeled PGK under high pressure inside living bacterial cells. The cells were subjected to a pressure of 1400 bar with an increment of 100 bar and wait time of approximately 8 mins was given at a particular pressure for equilibration. The midpoint pressure (Pm) obtained from the pressure titration was compared to *in vitro* pressure denaturation of PGK see Fig.7.2. I observed an increase in pressure denaturation midpoint for the FRET labeled PGK inside the bacterial cytoplasm. The protein melted at around 770 bar *in vitro* whereas inside cells the melting pressure was approximately 900 bar. It has also been reported earlier that PGK shows different thermal stabilities inside cells depending upon the its intracellular location [10]. PGK is thermally more stable in the nucleus compared to the cytoplasm. I have shown here that it is possible to monitor high

pressure unfolding of proteins inside living cells making it feasible to have a comparison of protein unfolding subjected to pressure and temperature inside cells.

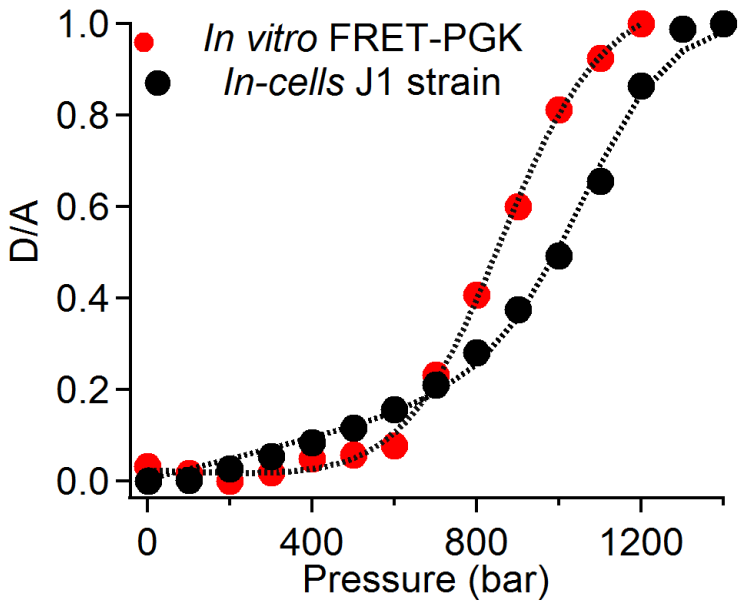

Fig.7.2: Normalized D/A vs Pressure plot. Comparison of *in vitro* and in-cell stability of PGK under pressure. Cells were excited at 475 nm (GFP) with the PMT set at 700 V and FRET was measured between AcGFP1 and ReAsH/mCherry to calculate donor to acceptor ratio.

7.2.3 Thermodynamic stability of PGK measured in different bacterial strains

As a step further in my investigation unfolding of PGK was conducted inside two different bacterial strains. The motivation behind was to understand how stability of PGK will change in cytoplasmic environment of the two different bacterial strains? We transformed the J1 (known pressure resistant strain) and MG1655 cells with GPGK-tc plasmid and then labelled them using ReAsH dye. The unfolding of PGK was triggered by increasing the temperature and pressure separately. A comparison of the melting temperature or pressure was made between the two strains see Table 7.1 and Fig.7.3. We observed that given the broad day to day experimental variation under high pressure conditions the stability of PGK in J1 strain is not significantly different than in wildtype MG1655. Interestingly, it was noticed that under

thermal stress PGK melted at a high temperature in the pressure resistant strain compared to the MG1655 (see Fig.7.3B).

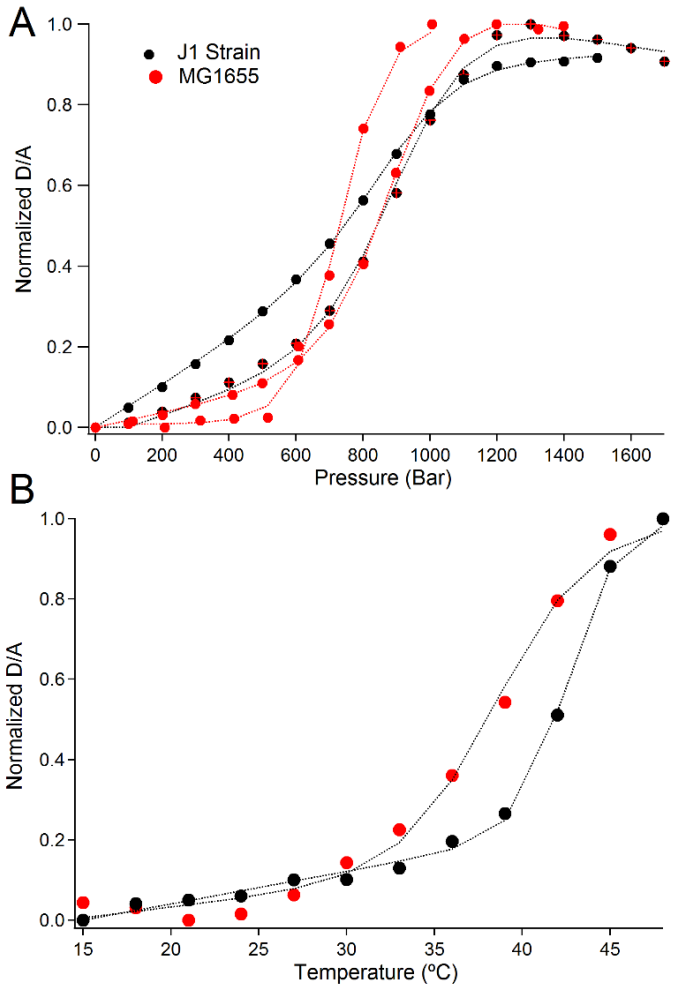

Fig.7.3: A) Comparison of FRET-PGK stability under pressure inside J1 (black circle) and MG1655 (red circle) strains. Experimental variation is shown by red and black bold line representing the sigmoidal unfolding curves obtained on different days. **B)** Thermal stability of FRET-PGK monitored in both strains J1 (black circle) and MG1655 (red circle). The cells were excited at 475 nm (GFP) with a PMT=700V and FRET was measured between AcGFP1 and ReAsH dye.

Table 7.1 Thermodynamic denaturation data for MG1655 and J1 strain

Strain	*Midpoint Temperature (Tm) °C	*Midpoint Pressure (Pm) Bar
MG1655	**39 (1)**	**964 (98)**
J1	**42 (0.5)**	**815 (120)**

*values were calculated by averaging the midpoints obtained from 3 different melts and standard deviation is 1σ

7.2.4 Change in colony morphology of pressurized MG1655 cells

In order to see the effect of high pressure on colony morphology of bacteria. The pressure treated MG1655 cells were platted and streaked on a LB+ Ampicillin plate along with the J1 and the wildtype MG1655 see Fig.7.4.It was discovered that cells that were subjected to pressure had smaller colony size along with well-defined boundaries compared to the wildtype or J1 strain. All these cells were grown for same time and in similar conditions (48 hrs in an incubator at 37 °C).

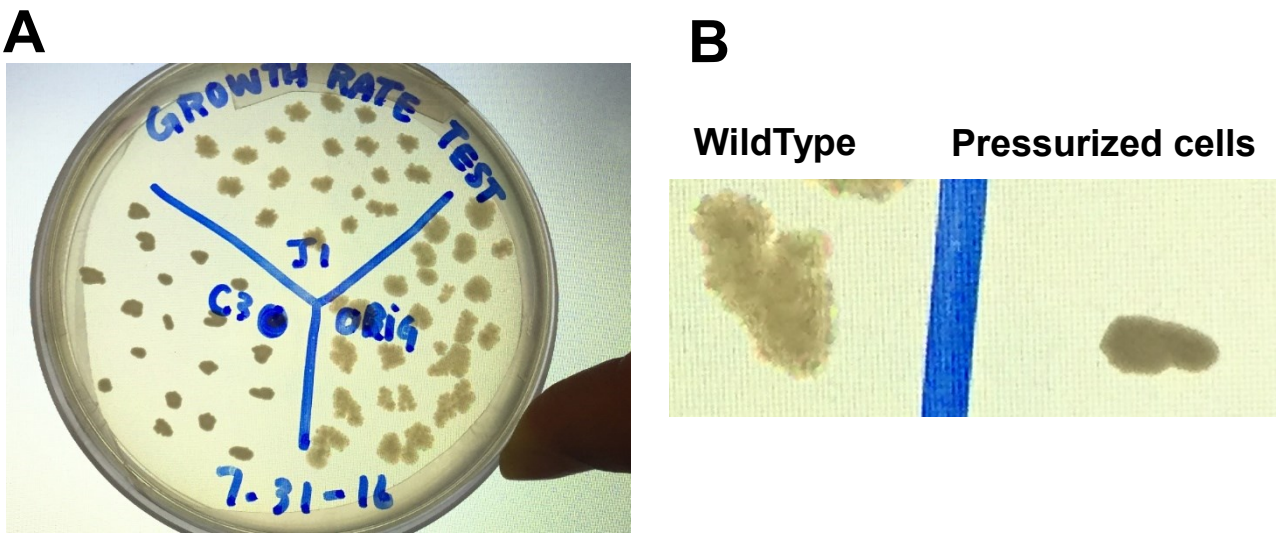

Fig.7.4: A) LB Ampicillin plate with colonies from the wildtype, J1 and pressure treated cells **B)** Smaller and more defined colonies were seen for the pressurized cells compared to MG1655.

7.3 Discussion and conclusion

In order to perform the thermodynamic denaturation experiments of PGK inside living bacterial cells I and Timothy Chen transformed both of the strains with AcGFP1-tc plasmid and then after induction labelled them with ReAsH dye. It was observed that J1 strain labeled every time the labeling experiment was performed and had 4 times more intensity than the MG1655 E.coli strain consistently see Fig. 7.1. I also performed phase contrast microcopy on these bacterial strains in Asst. Prof. Kulhman's lab at the physics department in UIUC to examine if there exists any size difference between these strains; possible reason for better labeling. The length and width analysis of the frames collected was done using Oufti software [18] with minor tweaks in the input files. The J1 cells were not significantly different in their length or width distribution from the Wildtype (see Appendix F). The tendency of J1 strain to yield higher signal is possibly attributed to better membrane permeability of the ReAsH dye. Pressure melts of PGK were performed *in vitro* and inside J1 strain as described in the method section. It was seen that cellular environment of the J1 bacterial cells has a stabilization effect on the protein PGK under high pressure stress (Fig. 7.2), similar results have been published for PGK stability when it was subjected to thermal stress inside mammalian cell [10,19].

In a step forward I investigated the thermodynamic stability of PGK inside different bacterial strains under both temperature and pressure stress. Surprisingly under pressure stress PGK stability as monitored by the FRET between ACGFP-1 and ReAsH dye didn't show a significant difference (see Table 7.1) but the protein seemed to be thermally stable inside the pressure resistant J1 strain by atleast ~2 degrees (see Fig. 7.3B).The pressure midpoints has a broader range of values due to higher day to day variation in pressure denaturation experiments. It is also worth mentioning that the protein tend to aggregate (abrupt decline in D/A) at much higher pressures (~1700 bar) compared to the MG1655 strain (~1450 bar); this can be attributed the inherit pressure tolerance of the J1 cells. The thermal stabilization of PGK in J1 strain indicate that the two thermodynamic parameters temperature and pressure possibly act differently on the structure giving rise to different unfolded state. Chaperoning or intercellular interactions might have preference towards one or the other of the unfolded structures. After looking into the colony morphology changes in the pressurized bacteria I performed laboratory pressure cycling of bacteria and sequenced the genome of these treated bacteria after first and ninth cycle along with the MG1655(control) and J1 strain (see Appendix F for more details).Genomic analysis revealed that C1 (first cycle) didn't have any genetic modification

when referenced against the MG1655 strain but interestingly cycle 9 showed a 1.4 kb insertion making the cyaA gene non-functional (see Appendix F for details). The non-functionality of cyaA gene was verified by measuring the growth curve for cycle 9 and MG1655 strains with and without cAMP (3',5'-cyclic adenosine monophosphate). The recovery in growth seen for cycle 9 with the addition of cAMP manifested the presence of mutated cyaA gene (see Appendix F). The pressure relevance of this mutation is still under investigation.

J1 strain's genome was mapped to the MG1655 genome available online (NCBI *E. coli* genome databank) gave a low overlap (~83%) and hence the mutations predicted by BRESEQ program [20] are not adequate, the reference genome for J1 strain is not known. In conclusion this chapter demonstrated the feasibility of performing pressure melts on protein inside living cells which opens the possibility of making comparison with temperature denaturation. This comparison will facilitate our understanding on how differently they acts on the proteins.

7.4 References

[1] A. Bird, A. Wolffe, Methylation-induced repression—belts, braces, and chromatin, Cell. (1999). http://www.sciencedirect.com/science/article/pii/S0092867400815329 (accessed May 15, 2017).

[2] T. Brummelkamp, R. Bernards, R. Agami, A system for stable expression of short interfering RNAs in mammalian cells, Science (80. (2002). http://science.sciencemag.org/content/296/5567/550.short (accessed May 15, 2017).

[3] T. Karve, A. Cheema, Small changes huge impact: the role of protein posttranslational modifications in cellular homeostasis and disease, J. Amino Acids. (2011). https://www.hindawi.com/journals/jaa/2011/207691/abs/ (accessed May 15, 2017).

[4] Y. Deribe, T. Pawson, I. Dikic, Post-translational modifications in signal integration, Nat. Struct. Mol. Biol. (2010). http://www.nature.com/nsmb/journal/v17/n6/abs/nsmb.1842.html (accessed May 15, 2017).

[5] T. Rapoport, Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes, Nature. (2007). http://www.nature.com/nature/journal/v450/n7170/abs/nature06384.html (accessed May 15, 2017).

[6] J. Rothman, F. Wieland, Protein sorting by transport vesicles, Science (80-.). (1996). http://search.proquest.com/openview/6d326bf594ebddf51b491acf1f0e1393/1?pq-origsite=gscholar&cbl=1256 (accessed May 15, 2017).

[7] J. Rothman, Mechanisms of intracellular protein transport, Nature. (1994). http://search.proquest.com/openview/1635d655e628dcd8ec19e32a35646599/1?pq-origsite=gscholar&cbl=40569 (accessed May 15, 2017).

[8] A. Varshavsky, Discovery of cellular regulation by protein degradation, J. Biol. Chem. (2008). http://www.jbc.org/content/283/50/34469.short (accessed May 15, 2017).

[9] The protein storage vacuole, J Cell Biol. (2001). http://jcb.rupress.org/content/155/6/991.abstract (accessed May 15, 2017).

[10] A. Dhar, K. Girdhar, D. Singh, H. Gelman, S. Ebbinghaus, M. Gruebele, Protein stability and folding kinetics in the nucleus and endoplasmic reticulum of eucaryotic cells, Biophys. J. 101 (2011) 421–430.

[11] V.N. Uversky, C.J. Oldfield, A.K. Dunker, Intrinsically disordered proteins in human diseases: Introducing the D(2) concept, Annu. Rev. Biophys. 37 (2008) 215–246. doi:10.1146/annurev.biophys.37.032807.125924.

[12] A.J. Wirth, M. Platkov, M. Gruebele, Temporal Variation of a Protein Folding Energy Landscape in the Cell, J. Am. Chem. Soc. 135 (2013) 19215–19221.

doi:10.1021/ja4087165.

[13] B. Volkman, T. Liu, F. Peterson, Lymphotactin structural dynamics, Methods Enzymol. (2009). http://www.sciencedirect.com/science/article/pii/S0076687909054032 (accessed May 15, 2017).

[14] J. Danielsson, X. Mu, L. Lang, H. Wang, A. Binolfi, F.-X. Theillet, B. Bekei, D.T. Logan, P. Selenko, H. Wennerström, M. Oliveberg, Thermodynamics of protein destabilization in live cells., Proc. Natl. Acad. Sci. U. S. A. 112 (2015) 12402–7. doi:10.1073/pnas.1511308112.

[15] A. Wirth, M. Gruebele, Quinary protein structure and the consequences of crowding in living cells: Leaving the test-tube behind, Bioessays. (2013). http://onlinelibrary.wiley.com/doi/10.1002/bies.201300080/full (accessed May 15, 2017).

[16] B. Klotz, P. Ma??as, B.M. Mackey, The relationship between membrane damage, release of protein and loss of viability in Escherichia coli exposed to high hydrostatic pressure, Int. J. Food Microbiol. 137 (2010) 214–220. doi:10.1016/j.ijfoodmicro.2009.11.020.

[17] P. Mañas, B.M. Mackey, Morphological and Physiological Changes Induced by High Hydrostatic Pressure in Exponential- and Stationary-Phase Cells of Escherichia coli : Relationship with Cell Death Morphological and Physiological Changes Induced by High Hydrostatic Pressure in Expo, Appl. Environ. Microbiol. 70 (2004) 1545–1554. doi:10.1128/AEM.70.3.1545.

[18] A. Paintdakhi, B. Parry, M. Campos, I. Irnov, J. Elf, I. Surovtsev, C. Jacobs-Wagner, Oufti: an integrated software package for high-accuracy, high-throughput quantitative microscopy analysis, Mol. Microbiol. 99 (2016) 767–777. doi:10.1111/mmi.13264.

[19] S. Ebbinghaus, A. Dhar, D. McDonald, M. Gruebele, Protein folding stability and dynamics imaged in a living cell, Nat. Methods. 7 (2010) 319–323. doi:10.1038/nmeth.1435.

[20] D.E. Deatherage, J.E. Barrick, Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq, Methods Mol. Biol. 1151 (2014) 165–188. doi:10.1007/978-1-4939-0554-6_12.

APPENDIXES

APPENDIX A

Supplementary information of high-resolution mapping of the folding transition state of a WW domain

A.1 Supplementary Methods

For proteins that (un)fold fast and reversibly upon perturbation by temperature, the Φ_T value offers a convenient reaction coordinate for locating the folding transition state (see Materials and Methods of main text for details). In this and previous Φ-analyses (see references in main text), we calculated the Φ_T value by using a Taylor series expansion of the free energy of activation around the midpoint of unfolding (T_m) (T_m-fit, eq. 1 below). It is also possible to expand about T_0 (the temperature of maximal stability), or to assume a constant heat capacity of folding DCP (the Gibbs-Helmholtz formula). We make the connections below.

Several (but not all) of the highly destabilized hPin1 WW variants, mostly within loop 2 or its immediate flanking residues, yielded unphysical ΦT > 1 (Fig. A.3). We pinpointed as the cause uncertainty in the curvature of $\Delta G(T)$ near Tm, given by the coefficient $\Delta G_f^{(2)}$ in the main text. An error in $\Delta G_f^{(2)}$ (due to fewer temperature points measured, or noisier kinetic traces) produces an error in ΦT when the derivative $\partial \Delta G/\partial T$ is calculated. This error can manifest itself as a physically unreasonable temperature of maximal stability, T_0. At maximal stability $\partial \Delta G/\partial T = 0$, so the "Tm-fit" predicts that

$$\partial \Delta G/\partial T = 0 = \partial/\partial \mathrm{T}\ [\Delta G_f^{(1)}\ (\mathrm{T}\text{-}\mathrm{T_m}) + \Delta G_f^{(2)}\ (\mathrm{T}\text{-}\mathrm{T_m})^2], \quad (1)$$

from which one can easily show that

$$T_0 = T_m - \Delta G_f^{(1)}/(2\Delta G_f^{(2)}) \quad (2)$$

T0 is generally expected to lie in a range near 0 °C for most proteins. (Below that range, cold denaturation occurs, above that range, heat denaturation occurs.) For some proteins, the "raw"

T0 value predicted from eq. (2) based on the "Tm-fit" is off by 100s of °C from the physical range (see Table 3). Therefore, an alternative "T0-fit" was used, which is briefly outlined here. Instead of Taylor-expanding about Tm, we can Taylor-expand

$$\Delta G(T)=\Delta G^{(0)'}+\Delta G^{(2)'}(T-T_0)^2 \quad (3)$$

about T0, where $\Delta G^{(0)'}$ is the free energy at maximal stability. This expansion has exactly the same number of adjustable parameters as the "Tm-fit." By plotting T0 from the "Tm-fit" vs. Tm, we found that most proteins produced physically reasonable T0 in the range from -40 °C to +30 °C. (Tm ranged from +38 °C to +78 °C T_0. We removed outliers, and fitted the correlation between T_m and T_0 to a straight line [T_0=a+b*T_m ; R^2 = 0.3 a=-72.4±16 ; b=1.03±0.3]. We then used this straight line to determine a smoothed T_0 for each protein (Table A.3, smoothed values from the linear correlation T_0=a+b*T_m). Finally, ΔG was re-fitted to eq. (3), and derivatives for ΦT were calculated using the fit to eq. (3). The difference between the "T_m-fit" and the "T_0-fit" is shown in Fig. 2.3A (main text). The distribution of ΦT values is generally very similar, but the outliers are removed. Thus we believe that the "T_0-fit" more accurately reflects the correct ΦT values.

A third useful expansion is the Gibbs-Helmholtz formula, which assumes a constant heat capacity of reaction. One can integrate $dH=\Delta C_P dT$ and $dS=\Delta C_P T/T_m$ (with Tm again as the reference temperature in this example) before assembling the free energy

$$\Delta G(T)= \Delta H-T\Delta S=\Delta H_m+\Delta C_P(T-T_m) - T\Delta S_m - T\Delta C_P \ln[T/T_m]. \quad (4)$$

This equation also has three adjustable parameters, and yields a fit of essentially the same quality as the three parameter Taylor expansions. Rewriting eq. (4) in the variable x=T-T_m (e.g. $\ln[T/T_m] = \ln[1+(T-T_m)/T_m] = \ln[1+x/T_m]$), and Taylor-expanding for comparison with equation (1) yields

$$\begin{aligned}\Delta G(T)&= (\Delta H_m - T_m\Delta S_m) - \Delta S_m(T-T_m) - (\Delta C_P/2_m)\cdot(T-T_m)^2\\ &= -\Delta S_m(T-T_m) - \Delta C_P/(2T_m)\cdot(T-T_m)^2. \end{aligned} \quad (5)$$

As expected for thermodynamic consistency, $-\partial\Delta G/\partial x|_{x=0} = \Delta S_m$ and $-T_m\partial^2\Delta G/\partial x^2|_{x=0} = \Delta C_P$. The relation between our Taylor parameters in Table 1 of the main text and the Gibbs-Helmholtz parameters is

$$\Delta H_m=-T_m\Delta G_f^{(1)},\ \Delta S_m=-\Delta G_f^{(1)}, \text{ and } \Delta C_P=-2T_m\,\Delta G_f^{(2)}. \quad (6)$$

It is worth noting that despite the equal quality fits and number of parameters, the Taylor expansion and Gibbs-Helmholtz parameters are not equivalent, as the Taylor expansion does

not assume that the heat capacity for reaction is independent of temperature, but truncates at second order, whereas the Gibbs-Helmholtz equation assumes constancy of ΔC_P, but the logarithm "expands" to infinite order.

The parameters of the "T0-fit" can also be related to the "Tm-fit" parameters by

$$0=\Delta G^{(0)'}+\Delta G^{(2)'}(T_0-T_m)^2,\ \Delta G^{(1)}=-2\Delta G^{(2)'}(T_0-T_m),\text{ and }\Delta G^{(2)}=\Delta G^{(2)'} \quad (7)$$

The second order expansion coefficients are the same when all fitting parameters are floated. However, if T0 is refitted and smoothed as discussed above and reported in Table A.3, then a corresponding value of $\Delta C_P=-2T_m\,\Delta G^{(2)'}$ may be calculated. Table A.4 shows the heat capacities from the "T0-fit," which we believe are more reliable. The ones from the "Tm-fit" are easily obtained using table 1 of the main text and eq. (6).

Table A.1: Transition state location calculated from the T_m-analysis and T_0-analysis

Variant	T_m-fit			T_0-fit		
	Φ_T (50 °C)	**Φ_T (55 °C)**	**Φ_T (60 °C)**	**Φ_T (50 °C)**	**Φ_T (55 °C)**	**Φ_T (60 °C)**
wt hPin1	0.35	0.45	0.53	0.35	0.45	0.53
K6A	0.36	0.39	0.42	0.39	0.40	0.42
K6M	0.49	0.51	0.53	0.52	0.52	0.52
L7A	0.60	0.65	0.71	0.52	0.54	0.56
L7I	0.50	0.55	0.61	0.50	0.50	0.51
L7V	0.62	0.68	0.74	0.58	0.61	0.63
P8A	0.40	0.43	0.46	0.40	0.44	0.47
P9A	0.49	0.53	0.57	0.50	0.53	0.56
G10A	0.46	0.53	0.60	0.45	0.51	0.57
W11F	0.86	1.02	1.18	0.68	0.75	0.80
E12A	0.49	0.58	0.67	0.50	0.57	0.63
E12Q	0.38	0.46	0.53	0.38	0.46	0.53
K13A	0.49	0.49	0.49	0.48	0.48	0.49
K13V	0.50	0.52	0.53	0.49	0.51	0.53
K13Y	0.35	0.36	0.37	0.31	0.32	0.32
R14A	0.50	0.61	0.73	0.45	0.54	0.61
R14F	0.34	0.44	0.54	0.33	0.43	0.51
R14L	0.44	0.54	0.62	0.44	0.53	0.61
M15A	0.41	0.49	0.56	0.41	0.49	0.56

Cont'd

S16A	0.49	0.55	0.60	0.49	0.55	0.66
S16G	0.52	0.61	0.69	0.51	0.59	0.49
S16T	0.36	0.43	0.49	0.36	0.43	0.49
R17A	0.49	0.54	0.57	0.50	0.54	0.57
R17G	0.59	0.63	0.66	0.59	0.63	0.66
S18A	0.37	0.51	0.64	0.38	0.52	0.63
S18G	0.35	0.49	0.61	0.36	0.49	0.60
S19G	0.37	0.46	0.54	0.38	0.46	0.54
G20A	0.61	0.63	0.64	0.62	0.63	0.65
R21A	0.37	0.40	0.43	0.37	0.39	0.41
R21H	0.38	0.41	0.42	0.38	0.39	0.40
R21L[1]	0.39	0.41	0.43	-	-	-
V22A	0.36	0.39	0.41	0.38	0.39	0.39
Y23A	0.49	0.53	0.57	0.44	0.46	0.48
Y23F	0.55	0.55	0.55	0.55	0.55	0.55
Y23L	0.52	0.54	0.56	0.51	0.52	0.53
Y24F	0.43	0.50	0.56	0.43	0.50	0.57
Y24W	0.33	0.43	0.52	0.33	0.43	0.51
F25A	0.57	0.59	0.62	0.48	0.48	0.48
F25L	0.52	0.55	0.58	0.51	0.54	0.57
N26D	0.57	0.62	0.67	0.49	0.52	0.54
H27A	0.49	0.52	0.55	0.49	0.52	0.54
H27G	0.40	0.44	0.47	0.40	0.42	0.45
I28A	0.32	0.41	0.49	0.33	0.41	0.48
I28G	0.54	0.61	0.68	0.53	0.58	0.62
I28V	0.34	0.42	0.49	0.34	0.42	0.50
T29A	0.53	0.58	0.62	0.51	0.54	0.56
T29D	0.58	0.65	0.73	0.53	0.56	0.59
T29G	0.88	0.96	1.04	0.72	0.74	0.76
T29S	0.45	0.50	0.55	0.45	0.49	0.53
N30A	0.72	0.76	0.80	0.73	0.76	0.78
A31G	0.61	0.66	0.70	0.60	0.64	0.67
A31S	0.22	0.31	0.38	0.22	0.31	0.38
S32G	0.43	0.46	0.49	0.43	0.46	0.48
S32T	0.14	0.19	0.24	0.13	0.19	0.24
Q33A	0.46	0.55	0.62	0.48	0.54	0.50

Cont'd

W34A	0.26	0.34	0.40	0.27	0.33	0.31
W34F	0.63	0.63	0.63	0.61	0.62	0.58
E35Q	0.56	0.59	0.62	0.56	0.59	0.55
R36A	0.26	0.31	0.36	0.27	0.31	0.29
S38A	0.32	0.39	0.45	0.33	0.40	0.37
S38G	0.56	0.58	0.60	0.54	0.58	0.54
S38T	0.53	0.58	0.62	0.52	0.57	0.62
S18G/S19G	0.41	0.46	0.50	0.42	0.57	0.43
S19G/G20S	0.38	0.42	0.45	0.38	0.46	0.39
I28N/T29G	1.09	1.19	1.29	0.94	0.97	0.92
Var1 (SADGR)	-	-	0.06[1]	-	-	0.06
Var2 (SSSGR)	-	-	0.38	-	-	0.40
Var3 (SNGR)	-	-	0.34	-	-	0.35
Var4 (SSGR)	0.42	0.46	0.48	0.43	0.46	0.48
Var5 (+1 Gly)	0.15	0.27	0.37	0.37	0.41	0.44
Var6 (+2 Gly)	0.30	0.29	0.29	0.64	0.67	0.69
K13k	0.49	0.45	0.42	0.47	0.42	0.37
S16s	0.59	0.57	0.54	0.52	0.47	0.42
R17r	0.56	0.61	0.66	0.56	0.59	0.62
V22v	0.76	0.78	0.80	0.75	0.78	0.80
H27h	0.58	0.66	0.72	0.59	0.66	0.73
S32s	0.98	0.98	0.97	0.90	0.87	0.83
W34w	0.46	0.61	0.74	0.46	0.61	0.74

[1]R21L the mutant was partially folded , [1]The values cannot be calculated accurately for all the temperatures

Table A.2: Φ_M values used in the calculation of the transition state maps at 50 °C and 60 °C

Residue	Mutation	Type [1]	Φ_M (50 °C)	Average Φ_M [2] (sc, 50 °C)	Φ_M (60°C)	Average Φ_M [2] (sc, 60 °C)
L7	L7A	sc	0.23 (0.02)	0.23	0.31 (0.04)	0.35
	L7V	sc	0.23 (0.02)		0.37 (0.02)	
G10	G10A	sc	0.52 (0.02)	0.52	0.61 (0.03)	0.61
E12	E12A	sc	0.15 (0.12)	0.17	0.36 (0.08)	0.34
	E12Q	sc	0.22 (0.35)		0.25 (0.41)	
	K13k	hb	0.79 (0.01)		0.77 (0.01)	
R14	R14A	sc	0.72 (0.01)	0.75	0.82 (0.02)	0.80
	R14F	sc	0.76 (0.01)		0.84 (0.02)	
	R14L	sc	0.77 (0.01)		0.73 (0.02)	
M15	M15A	sc	0.81 (0.02)	0.81	0.85 (0.03)	0.85
S16	S16G	sc	1.13 (0.01)	1.13	1.25 (0.02)	1.25
	S16s	hb	0.91 (0.01)		0.97 (0.02)	
	R17r	hb	1.08 (0.03)		1.19 (0.02)	
S18 [3]	S18G/S19G	sc	1.36 (0.02)	1.36	1.36 (0.03)	1.36
S19	S19G	sc	1.38 (0.04)	1.38	1.41 (0.05)	1.41
	R17r	hb	1.07 (0.03)		1.19 (0.02)	
G20	G20A	sc	1.33 (0.01)	1.33	1.50 (0.01)	1.50
R21	R21A	sc	1.00 (0.02)	0.93	0.94 (0.03)	0.89
	R21H	sc	0.86 (0.02)		0.83 (0.03)	
	S16s	hb	0.91 (0.01)		0.97 (0.02)	
Y23	Y23A	sc	0.55 (0.01)	0.65	0.58 (0.01)	0.66
	Y23L	sc	0.74 (0.01)		0.84 (0.02)	
Y24	Y24F	sc	0.64 (0.03)	0.64	0.71 (0.03)	0.71
	W34w	hb	0.39 (0.01)		0.57 (0.02)	
F25	F25A	sc	0.72 (0.01)	0.67	0.79 (0.03)	0.72
	F25L	sc	0.62 (0.01)		0.68 (0.02)	
	K13k	hb	0.79 (0.01)		0.77 (0.01)	
N26	N26D	sc	0.42 (0.01)	0.42	0.50 (0.03)	0.50
	H27h	hb	0.46 (0.02)		0.57 (0.02)	
H27	H27G	sc	0.53 (0.02)	0.53	0.53 (0.02)	0.53
I28	I28A	sc	0.17 (0.22)	0.45	0.08 (0.50)	0.57
	I28V	sc	0.53 (0.12)		0.50 (0.16)	
	I28G	sc	0.46 (0.01)		0.60 (0.02)	
T29	T29A	sc	0.44 (0.01)	0.44	0.53 (0.02)	0.55
	T29S	sc	0.65 (0.03)		0.72 (0.05)	
	T29D	sc	0.38 (0.01)		0.51 (0.02)	
N30	H27h	hb	0.46 (0.02)	-	0.57 (0.02)	-
A31	A31G	sc	0.58 (0.01)	0.58	0.70 (0.01)	0.70
S32	S32G	sc	0.29 (0.03)	0.29	0.30 (0.05)	0.38
Q33	Q33A	sc	0.50 (0.04)	0.50	0.70 (0.06)	0.70
	W34w	hb	0.39 (0.01)		0.57 (0.02)	
W34	W34A	sc	0.43 (0.06)	0.43	0.24 (0.13)	0.24
E35	E35A	sc	0.70 (0.06)	0.71	0.79 (0.08)	0.87
	E35Q	sc	0.72 (0.09)		0.96 (0.10)	

[1] Type of mutation: side chain (sc), backbone H-bond (hb). [2] Error weighted average Φ_M-value for residues probed by multiple mutations. [3] Φ_M-value of the S18G/S19G mutant was assigned to S18. [4] No Φ_M-value calculated, because of large error.

Table A.3: Alternative free energy fits: "Raw" and smoothed T_0 values of mutants that qualify for Φ_M-value analysis.

Mutant	"Raw" T_0 values (°C)	Smoothed T_0 values (°C) [1]
L7A	-646.0	-33.5
L7I	394.0	-21.6
L7V	-347.0	-27.1
P8A	-14.2	-23.6
G10A	-66.0	-21.9
W11F	349.0	-36.4
E12A	-126.0	-18.2
E12Q	-7.1	-15.4
K13k	-158.6	-24.6
R14A	-195.0	-32.0
R14F	-54.0	-25.9
R14L	-30.6	-23.2
M15A	-13.9	-19.1
S16G	-47.0	-23.4
S16A	-6.7	-16.9
S16T	-8.0	-17.6
S16s	442.2	-28.9
R17r	-75.9	-21.8
S19G	-22.6	-15.9
S18G/S19G	-64.2	-17.8
G20A	-16.8	-22.1
R21A	-77.0	-20.0
R21H	-88.0	-20.9
V22A	-119.5	-16.6
Y23A	-133.0	-37.5
Y23L	-56.0	-25.8
Y23F	-21.2	-18
Y24F	-13.7	-19.5
Y24W	-24.7	-17.9
F25A	-343.0	-38.9
F25L	-41.0	-28.6
N26D	-335.0	-35.3
H27G	-90.0	-20.4
H27h	-29.0	-32.6
I28A	-60.0	-16.6
I28G	-125.0	-23.8
I28V	-2.8	-15.4
I28N/T29G	-696.9	-34.9
T29A	-114.0	-26.8
T29D	-1834.0	-28.2
T29G	2291.0	-37.0
T29S	-66.0	-20.1
N30A	-36.1	-17.5
A31G	-50.0	-30.3
S32G	-33.7	-20.8
S32s	-213.5	-29.7
Q33A	-108.0	-17.7
W34A	-235.0	-17.9
W34w	-17.7	-21.4
E35A	-14.9	-20.5
E35Q	-14.76	-17.7

Table A.4: "T0-fit" parameters and corresponding ΔC_P

Variants	T_0 (°C)	T_m (°C)	$\Delta G^{(0)'}$, kJ/mol	$\Delta G^{(2)}$, kJ/mol/K^2	ΔC_P, kJ/mol/K
Wildtype hPin1	-12.0	58.6	-14.24	0.00285	-0.334
K6A	-11.2	59.4	-14.20	0.00282	-0.336
K6M	-12.5	58.1	-14.69	0.00292	-0.340
L7A	-33.5	37.8	-10.83	0.00210	-0.159
L7I	-21.6	49.3	-11.43	0.00223	-0.220
L7V	-27.1	44	-11.51	0.00225	-0.198
P8A	-23.6	47.4	-12.79	0.00254	-0.241
P9A	-14.7	56	-14.07	0.00280	-0.314
G10A	-21.9	49	-12.39	0.00245	-0.240
W11F	-36.4	35	-11.14	0.00215	-0.150
E12A	-18.2	52.6	-13.30	0.00263	-0.276
E12Q	-15.4	55.4	-13.59	0.00272	-0.301
K13A	-11.0	59.6	-13.58	0.00272	-0.324
K13V	-7.7	62.8	-14.12	0.00284	-0.357
K13Y	-19.1	51.7	-12.02	0.00283	-0.292
R14A	-32.0	39.2	-12.45	0.00243	-0.190
R14F	-25.9	45.2	-13.83	0.00272	-0.246
R14L	-23.2	47.8	-13.03	0.00258	-0.247
M15A	-19.1	51.8	-13.45	0.00268	-0.277
S16A	-16.7	54	-13.42	0.00268	-0.289
S16G	-23.4	47.6	-13.13	0.00259	-0.247
S16T	-17.6	53.2	-14.06	0.00281	-0.299
R17A	-11.8	58.8	-13.83	0.00276	-0.325
R17G	-13.3	57.3	-13.21	0.00264	-0.303
S18A	-12.2	58.4	-14.11	0.00281	-0.328
S18G	-14.2	56.5	-15.60	0.00310	-0.351
S19G	-15.9	54.8	-13.60	0.00271	-0.297
G20A	-22.1	48.9	-12.58	0.00250	-0.244
R21A	-20.0	50.9	-13.14	0.00260	-0.264
R21L	-14.8	55.9	678.60	-	-
R21H	-20.9	50	-12.80	0.00252	-0.252
V22A	-16.6	54.2	-14.36	0.00284	-0.308
Y23A	-37.5	33.9	-11.78	0.00229	-0.155
Y23F	-18.0	52.8	-13.32	0.00265	-0.280
Y23L	-25.8	45.3	-11.15	0.00220	-0.199
Y24F	-19.5	51.4	-12.85	0.00256	-0.263
Y24W	-17.9	52.9	-12.65	0.00252	-0.266
F25A	-38.9	32.5	-11.39	0.00220	-0.143

Cont'd

F25L	-28.6	42.5	-12.11	0.00238	-0.203
N26D	-35.3	36	-11.77	0.00228	-0.164
H27A	-12.9	57.7	-13.72	0.00274	-0.316
H27G	-20.4	50.5	-13.08	0.00258	-0.261
I28A	-16.6	54.2	-13.47	0.00267	-0.289
I28G	-23.8	47.2	-12.97	0.00255	-0.241
I28V	-15.4	55.4	-13.48	0.00270	-0.299
T29A	-26.8	44.3	-11.33	0.00222	-0.197
T29D	-28.2	42.9	-12.15	0.00237	-0.203
T29G	-37.0	34.4	-11.41	0.00221	-0.152
T29S	-20.1	50.8	-13.28	0.00262	-0.267
N30A	-17.5	53.3	-13.20	0.00262	-0.279
A31G	-30.3	40.9	-12.81	0.00252	-0.206
A31S	-12.9	57.7	-13.45	0.00269	-0.311
A31V	-7.9	62.6	-14.75	0.00296	-0.370
S32G	-20.8	50.1	-11.89	0.00236	-0.236
S32T	-8	61.7	-13.99	0.00282	-0.348
Q33A	-17.7	53.1	-11.83	0.00234	-0.248
W34A	-17.9	52.9	-13.79	0.00272	-0.287
W34F	-12.6	58	-14.07	0.00282	-0.327
E35Q	-17.7	53.1	-13.44	0.00268	-0.284
R36A	-14.0	56.7	-12.63	0.00252	-0.286
S38A	-11.5	59.1	-13.92	0.00278	-0.328
S38G	-12.4	58.2	-14.47	0.00290	-0.338
S38T	-12.4	58.2	-13.75	0.00276	-0.321
S18G/S19G	-17.8	53	-13.59	0.00269	-0.285
S19G/G20S	-14.0	56.7	-13.89	0.00277	-0.315
I28N/T29G	-34.9	36.4	-12.68	0.00246	-0.179
var1 (FiP)	7.4	77.5	-15.00	0.00305	-0.473
var2	-1.1	69.2	-15.03	0.00302	-0.418
var3	-2.2	68.1	-14.87	0.00299	-0.407
var4	-8.5	62	-13.89	0.00278	-0.345
var5 (+1G)	-23.2	47.7	-13.41	0.00264	-0.252
var6 (+2G)	-19.9	50.9	-12.92	0.00258	-0.262
K13k	-24.6	46.4	-14.67	0.00288	-0.267
S16s	-28.9	42.2	-14.42	0.00280	-0.236
R17r	-21.8	49.1	-14.26	0.00281	-0.276
V22v	-14.0	56.7	-14.82	0.00297	-0.336
H27h	-32.6	38.7	-14.95	0.00294	-0.228
S32s	-29.7	41.5	-18.29	0.00357	-0.297
W34w	-21.4	49.5	-15.24	0.00303	-0.300

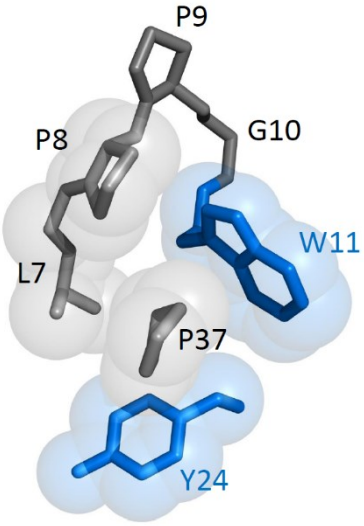

Fig. A.1: Side chain packing in hydrophobic core 1. Side chain packing of hydrophobic core residues L7, P8, W11, Y24 and P37, with side chains shown in stick mode and overlaid van der Waals surfaces. In native hPin1 WW, the absolutely conserved C-terminal Pro37 intercalates between the side chains of absolutely conserved Trp 11 (β strand 1) and highly conserved Tyr 24 (β-strand 2). The side chains of Leu7 and Pro8 are not strongly conserved among WW domains and contribute only peripherally to the hydrophobic core.

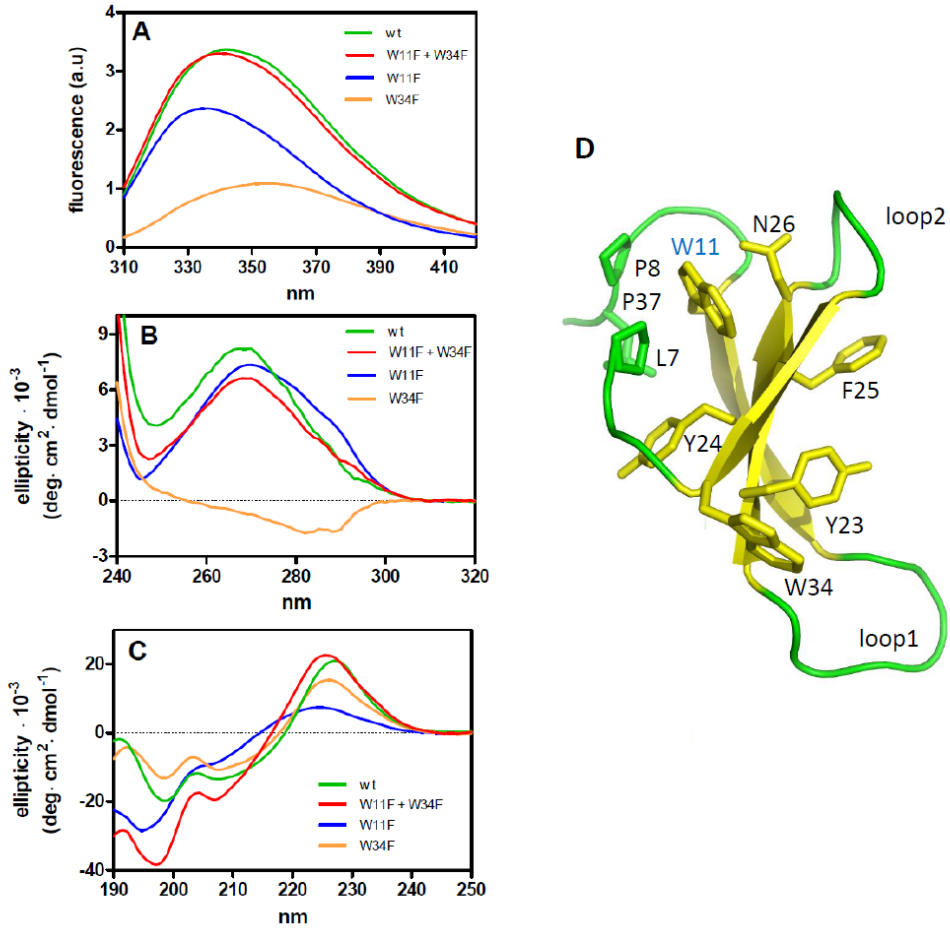

Fig. A.2: Structural assessment of hydrophobic core 1 mutant W11F. (A) Tryptophan fluorescence emission spectra (ex: 295 nm, 2 µM [protein]). W11 is largely excluded from solvent in folded hPin1 WW that explains the blue-shifted fluorescence emission maximum (332 nm) of mutant W34F. W34 is largely solvent-exposed, consistent with the red-shifted fluorescence emission maximum (352 nm) of mutant W11F. The aggregate spectrum (W11F + W34F) agrees well with the wild type spectrum, ruling out major structural changes upon W11F and W34F mutation. **(B)** Near-UV CD spectra (40 µM [protein]). Mutant W34F exhibits two bands with negative ellipticity at 282 and 287 nm that most likely result from L_{1b}-transitions within the indole ring of W11 (see, for example, chapter 4 in "Circular Dichroism and the Conformational Analysis of Biomolecules", 1996 Plenum Press, NY). Mutant W11F shows a rather broad and featureless spectrum with strong positive ellipticity. As Trps are the dominant chromophores in the near-UV, this band most likely originates from L_a-transitions within the indole ring of W34. The side chain of W34, although largely solvent exposed, must thus be in an asymmetric environment, possibly mediated by the clamp-like interaction of W34 with the side chain of Y23 (panel D). As the positive ellipticity is retained in wild type hPin1

Figure A.2 (cont.)

WW and only slightly higher in magnitude, this interaction cannot be significantly weakened in the W11F variant, thus further arguing against significant tertiary structural changes within hPin1 WW upon W11F mutagenesis. The minor differences between the wild type and aggregated spectra (W11F + W34F) might originate from coupling interactions of the indole ring of W11 with Y24 and the indole ring of W34 with Y23, as well as weak coupling between the two indole rings, which are separated less than 15 Å in folded hPin1 WW. **(C)** The far-UV CD spectra of wild type hPin1 WW and mutants W11F and W34F (16 μM [protein]) are atypical for an all-β-sheet protein and exhibit a strong positive band around 226 nm. Both Trps contribute to the ellipticity at 226 nm, and the ellipticity of the aggregate spectrum (W11F + W34F) almost quantitatively agrees with ellipticity of wild type hPin1 WW. This suggests that far-UV CD, like Trp-fluorescence and near-UV CD, predominantly monitors changes in tertiary structure rather than secondary structure. More significant deviations between the aggregate and wild type spectra are manifest at wavelengths below 210 nm, where Phe, Tyr and Trp residues absorb significantly. As for near-UV CD, these deviations likely result from non-additive side chain chromophore couplings. **(D)** Structural cartoon of hPin1 WW (residues 7-37, pdb: Pin1) with the side chains of L7, P8, W11, Y23, Y24, F25, N26 and P37 shown explicitly in stick mode presentation.

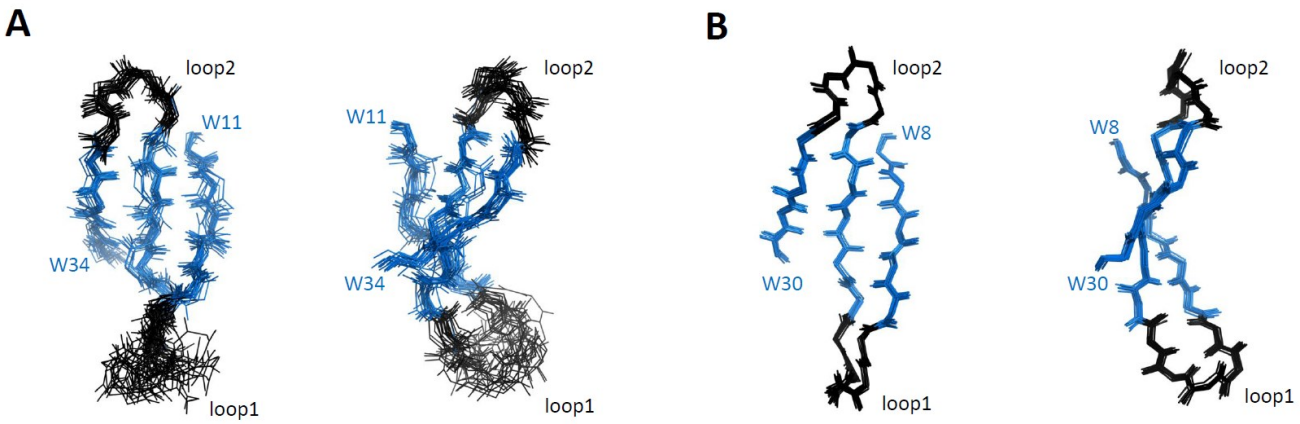

Fig. A.3: Native state dynamics of loop 1 in hPin1 WW and and FBP28 WW. (A) Superposition of the 15 lowest-energy solution conformations of the apo-form of the isolated hPin1 WW domain (residues W11-W34, pdb-file: 2KCF). (B) Superposition of the eight lowest-energy solution conformations of the isolated FBP28 WW domain (residues W8-W30, pdf-file: 1EP0). β strands and loop substructures are color coded blue and black, respectively. Increased local backbone dynamics is clearly visible within loop 1 of hPin1 WW, while the thermodynamically and kinetically optimized 5-residue type-I G-bulge turn of FBP28 WW appears to be more ordered.

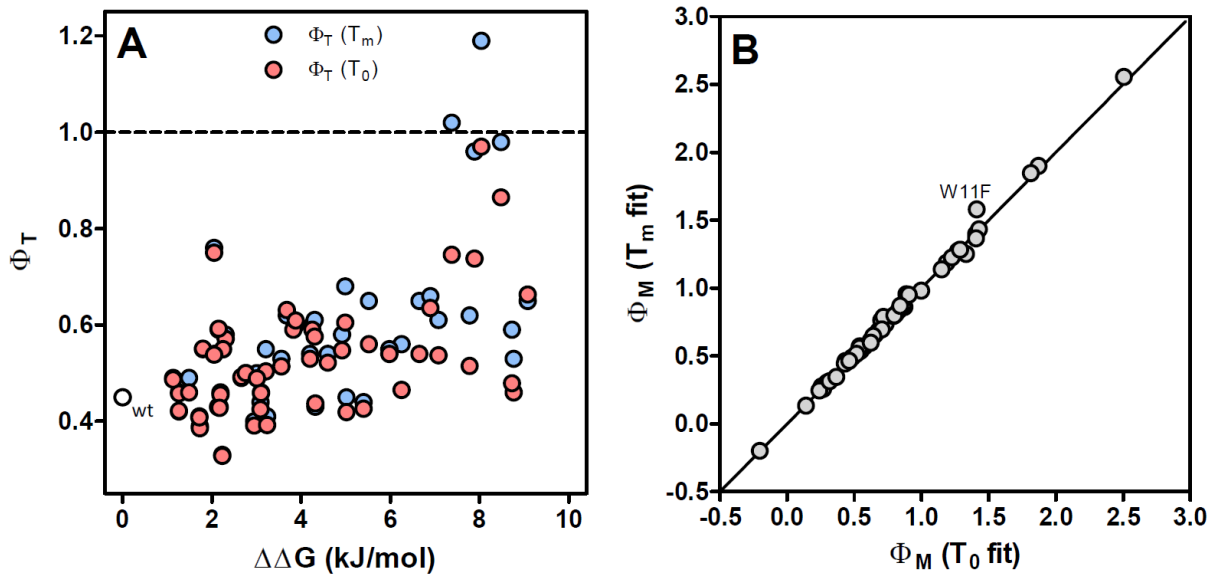

Fig. A.4: Dependence of Φ_T and Φ_M values on fitting model used. (A) Φ_T values from the T_m and T_0 fit (filled blue and red circles, respectively) for the 39 cross-validated consensus and 9 perturbing/outlier mutants (see Fig. A.5) that fulfill the requirements for reliable Φ_M-value analysis ($\Delta\Delta G_f$ < 1 kJ/mol, ΔT_m < 2.5 °C, with a typical error in T_m of 0.5 – 1 °C). While the more stable and moderately destabilized variants do not exhibit a significant shift in Φ_T, some noticeable differences are manifest for several unstable variants. (B) Plot of Φ_M-values calculated by the T_0-fit against corresponding values from the T_m-fit. Unlike their Φ_T-value counterparts, Φ_M-values are more robust and depend only marginally on the particular energy function used.

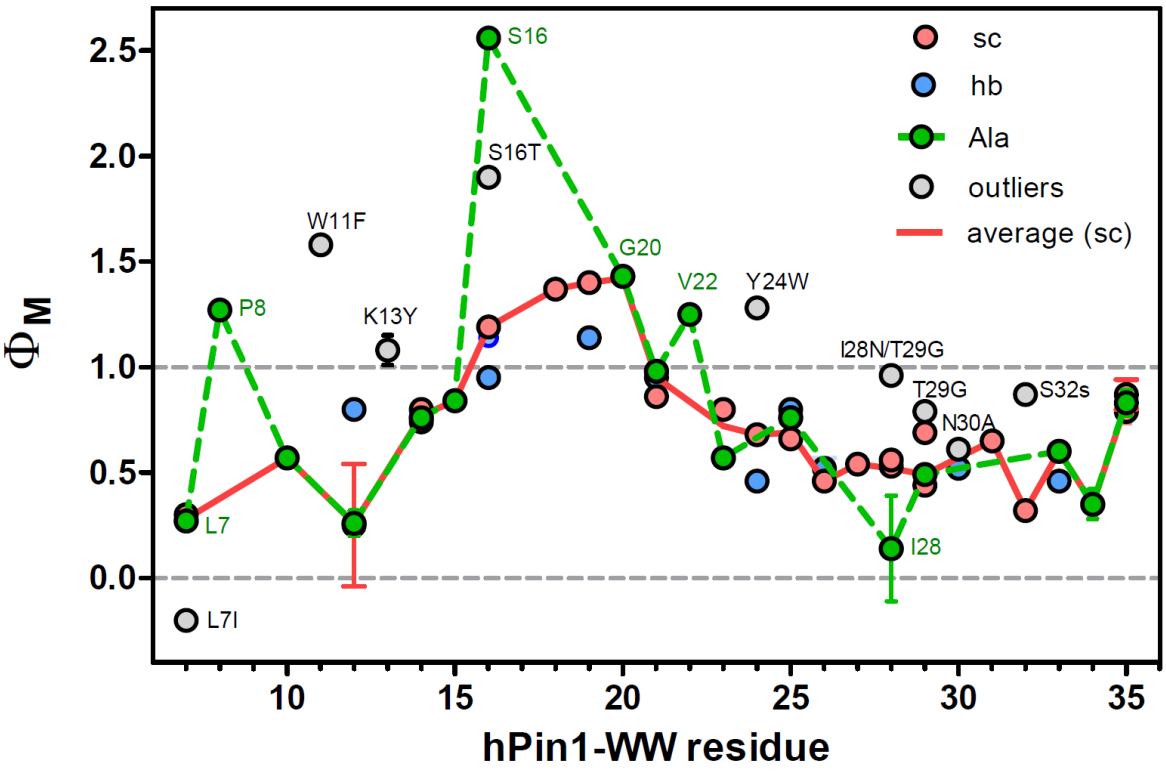

Fig. A.5: Extended Φ_M vs. sequence plot. Overall, we find good agreement between the Φ_M values from Ala mutants and Φ_M values from non-Ala mutations. Mutants P8A, S16A and V22A, however, yield Φ_M values that cannot be cross-validated by structural context. Filled red circles show the Φ_M-values calculated from consensus non-Ala mutations (for clarity, no error bars are shown), while the filled blue circles represent the Φ_M values from backbone H-bond mutants (for clarity no error bars are shown) (for errors, see Fig. 2.4 main text). Perturbing non-Ala mutants that excessively shift the transition state ensemble more towards the native state and mutants with outlier Φ_M values are depicted as filled grey circles. The solid red line is an error-weighted average side chain trend that includes the Ala mutants (Table 2 main text). While some perturbing mutations (e.g. P8A, W11F, S16A/T, Y24W) are readily identified in the plot, others (e.g. T29G, N30A, S32s) are more difficult to spot without considering data from the accompanying Φ_T-value analysis (Fig. 2.3, main text; Fig. A.4).

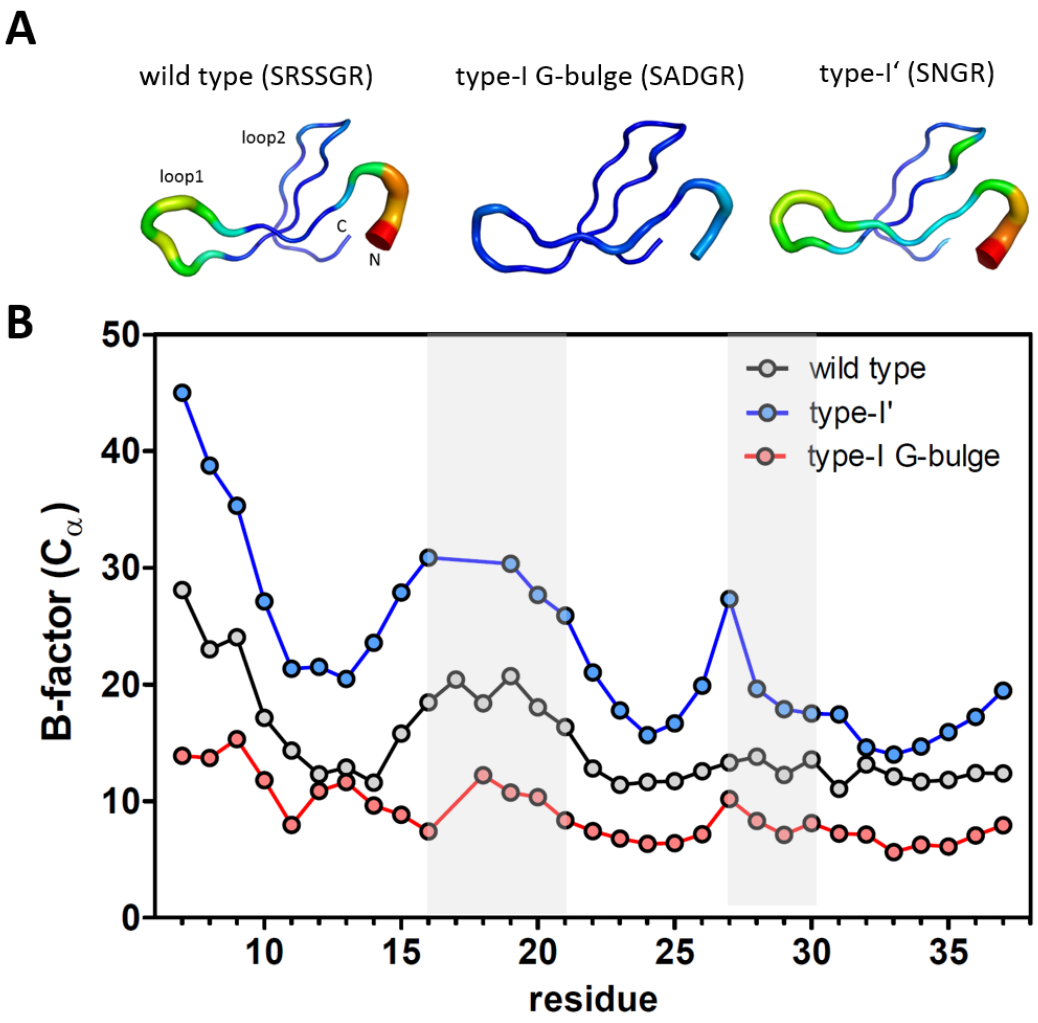

Fig. A.6: Thermal B-factors of stabilized loop 1 deletion variants. (A) Tube plot with the thermal B factors superimposed onto the X-ray structure of wild type hPin1 WW (pdb-file: 1PIN, loop 1: SRSSGR, left), stabilized type-I G-bulge variant 1 (FiP) (PDB-file: 1zcn, loop 1: SADGR, middle) and stabilized type-I' variant 3 (PDF-file: 2f21, loop 1: SNGR, right). **(B)** Plot of the thermal B factors vs. the sequence for wild type hPin1 WW (filled grey circles), type-I G-bulge variant 1 (FiP) (filled red circles) and type-I' variant 3 (filled blue circles). Loop 1 and loop 2 residues are highlighted in light grey color. Residue numbering is that of wild type hPin1 WW. While the differences in absolute B factors may result from crystal packing variations, loop 1 in both wild type and variant 3 appears to be more disordered than the embedding β sheet and clearly stand out as local maxima, while loop 1 in variant 1 (FiP) is conformationally more rigid, consistent with this loop in its natural context, the FBP28 WW domain (Fig. A.3).

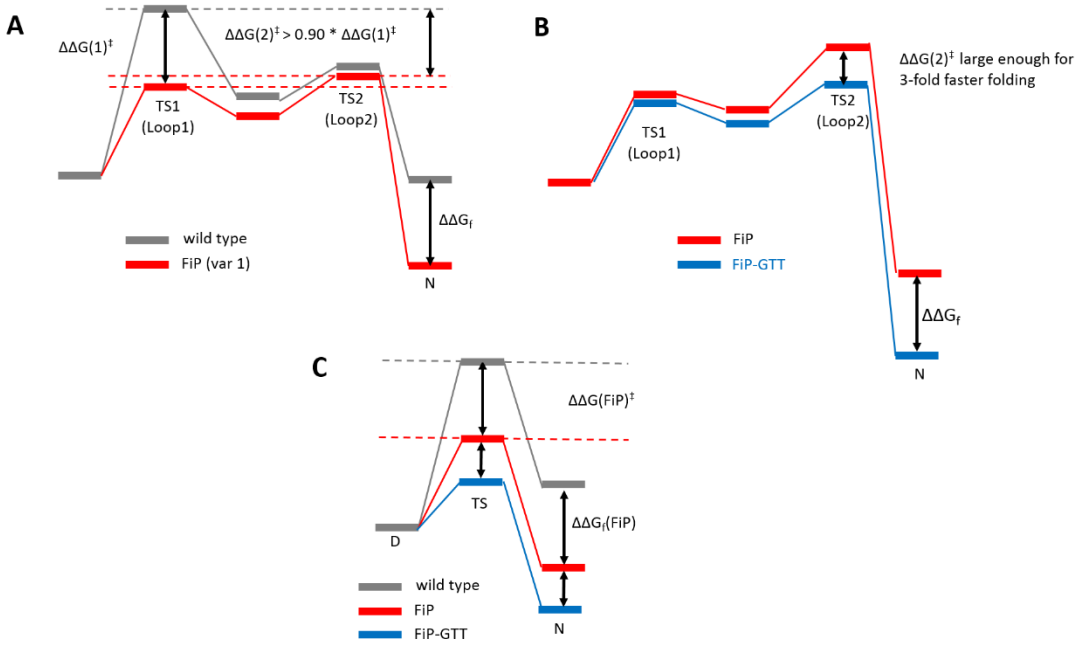

Fig. A.7: Simplified 1D-energy landscape models for loop 1 variant FiP. (A) Schematic energy landscapes of wild type (grey) and loop 1 deletion variant FiP (red), assuming a sequential folding proposed by Shaw, Maragakis et al. (see reference in main text). Folding of wild type is rate-limited by loop 1 nucleation (first barrier, highest energy) masking the (unperturbed) second barrier (loop 2 nucleation). Optimizing loop 1 in FiP lowers the first barrier (loop 1 nucleation) by $\Delta\Delta G(1)^{\ddagger}$, with a smaller effect on the second barrier. For the FiP Φ_M value to be > 0.90, the difference in the free energy barrier of the first and second transition ($\Delta\Delta G(2)^{\ddagger}$) must be on the order of $0.90 \cdot \Delta\Delta G(1)^{\ddagger}$. **(B)** Corresponding energy landscape of FiP that accounts for the approximately 3-fold acceleration of folding observed with the FiP-GTT variant containing the loop 2 stabilizing mutation N30G. The mere fact that loop 2 stabilization hastens FiP folding (13 µs vs. 4 µs folding rate) must imply that in FiP, loop 2 nucleation is rate-limiting for folding, and therefore the second barrier must be higher in free energy than the first barrier, which is difficult to reconcile with landscape (A). **(C)** Schematic energy landscape for wild type and the FiP variant obeying a simple two-state folding mechanism. Both stabilizing loop 1 and loop 2 mutations can act independently and/or additively on a single transition barrier, thus avoiding the above-mentioned inconsistency of the sequential folding model.

A.2 W34A mutant response

The response of the Ala-mutant W34A in Fig. 2.6C of the main text to temperature-tuning is unusual in that it is the only mutant that shows a decrease in Φ_M with temperature beyond experimental uncertainty. Such a trend that has also been reported for the analogous W39F mutant in the hYap65 WW domain. In both WW domains, the C-terminal Trp is largely surface-exposed and makes only one significant side chain contact in the folded protein - a clamp-like structure with the side chain of Y23 (Y28 in hYap65) in β strand 2 (Fig. A.2). The decrease in Φ_M of W34A (W39F in hYap65) might suggest that the interaction between W34 and Y23 is weakened in the folding transition state at higher temperature. Hydrophobic interactions, however, should strengthen at elevated temperature, and the Φ_M values of Y23A/L in hPin1 WW (slightly) increase with temperature, which argues against this hypothesis. Molecular dynamics simulations on WW domains reveal that β strand 3 is prone to forming transient, non-native interactions (main text references [9, 14]). As Trp residues are often found in helical structures, one plausible explanation for our observation is that in both hPin1 WW and hYap65 WW, the bulky, hydrophobic side chain of the C-terminal Trp engages in such transient and temperature-sensitive, non-native interactions that nevertheless must speed up folding, and that are disrupted by Ala (or Phe) mutations. Importantly, as the Φ_M value of W34A of hPin1 WW blends in well with the Φ_M values of other hairpin 2 mutants (main text figures Fig. 2.2A, Fig. 2.4A, Fig. 2.5A), its unusual temperature dependence becomes apparent only upon a more elaborate temperature-dependent Φ_M-value analysis.

APPENDIX B

Supplementary information of eliminating a protein folding intermediate by tuning a local hydrophobic contact

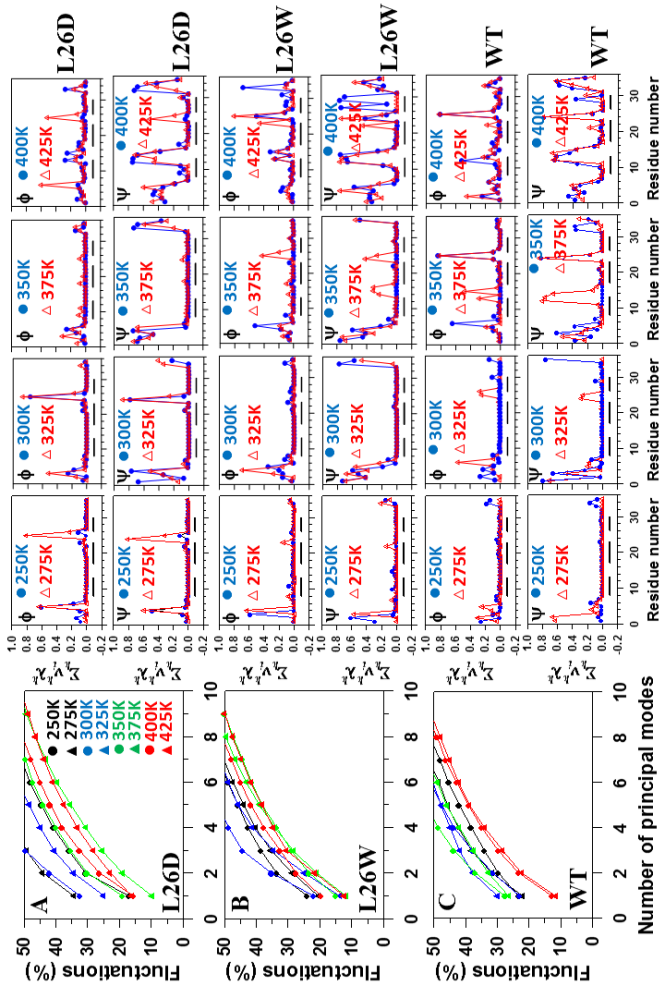

Fig. B.1: Percentages of the total fluctuations captured by the principal components for Leu26Asp (panel A), Leu26Trp (panel B) and the wild type (panel C) at eight different temperatures. The panels on the right represent contributions of the first *k* collective modes (*k* is the number of modes capturing at least 40% of the total fluctuations) to the MSF along the angles ϕ and ψ at eight different temperatures for Leu26Asp, Leu26Trp and the wild type. The black bars above each x-axis label the β-strand locations.

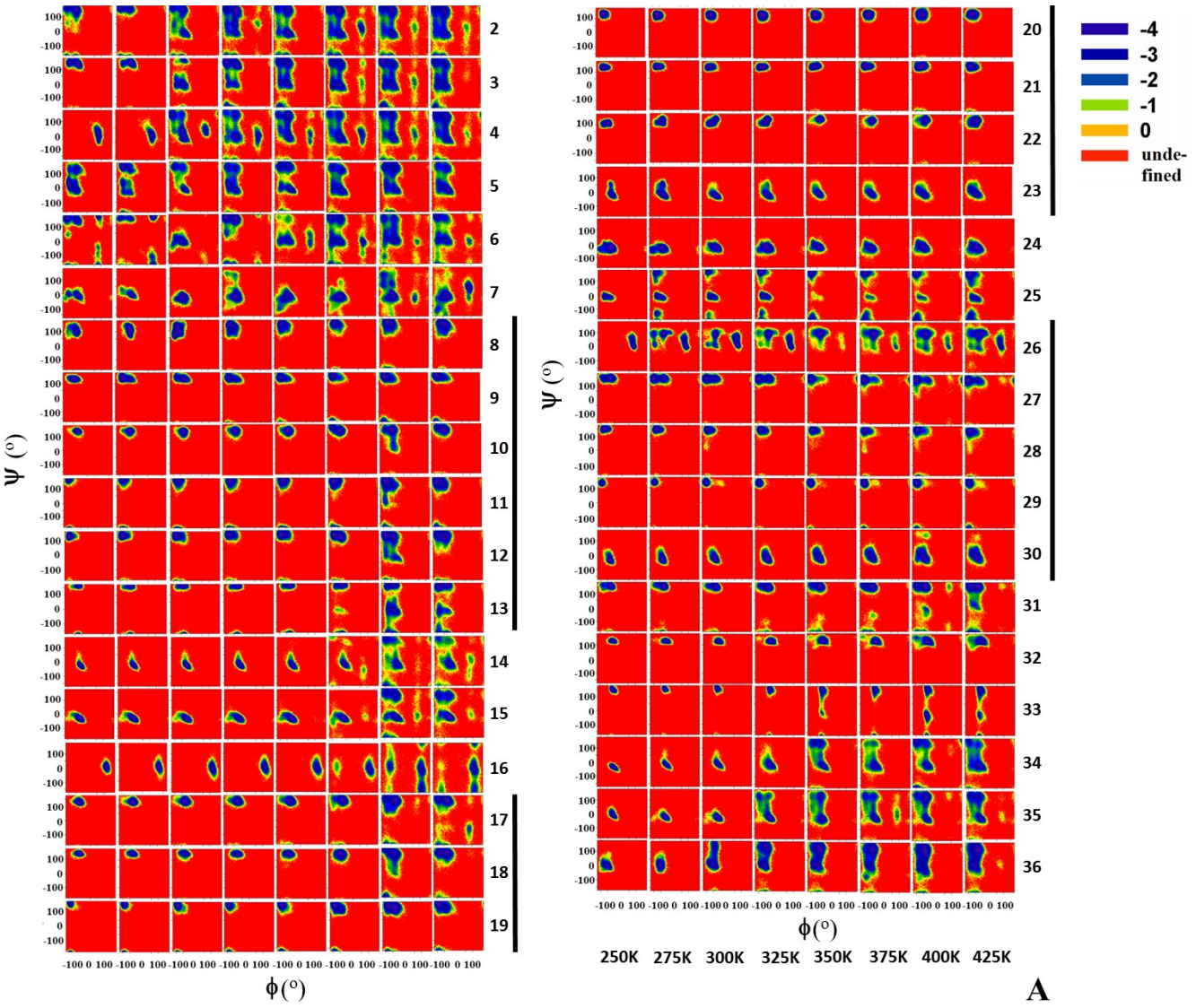

Fig. B.2: (panel A).

Figure B.2 (cont.)

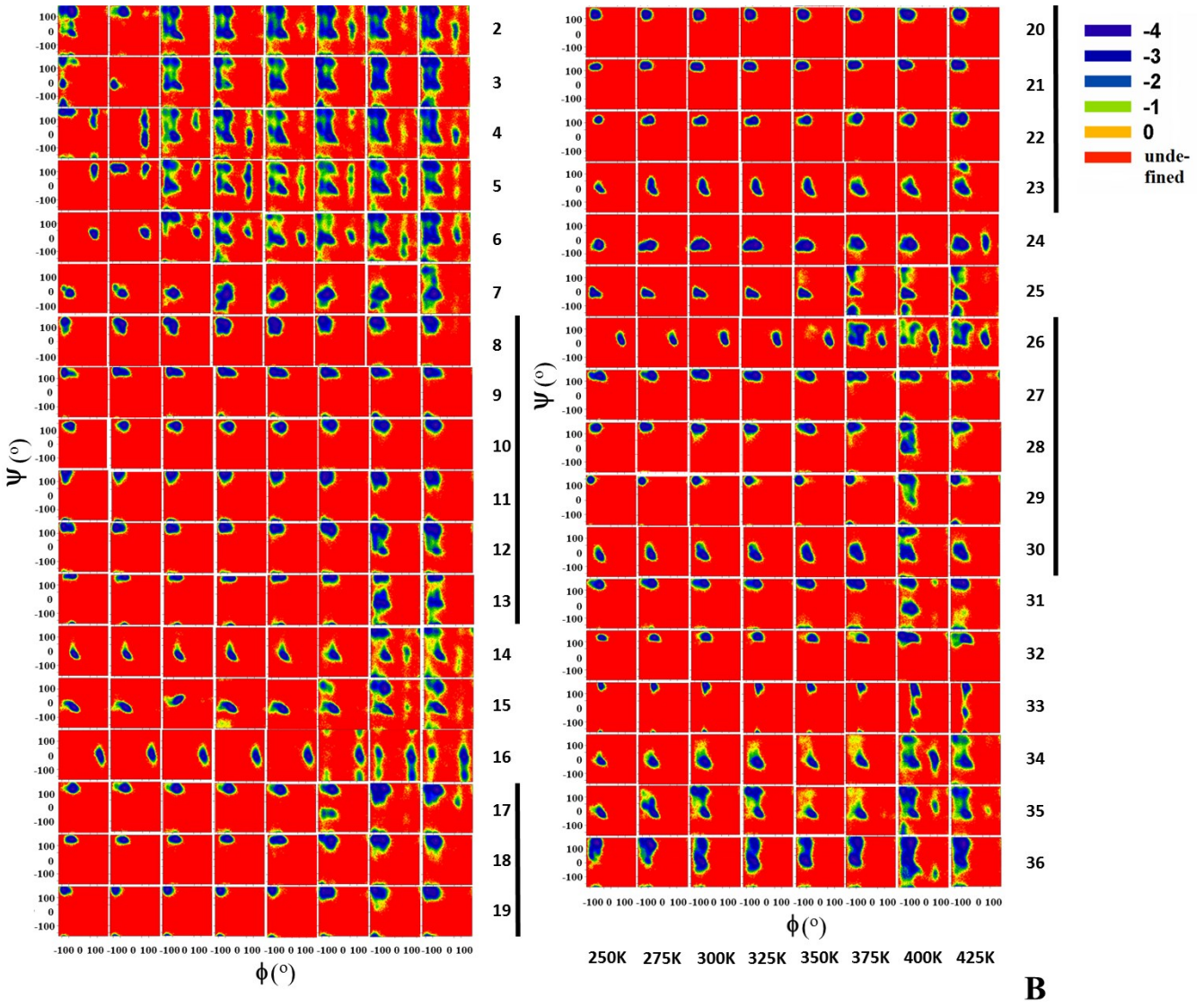

Fig. B.2: (panel B).

Figure B.2 (cont.)

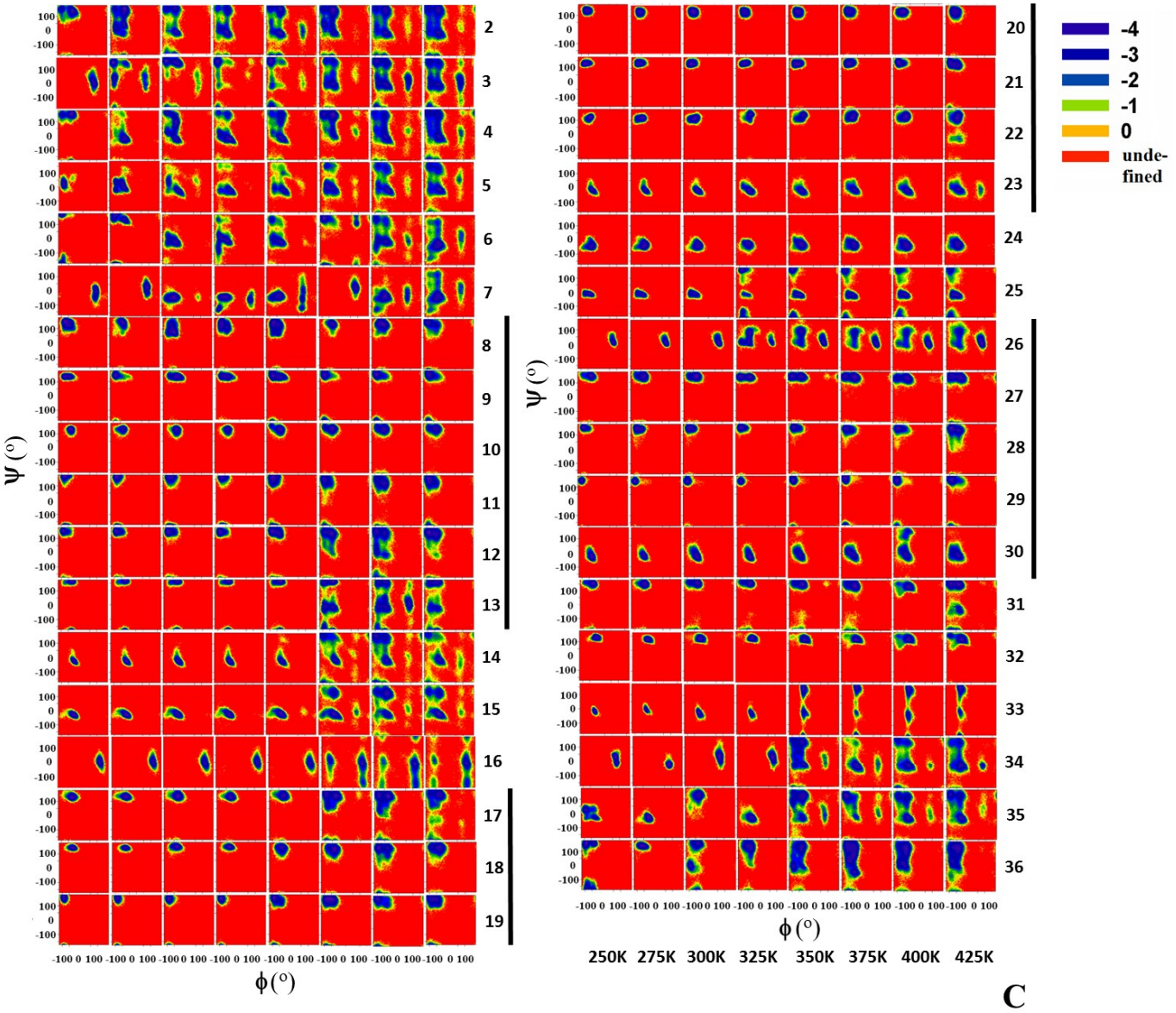

Fig. B.2: (Panel C). Each square represents a free-energy landscape (kcal/mol) along angles ϕ_i and ψ_i (in degrees) of Leu26Asp (panel A), Leu26Trp (panel B) and wild type (panel C) at 250K, 275K, 300 K, 325K, 350K, 375K, 400K, and 425K. The vertical black lines on each panel correspond to the β-strand regions. The numbers on the right are the residue numbers. The numbers on the X and Y axes are from the -180° to +180° regions of the ϕ and ψ angles, respectively. The colors on the upper-right side define the regions explored by the angles ϕ_i and ψ_i.

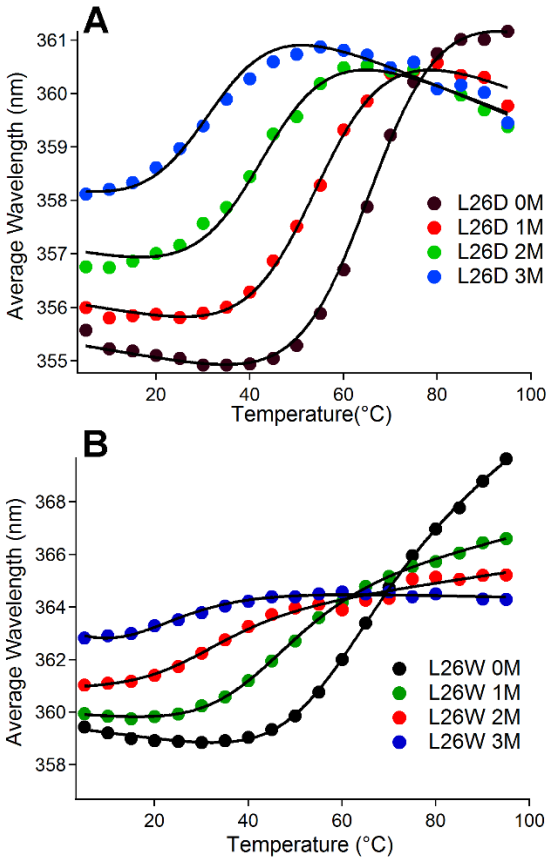

Fig. B.3: Global fits for both mutants given in Table 1 of the main text.

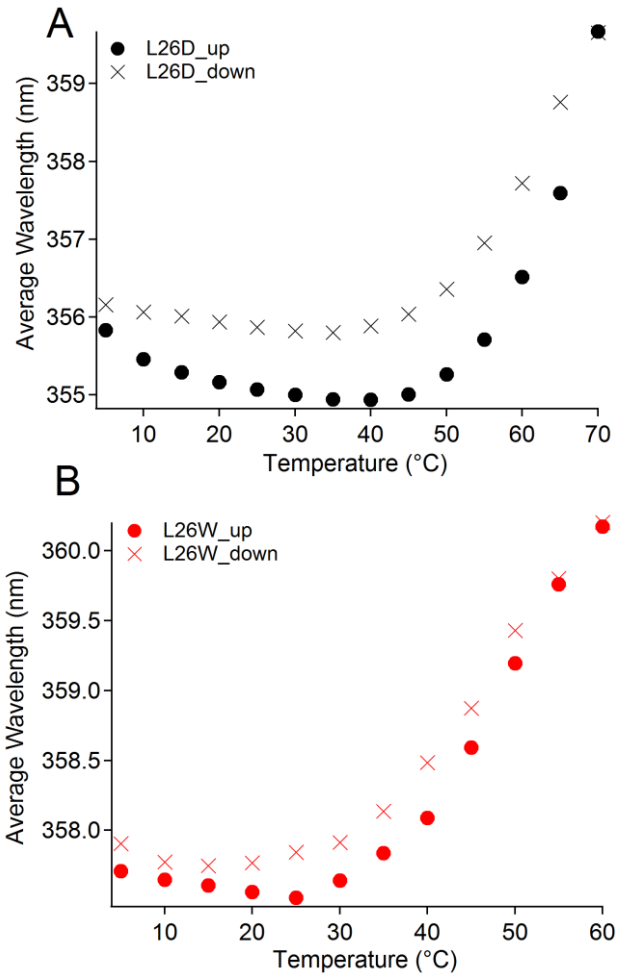

Fig. B.4: Reversible thermal melts of both Leu26Asp (A) and Leu26Trp (B)

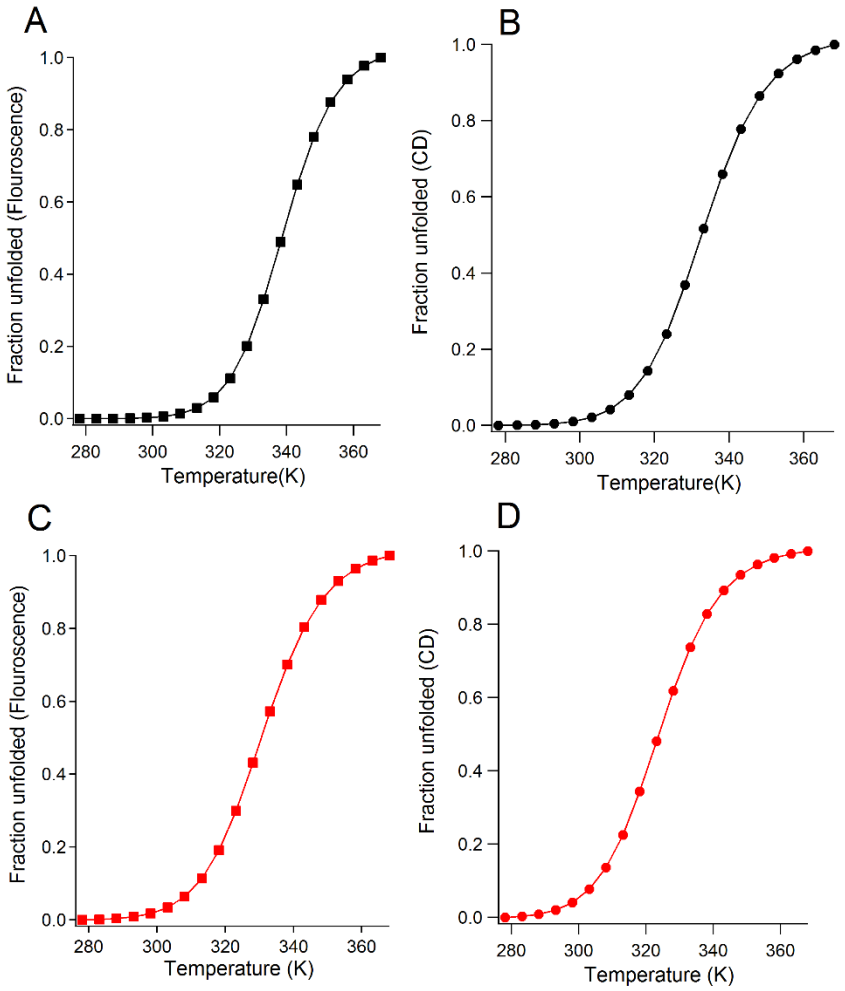

Fig. B.5: Fraction unfolded for the thermodynamics of both Leu26Asp (A, B) and Leu26Trp (C, D).

APPENDIX C

Supplementary information of the effect of fluorescent protein tags on phosphoglycerate kinase stability is non-additive

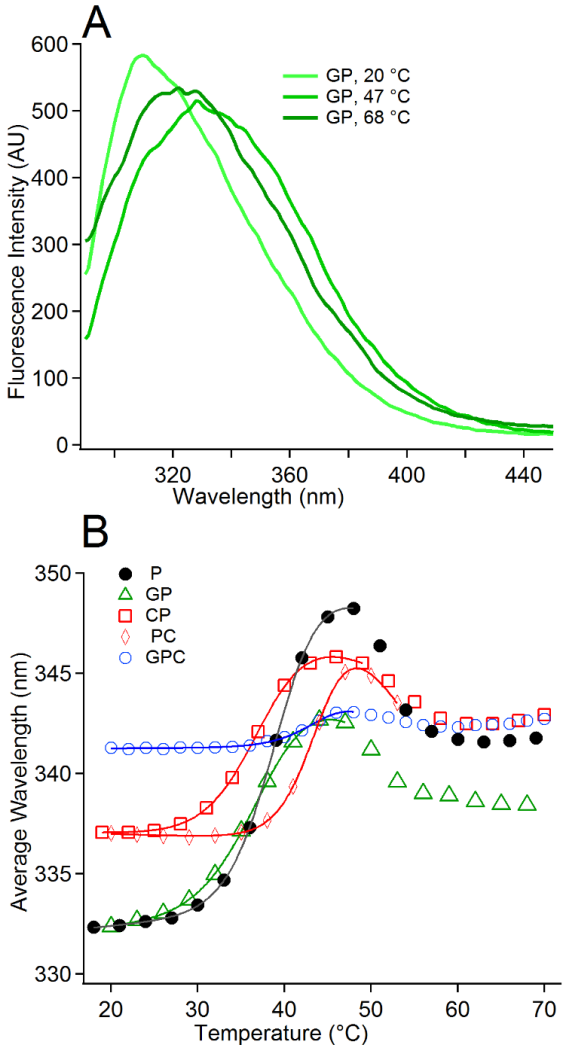

Fig. C.1: Thermal denaturation of the protein constructs as monitored by tryptophan. **A**. Emission spectrum of GP at 20 °C, 47 °C and 68 °C . **B.** Average wavelength to monitor the unfolding midpoint for GP (green), CP (red), PC (red diamonds),GPC (blue) and P (black).

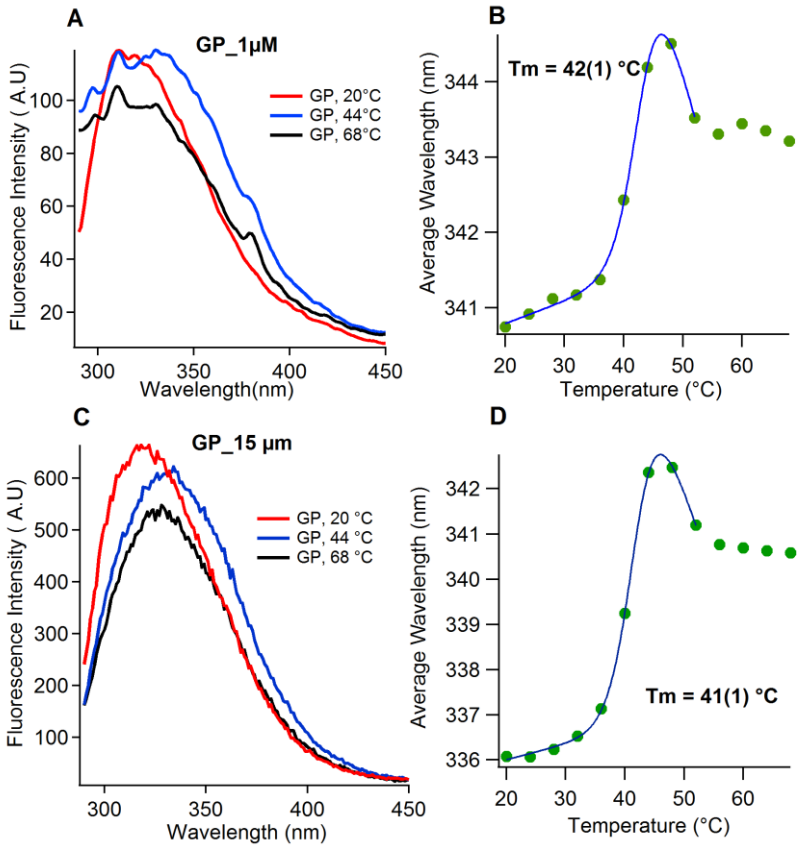

Fig. C.2: The hyper-fluorescent intermediate is not concentration dependent. (**A**) Representative tryptophan emission spectra at three different temperatures for 1 μM (GP) show a blue shift from 44 °C to 68 °C. (**B**) Thermal denaturation curves show melting point of 42 ± 1 °C for 1 μM (**B**). (**C**) Tryptophan emission spectra at three different temperatures for 15 μM GP again show a blue shift from 44 °C to 68 °C. **D**) Thermal denaturation curves show melting point of 42 ± 1 °C for 15 μM.

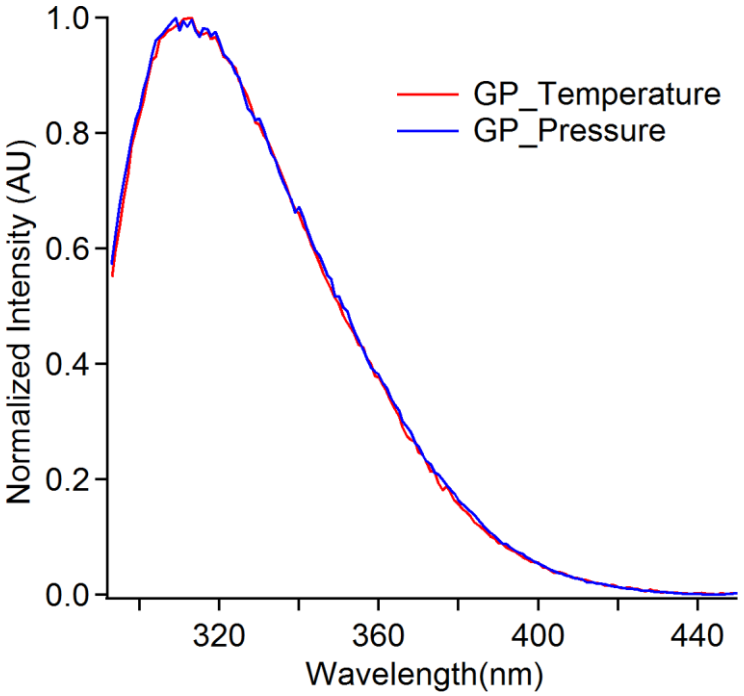

Fig. C.3: Comparison of initial pressure and temperature fluorescence emission spectrum at 1 bar and 23 °C from two different samples.

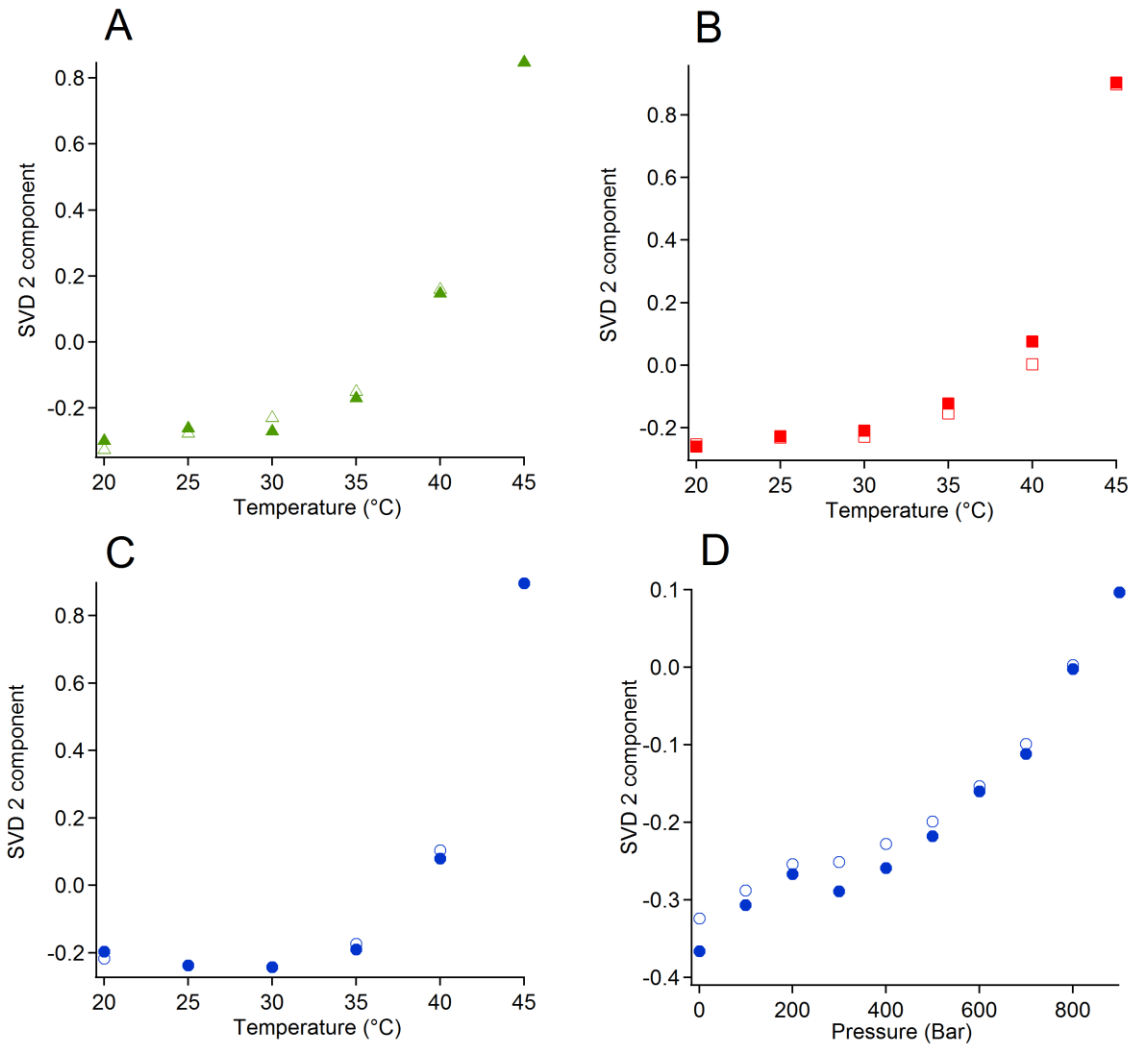

Fig. C.4: Singular value analysis reversibility plots for temperature and pressure measurements (see also Figs. C.11 and C.12). **A**. GP construct forward (filled green marker) and backward melt (open green marker) overlap well as seen in the second principal component *vs.* temperature plot. Similarly CP **(B)** and GPC **(C)** constructs are also reversible. **D.** Representative Plot for pressure reversibility of GPC under pressure forward (filled blue marker) and backward (open blue marker).

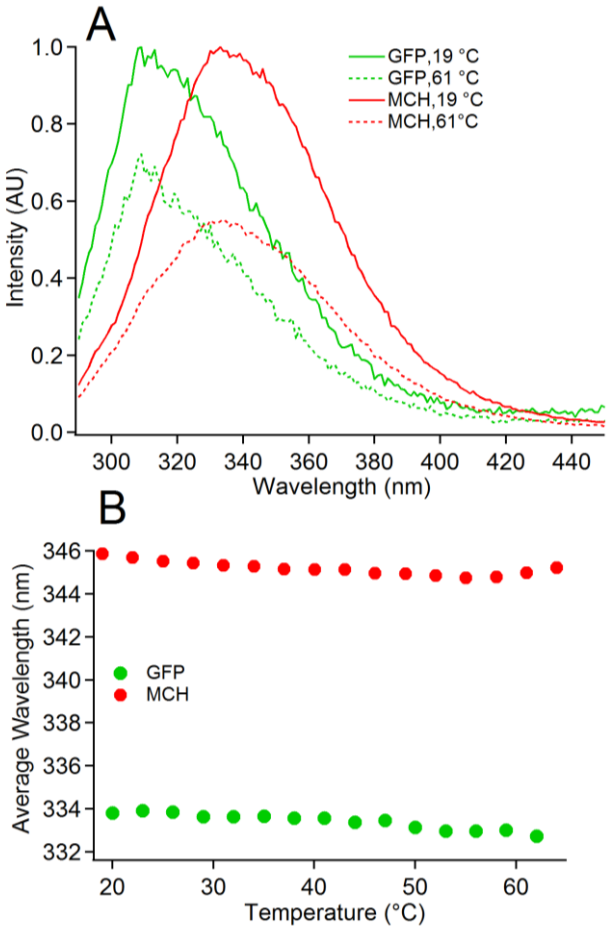

Fig. C.5: A. Tryptophan fluorescence spectrum monitored during a temperature melt for mCherry and AcGFP1. **B.** Average tryptophan emission wavelength calculated for AcGFP1 and mCherry. Neither AcGFP1 nor mCherry shows significant wavelength shift or a co-operative transition in the experimental temperature range.

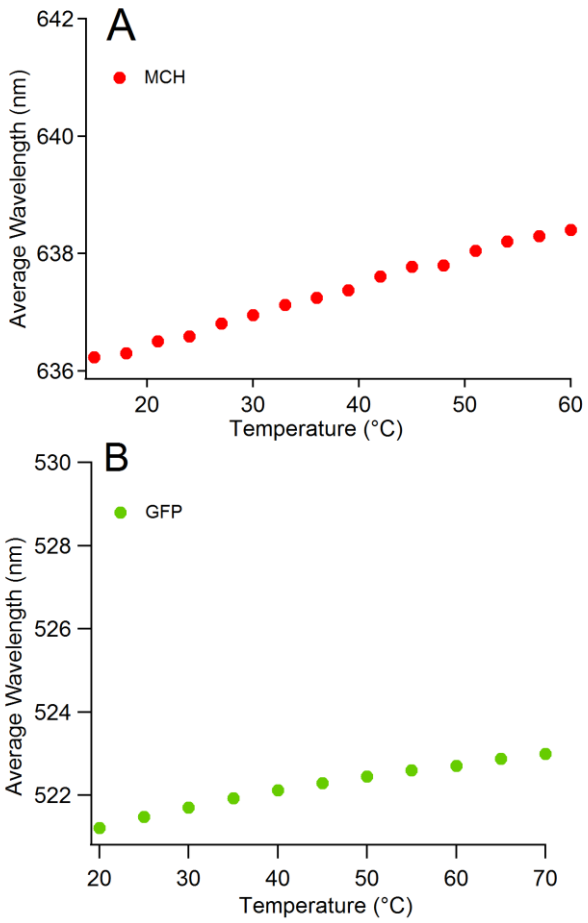

Fig. C.6: Average wavelength vs Temperature plots for AcGFP1 and mCherry excited at 475 and 585 nm respectively. Neither mCherry (**A)** or AcGFP1 (**B)** showed a significant wavelength shift or co-operative transition from 10 °C – 60 °C.

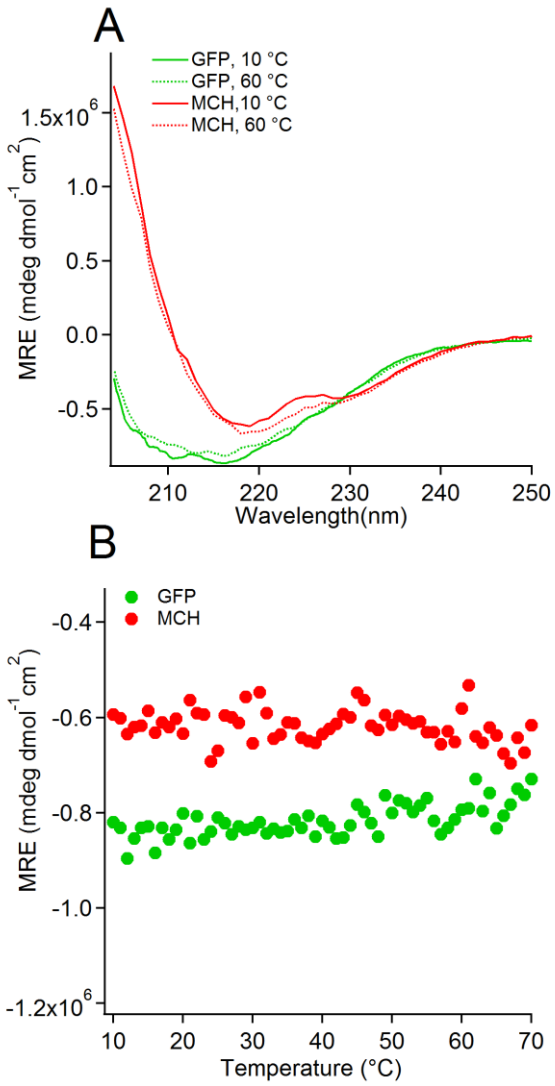

Fig. C.7: A. CD spectra of mCherry (red) and AcGFP1 (green) at low (solid lines) and high (dotted lines) temperatures show no change in secondary structure. **B.** Continuous CD measurement; MRE at 222 nm *vs.* temperature for AcGFP1 and mCherry showing no change in secondary structure.

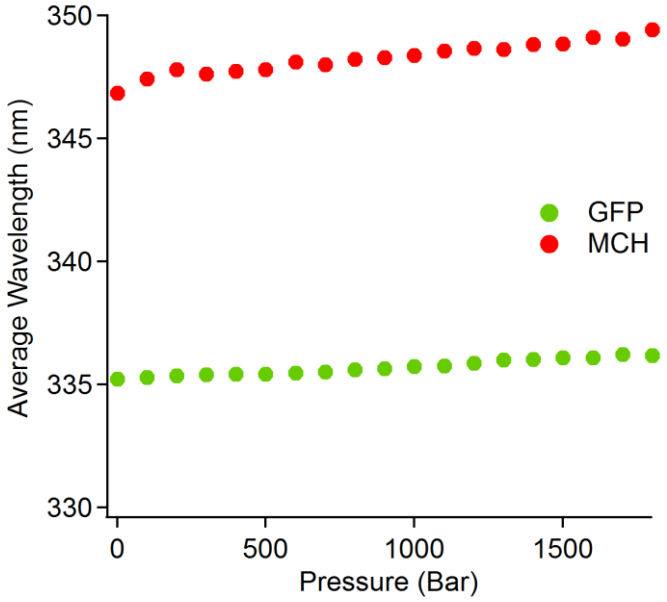

Fig. C.8: Tryptophan detected pressure titration of AcGFP1 (green) and mCherry (red). The average wavelengths show only a linear change in the pressure regions where the tagged PGK constructs undergo cooperative transitions.

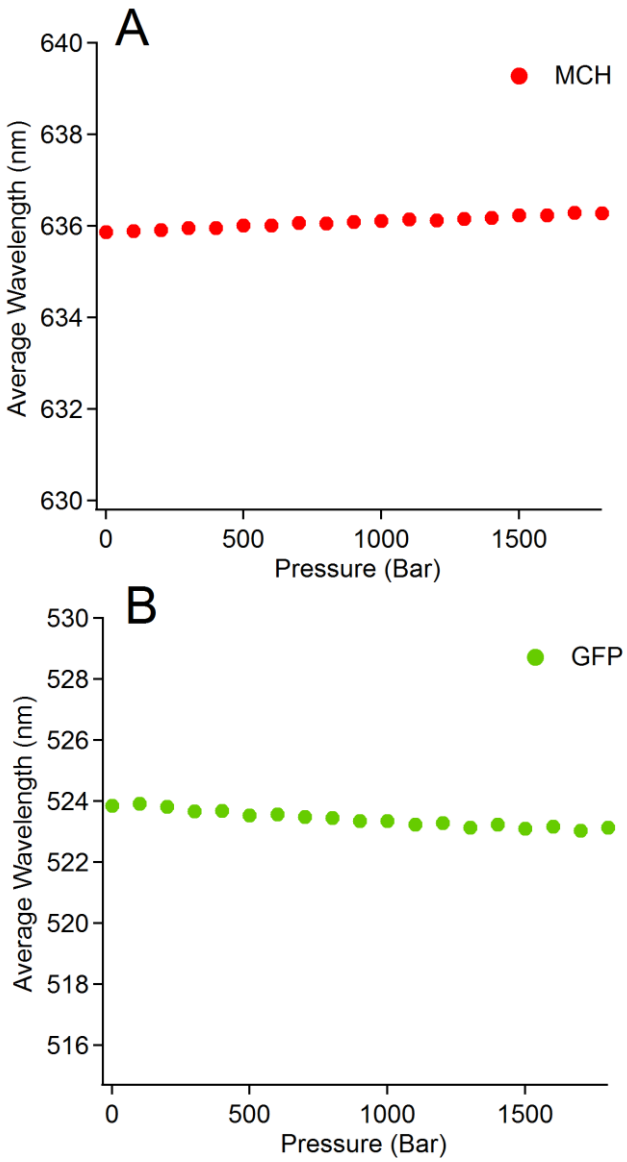

Fig. C.9: Average wavelength analysis for AcGFP1 and mCherry excited directly at 475 nm and 585 nm respectively shows no significant change in the experimental pressure range.

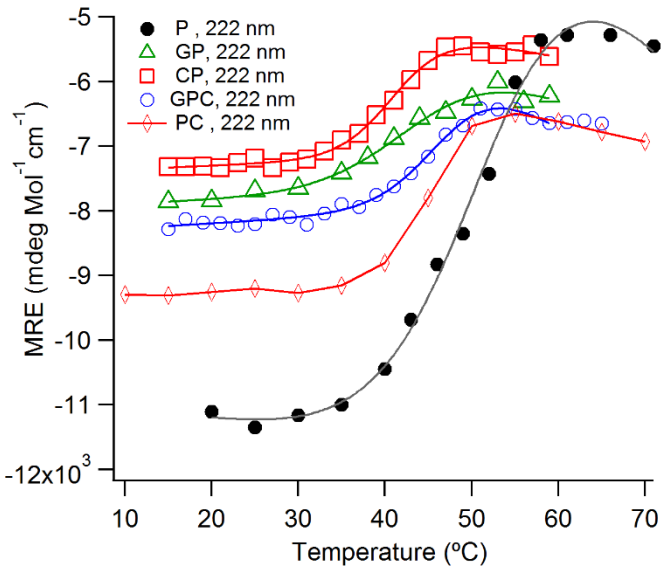

Fig. C.10: Absolute MRE *vs.* temperature for all protein constructs; GP (green), P (black), GPC (blue) and CP (red) PC (red diamonds). The tags are thermally stable, so the melting curve monitors PGK denaturation with or without various tags. For this reason P has a much larger MRE change.

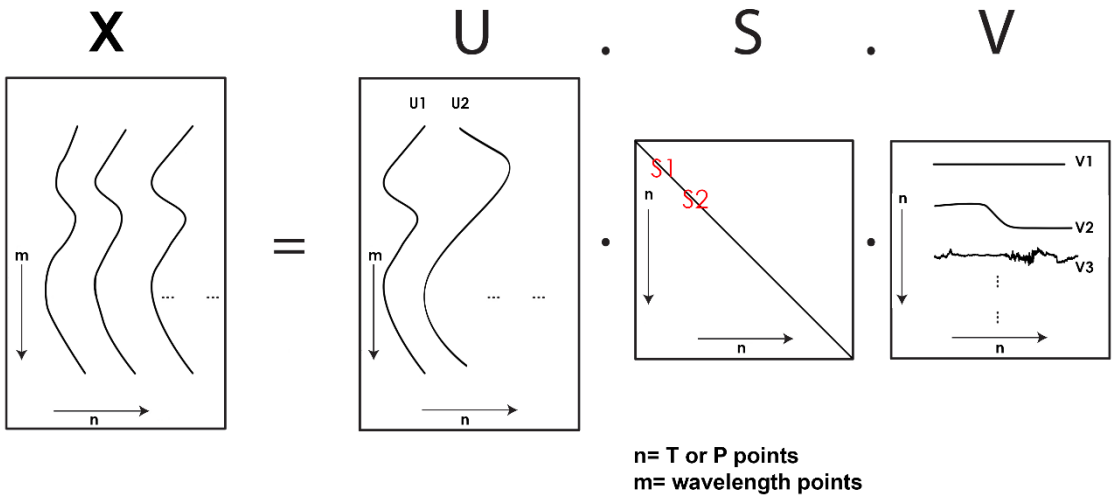

Fig. C.11: Top row: Singular value decomposition (SVD) was applied to analyze the data obtained from both thermal and pressure melts. In SVD, a data matrix (X) is decomposed uniquely into three matrices. Each column of the data matrix contains a spectrum, and the temperature or pressure changes as one goes right from column to column. On the right hand side of the equality are are orthogonal SVD basis vectors U that represent the basis spectra, singular values S that represent the importance of each basis spectrum to reconstruct the original spectra, and a trend matrix V that shows how each basis spectrum contributes as a function of temperature or pressure. We conducted SVD analysis to emphasize that in the temperature and pressure range of our experiment, we observed a quasi two-state transition, as shown by the "V2" component on the far right.

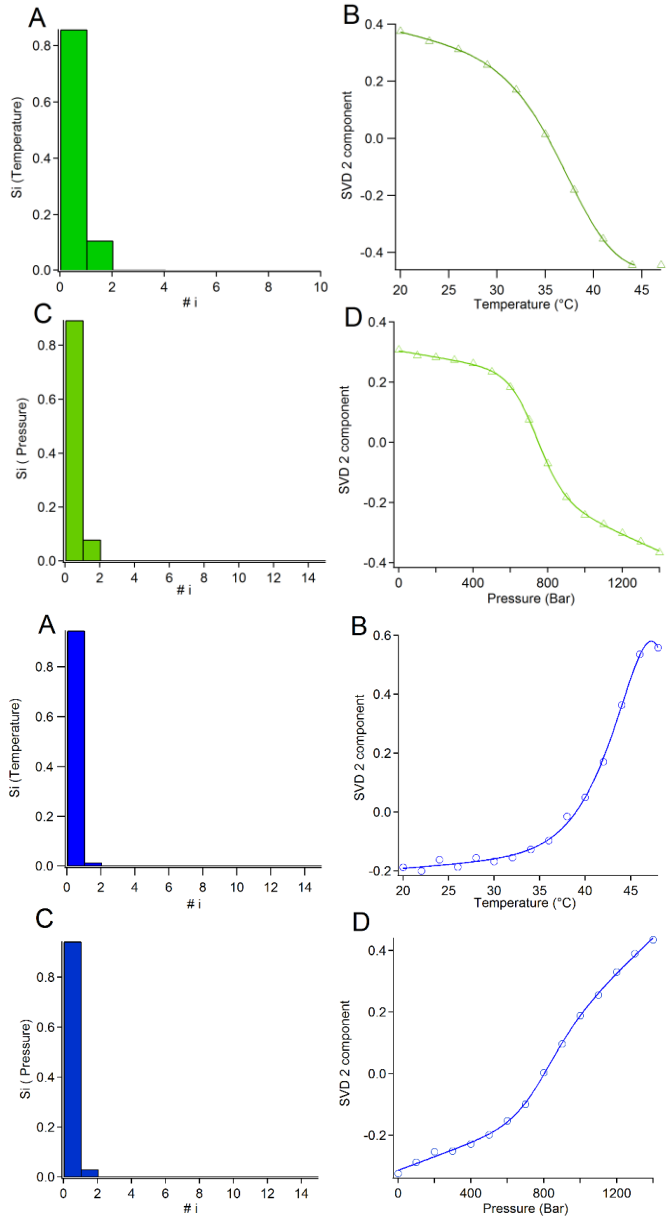

Fig. C.12: Top four panels (green): Singular value decomposition (SVD) analysis of the tryptophan emission spectra from 20 to 50 °C and from 0 to 1400 bar for the GP construct. 95-98% of the signal change is accounted for by the first two principal components for temperature **(A)** and pressure **(C).** The second principal component undergoes a transition at melting temperature T_m= 38 (±1) **(B)** and 770 (±10) for pressure **(D).**

Bottom four panels (blue): Singular value decomposition (SVD) analysis of the tryptophan emission spectra from 20 to 50 °C and from 0 to 1400 bar for the GPC construct. 95-98% of the signal change is accounted for by the first two principal components for temperature **(A)** and pressure **(C).** Second principle component undergoes a transition at melting temperature T_m = 44 (±1) **(B)** and P_m = 780 (±10) for pressure **(D).**

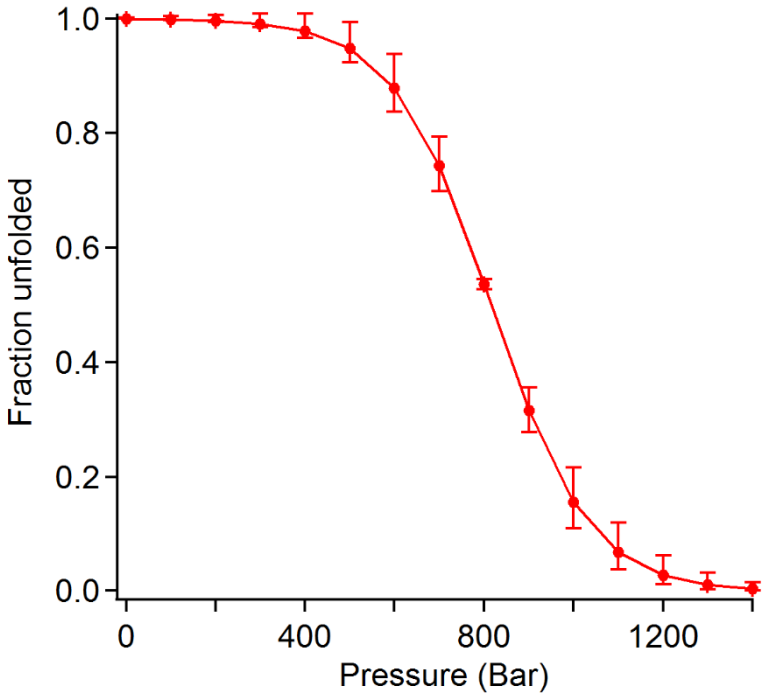

Fig. C.13: Fraction folded *vs.* pressure plot for CP construct. Error bar represent the variation in the signal based on the errors in fitted P_m and V_0. (See also Appendix C tables of fitting parameters.)

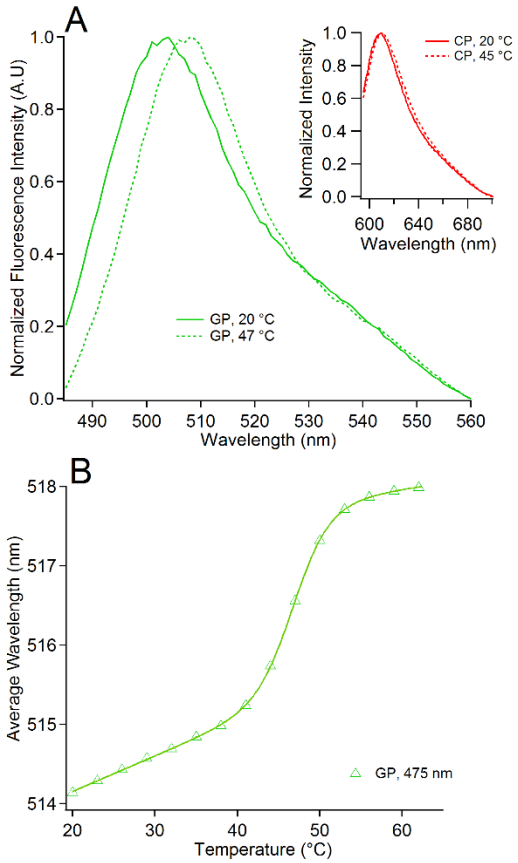

Fig. C.14: Thermal denaturation shift in AcGFP1 emission spectrum **A.** Normalized fluorescence intensity *vs.* wavelength plot for GP at 20 °C (dashed green line), 47 °C (dotted green line). *Inset:* Normalized fluorescence intensity *vs.* wavelength for CP showing no significant shift **B.** Average GFP emission wavelength of GP (green) is sensitive to GP unfolding showing a cooperative transition.

Table C.1: Cooperativity parameters (g_X in eq. 2 in the main text) for the protein constructs corresponding to Table 4.1 in the main text. Bottom: ΔΔG= =gX . [Tm(X)-Tm(P)] referenced to the melting temperature of P. This illustrates non-additivity directly in terms of a free energy parameter, rather than the midpoint quantities Pm and Tm given in the main text.

Protein	Thermal Denaturation ΔΔG(kJ/mol)		Pressure Denaturation ΔΔG(kJ/mol) (280 nm excitation)
	Measured via Fluorimeter (280 nm excitation)	Measured via Circular Dichroism (CD)	
P	**0**	**0**	**0**
GP	**-1.5**	**-4.6**	**-6.5**
CP	**-1.7**	**-8.5**	**-6.4**
PC	**3.5**	**-6.9**	**-5.7**
GPC	**4.6**	**-5.2**	**-8.0**

Protein	Thermal denaturation		Pressure denaturation (280 nm excitation) g_P (L mol^{-1})
	Fluorimeter (280 nm excitation) g_T (J mol^{-1} K^{-1})	Circular Dichroism (CD) g_T (J mol^{-1} K^{-1})	
P	930 ± 30	450 ± 110	0.25 ± 0.03
GP	730 ± 60	510 ± 240	0.4 ± 0.1
CP	830 ± 30	850 ± 120	0.28 ± 0.06
PC	1180 ± 130	990 ± 60	0.26 ± 0.02
GPC	1160 ± 110	740 ± 80	0.23 ± 0.04

Table C.2: Cooperativity parameters (g_X in eq. 2 of the main paper) for constructs tagged with GFP or AcGFP1 and mCherry and monitored by fluorescence excited at 475 nm, or by FRET Donor/Acceptor ratio for the corresponding Table 4.2 in the main text.

Protein	**Thermal denaturation g_T (J mol^{-1} K^{-1}) 475 nm excitation**	**Pressure denaturation g_p (L mol^{-1}) 475 nm excitation**
GP	1080 ± 30	0.36 ± 0.05
GPC	790 ± 40	0.30 ± 0.03

Table C.3: Signal baselines $S_{U,F}$ for the protein constructs for the corresponding Table 4.2 in main text
$S_U = c+d(T-T_m)$; $S_F = a+b(T-T_m)$ for pressure, replace T by P and T_m by P_m

Protein	Temperature parameter $S_{U,F}$		Pressure parameter $S_{U,F}$ (280 nm excitation)
	Fluorimeter (280 nm excitation)	Circular Dichroism (CD)	
P	a=333.2 ± 0.2;b=0.040 ± 0.009 c=351.7 ± 0.6;d=-0.31 ± 0.06	a=-11831 ± 1;b=-20 ± 40 c=-1884.8 ± -0.1;d=-180 ± 125	a=331 ± 1;b=0.0077 ± 0.0008 c=326.1 ± 0.2;d=0.0018 ± 0.0006
GP	a=332.9 ± 0.6;b=0.034 ± 0.03 c=348.4 ± 0.3;d=-0.4 ± 0.1	a=-7696 ± 511;b=6 ± 18 c=-5452 ± 961;d=-50 ± 68	a=336.7 ± 0.2;b=0.0038 ± 0.0001 c=335.7 ± 0.2;d=0.0023 ± 0.0005
CP	a=336.9 ± 0.2;b=-0.008 ± 0.012 c=346.8 ± 1;d=-0.23 ± 0.03	a=-7183 ± 108;b=-5.7 ± 5.2 c=-5153 ± 179;d=-25 ± 12	a=341.2 ± 0.2;b=0.0023 ± 0.0001 c=339.9 ± 0.1;d=0.0018 ± 0.0003
PC	a= 336.7 ± 0.1;b= -0.0118 ± 0.005 c= 349.4 ± 0.2;d= -0.66 ± 0.03	a=- 6068 ± 38;b= 36 ± 2 c= -9169 ± 47;d= 3.8 ± 1.2	a=343.4 ± 0.6;b=0.0013 ± 0.0001 c=337.7 ± 0.1;d=0.0028 ± 0.0003
GPC	a=341.3 ± 0.1;b=0.005 ± 0.003 c=343.9 ± 0.1;d=-0.13± 0.01	a=-5618 ± 223;b=-84 ± 25 c=-7978 ± 91;d=8 ± 3	a=338.7 ± 0.1;b=0.0030 ± 0.0001 c=338.3 ± 0.1;d=0.0017 ± 0.0003

Table C.4: Signal baselines $S_{U,F}$ for constructs tagged with GFP or AcGFP1 and mCherry and monitored by fluorescence excited at 475 nm, or by FRET Donor/Acceptor ratio for the corresponding Table 4.2 in main text

Protein	Temperature $S_{U,F}$ 475 nm excitation	Pressure $S_{U,F}$ 475 nm excitation
GP	a=518.6 ± 0.1;b=-0.002 ± 0.011 c=515.5 ± 0.1;d=0.0431 ± 0.0001	a=519.7 ± 0.1;b=0.00015 ± 0.00001 c=518.4 ± 0.1;d=-3.755e-005 ± 0.00016
GPC	a=0.3184 ± 0.0107;b=0.011441 ± 0.00047 c=0.95384 ± 0.0225;d=0.0025878 ± 0.00141	a=0.57 ± 0.2;b=0.00035 ± 0.00014 c=-0.002 ± 0.001;d=0.00021799 ± 3.87e-005

$S_U = c+d(T-T_m)$; $S_F = a+b(T-T_m)$ for pressure, replace T by P and T_m by P_m

APPENDIX D

Supplementary information of environmental fluctuations and stochastic resonance in protein folding

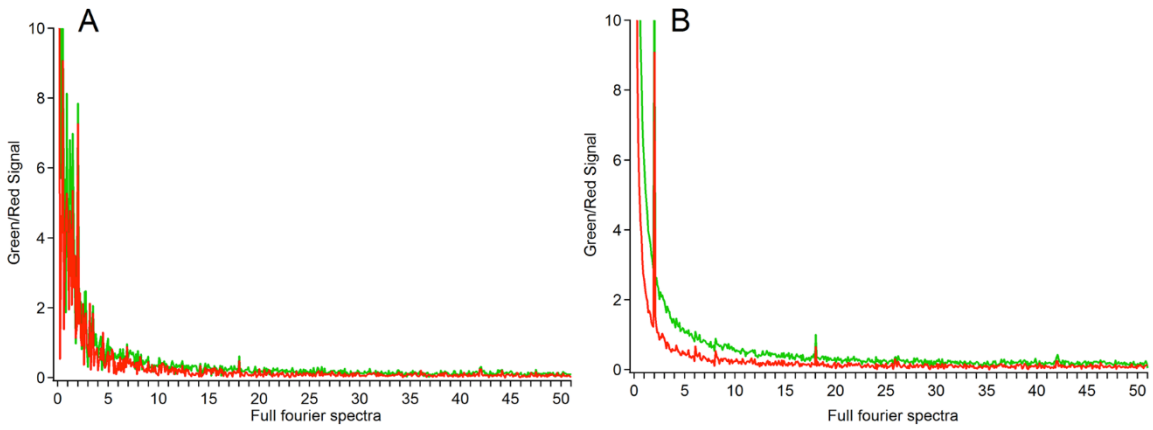

Fig. D.1: A. Full Fourier spectra for 2 Hz modulation as the noise is increased. The spectra has multiple peaks and the signal is beginning to swamp by high noise level. **B.** Showing full Fourier spectra for the 2 Hz modulation. A clear peak at 2 Hz is seen at lower noise level. Part of the peak is background signal due to quantum yield modulation, part stochastic resonance, as seen by the plot as a function of noise amplitude in Fig. 5.7 of the main paper.

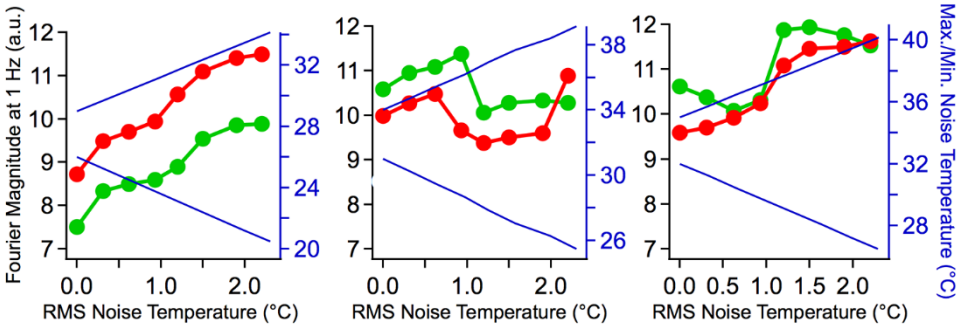

Fig. D.2: Stochastic resonance, detected by Fourier transform magnitude of the donor (green) and acceptor (red) signals grows in when a root-mean-squared (RMS) temperature noise of ca. 1.2 °C is superimposed on the sub-threshold sine wave modulation at 1Hz (τ<k_{obs}). (A) T0 = 28 °C. (B) T0=32 °C. (C) T0=33 °C. Stochastic resonance grows in at $\sqrt{\delta(T-T_0)^2} \approx 1.25$ °C as the average temperature T0 is increased, but the signal is weaker than at 2 Hz. The blue decreasing line is the minimal noise temperature applied on the protein at each RMS temperature, and the blue increasing line is the maximal noise temperature thereof.

D.1 FORTRAN Code

FORTRAN Code Named "FRETMODULATE" which integrates the kinetic equations of a protein subjected to external temperature modulation; detected by FRET labeling. The code is similar to a modulation model by Lemarchand, JCP 138, 244109 (2013), with addition of temperature-dependent quantum yield of the chromophores that can interfere with the fluorescence changes due to chemical reaction. The code should be compiled with the Intel F90 compiler on Macintosh OS X 9.8 or later. Use of static flag or equivalent is required to avoid problems during dynamic variable space allocation. This code was tested and run with the Intel FORTRAN Composer XE for OS X 9.8. The code can perform a non-linear least squares fit to optimize parameters by setting fitting flags in its fort.1 input file equal to 1 for each parameter to be fitted.

1) Sample input file (fort.1); three parameters (preceding bold fitting flags) are fitted.
311,0 ! T0 for all states and its fitting flag: fitflag=0 (no fit) or 1 (fit)
0,0,0,0 ! Free energy coefficients for state 1 (folded) and fitting flags
0,-0.204,0,0 ! Free energy coefficients for state 2 (denatured) and fitting flags
1e5,22.64,0,0,**1**,0 ! Prefactor, barrier coefficients between states 1 and 2 and fitting flags
0.4965,0.5,**1**,**1** ! Green and red signals for state 1 and fitting flags
0.52,0.3,0,0 ! Green and red signals for state 2 and fitting flags
-0.011,-0.016,310,0,0,0 ! Green and red signal slopes per degree and reference T (i.e. -0.02 means quantum yield drops by 2% for every degree away from ref. T) and fitting flags
14,1224,311,0.02,0 !Number of temperature waveforms, seed, average temperature, printing interval time, and printflag

```
4.00 3 1 0 0.52 1
2.50 3 1 0 2.67 1
2.00 3 1 0 4.10 1
1.25 3 1 0 6.18 1
1.00 3 1 0 7.28 1
0.80 3 1 0 8.51 1
0.625 3 1 0 10.11 1
0.50 3 1 0 11.36 1
0.40 3 1 0 12.80 1
0.25 3 1 0 15.70 1
0.125 3 1 0 17.43 1
0.10 3 1 0 16.94 1
0.08 3 1 0 15.95 1
0.078125 3 1 0 15.89 1
```

2) Sample output file (fort.7): This file outputs average modulation temperature, chi-squared of the fit, the three parameters from the fort.1 file that were fitted plus uncertainties, fitted parameter correlation matrix, an output of all fitted and unfitted parameters that can be pasted into fort.1 for another fit, and the result: modulation frequency, observed phase, uncertainty, calculated phase, fitting error.

```
Average temperature    311.000000000000
 Chi-squared=    1.03916832386365
 #, Parameter, Parameter uncertainty:
         1   22.9014183110495       0.103484887339052
         2  0.511808045002848       3.933723237818611E-003
         3  0.503817459108459       8.208087196948434E-003
 Correlation matrix:
  1.00000 -0.20676 -0.29079
 -0.20676  1.00000  0.60584
 -0.29079  0.60584  1.00000

 New values of all input parameters and fitting flags:
   311.000000000000              0
 0.000000000000000E+000  0.000000000000000E+000          0          0
 0.000000000000000E+000 -0.204000000000000              0          0
  100000.000000000       22.9014183110495       0.000000000000000E+000
          0          1          0
 0.511808045002848       0.503817459108459              1          1
 0.520000000000000       0.300000000000000              0          0
-1.100000000000000E-002 -1.600000000000000E-002   310.000000000000
          0          0          0

 Nu(Hz) Obs. phase(deg) Calc. phase(deg)
  0.250   0.520   1.000   1.562  -1.042
  0.400   2.670   1.000   2.500   0.170
  0.500   4.100   1.000   3.094   1.006
  0.800   6.180   1.000   4.900   1.280
  1.000   7.280   1.000   6.062   1.218
  1.250   8.510   1.000   7.500   1.010
  1.600  10.110   1.000   9.300   0.810
  2.000  11.360   1.000  11.125   0.235
  2.500  12.800   1.000  13.125  -0.325
  4.000  15.700   1.000  17.000  -1.300
  8.000  17.430   1.000  18.000  -0.570
 10.000  16.940   1.000  16.875   0.065
```

```
 12.500  15.950   1.000  15.625   0.325
 12.800  15.890   1.000  14.400   1.490
```

Code below :

```
module dimensio
implicit none
integer(4), parameter :: predim=40000,dim=400000,statedim=5,wavedim=20
end module dimensio

module thermokin
use dimensio
implicit none
real(8) :: T0, gcoef(statedim,0:2),g(statedim)
real(8) :: km, gdcoef(statedim,statedim,0:2),gd(statedim,statedim)
integer(4) :: T0flag,gcflag(statedim,0:2)
integer(4) :: kmflag,gdcflag(statedim,statedim,0:2)
end module thermokin

module waves
use dimensio
implicit none
integer(4) :: m_waves,itime,maxfac,time_points,use_points
real(8) :: Speriod(wavedim),Samplitude(wavedim),Somega(wavedim)
real(8) :: Nperiod(wavedim),Namplitude(wavedim)
real(8) :: Taverage,temperature(0:dim),tstep,odestep,dtprint
end module waves

module sigparameters
implicit none
real(8) :: s1red,s1green,s2red,s2green,sred_slope,sgreen_slope,Tsignal
integer(4) :: s1rflag,s1gflag,s2rflag,s2gflag,sr_slopeflag,sg_slopeflag,Tsflag
end module sigparameters

module popsigoutputs
use dimensio
implicit none
real(8) :: popmaxtime(statedim,wavedim),popphase(statedim,wavedim)
real(8) :: sigmaxtime(statedim,wavedim),sigphase(statedim,wavedim)
end module popsigoutputs

!    VARIABLE DECLARATION FOR NLLSQ

module masterdim
implicit none
integer(4), parameter ::  odim=1000, padim=49, jdim=50
end module masterdim

module obscalc
use masterdim
implicit none
!    EXCEPT FOR CALC(), all of these parameters must be read by the main program and be made available
!    to NLLSQ before the first subroutine call. See Subroutine NLLSQ for what typical values are.
integer(4) :: err,debug,maxiter
integer(4) :: onum, panum, paf(padim)
real(8) :: marq,delchi,grad,delgrad,chsq
real(8) :: obs(odim),osig(odim)
real(8) :: pa(padim),pasig(padim)
real(8) :: calc(odim)
end module obscalc

!    Least squares routines
!    M. GRUEBELE upgraded to F95 2000; note that nllsq must
!    be provided with all the input parameters listed in bold
!    in the subroutine declaration, including those which are
!    not passed directly but only declared in module nllsqfit,
!    obscalc or masterdim; obscalc can be used to pass fitting
```

```
!     parameters, obs. and calc. values to the main program and
!     subroutine cal(icount,ifail), which must be provided to
!     evaluate the array calc() given the array pa(); icount
!     sends the iteration count starting at 1 to cal, ifail
!     returns a flag in case cal fails [although not many checks
!     in nllsq are currently implemented].



!     Finds the inverse of a matrix by Gauss-Jordan elimination,
!     n is the actual matrix size, np the storage size (see
!     numerical recipes gaussj)
module invert
implicit none
integer(4),parameter :: nmax=100 !max dimension of matrix

contains

subroutine matinv(a,np,n,err)
implicit none
real(8) :: a(np,np),big,dum,pivinv
integer(4) :: ipiv(nmax),indxr(nmax),indxc(nmax),i, &
icol,irow,j,k,l,ll,n,np,err

err=0
if(n > np .or. n > nmax) then
write(6,*) "Dimensions exceeded in matinv."
stop
endif
do j=1,n
ipiv(j)=0
enddo
do i=1,n
big=0
do j=1,n
if(ipiv(j) /= 1) then
do k=1,n
if (ipiv(k) == 0) then
if (abs(a(j,k)) >= big)then
big=abs(a(j,k))
irow=j
icol=k
endif
else if (ipiv(k) > 1) then
err=4
return
endif
enddo
endif
enddo
ipiv(icol)=ipiv(icol)+1
if (irow /= icol) then
do l=1,n
dum=a(irow,l)
a(irow,l)=a(icol,l)
a(icol,l)=dum
enddo
endif
indxr(i)=irow
indxc(i)=icol
if (a(icol,icol) == 0) then
err=4
return
endif
pivinv=1./a(icol,icol)
a(icol,icol)=1
do l=1,n
a(icol,l)=a(icol,l)*pivinv
enddo
do ll=1,n
```

```
if(ll /= icol) then
dum=a(ll,icol)
a(ll,icol)=0
do l=1,n
a(ll,l)=a(ll,l)-a(icol,l)*dum
enddo
endif
enddo
enddo
do l=n,1,-1
if(indxr(l) /= indxc(l)) then
do k=1,n
dum=a(k,indxr(l))
a(k,indxr(l))=a(k,indxc(l))
a(k,indxc(l))=dum
enddo
endif
enddo
end subroutine matinv

end module invert

!    Non-linear least squares fitting module
module nllsqfit
use masterdim  !From this, needs padim and odim
use obscalc    !From this, needs obs,osig,calc,pa,paf,
!pasig,panum,onum,marq,delchi,grad,delgrad,chsq
use invert     !Needs this for matrix inversion calls
implicit none
integer(4) :: xpos(padim),xnum
real(8) :: ocalc(odim),weight(odim), &
beta(padim),alpha(padim,padim),alphin(padim,padim), &
deriv(padim,odim)

contains

subroutine nllsq(icount,ifail)
implicit none
integer(4) :: fnum,icount,ifail,i1,i2,i3
real(8) :: ograd,chold,x(padim),nextx(padim)

!    Variables that must be specified before subroutines in module are called by main program:

!    DEBUG     =0 or 1; 1 OUPUTS ADDITIONAL DEBUG INFO TO CONSOLE
!    OBS       ARRAY OF OBSERVED FUNCTION VALUES
!    OSIG      UNCERTAINTIES OF OBSERVABLES
!    ONUM      NUMBER OF OBSERVED PARAMETERS
!    PANUM     NUMBER OF FITTING PARAMETERS.
!    PA        ARRAY OF FITTING PARAMETERS
!    PAF       FITTING FLAGS;=0 FOR PARAMETERS HELD CONSTANT,
!                    =1 FOR FITTED PARAMETERS
!    MARQ      MARQUARD PARAMETER; LARGE VALUE INDICATES
!              STEEPEST DESCENT STEP, SMALL VALUE NEWTON
!              (LINEARIZED CHISQ) STEP. SHOULD BE SET TO 0.001
!              INITIALLY
!    DELCHI    IF TWO SUCCESIVE CHSQ AGREE WITHIN DELCHI, THE
!              FIT IS TERMINATED; TYPICAL VALUE: 0.01
!    DELGRAD   IF THE GRADIENT OF CHSQ FALLS BELOW DELGRAD, THE
!              FIT IS TERMINATED; TYPICAL VALUE: 0 IF DELCHI≠0
!    ERR       ERROR CODE;SHOULD BE SET TO ZERO INITIALLY.
!              ERR=1: NO PARAMETERS FITTED;ONLY CHSQ IS RETURNED
!              ERR=2: MORE PARAMETERS THAN OBSERVABLES FITTED
!              ERR=3: MATRIX INVERSION FAILED;JACOBIAN SINGULAR
!              ERR=4: RECOVERY FROM SINGULAR JACOBIAN FAILED
!              ERR=5: NUMBER OF ITERATIONS IN ITER EXCEEDED
!              ERR=6: MARQ EXCEEDED 10**10; SSQ CANNOT BE MINI
!                   MIZED BECAUSE GRADIENTS TO STEEP OR DELCHI
!                   SET UNREALISTICALLY SMALL
!    MAXITER   MAXIMUM NUMBER OF ITERATIONS(CALLS OF DERIVATIVE)
!
```

```
!    Variables that are output by module (in addition to ones that are overwritten with
!    with final values, such as "PA":
!    CALC       ARRAY OF CALCULATED FUNCT. VALUES RETURNED BY CAL()
!    PASIG      ARRAY THAT RETURNS UNCERTAINTIES IN PARAMETERS
!    CHSQ       CHI**2 OF FIT
!    XNUM       NUMBER OF ACTUALLY FITTED PARAMETERS
!
!    THE FOLLOWING PARAMETERS MUST NOT BE SET TO ANYTHING INITIALLY,
!    BUT ARE USEFUL FOR DEBUGGING OR ADDITIONAL INFORMATION ABOUT
!    THE FIT; DIMENSIONING IS AS FOR OBS AND PA
!
!    X          ARRAY OF THOSE PA WHICH ARE FITTED
!    GRAD       NORM OF THE GRADIENT OF CHSQ;SHOULD BE CLOSE TO
!               ZERO NEAR THE MINIMUM
!    NEXTX      ARRAY OF FITTING PARAMETERS BEFORE TESTING FOR
!               ITS VIABILITY IN DECREASING CHSQ
!    ALPHA      MATRIX THAT CONTAINS THE JACOBIAN TRANSPOSE TIMES
!               THE JACOBIAN
!    BETA       GRADIENT OF SSQ
!    DERIV      MATRIX OF DERIVATIVES OF ALL OBS W/R TO ALL PARA-
!               METERS
!    ALPHIN     ON OUTPUT, CONTAINS PARAMETER CORRELATIONS

!    DETERMINE CONSTANTS TO BE FIT AND THEIR NUMBER

icount=1
ograd=0d0
xnum=0
do i1=1,panum
if(paf(i1) /= 0) then
xnum=xnum+1
xpos(xnum)=i1
x(xnum)=pa(i1)
endif
enddo
if(xnum == 0) then
call chisq(chsq,x,icount,ifail)
err=1
return
endif

!    EVALUATE DEGREES OF FREEDOM

fnum=onum-xnum
if (fnum < 1) then
err=2
return
endif

!    CALCULATE WEIGHTS

do i1=1,onum
weight(i1)=1d0/(osig(i1)*osig(i1))
enddo

!    EVALUATE INITIAL CHSQ; NOTE THAT THIS ALSO CALCULATES
!    CALC FOR THE PARAMETER SET X

call chisq(chold,x,icount,ifail)

!    CALCULATE INITIAL GRADIENT OF CHISQ
!
do i1=1,xnum
beta(i1)=0d0
do i2=1,i1
alpha(i1,i2)=0d0
enddo
enddo

icount=icount+1
```

```
call der(icount,ifail)
do i1=1,onum
do i2=1,xnum
beta(i2)=beta(i2)+weight(i1)*(obs(i1)-calc(i1))* &
deriv(i2,i1)
do i3=1,i2
alpha(i2,i3)=alpha(i2,i3)+weight(i1)*deriv(i2,i1) &
*deriv(i3,i1)
enddo
enddo
enddo
do i1=1,xnum
do i2=1,i1
alpha(i2,i1)=alpha(i1,i2)
enddo
enddo

!    CALCULATE PARAMETER INCREMENTS AS DELX=BETA*(MARQ*DIAGONAL(
!    ALPHA)+ALPHA)**-1 AND ADD TO X TO GIVE NEXTX, THE NEW TRIAL
!    SET OF PARAMETERS. NOTE THAT A SCALED ALPHA IS INVERTED, TO
!    IMPROVE ACCURACY, AND THEN RESCALED

DO !Start of main fitting loop
do i1=1,xnum
do i2=1,xnum
alphin(i1,i2)=alpha(i1,i2)/dsqrt(alpha(i1,i1)*alpha(i2,i2))
enddo
alphin(i1,i1)=1d0+marq
enddo
call matinv(alphin,padim,xnum,err)
if(err == 0) then
do i1=1,xnum
nextx(i1)=x(i1)
do i2=1,xnum
nextx(i1)=nextx(i1)+beta(i2)*alphin(i1,i2)/ &
dsqrt(alpha(i1,i1)*alpha(i2,i2))
enddo
enddo
call chisq(chsq,nextx,icount,ifail)
else
return
endif

!    CALCULATE NEW TRIAL CHSQ AND CHECK IF IT INCREASED OR DE-
!    CREASED. IF IT DECREASED, NEXTX BECOMES X. SINCE CHSQ HAS
!    ALREADY EVALUATED CALC(NEXTX), THE NEXT ITERATION CAN BE
!    CONTINUED BY CALCULATING A NEW ALPHA AND BETA.

if(chold-chsq >= 0d0.and.ifail == 0) then
grad=0d0
do i1=1,xnum
x(i1)=nextx(i1)
grad=grad+beta(i1)*beta(i1)
enddo
grad=sqrt(grad)
if(dabs(chold-chsq) < delchi.or.dabs(ograd-grad) &
< delgrad) then
do i1=1,xnum
do i2=1,xnum
alphin(i1,i2)=alpha(i1,i2)/dsqrt(alpha(i1,i1)*alpha(i2,i2))
enddo
enddo
call matinv(alphin,padim,xnum,err)
if(err /= 0) then
err=8
endif
do i1=1,xnum
pasig(xpos(i1))=dsqrt(alphin(i1,i1)/alpha(i1,i1))
enddo
do i1=1,xnum
```

```
do i2=1,xnum
alphin(i1,i2)=alphin(i1,i2)/dsqrt(alpha(i1,i1)* &
alpha(i2,i2))/(pasig(xpos(i1))*pasig(xpos(i2)))
enddo
enddo
do i1=1,xnum
pasig(xpos(i1))=dsqrt(alphin(i1,i1)/alpha(i1,i1))
enddo
return
endif
marq=marq/10d0
ograd=grad
chold=chsq
if(debug == 1) then
write(6,fmt=' (" ****"/"CHI**2= ",e15.7/ &
&" PREV.GRAD= ",e15.7/"MARQ= ",e10.2) ') chsq,grad,marq
write(6,*) "ITERATION DECREASED CHI**2; TRYING SMALLER MARQ"
do i1=1,xnum
write(6,fmt=' (" #",i2," X= ",e15.8," PREV.GRAD= ",&
& e15.7) ') i1,x(i1),beta(i1)
enddo
endif

!    CALCULATE NEW ~(J)*F*dF/dX , NEW GRADIENT OF CHSQ AND THE
!    ~J*J MATRIX, WHERE F=(O-C)/OSIG AND ~ MEANS TRANSPOSE
do i1=1,xnum
beta(i1)=0d0
do i2=1,i1
alpha(i1,i2)=0d0
enddo
enddo

icount=icount+1
call der(icount,ifail)
do i1=1,onum
do i2=1,xnum
beta(i2)=beta(i2)+weight(i1)*(obs(i1)-calc(i1))* &
deriv(i2,i1)
do i3=1,i2
alpha(i2,i3)=alpha(i2,i3)+weight(i1)*deriv(i2,i1) &
*deriv(i3,i1)
enddo
enddo
enddo
do i1=1,xnum
do i2=1,i1
alpha(i2,i1)=alpha(i1,i2)
enddo
enddo

!    IF CHSQ INCREASED, THE MARQUARDT PARAMETER MUST BE INCREA-
!    SED TO FORCE DESCENT IN CHSQ. NEXTX IS DISCARDED AND A
!    SMALLER STEP AWAY FROM THE ORIGINAL X IS TRIED

else
if(maxiter < icount) then
err=5
return
endif
marq=max(marq*10d0,0.001d0)
if(marq.gt.1d3) then
err=6
return
endif
if(debug == 1) then
write(6,fmt='(" ****"/"CHI**2= ",e15.7/" GRAD= ", &
&e15.7/"MARQ= ",e10.2/" ITERATION INCREASED CHI**2; TRYING LARGER MARQ")') chsq,grad,marq
do i1=1,xnum
write(6,fmt='(" #",i2," NEXTX= ",e15.8," GRAD= ", &
&e15.7)') i1,nextx(i1),beta(i1)
```

```
enddo
endif
do i1=1,onum
calc(i1)=ocalc(i1)
enddo
endif
ENDDO !End of main fitting loop

end subroutine nllsq

!    CHISQ RETURNS THE REDUCED CHI**2 AFTER CALLING THE ROUTINE
!    OBSERVED WHICH SHOULD RETURN THE OBSERVED VALUES

subroutine chisq(chsq,x,icount,ifail)
use masterdim
use obscalc
implicit none
real(8) :: chsq,x(padim)
integer(4) :: i1,ifail,icount
do i1=1,xnum
pa(xpos(i1))=x(i1)
enddo
call cal(icount,ifail)
if(ifail /= 0) then
write(6,*) "Warning; subroutine cal() has ifail= ",ifail
endif
chsq=0d0
do i1=1,onum
chsq=chsq+(obs(i1)-calc(i1))*(obs(i1)-calc(i1))*weight(i1)
enddo
chsq=chsq/(onum-xnum)
end subroutine chisq

!    DER CALCULATES THE DERIVATIVES OF ALL OBS W/R TO ALL X
!    IT ALSO SAVES CALC FOR RECOVERY SHOULD SSQ NOT DECREASE

subroutine der(icount,ifail)
use masterdim
use obscalc
implicit none
real(8) :: save
integer(4) :: ifail,i1,i2,icount
do i1=1,onum
ocalc(i1)=calc(i1)
enddo
do i1=1,xnum
save=pa(xpos(i1))
pa(xpos(i1))=pa(xpos(i1))*1.01d0
if(pa(xpos(i1)).eq.0) then
pa(xpos(i1))=1d-2
endif
call cal(icount,ifail)
if(ifail /= 0) then
write(6,*) "Warning; subroutine cal() has ifail= ",ifail
endif
do i2=1,onum
if(save.ne.0) then
deriv(i1,i2)=(calc(i2)-ocalc(i2))*1d2/save
else
deriv(i1,i2)=(calc(i2)-ocalc(i2))*1d2
endif
enddo
pa(xpos(i1))=save
enddo
end subroutine der

end module nllsqfit

!    Main program: reads fort.1 input file, calls least squares fit if any
!    fitting flags are =1, or calls thkn() subroutine directly if no
```

```
!     fitting is done, outputs results to fort.3 and fort.7
program FRETMODULATE
use dimensio
use masterdim
use obscalc
use thermokin
use waves
use sigparameters
use popsigoutputs
use nllsqfit
implicit none
integer(4) :: i,j,k,n,m,ij,seed,iminus,iplus,printflag,icount,ifail
real(8) :: x(statedim),sum
real(8) :: tauobs,maxtime,mintime,time,tinit
real(8) :: tempnoise(-predim:dim),avg_noise,square_noise(0:dim),avg_noise2,rms_noise
real(8), parameter :: Pi=3.141592653589793d0
real(8) :: popsum,gauran,pop(statedim,0:dim),s(statedim),sig(statedim,0:dim)
real(8) :: xsum,relerr
real(8) :: xcpop(statedim,0:dim),xcsig(statedim,0:dim),norm

!     Protein input parameters for simulation:
!     Currently only a two-state folder is supported
read(1,*) T0, T0flag !Reference temperature for all protein states in Kelvin
read(1,*) gcoef(1,0),gcoef(1,1), gcflag(1,0),gcflag(1,1) !Free energy g=gcoef(0,1)+gcoef(1,1)*(temperature-T0) for state 1
read(1,*) gcoef(2,0),gcoef(2,1), gcflag(2,0),gcflag(2,1) !Same for state 2, also in kJ/mole
read(1,*) km,gdcoef(1,2,0),gdcoef(1,2,1), kmflag,gdcflag(1,2,0),gdcflag(1,2,1) !Transition state energy in kJ/mole
read(1,*) s1green,s1red, s1gflag, s1rflag !Red and green FRET signal values for state 1
read(1,*) s2green,s2red, s2gflag, s2rflag !Same for state 2
read(1,*) sred_slope,sgreen_slope,Tsignal, sr_slopeflag, sg_slopeflag, Tsflag !Red and green signal T-dependent quantum yields
!     The program will set sred=(s1red*x(1)+s2red*x(2))*(1+sred_slope*(temp-Tsignal)), similar for signal 2 (green)

!     Driving waveform for simulation
read(1,*) m_waves, seed, Taverage, dtprint, printflag !The # Driving waves, random number seed,
!     avg T of driving waveform, time intervals for printout to fort.2
write(7,*) "Average temperature ",Taverage
do m=1,m_waves
!     Read in Sine periods and amplitudes, noise 1/e time and amplitudes
!     Note that Nperiod is the inverse bandwidth of the noise, not its 'period.'  The period of the noise is the same as of
!     the sine waveform, i.e. the noise pattern repeats together with the sine waveform.
!     obs() and osig() are optional arrays of observed phase differences between red and green channels, and uncertainties, that
!     need to be read only if data is to be fitted. Set to 0 otherwise
read(1,*) Speriod(m),Samplitude(m),Nperiod(m),Namplitude(m),obs(m),osig(m)
enddo
!     Find overall dynamic range
tauobs=1/(km*dexp(-gdcoef(1,2,0)/(0.00831*T0)))
maxtime=0
mintime=tauobs
do m=1,m_waves
maxtime=max(maxtime,Speriod(m))
mintime=min(mintime,Speriod(m))
enddo
if(maxtime == 0) then
write(6,*) "A nonzero period or relaxation time must be specified."
stop
endif
tstep=maxtime/720 !Allow at least 1/2 degree of phase resolution for the slowest period
maxfac=3
time_points=720*maxfac !Evaluate data out to three times the slowest driving period
write(6,*) "Time dynamic range: ",mintime," to ",maxtime,". Unadjusted step: ",tstep
!     Decrease time step if fast dynamics requires it to avoid aliasing
do i=1,8
if(mintime/tstep < 256) then
tstep=tstep/2
time_points=time_points*2
else
exit
endif
enddo
if(i > 8 .or. time_points>dim) then
write(6,*) "Dynamic range of fastest rate to slowest driving period is too large."
```

```
stop
endif
write(6,*) "Adjusted step: ",tstep
odestep=tstep/10 !Make differential equation solver step 10x smaller than sampling step
!     Perform least squares fit if desired; only phase difference is currently supported as an observable.
onum=m_waves
panum=0
!     Determine number of fitted parameters "panum," and copy parameters to "pa" for use by subroutine nllsq
if(T0flag.eq.1) then
panum=panum+1
pa(panum)=T0
paf(panum)=1
endif
if(gcflag(1,0).eq.1) then
panum=panum+1
pa(panum)=gcoef(1,0)
paf(panum)=1
endif
if(gcflag(1,1).eq.1) then
panum=panum+1
pa(panum)=gcoef(1,1)
paf(panum)=1
endif
if(gcflag(2,0).eq.1) then
panum=panum+1
pa(panum)=gcoef(2,0)
paf(panum)=1
endif
if(gcflag(2,1).eq.1) then
panum=panum+1
pa(panum)=gcoef(2,1)
paf(panum)=1
endif
if(kmflag.eq.1) then
panum=panum+1
pa(panum)=km
paf(panum)=1
endif
if(gdcflag(1,2,0).eq.1) then
panum=panum+1
pa(panum)=gdcoef(1,2,0)
paf(panum)=1
endif
if(gdcflag(1,2,1).eq.1) then
panum=panum+1
pa(panum)=gdcoef(1,2,1)
paf(panum)=1
endif
if(s1gflag.eq.1) then
panum=panum+1
pa(panum)=s1green
paf(panum)=1
endif
if(s1rflag.eq.1) then
panum=panum+1
pa(panum)=s1red
paf(panum)=1
endif
if(s2gflag.eq.1) then
panum=panum+1
pa(panum)=s2green
paf(panum)=1
endif
if(s2rflag.eq.1) then
panum=panum+1
pa(panum)=s2red
paf(panum)=1
endif
if(sr_slopeflag.eq.1) then
panum=panum+1
```

```
pa(panum)=sred_slope
paf(panum)=1
endif
if(sg_slopeflag.eq.1) then
panum=panum+1
pa(panum)=sgreen_slope
paf(panum)=1
endif
if(Tsflag.eq.1) then
panum=panum+1
pa(panum)=Tsignal
paf(panum)=1
endif
!    Fitting info variables are currently hardwired
err=0
debug=1
maxiter=100000
marq=0.01
delchi=0.001
delgrad=0
if (panum /= 0) then
call nllsq(icount,ifail)
write(6,*) "Fit completed, ifail= ",ifail
write(7,*) "Chi-squared= ",chsq
write(7,*) "#, Parameter, Parameter uncertainty:"
do i=1,panum
write(7,*) i,pa(i),pasig(i)
enddo
write(7,*) "Correlation matrix:"
!    Output correlation matrix with scaled diagonal
do i=1,panum
write(7,fmt='(100(f9.5))') (alphin(i,j), j=1,panum)
enddo
write(7,*) " "
write(7,*) "New values of all input parameters and fitting flags:"
write(7,*) T0, T0flag !Reference temperature for all protein states in Kelvin
write(7,*) gcoef(1,0),gcoef(1,1), gcflag(1,0),gcflag(1,1) !Free energy g=gcoef(0,1)+gcoef(1,1)*(temperature-T0) for state 1
write(7,*) gcoef(2,0),gcoef(2,1), gcflag(2,0),gcflag(2,1) !Same for state 2, also in kJ/mole
write(7,*) km,gdcoef(1,2,0),gdcoef(1,2,1), kmflag,gdcflag(1,2,0),gdcflag(1,2,1) !Transition state energy in kJ/mole
write(7,*) s1green,s1red, s1gflag, s1rflag !Red and green FRET signal values for state 1
write(7,*) s2green,s2red, s2gflag, s2rflag !Same for state 2
write(7,*) sred_slope,sgreen_slope,Tsignal, sr_slopeflag, sg_slopeflag, Tsflag !Red and green signal T-dependent quantum yields
write(7,*) " "
endif
!    Loop through waveforms to output all output signals and write to output files.
!    fort.3 has the complete information, fort.7 the parameters and
write(3,*) "wv_n Per  nu  xpmt1 xpph1   xpmt2 xpph2   xsmt1 xsph1   xsmt2 xsph2  xsph2m1"
write(7,*) " Nu(Hz) Obs. phase(deg) Calc. phase(deg)"
!    x=cross-crorrelation; p=population; mt=time of maximum; 1,2=state or signal; ph=phase, s=signal
printflag=0
do m=1,m_waves
call thkin(m,printflag)
!    Save times and phases (relative to Speriod for each wave)
!    Note: The "360-" is a question of how the phase is defined, lagging or advanced.
write(3,'(i2,20(1x,f7.3))') m, Speriod(m), 1.0/Speriod(m), &
popmaxtime(1,m), 360-popphase(1,m), popmaxtime(2,m), 360-popphase(2,m), &
sigmaxtime(1,m), 360-sigphase(1,m), sigmaxtime(2,m), 360-sigphase(2,m), &
sigphase(1,m)-sigphase(2,m)
relerr=(obs(m)-sigphase(1,m)+sigphase(2,m))/osig(m)
write(7,'(20(1x,f7.3))')  1.0/Speriod(m), obs(m), osig(m), sigphase(1,m)-sigphase(2,m), relerr
enddo

end program FRETMODULATE

subroutine derivs(n,time,x,xp)
use thermokin
use waves
implicit none
real(8) :: x(statedim),time,xp(statedim),tinit,tfinal
integer(4) :: n,j
```

```
real(8) :: deltag,k(statedim,statedim),frac,temp
real(8), parameter :: round=1e-4
!     Solve the coupled DEQ for a two-state system

!     First, linearly interpolate temperature to DQE solver time
tinit=(itime-1)*tstep
tfinal=itime*tstep
frac=(time-tinit)/tstep
if(frac > 1+round .or. frac < -round ) then
write(6,*) "Interpolation of temperature in DERIVS falls"
write(6,*) "significantly outside the range of time points."
write(6,*) itime, tstep, tinit, tfinal, frac
stop
endif
temp=frac*temperature(itime)+(1-frac)*temperature(itime-1)
!     Compute 2-state equilibrium constant; for now, only n=2 is implemented
do j=1,n
g(j)=gcoef(j,0)+gcoef(j,1)*(temp-T0)
enddo
deltag=g(2)-g(1)
!      Compute barriers and forward/backward rate coefficients
gd(1,2)=gdcoef(1,2,0)+gdcoef(1,2,1)*(temp-T0) + deltag/2
gd(2,1)=gdcoef(1,2,0)+gdcoef(1,2,1)*(temp-T0) - deltag/2
k(1,2)=km*dexp(-gd(1,2)/(0.00831*temp))
k(2,1)=km*dexp(-gd(2,1)/(0.00831*temp))
!      Compute derivatives
xp(1)=-k(1,2)*x(1)+k(2,1)*x(2)
xp(2)=+k(1,2)*x(1)-k(2,1)*x(2)
end subroutine derivs
!
subroutine signal(n,x,s)
use dimensio
use waves
use sigparameters
implicit none
real(8) :: x(statedim),s(statedim),temp
integer(4) :: n
!     Compute green (1) and red (2) signals

!     Evaluate temperature at beginning of interval
temp=temperature(itime-1)
!     Compute signal at average temperature for the time step
s(1)=(s1green*x(1)+s2green*x(2))*(1+sgreen_slope*(temp-Tsignal))
s(2)=(s1red*x(1)+s2red*x(2))*(1+sred_slope*(temp-Tsignal))
end subroutine signal
!
!     This differential equations solver calls subroutine derivs, which
!     provides it with the kinetic equations
!     Note: simple Runge-Kutta forward propagation is used here: the fastest
!     time scale of the differential equation is given by the largest k, and
!     integration simply must me reasonably smaller steps than this fastest
!     time scale.  Adaptive methods simply fail when least-squares parameters
!     are adjusted to crazy values, whereas this provides a bad answer, which
!     is OK because it produces a large obs-calc error!
!
subroutine odesolve(n,y,xinit,xfinal,odestep)
use dimensio
implicit none
integer n
real(8) :: xinit,y(statedim),xfinal,x,dx,dydx(statedim),odestep
real(8), parameter :: round=1e-14
!
!     xinit    initial x
!     xfinal   final x
!     mindx     stepsize used, in units of x
!
!     Initialize x value, step size, and derivatives; RK4 is modified to
!     update y, the derivative dydx  and x to the final value x+dx, ready
!     for the next step
x=xinit
```

```
dx=odestep
call derivs(n,x,y,dydx)
!
!    WHILE loop to step from tinit to tfinal using
!    Runge-Kutta over the rest interval
!
do
if ( x >= xfinal*(1-round)) exit
if(x+dx.gt.xfinal) then
dx=xfinal-x
endif
call rk4(n,y,dydx,x,dx)
enddo
end subroutine odesolve
!
subroutine rk4(nv,y,dydx,x,h)
use dimensio
implicit none
integer(4) :: i,nv
real(8) :: y(statedim),dydx(statedim),yt(statedim)
real(8) :: h,dym(statedim),hh,h6,xhh,xh,dyt(statedim),x
hh=h*0.5d0
h6=h/6d0
xh=x+h
xhh=x+hh
do i=1,nv
yt(i)=y(i)+hh*dydx(i)
enddo
call derivs(nv,xhh,yt,dyt)
do i=1,nv
yt(i)=y(i)+hh*dyt(i)
enddo
call derivs(nv,xhh,yt,dym)
do i=1,nv
yt(i)=y(i)+h*dym(i)
dym(i)=dyt(i)+dym(i)
enddo
call derivs(nv,xh,yt,dyt)
do i=1,nv
!    Update y, its derivative, and x to value at final point
y(i)=y(i)+h6*(dydx(i)+dyt(i)+2d0*dym(i))
dydx(i)=dyt(i)
x=xh
enddo
end subroutine rk4
!
real(8) function gauran(hwhm,seed)
implicit none
real(8) :: y,ran,hwhm,width
integer(4) :: seed
integer(4), parameter :: ia=7141,ic=54773,im=259200
real(8), parameter :: numstd=3d0,f2=3.85802469d-6,f1=f2*numstd*2d0
y=0d0
ran=0d0
width=0.8325546d0/hwhm
do while (y.le.ran)
seed=mod(seed*ia+ic,im)
gauran=(dfloat(seed)*f1-numstd)
y=dexp(-(gauran*width)**2)
seed=mod(seed*ia+ic,im)
ran=dfloat(seed)*f2
enddo
return
end function gauran

!!!! NOTE: the Absoft Fx3 debugger needs a ^M line return character to recognize the
!!!! end of the code; if this is missing, the debugger will just not let you
!!!! open the file to debug.

subroutine thkin(m,printflag)
```

```
use dimensio
use thermokin
use waves
use sigparameters
use popsigoutputs
implicit none
integer(4) :: i,j,k,n,m,ij,seed,iminus,iplus,printflag
integer(4) :: presteps,drop_points
real(8) :: x(statedim),sum
real(8) :: maxtime,mintime,time,tinit,tprint
real(8) :: tempnoise(-predim:dim),avg_noise,square_noise(0:dim),avg_noise2,rms_noise
real(8), parameter :: Pi=3.141592653589793d0
real(8) :: popsum,gauran,alpha,pop(statedim,0:dim),s(statedim),sig(statedim,0:dim)
real(8) :: xsum
real(8) :: xcpop(statedim,0:dim),xcsig(statedim,0:dim),norm
real(8) :: val,valplus,valminus,t,tplus,tminus
integer(4) :: tem(0:dim)
if(printflag == 1) then
!       write(2,*) "t_sec T_Kelvin Sig1 Sig2 Pop1 Pop2 Popsum"
!       write(4,*) "Wave ",m
!       write(4,*) "t_xcor  sig1_xcor sig2_xcor pop1_xcor pop2_xcor"
endif
!     Create waveform
Somega(m)=2*Pi/Speriod(m)
do i=0,time_points-1
time=tstep*i
temperature(i)=Taverage+Samplitude(m)*dsin(Somega(m)*time)
enddo
if(Namplitude(m) /= 0) then
alpha=tstep/(tstep+Nperiod(m))
presteps=maxfac/alpha
if(presteps > predim) then
presteps=predim
write(6,*) "Warning: noise time constant too long; increase predim"
endif
do i=-presteps,time_points-1
tempnoise(i)=gauran(1d0,seed)
enddo
!       Filter noise to lowpass period Tnoise_taufilter; alpha=1 no filtering, alpha->0 strong low-pass filtering
do i=-presteps+1,time_points-1
tempnoise(i)=tempnoise(i-1)+alpha*(tempnoise(i)-tempnoise(i-1))
enddo
!       Shift noise so noise waveform starts at 0 amplitude
do j=0,presteps-1
if(dsign(1d0,tempnoise(j)) /= dsign(1d0,tempnoise(j+1))) then
tempnoise(j)=0
exit
endif
enddo
if (j == presteps) then
write(6,*) "Error in noise generation: no zero crossing was found."
write(6,*) "Try a different random number seed."
stop
endif
do i=0,time_points-1
tempnoise(i)=tempnoise(i+j)
enddo
!       Scale noise so it has the desired rms amplitude
avg_noise=sum(tempnoise(0:time_points-1))/time_points
square_noise(0:time_points-1)=tempnoise(0:time_points-1)*tempnoise(0:time_points-1)
avg_noise2=sum(square_noise(0:time_points-1))/time_points
rms_noise=dsqrt((avg_noise2-avg_noise**2))
tempnoise(:)=tempnoise(:)*Namplitude(m)/rms_noise
!       Add noise to temperature waveform
do j=1,maxfac
temperature((j-1)*time_points/maxfac:(j*time_points/maxfac)-1)= &
temperature((j-1)*time_points/maxfac:(j*time_points/maxfac)-1)+tempnoise(0:time_points/maxfac-1)
enddo
endif
```

```
!     Start simulation of modulated kinetics
n=2 !2-state system hardwired for now
tprint=0
!     Compute initial state population fraction x(j) in each state "j"
popsum=0
do j=1,n
g(j)=gcoef(j,0)+gcoef(j,1)*(Taverage-T0)
x(j)=dexp(-g(j)/(0.00831*Taverage))
popsum=popsum+x(j)
enddo
xsum=0
do j=1,n
x(j)=x(j)/popsum
pop(j,0)=x(j)
xsum=xsum+x(j)
enddo
!     Propagate in time
do itime=1,time_points
tinit=(itime-1)*tstep
time=itime*tstep
!     Compute signal s from population x, save populations and signals for later analysis
call signal(n,x,s)
xsum=0
!     Save populations and signals.  Note that j counts states (Native =1, denatured=2) for populations,
!     Green (1) and Red (2) fluorescence for signals
do j=1,n
xsum=xsum+x(j)
pop(j,itime)=x(j)
sig(j,itime)=s(j)
enddo
if(tinit >= tprint .and. printflag ==1) then
write(2,'(f7.4,1x,f7.2,1x,2(f7.4,1x),3(f6.3,1x))') tinit, &
temperature(itime-1),s(1),s(2),x(1),x(2),xsum
tprint=tinit+dtprint
endif
call odesolve(n,x,tinit,time,odestep)
xsum=xsum
enddo
if(printflag == 1) then
write(2,*) " "
endif

!     Calculate cross-correlation functions of populations and signals
!     with respect to driving temperature waveform
!     Use only one period from end of data (three long periods were calculated if maxfac=3).
use_points=Speriod(m)/tstep
drop_points=time_points-use_points
norm=1d0
do i=0,use_points-1
do k=1,n
pop(k,i)=pop(k,i+drop_points)
sig(k,i)=sig(k,i+drop_points)
enddo
tem(i)=temperature(i+drop_points)
enddo
xcpop(:,:)=0
xcsig(:,:)=0
do k=1,n
do i=0,use_points-1
do j=0,use_points-1
!      Wrap the index j around if it exceeds use_points, to calculate circular autocorrelation
if(i+j >= use_points) then
ij=i+j-use_points
else
ij=i+j
endif
if(ij >= use_points) then
write(6,*) "If the code executed this statement, woe unto you!"
stop
endif
```

```
!     A simple box formula is used; this causes a phase error if Speriod(m) is not an
!     exact multiple of tstep
xcpop(k,i)=xcpop(k,i)+pop(k,j)*(tem(ij)-Taverage)
xcsig(k,i)=xcsig(k,i)+sig(k,j)*(tem(ij)-Taverage)
enddo
xcpop(k,i)=xcpop(k,i)*norm
xcsig(k,i)=xcsig(k,i)*norm
enddo
enddo
tprint=0
do i=0,use_points-1
time=i*tstep
if(time >= tprint .and. printflag==1) then
write(4,'(f8.4,4(1x,f9.2))') time,xcsig(1,i),xcsig(2,i),xcpop(1,i),xcpop(2,i)
tprint=time+dtprint
endif
enddo
!     Find first maxima in correlation functions (could be at i=0 or i=use_points-1)
!     Note: crude 3-point search with parabolic interpolation of three points near max.;
!     requires small enough time step for adequate sampling.
do k=1,n
do i=0,use_points-1
iminus=i-1
iplus=i+1
if(i == 0) iminus=use_points-1
if(i == use_points-1) iplus=0
valminus=xcpop(k,iminus)
val=xcpop(k,i)
valplus=xcpop(k,iplus)
if(val > valminus .and. val >= valplus) exit
enddo
if(i == use_points) then
write(6,*) "No maximum found in autocorrelation."
stop
endif
!     "i" is position of sampled max.  Use parabolic interpolation to get better position
!       popmaxtime(k,m)=i*tstep
tminus=(i-1)*tstep
t=i*tstep
tplus=(i+1)*tstep
popmaxtime(k,m)=0.5*( (val-valplus)*(tminus+t) - (valminus-val)*(t+tplus) )/(2*val-valminus-valplus)
if(popmaxtime(k,m) < 0 ) popmaxtime(k,m)=popmaxtime(k,m)+(use_points-1)*tstep
if(popmaxtime(k,m) >= (use_points-1)*tstep ) popmaxtime(k,m)=popmaxtime(k,m)-(use_points-1)*tstep
popphase(k,m)=popmaxtime(k,m)/Speriod(m)*360
valminus=xcsig(k,use_points-1)
valplus=xcsig(k,1)
do i=0,use_points-1
if(xcsig(k,i) > valminus .and. xcsig(k,i) > valplus) exit
valminus=xcsig(k,i)
if(i<use_points-2) then
valplus=xcsig(k,i+2)
else
valplus=xcsig(k,0)
endif
enddo
if(i == use_points) then
write(6,*) "No maximum found in autocorrelation."
stop
endif
sigmaxtime(k,m)=i*tstep
sigphase(k,m)=sigmaxtime(k,m)/Speriod(m)*360
enddo
return
end subroutine thkin

!     Subroutine cal() uses thkin() to calculate phases for each of the m observed frequencies/periods
!     for least-squares fitting; it renames "pa" back to the model parameters for use by subroutine thkin()
subroutine cal(icount,ifail)
use masterdim
use obscalc
```

```
use thermokin
use sigparameters
use popsigoutputs
implicit none
integer(4) :: i,icount,ifail,m,printflag
ifail=0
icount=icount+1
printflag=0
!     Rename fitting parameters sent by nllsq()
i=0
if(T0flag.eq.1) then
i=i+1
T0=pa(i)
endif
if(gcflag(1,0).eq.1) then
i=i+1
gcoef(1,0)=pa(i)
endif
if(gcflag(1,1).eq.1) then
i=i+1
gcoef(1,1)=pa(i)
endif
if(gcflag(2,0).eq.1) then
i=i+1
gcoef(2,0)=pa(i)
endif
if(gcflag(2,1).eq.1) then
i=i+1
gcoef(2,1)=pa(i)
endif
if(kmflag.eq.1) then
i=i+1
km=pa(i)
endif
if(gdcflag(1,2,0).eq.1) then
i=i+1
gdcoef(1,2,0)=pa(i)
endif
if(gdcflag(1,2,1).eq.1) then
i=i+1
gdcoef(1,2,1)=pa(i)
endif
if(s1gflag.eq.1) then
i=i+1
s1green=pa(i)
endif
if(s1rflag.eq.1) then
i=i+1
s1red=pa(i)
endif
if(s2gflag.eq.1) then
i=i+1
s2green=pa(i)
endif
if(s2rflag.eq.1) then
i=i+1
s2red=pa(i)
endif
if(sr_slopeflag.eq.1) then
i=i+1
sred_slope=pa(i)
endif
if(sg_slopeflag.eq.1) then
i=i+1
sgreen_slope=pa(i)
endif
if(Tsflag.eq.1) then
i=i+1
Tsignal=pa(i)
endif
```

APPENDIX E

Supplementary information of tethered WW domains from monomer to tetramer: folding competing with aggregation

E.1: Primary sequence for all of the protein constructs

Mfip35 (Monomer)

K L P P G W E K R M S R D G R V Y Y F N H I T N A S Q F E R P S G

Dfip35 (Dimer)

K L P P G W E K R M S R D G R V Y Y F N H I T N A S Q F E R P S G **G G S G G S G G S G** K L P P G W E K R M S R D G R V Y Y F N H I T N A S Q F E R P S G

Tfip35 (Trimer)

K L P P G W E K R M S R D G R V Y Y F N H I T N A S Q F E R P S G **G G S G G S G G S G** K L P P G W E K R M S R D G R V Y Y F N H I T N A S Q F E R P S G **G G S G G S G G S G** K L P P G W E K R M S R D G R V Y Y F N H I T N A S Q F E R P S G

Qfip35 (tetramer)

K L P P G W E K R M S R D G R V Y Y F N H I T N A S Q F E R P S G **G G S G G S G G S** G K L P P G W E K R M S R D G R V Y Y F N H I T N A S Q F E R P S G **G G S G G S G G S G** K L P P G W E K R M S R D G R V Y Y F N H I T N A S Q F E R P S G **G G S G G S G G S G** K L P P G W E K R M S R D G R V Y Y F N H I T N A S Q F E R P S G

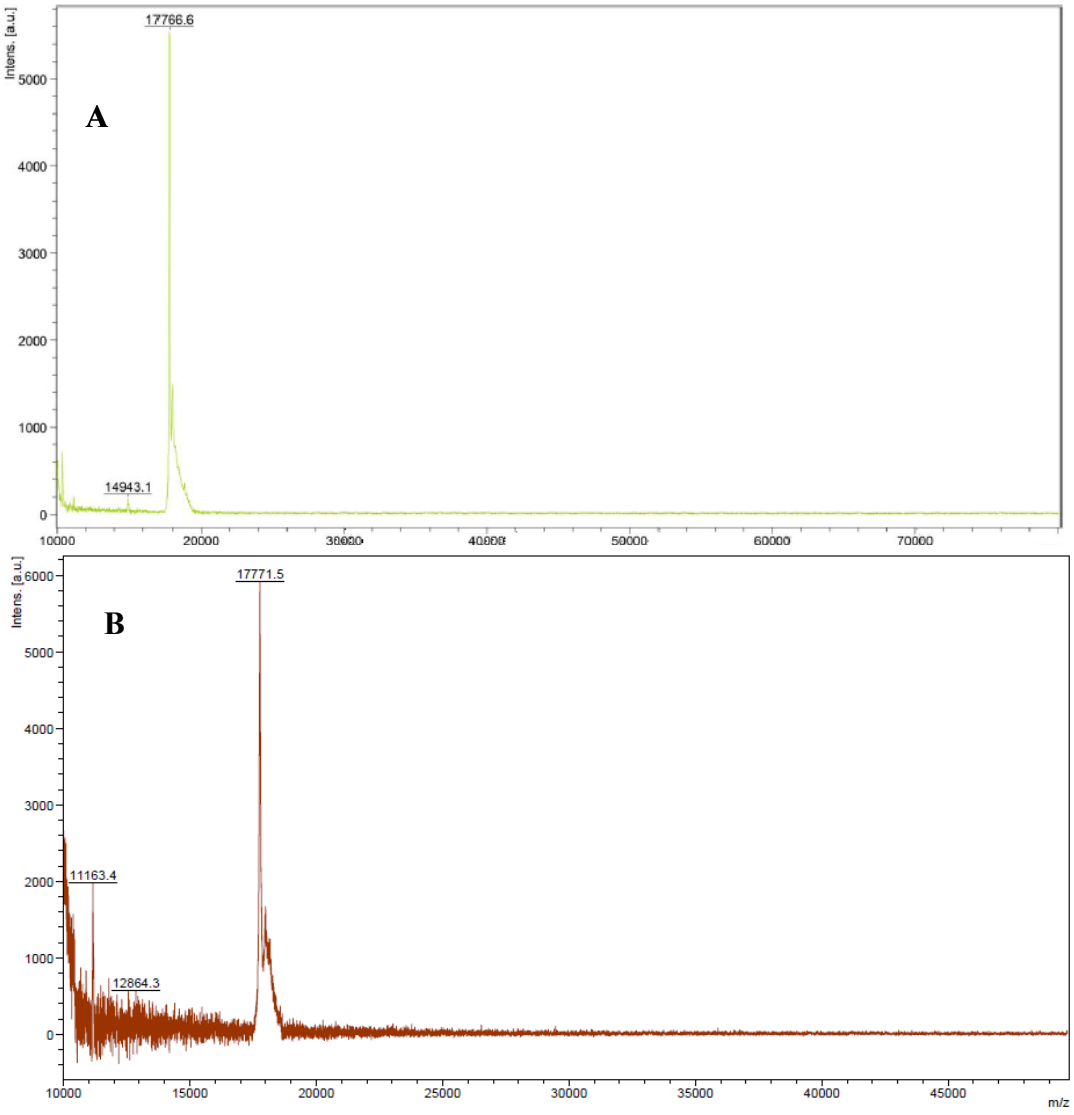

Fig. E.1: Mass spectrometry results of Qfip35 protein purified using **A)** GST and **B)** His tag both showing a peak at m/z = 17.766 K Da and 17.771 K Da respectively.

E.2: Multimeric network model code

E.2.1 Parameters explanation in the model

The parameters mentioned in Table 6.2 in the main text are described below in the model.

N (folded form) = 1; M (Intermediate form) = 2; U (unfolded form) = 3

dG = howmanyN*(g31*(T-Tm)+ gg31*GHCL + xn*gnn)+ howmanyM*(g32*(T-Tm)+ go32+ gg32*GHCL +xm*gmm) + howmanyU*0 ---------- (1)

The free energy for any state NN OR NM or UN is written by the general equation (1) where

dG = Free energy

Tm = melting temperature

g31= co-operatively factor (going from Unfolded to folded)

gg31= guanidine dependence

xn = number of NN pair present in any species for eg: NN will have only one pair

gnn= interaction term for NN pairs

g32= co-operatively factor (going from Unfolded to intermediate)

gg32= guanidine dependence

xm = number of MM pair present in any species for eg: NMM will have only one pair

gmm= interaction term for MM pairs

Su = bu+mu*(T-Tm) ---------- (2)

Su = unfolded baseline

bu = unfolded intercept

mu = unfolded slope

Similarly folded intercept and baseline and slope are represented as Sf , bf and mf.

Gk13k = kinetic barrier going from folded to unfolded

Gk23k = kinetic barrier going from intermediate to unfolded

In the model the starting point for the experimental data and simulated data was matched in order to form a correct representation for chi. The assumption made here is that there exist no ultra-fast phase.

E.2.2 Thermodynamic representation

```
Function[S,dG,Keq,Si]=ThermoFit(Mer,TotalStates,GHCL,TRange,Tm,g31,g32,go32,gg31,
gg32,gnn,gmm,bu,mu,bf,mf,bm,mm)
%
NumStates = size(TotalStates,1);      % number of all possible species
%%
for p = 1:numel(TRange)
   T= TRange(p);
   % Signals for indivdual N, M and U
   Su = bu+mu*(T-Tm)+ 5*(Mer==1)- 4*(Mer==4);   %unfolded baseline%%%%%%%% addtional 5nm for monomer
   Sf = bf+mf*(T-Tm)+0.5*(Mer==1)- 4*(Mer==4);   %folded baseline %%%%%%%%%% additional 1nm for monomer
   Sm = bm+mm*(T-Tm);  %misfolded baseline
   % Thermodynamic delta G for transitions
   % Here each of species is a separate state with associated G
   % U/UU/UU/UUUU is the ground/ref state with G ==0
   deltaG  = zeros(1,NumStates); % Initialize
   for i =1:NumStates % This loop will calculate signal & k_eq for each species coming from ground species
      howmanyN = numel(find(TotalStates(i,:)==1)); % how many N are there in order to make signal
      howmanyM = numel(find(TotalStates(i,:)==2)); % how many M are there in order to make signal
      howmanyU = numel(find(TotalStates(i,:)==3)); % how many U are there in order to make signal
      Si(p,i) = (howmanyN*Sf+howmanyM*Sm+howmanyU*Su)/Mer; % generate signal for all the species
      %
      speciesstr = sprintf('%u',TotalStates(i,:)); % change species into a string
      xn = numel(findstr(speciesstr, '11')); % find the pair MM in the species
      xm = numel(findstr(speciesstr, '22')); % find the pair MM in the species
      xu = numel(findstr(speciesstr, '33')); % find the pair MM in the species
```

```
    dG(p,i) = howmanyN*(g31*(T-Tm)+ gg31*GHCL + xn*gnn)+ howmanyM*(g32*(T-
Tm)+ go32+ gg32*GHCL +xm*gmm) + howmanyU*0;
    Keq(p,i) =exp(-dG(p,i)/8.31/(T+273.15)); % equilbrium rate for all the species i
  end
  S(p,1) = Si(p,:)*Keq(p,:)'/sum(Keq(p,:),2); % generating signal for thermodynamics
end
end
```

E.2.3 Kinetic representation

```
function[Chi,Time,Conc,ConcEq,TransMatrix]=KinFit(Mer,TotalStates,GHCL,T,T_fin,...
Tm,g31,g32,go32,gg31,gg32,gnn,gmm,Gk13k,Gk23k,tspan,ExpData)
NumStates = size(TotalStates,1);      % number of all possible species
% Calculate the barriers and kinetic parameters
W = 20; % prefactor [1/us]
%% Solve ODE at T = temp
TransMatrix = zeros([NumStates, NumStates]);
% TransMatrix(i,j) is rate of reaction of species i going to species j
for i = 1:NumStates
  howmany(i,1) = numel(find(TotalStates(i,:)==1)); % how many N are there in order to
make signal
  howmany(i,2) = numel(find(TotalStates(i,:)==2)); % how many M are there in order to
make signal
  howmany(i,3) = numel(find(TotalStates(i,:)==3)); % how many U are there in order to
make signal
  for j = 1:NumStates
  speciesstr1 = sprintf('%u', TotalStates(i,:)); % change reactant species into a string
  speciesstr2 = sprintf('%u',TotalStates(j,:)); % change product species into a string
  xn = numel(strfind(speciesstr2, '11'))- numel(strfind(speciesstr1, '11')); % find effective
change in pairs
  xm = numel(strfind(speciesstr2, '22'))- numel(strfind(speciesstr1, '22'));
  xu = numel(strfind(speciesstr2, '33'))- numel(strfind(speciesstr1, '33'));
  %% Thermodynamic delta G for transitions
  % G31 is defined outside the for loop as it is NOT dependent on x
  G31     = g31*(T-Tm)+ gg31*GHCL + xn*gnn;
```

```
G32     = g32*(T-Tm)+ go32+ gg32*GHCL + xm*gmm; % uses x
%%% Remaining kinetics from here
%Important parameter to play with
Gk13=(Gk13k-0.5*G31);
Gk31=(Gk13k+0.5*G31);
Gk23=(Gk23k-0.5*G32);
Gk32=(Gk23k+0.5*G32);
%
kmatrix=zeros([3,3]); % kmatrix initiation
kmatrix(1,3)=W*exp(-Gk13/(8.31*(T+273.15))); % units would be microsec inverse
kmatrix(3,1)=W*exp(-Gk31/(8.31*(T+273.15)));
kmatrix(2,3)=W*exp(-Gk23/(8.31*(T+273.15)));
kmatrix(3,2)=W*exp(-Gk32/(8.31*(T+273.15)));
% when monoMer
if(Mer==1)
  if(i==j)
   TransMatrix(i, j) = 0;
  else
    TransMatrix(i, j) = kmatrix(TotalStates(i), TotalStates(j));% filling up transmatrix from the kmatrix which is created in kinetic_nMer script
  end
end
% when polyMer more than monoMer system
if(Mer>1)
  transformInd=[];flipMer=[];beforeSwitch=[];afterSwitch=[];
  subtract1 = TotalStates(i,:)- TotalStates(j,:);        % substraction of rows in order to determine if only one of the N,M,U is switching
  subtract2 = fliplr(TotalStates(i,:))- TotalStates(j,:); % flipping the sequence 123-322 makes it seems like 2 places are changed but if we flip 123 to 321-322 only one place is changed and hence it should be allowed
  if(nnz(subtract1)==1)                    % if only subtraction lead to one non-zero entry then do the below loop
      transformInd = find(subtract1~=0);          % what is the position/index where the switch is happening
```

```
        beforeSwitch = TotalStates(i,transformInd);       % what was it (N=1,M=2,U=3) that
switched
        afterSwitch = TotalStates(j,transformInd);        % what was it (N=1,M=2,U=3) that it
switched to
        TransMatrix(i, j) =  kmatrix(beforeSwitch, afterSwitch);   % picking the rates from
kinetic Mer kmatrix and filling in trans matrix
    elseif(nnz(subtract2)==1)                             % for the flipping case doing the same
thing
        transformInd = find(subtract2~=0);
        flipMer = fliplr(TotalStates(i,:));
        beforeSwitch = flipMer(1,transformInd);
        afterSwitch = TotalStates(j,transformInd);
        TransMatrix(i, j) =  kmatrix(beforeSwitch, afterSwitch);
    else
      TransMatrix(i, j) = 0;
    end
   end
  end
 end
 ratematrix = TransMatrix'; % Transpose of the transmatrix should give us ratematrix for
make differential equation
   for i = 1:NumStates
    for j = 1:NumStates
     if (i==j)
       ratematrix(i,j) = - sum(TransMatrix(i, :)); % making the ratematrix from Transition
matrix
      end
    end
   end
   conc0   = zeros([NumStates, 1]); % initial conc initialization for all the states to be zero
   conc0(1)= 40e-6;  % initial concentration of nn nnn nnnn
   options = odeset('RelTol',1e-8,'AbsTol',1e-14,'Stats','off',...
  'NormControl','on','NonNegative',numel(conc0),'Refine',1,...
  'MStateDependence','weak','MassSingular','maybe','BDF','off');
```

```
    [TEq,ConcEq] =
ode15s(@(t,conc)myODE(t,conc,ratematrix),linspace(0,1e4,1e2),conc0,options);
    %% Calulation of kinetic rates
    TransMatrix = zeros([NumStates, NumStates]); ratematrix = [];
    % TransMatrix(i,j) is rate of reaction of species i going to species j
    for i = 1:NumStates
      for j = 1:NumStates
        speciesstr1 = sprintf('%u', TotalStates(i,:)); % change reactant species into a string
        speciesstr2 = sprintf('%u',TotalStates(j,:)); % change product species into a string
        xn = numel(strfind(speciesstr2, '11'))- numel(strfind(speciesstr1, '11')); % find effective
change in pairs
        xm = numel(strfind(speciesstr2, '22'))- numel(strfind(speciesstr1, '22'));
        xu = numel(strfind(speciesstr2, '33'))- numel(strfind(speciesstr1, '33'));
        %% Thermodynamic delta G for transitions
        % G31 is defined outside the for loop as it is NOT dependent on x
        G31     = g31*(T-Tm)+ gg31*GHCL + xn*gnn;
        G32     = g32*(T-Tm)+ go32+ gg32*GHCL + xm*gmm; % uses x
        %% Remaining kinetics from here
        %Important parameter to play with
        Gk13=(Gk13k-0.5*G31);
        Gk31=(Gk13k+0.5*G31);
        Gk23=(Gk23k-0.5*G32);
        Gk32=(Gk23k+0.5*G32);
        %
        kmatrix=zeros([3,3]); % kmatrix initiation
        kmatrix(1,3)=W*exp(-Gk13/(8.31*(T_fin+273.15))); % units would be microsec inverse
        kmatrix(3,1)=W*exp(-Gk31/(8.31*(T_fin+273.15)));
        kmatrix(2,3)=W*exp(-Gk23/(8.31*(T_fin+273.15)));
        kmatrix(3,2)=W*exp(-Gk32/(8.31*(T_fin+273.15)));
        % when monoMer
        if(Mer==1)
          if(i==j)
           TransMatrix(i, j) = 0;
          else
```

```
        TransMatrix(i, j) = kmatrix(TotalStates(i), TotalStates(j));% filling up transmatrix
from the kmatrix which is created in kinetic_nMer script
      end
    end
    % when polyMer more than monoMer system
    if(Mer>1)
      transformInd=[];flipMer=[];beforeSwitch=[];afterSwitch=[];
      subtract1 = TotalStates(i,:)- TotalStates(j,:);       % substraction of rows in order to
determine if only one of the N,M,U is switching
      subtract2 = fliplr(TotalStates(i,:))- TotalStates(j,:); % flipping the sequence 123-322
makes it seems like 2 places are changed but if we flip 123 to 321-322 only one place is
changed and hence it should be allowed
      if(nnz(subtract1)==1)                    % if only subtraction lead to one non-zero
entry then do the below loop
        transformInd = find(subtract1~=0);          % what is the position/index where
the switch is happening
        beforeSwitch = TotalStates(i,transformInd);      % what was it (N=1,M=2,U=3)
that switched
        afterSwitch = TotalStates(j,transformInd);       % what was it (N=1,M=2,U=3) that
it switched to
        TransMatrix(i, j) =  kmatrix(beforeSwitch, afterSwitch);   % picking the rates from
kinetic Mer kmatrix and filling in trans matrix
      elseif(nnz(subtract2)==1)                    % for the flipping case doing the
same thing
        transformInd = find(subtract2~=0);
        flipMer = fliplr(TotalStates(i,:));
        beforeSwitch = flipMer(1,transformInd);
        afterSwitch = TotalStates(j,transformInd);
        TransMatrix(i, j) =  kmatrix(beforeSwitch, afterSwitch);
      else
        TransMatrix(i, j) = 0;
      end
    end
```

```
    end
   end
   ratematrix = TransMatrix'; % Transpose of the transmatrix should give us ratematrix for
make differential equation
   for i = 1:NumStates
    for j = 1:NumStates
     if (i==j)
      ratematrix(i,j) = - sum(TransMatrix(i, :)); % making the ratematrix from Transition
matrix
     end
    end
   end
   options = odeset('RelTol',1e-8,'AbsTol',1e-14,'Stats','off',...
  'NormControl','on','NonNegative',numel(conc0),'Refine',1,...
  'MStateDependence','weak','MassSingular','maybe','BDF','off');
   [Time, Conc] = ode15s(@(t,conc)myODE(t,conc, ratematrix), tspan, ConcEq(end,:)',
options);
   %% Formulating the X(but not sure at this point)
   concN=zeros(size(Conc(:,1)));
   concM=zeros(size(Conc(:,1)));
   concU=zeros(size(Conc(:,1)));
   for i = 1:NumStates
   concN = concN + Conc(:,i)*howmany(i,1)/Mer;
   concM = concM + Conc(:,i)*howmany(i,2)/Mer;
   concU = concU + Conc(:,i)*howmany(i,3)/Mer;
   end
   sumconc = concN+concM+concU;
   concN=concN./sumconc;
   concM=concM./sumconc;
   concU=concU./sumconc;
```

APPENDIX F

Supplementary information of folding under high pressure inside the bacterial cytoplasm

F.1 Procedure for ReAsh labeling

*A 1-5mg/mL stock of lysozyme should be prepared prior to experiment. This stock can be used 1-2 weeks after preparation. A plate should also be streaked prior to Day 1. Ideally the day directly before for best results but a 1-week old plate is acceptable.

**Any pipetting involving cell should be done with the wide orifice pipettes. Pipetting off supernatant can be done with normal pipettes.

Day 1

1) In the morning, start a 2 mL culture of cells from 1 colony (falcon tube). Add appropriate be started (each from a separate colony). The rest of the procedure is then done in parallel. Note it takes about 10 hours from starter culture to induction.
2) Allow the culture to grow until it is cloudy (~4-6 hours).
3) Make a 1:100 dilution of cells (note once the dilution is done, it will take around 5-6 hours until the induction step) into 2mL of LB (20µL cells) and add antibiotic (2µL for 1000x ampicillin) – done in a falcon tube.
4) Put in shaker at 37°C and monitor until OD600 reaches 0.5 – 0.7 (higher end of range may produce better results, takes ~2-3 hours to reach 0.5-0.6). Use plastic disposable cuvettes (1.5mL size). Baseline the UV-vis using LB from 650nm to 550nm. Pipette 1mL of cells using wide orifice pipettes. Once the measurement is taken, pipette the cells back into the falcon tube.
5) Add lysozyme to the cells in the falcon tube for a final lysozyme concentration of 50µg/mL (ε=36000 and MW=14,307 g/mol) (Gently shake). Place on ice for 10 minutes. At this point, the water bath next to the shaker in A229 should be set to 10°C so it can be ready by induction (see step 14).

6) After 10 minutes, transfer the cells in the falcon tube to a round bottom eppendorf tube and spin down the cells for 10 minutes at 10,000g. A convenient way to perform this is to fill two tubes with 500µL of cells each for balancing. Keep the excess of cells as a backup. If using two cultures, use 500µL from one and 500µL from the other.
7) Gently pipette off the supernatant and resuspend the cell pellet using LB at the original volume (500µL of LB for each tube if doing the "convenient way"). Spin down the cells again with the same settings. Note that there should be little to no delay between the end of spindown and pipetting as the pellet dissolves quickly.
8) Pipette off the supernatant again (pipette supernatant for all spin steps to reduce pellet loss) and resuspend using LB at the original volume (500µL per tube). Add antibiotic to each tube (0.5µL of 1000x amp for each tube).
9) During the second spin, poke holes (~3) (with small needle) in the top of two new round bottom eppendorf tubes. Take 100µL of the resuspended cells and place them in one of these tubes.

   Then do step 12
10) Add 1µL of ReAsh stock (final concentration=20µM) to these 100µL of cells. The ReAsh will need to be thawed for a minute or two before it can be pipetted. Completely wrap the ReAsh tube with aluminum foil and let it thaw at room temperature for 2 min.

    Note:

    + Good ReAsh has red color, never use the blue one (bad ReAsh).

    + Never touch the bottom of the ReAsh tube, it is best to turn off the light when handling ReAsh.

    + Thaw the cell carefully on ice before adding ReAsh to the tube.

    + The ReAsh tubes should be collected into the desired bag in room A223.
11) Pipette up and down a few times gently to mix (recommended to use ~50µL volume on 100 or 200µL pipette for mixing).
12) From the tube that the 100µL cells were taken out, take ~300µL of the remaining cells and place them into the other round bottom eppendorf tube with holes poked in the top (these cells could have also come from the other 500µL tube).
13) Shake the 100µL labeled and 300µL unlabeled cells at 37°C. Cover the top of the tubes loosely with aluminum foil to prevent contaminants from falling through holes. Monitor the OD600 with the unlabeled cells to prevent loss of ReAsh labeled cells. Make sure the foil is loose enough to allow air into the tube but tight enough to not fall off while the tube shakes.

14) When the OD600 reaches 1 (around 1.5-2 hours after lysozyme addition/spin down), induce both tubes with IPTG at 500µM (stock IPTG is usually 1M, for a 100mM stock use 0.5µL/100µL cells). If the OD600 still has not reached 1 after 3 hours, induce the cells anyway. A more dilute sample of IPTG may be desired to avoid needing to pipette extremely small volumes. Use MQ water to dilute the IPTG stock.
15) Leave the cultures to induce overnight at 25°C. Note to use lower temperatures with the shaker in A229, the bath should be set 15°C below the desired temperatures. In this case, it should be set to 10°C.

Day 2

1) 12-13 after induction (12 hours is best) (aim for lower end of range especially for imaging), spin down the cells at 10,000g for 10 minutes. If desired, 100µL of cells from the 300µL tubes can be placed in a separate tube, and that tube can be spun down. This way you can be sure it's 100µL since there's less than 300µL in the tube from volume loss (from things such as checking OD). All 100µL of the ReAsh cells can be also placed in a new round bottom eppendorf tube if one wants to avoid spin down in tubes with holes.

   Note: Spin down labeled and unlabeled cell with new round bottom eppendorf tubes. Do not use the poked ones.
2) Pour off the supernatant and resuspend at a 1:4 dilution in ice cold PBS (400µL PBS/100µL cells). Spin down again at 10,000g for 10 minutes. Pour off the supernatant and resuspend at a 1:4 dilution in cold PBS. Prepare for imaging or other experiments. Store the 1:4 stocks on ice until they are needed.
   a. For imaging, use a final dilution of ~1:20 (more dilutions may be necessary depending how crowded the cells look under the microscope).
   b. For performing a melt, use a final dilution of ~1:8 (more dilutions may be necessary if there appears to be significant scattering).

F.2 Whole Genome sequencing and analysis of pressurized bacteria

Whole genome sequencing was performed using the Illumina platform. For all sequencing, cultures were grown by inoculating fresh medium from frozen stocks made after pressure treatment and growing to saturation at 37 °C. Sequencing was performed on a locally operated Illumina MiSeq system present at Center for the physics of living center at UIUC. For MiSeq runs DNA was extracted with MoBio Ultraclean Microbial DNA isolation kit.DNA was quantified by qubit and Bioanalyzer and libraries were prepared using the NexteraXT kit from Illumina. MiSeq runs were demultiplexed and trimmed using the onboard Illumina software. Analysis was performed using the online breseqplatform in polymorphism mode (http://barricklab.org/twiki/bin/view/Lab/ToolsBacterialGenomeResequencing). Breseq uses an empirical error model and a Bayesian variant caller to predict polymorphisms at the nucleotide level. The algorithm uses a threshold on the empirical error estimate (E-value) to call variants (Barrick and Lenski, 2009). All other parameters were set to their default values. Reads were aligned to the MG1655 genome (INSDC U00096.3) to predict possible mutations.

F.3 Laboratory pressure cycling of Mg1655 strain

Mg1655 strain was subjected to maximum pressure of 1900 bar with an increment of 100 bar with a wait time of 3 mins at a particular pressure. These pressure treated cells were regrown in LB media and stored in frozen vials. The cells were grown overnight at 37 °C to be used for the next pressure cycling experiment. After nine repeated cycles of pressurization the obtained strain's genome was sequenced along with MG1655.It was discovered that ~ 1.4 kb IS4 element was inserted at location 3,991,653 which led to a dysfunctional cyaA gene. cyaA gene product is an enzyme adenylate cyclase which catalyzes the formation of second messenger cAMP (cyclic AMP) from ATP. In order to confirm the mutation in cyaA gene in cycle 9 MG1655 (WT) and cycle 9 were grown in the presence and absence of cAMP in minimal media M63 + 5% glycerol (see Fig. F.2). It was seen that in the absence of cAMP cycle 9 had an interesting oscillating growth curve whereas WT had a normal growth curve with doubling time of ~ 57 mins typical for MG1655 strain. In the presence of 10mM cAMP the growth curve of cycle 9 recovered but the doubling time was around ~114 mins (see Fig.F.2).

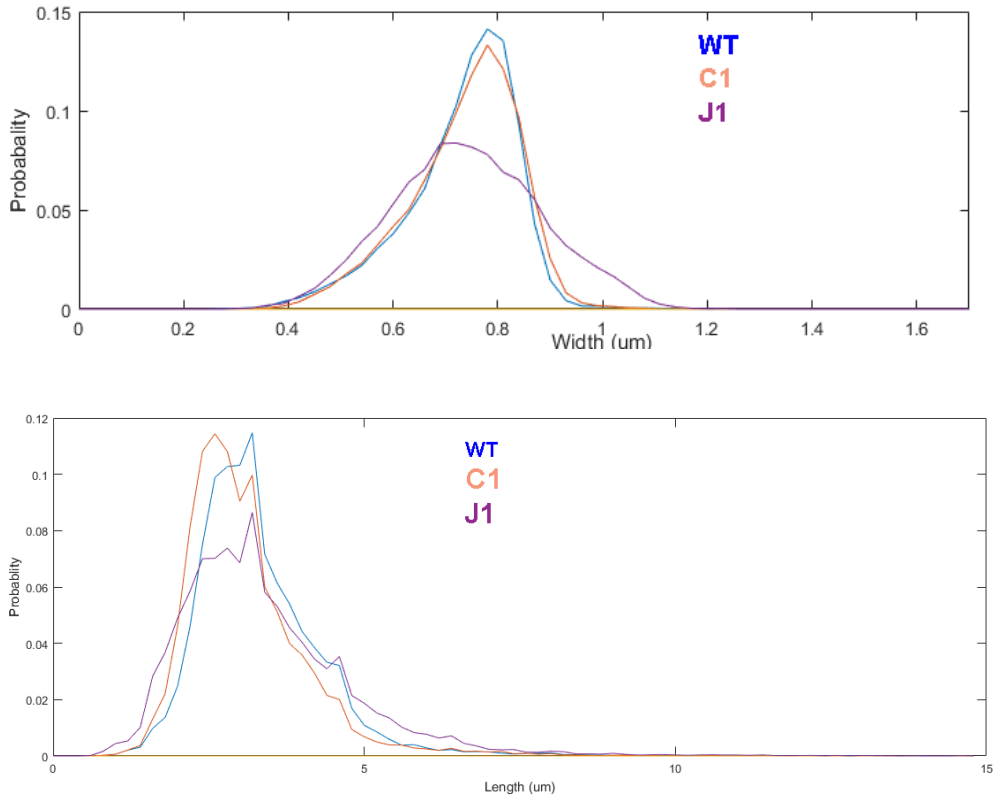

Fig. F.1: The width and length distributions for MG1655 (WT), cycle 1 (C1) and J1 pressure resistant strain. The plots shows no significant difference in the size of these strains. Length-width analysis on the imaging frames using oufti software resulted in length= 2.5 µm and width =0.8 µm.

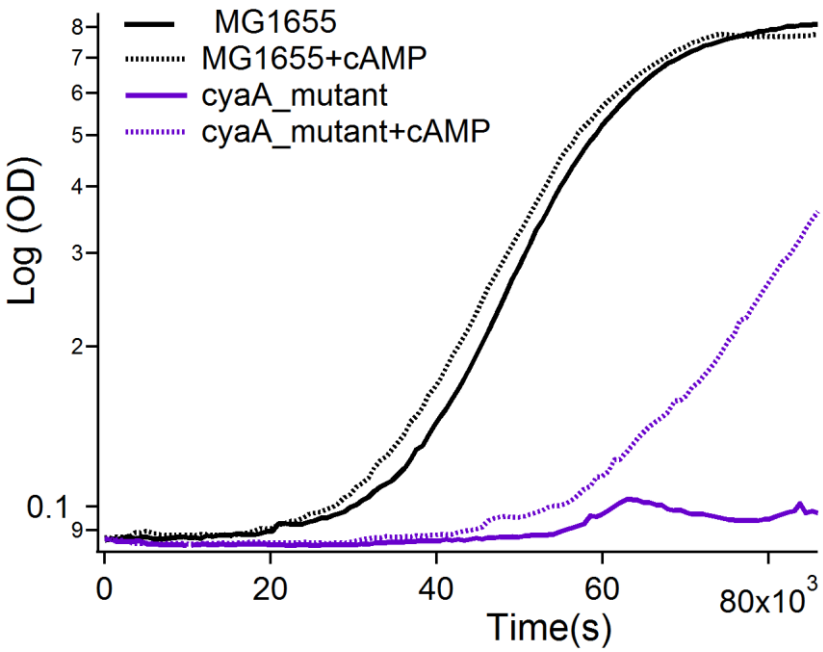

Fig. F.2: Log (OD) vs time plot for both MG1655 and cycle 9 (cyaA mutant) with and without cAMP in minimal media M63 + 5% glycerol. The purple curves of the cyaA mutant was repeated six times to confirm the oscillating growth behavior.

APPENDIX G

Future ideas

G.1 Tethered WW Domains

G.1.1 Pressure Denaturation experiments on tethered WW domains

Temperature and pressure are two thermodynamic parameters that acts differently on protein. Temperature denaturation works via change in entropy whereas pressure denaturation occur due to volume change. Temperature denaturation of the tethered WW domains are studied in detail in my thesis. It would be interesting to see if under pressure perturbation similar or dissimilar trend in thermodynamic stability (decrease as more monomer units are added) would be observed. In particular the tetramer have shown a greater tendency to form aggregates hence making it an interesting system for pressure titrations (any intermediates are populated).

G.1.2 Coarse grained MD simulations

In order to reveal the nature of interaction between the domains tethered together, it is possible to perform coarse-grained MD simulations on these fast folding proteins. It would provide evidence that domains with same sequence has lower or higher tendency to form domain swapped aggregates. I have already made a working box for the dimer and also have also equilibrated it for around 1 ns.

G.1.3 Mutated Fip35 L7A

I have performed site directed mutagenesis on the Fip35, Dfip35 plasmids in order to generate Fip35 L7A and single monomer unit mutated in dimer. The aim was to generate a construct which folds faster but also have a lower melting temperature. I was successful in getting ~ 14 degree destabilization in Fip35 by making a single L7A point mutation. This makes it a good system to perform kinetics at relatively low temperatures, avoiding problems of cavitation and low signal noise at high temperatures. I have collected kinetics on the same.

G.1.4 Comparison of folding rates with different probes

My tethered WW domains construct may serve as a good model for performing force pulling experiments. Marqusee and co-workers have reported to observe parallel folding pathways being for SH3 domain. It is intriguing that whether similar or different folding rates will be observed for the tethered WW domain constructs by fluorescence T-jump and force pulling experiments.

G.2 Pressure Denaturation of protein inside living cells

G.2.1 How does stability of the protein PGK is perturbed in presence of co-solutes like TMAO under high pressure stress inside living cells?

I have reported in my doctorate thesis work that it is possible to perform pressure unfolding experiments in living bacterial cells using ReAsH labeling scheme. It is also known that co-solutes like TMAO stabilizes proteins *in vitro*. TMAO is also found in in large amounts in fishes and deep sea organisms which are subjected to high pressure. Thus, adding ~ mM concentration range of TMAO and performing the pressure titration in both the pressure resistant J1 strain and wildtype MG1655 strains would be experiments of prime interest.

In a step further it would be great to perform these high pressure denaturation experiments in mammalian cells to draw comparison with the bacterial cells. Recently Oliverberg and co-workers have shown that protein stability is different in cellular environment (bacterial vs mammalian cell)

G.2.2 P-T phase Diagram inside living bacterial cells

With my semi-automated pressure generator it is possible to perform pressure and temperature denaturation experiments efficiently (atleast 2 experiments in a day) to get a P-T phase diagram inside living cells.