An interdisciplinary study of the mechanical and dynamic properties of α-solenoid repeat proteins

Marie Synakewicz

Jesus College
University of Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy

October 2018

"Reserve your right to think,
for even to think wrongly is better than not to think at all."
Hypatia of Alexandria

"If it is a good idea, go ahead and do it."
Grace Hopper

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text

It does not exceed the prescribed word limit of 60,000 words.

Marie Synakewicz

7 August 2019

Abstract

Tandem-repeat proteins differ from globular proteins, both in their biophysical characteristics and in how they interact with their respective partners, yet they comprise nearly one third of the human proteome and are central to many cellular processes and disease phenotypes. Repeat proteins have been shown to behave like nano-sized biological springs: they are flexible, dynamic and elastic.

Using coarse-grained models, I discuss how intrinsic flexibility may arise in repeat proteins and how it could be crucial for the biological function of two systems: PR65, the scaffold protein of the protein phosphatase 2A, and Rap proteins, which are involved in quorum sensing. To interrogate α-solenoids at physiologically relevant forces, I performed force spectroscopy experiments using a dumbbell optical tweezers set up for which it is necessary to attach the relevant protein to DNA. As PR65 is not amenable to current DNA-protein attachment methods, I developed a protocol that allows the cross-linking of DNA oligos to proteins using bio-orthogonal chemistry. I then explored the mechanics of the natural repeat protein, PR65, and a series of designed TPR proteins. I find that these proteins respond to forces in a novel manner which is significantly different to what has been previously reported. TPRs unfold and refold in quasi-equilibrium at constant force without energy loss. In contrast, PR65 unfolds in separate domains and refolds along an entirely different pathway.

In conclusion, my doctoral studies explore the physical characteristics of repeat proteins in more detail. Using both experimental and computational techniques, I provide unique perspectives on different aspects of their mechanical and dynamic capabilities. This work provides the basis for future investigations of how such interesting mechanical behaviour relates to biological function. Are repeat proteins simply a molecular recognition platforms for their multitude of binding partners, or do their mechanics matter in a biological context?

Acknowledgements

First and foremost, I would like to thank Laura Itzhaki for providing me with the opportunity to work on this particular project for my PhD. Throughout, she has given me freedom to be creative and to explore ideas even if they lead to nothing eventually. She also provided excellent guidance whenever I needed it and she has always supported me when I sought for help or expertise beyond our own lab.

Next, I would like to thank all members of the Itzhaki lab that I had the fortune to work with. Whether it is discussing family or politics over a cup of tea/coffee/water/squash and cake, or whether it is to bounce ideas of each other, this group makes it possible to motivate yourself on days when your molecular cloning has not worked for the fifth time in a row.

I have always been intrigued by single-molecule force spectroscopy and therefore, I am very grateful to Matthias Rief for agreeing to collaborate with us. A great many thanks also go to Daniela Bauer as well as the rest of the Rief lab for their patience with my never-ending stream of questions, frequent calls for help from the cellar and for providing constant entertainment (Igor - that should say it all).

Most of the work on amber suppression would not have been possible without the continuous help from Kaihang Wang and Nicolas Huguenin from the Chin Lab. They shared tips and tricks, some of which were unpublished, and took the time to answer the multitude of questions I had while I was trouble-shooting.

There are a many people, both in Cambridge and abroad, who I would like to thank for simply being friends in good times and in bad ones. You know who you are - but most of you are unlikely to ever read this.

Last but not least, I want to thank those people closest to me. Without my parents' continued support and advice I would not have been able to do what I wanted to do over the past ten years, and I promise that the mud of the Rhinluch will always stick to my boots wherever I go. And then ... there is this guy who wants to be named in here but just because of that wont be. Thank you for making me try out rowing. Thank you for being strong when I can't be. But most of all, thank you for making me laugh every day.

Contents

List of Figures

Copyright

All figures in Section 3.4.1 were reused from Perez-Riba *et al.* [1], as authors are allowed to re-use parts of their own work without seeking the Royal Society's permission. The full publication is attached in Appendix C.

Reuse licenses or permissions have been obtained for the following figures:

Figure	**Publisher**	**License number**
1.2	Annual Reviews, Inc	4601410938289
1.6a	Springer Nature	4457111182585
1.6b	AAAS	4457111356752
1.6c	AAAS	4457111476628
1.7	Elsevier	4457201430746
3.4	Elsevier	4404171235216
3.3a	The Royal Society	Permission granted by email
5.3a	ACS Publications	Permission granted by CCC
5.3b	Springer Nature	4431280692138
7.9a	Springer Nature	4437250816027
7.9b	American Physical Society	RNP/18/SEP/008105
7.9c	Springer Nature	4437541075191
8.2	-	Permission granted by L. S. Itzhaki

List of Tables

List of Abbreviations

aaRS	Aminoacyl tRNA synthetase
AD	Alzheimer's Disease
(C)ANK	(Consensus) Ankyrin
(C)ARM	(Consensus) Armadillo
CLT	Contour-length transformation
CoA	Co-enzyme A
CuAAC	Copper(I)-catalyzed azide-alkyne 1,3-dipolar cycloaddition
DHR	Designed helical repeat
DIBO/DBCO	Dibenzocyclooctyne
EAR	Extra-ordinary acoustic Raman
ENM	Elastic network model
EDA	Essential dynamics analysis
ESI	Electrospray ionisation
FRET	Förster resonance energy transfer
HDX	Hydrogen-deuterium exchange
HEAT	Huntingtin, EF3, A subunit of PP2A, Tor1
IED-DA	Inverse-electron demand Diels-Alder
IMAC	Immobilized metal affinity chromatography
LRR	Leucine-rich repeat
MALDI	Matrix-assisted laser desorption/ionization
MD	Molecular dynamics
NaAsc	Sodium ascorbate
NMA	Normal mode analysis
NMR	Nuclear magnetic resonance
PBS	Phosphate-buffered saline
PCA	Principal component analysis
PP2A	Serine/threonine-protein phosphatase 2A
PR65	PP2A 65 kDa regulatory subunit A alpha isoform
PrK	Propargyl-lysine
RF1	Release factor 1

RMSD	Root-mean-square deviation
RMSIP	Root-mean-square inner product
RTH-SDM	Round-the-horn site-directed mutagenesis
SAXS	Small-angle X-ray scattering
SDS-PAGE	Sodium dodecyl sulfate olyacrylamide gel electrophoresis
Sfp	Sfp 4'-phosphopantetheinyl transferase
SMFS	Single-molecule force spectroscopy
SPAAC	Strain-promoted azide-alkyne cycloaddition
(C)TALE	(Consensus) Transcription activator-like effector
TCO	Trans-cyclooctyne
THPTA	Tris(3-hydroxypropyltriazolyl-methyl)amine
(C)TPR	(Consensus) Tetratricopeptide repeat
UAA	Unnatural amino acid
WLC	Worm-like chain

Chapter 1

General introduction

When considering the variety of ongoing inter-disciplinary research, biology has probably profited the most from all other disciplines. Whether it is the computer sciences that aid data mining of a vast number of DNA and protein sequences, technological advances in the development of biomedical prostheses, or the use of statistical physics to model biomolecular processes. The work presented here would not have been possible without the application of linear algebra, chemistry, and statistical and polymer physics. Each of them has provided a unique point of view with which to gain novel insights into the properties of spring-like molecules that are crucial to the functioning of life; and only, when taken together, do they have the potential to connect different characteristics to yield a holistic picture. To note, the development of optical tweezers and directed evolution have been honoured with Nobel Prizes this year: Arthur Ashkin (Physics) and Frances Arnold (Chemistry). Without their work and that of their colleagues, an integral part of my research would not have been possible.

1.1 The world of tandem-repeat proteins

A random search for 3D structures of biomolecules in the Protein Data Bank (PDB), will yield a huge variety of shapes and sizes: some are elongated or fibrillar, others approximately round; some are single domains and others again form incredible multi-molecular assemblies. A subset of these structures is classified by the repetition of relatively small repeats; together they form the tandem-repeat protein class. The many variations present in this class can be distinguished according to the size of their repeats [2].

The smallest repeats only contain 1-2 amino acids, are very hydrophilic and form large crystalline aggregates that are cytotoxic. Repeats containing 3-5 amino acids form long fibrous structures (e.g. collagen or coiled coils), in which one repeat constitutes one turn of the secondary structure and stabilization from inter-chain interactions is required. The next subset of repeats also arranges into elongated structures that are built from 5-40 residues. However, the array is stabilised by interactions between the individual repeating units that each can consist of 1-4 segments of secondary structure. One generally distinguishes between solenoids, in which the amino acid backbone wraps around one central axis of the protein, and non-solenoids, in which β-strands form different variations of elongated sheets or hairpins. In yet another subset of repeat proteins, the individual modules circularize to form toroid-like shapes and barrels, which are often stabilized by a small part of one repeat forming part of a neighbouring repeat. Finally, there are repeat proteins in which the repetitive unit folds as independently stable domains of 30-130 amino acids [2].

Together, these sub-classes of repeat proteins span a large variety of biological function and make up approximately a third of the human proteome [3]. Here, I will focus on the solenoid class of repeat proteins, which mainly interact with other proteins and DNA, and are therefore central to many cellular signalling pathways [4].

1.1.1 Structure and function of solenoids

Nearly two decades ago, Kobe and Kajava [5] compared the structural building blocks of a variety solenoid proteins and classified the individual repeats into 18 types. Examples are shown in Figure 1.1 that highlight both structural similarity and structural differences. All of them contain 1-4 segments of either α-helices, 3_{10}-helices or β-strands which are connected by short loops. A given repeat can contain segments of only one type of secondary structure, or a mixture of at most two. The individual building blocks then stack into elongated repeat arrays which can form a variety of supra-molecular shapes such has spirals, helices, rods, horseshoes and tubes (Figure 1.1). Due to the regular arrangement of the individual building blocks, repeat proteins can be described geometrically either by defining angles between repeat planes [6] or by describing their supra-molecular shape using the description of a helix [7].

HEAT repeats are approximately 37-50 residues and consist of two helices that are of approximately equal size [15]. They are stabilised by an extensive, conserved hydrophobic core and by electrostatic interactions on the convex face of the array which arise from conserved arginine and aspartate residues [16]. As sequence analysis has shown, HEAT repeats can be further subdivided into three classes: the first contains Importin-β and other karyopherins, the second comprises clathrin and related proteins involved in vesicle formation, while the third contains multiple proteins with a variety of functions, such as a phosphatase subunit (e.g. PR65), a kinase (e.g. mTOR) and an elongation factor [17].

TPRs also contain a helix-turn-helix motif. They are 34 amino acids in length, and can be found in arrays of 3-16 repeats [18]. The stabilization of the hydrophobic core occurs between three helices, whereby the A-helix of a given repeat stacks against both

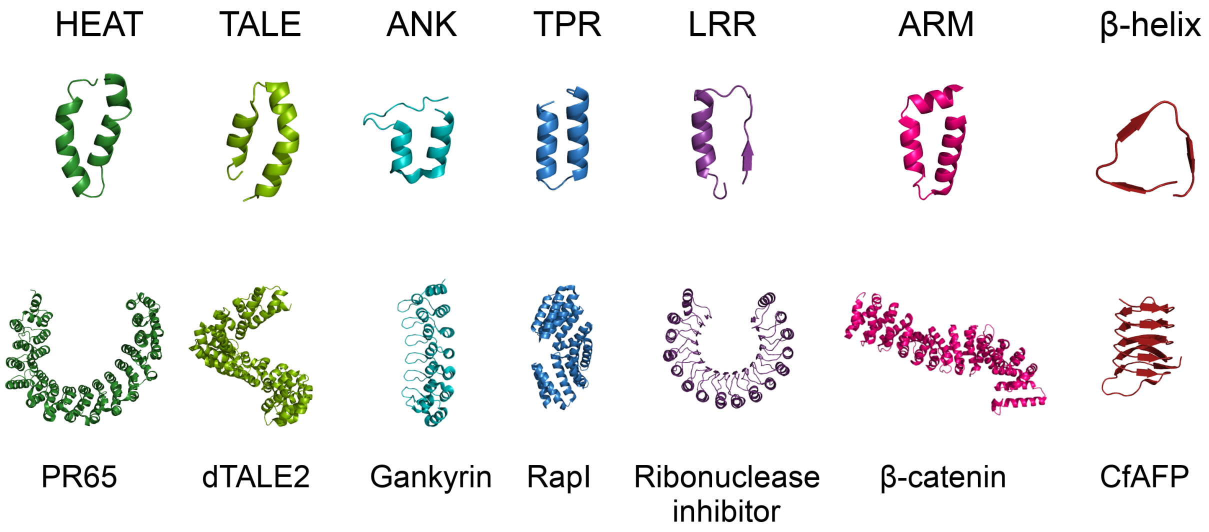


Figure 1.1: Examples of different solenoid proteins and their repeat type (PDB identifiers in parenthesis): HEAT - Huntington, elongation factor 3, protein phosphatase 2A, and yeast kinase TOR1 (2iae, [8]); TALE - transcription activator-like effectors (4hpz, [9]); ANK - ankyrin (1uoh, [10]); TPR - tetratricopeptide repeat (4i1a, [11]); LRR - leucine rich repeat (2bnh, [12]); ARM - armadillo (2z6h, [13]); β-helix (1l0s, [14]). The N-termini are at the top/left of each structure.

A- and B-helices of the preceding repeat, giving rise to an overall superhelical structure [19]. Many proteins contain TPRs and their functions are diverse, ranging from regulating molecular chaperones to mediating bacterial quorum sensing [18, 20]. In fact, TPRs can be sub-divided into 21 families, some of which are very similar in sequence, and others that vary in sequence and length [21].

TALE repeats are 34 amino acids long, The shorter of the two α-helices forms the concave surface and the longer helix forms the convex surface of a superhelix [22]. The central domain of TALE proteins is highly conserved and contains near-identical repeats [9]. Amino acids at two specific positions of the concave side of each repeat allow TALE proteins to interact with a specific DNA sequence [9]. Their function is therefore highly specialised and they are used by bacterial plant pathogens to mimic eukaryotic transcription [23].

ARM repeats are related to HEAT repeats but consist of three helices that arrange in a triangular fashion [17]. A conserved glycine between helix 1 and helix 2 initiates the turn. The third helix, also very conserved, form a concave ligand binding surface. ARM repeats are part of nuclear transport proteins, cell-adhesion sites and signalling pathways [17].

ANK repeats are the most well studied repeat proteins. They consist of approximately 33 amino acids which form a β-turn, followed by an anti-parallel pair of α-helices and a loop that connects to the next repeat [24]. ANK domains of proteins are at least 4 repeats in length but can contain up to 24 repeats [25, 26]. They are stabilised partially by a hydrophobic core between the α-helices but mostly by hydrogen-bond formation between the β-turns. ANK repeats occur in all species and are often part of larger proteins, including of enzymes, toxins and transcription factors [25]. The interaction with other proteins occurs largely *via* their β-turns [24].

LRRs are on average 24 residues in length and contain one β-strand followed by an α- or 3_{10}-helix [27]. Similar to ANK repeats the inter-repeat interaction is hydrophobic as well as due to hydrogen-bonds in the concave β-sheet. LRRs can be subdivided into 6 families, each of which differs in length and consensus sequence. They are often part of hormone receptors, are enzyme inhibitors, or have ribosome binding functions [27].

β-solenoids are entirely made up of β structures, although they may contain N- and/or C-terminal α-helical capping domains. The length varies between 5-30 residues per repeat and as many of the other repeats, they are also highly diverse in sequence and structure [2]. However, their function as virulence factors, toxins and allergens suggests that they are primarily involved in diseases [28].

The variety of structures of different repeat types on the one hand, combined with the structural similarity of a single repeat type on the other, makes solenoid repeat proteins ideal candidates to gain detailed insight into different folding mechanisms.

1.2 A brief overview of protein folding

As a protein is produced as a one-dimensional heteropolymer there must be rules that define its three-dimensional arrangement, independent of whether such a chain adopts traditional α-helical and/or β-strand secondary structures or whether it remains disordered. Otherwise, if a heteropolymer sampled all physically possible conformations, it would take longer than the current estimated age of the universe until most proteins found their energetic optimum [29]. In fact, some proteins can fold on the orders of 10-100 μs [30–32], and a small α-helical domain, the villin headpiece, can fold in under one microsecond [33].

The thermodynamics of the protein folding process are driven by an intricate balance of enthalpic and entropic contributions to the free energy of the system. Both contributions are of similar magnitude but opposite values, since enthalpy favours the formation of inter-residue contacts whereas entropy favours expansion of an amino acids chain and thereby opposes the folding process [34]. The enthalpic and entropic contributions from the solvent must also be considered. Graphically, the energy landscape of a protein can be represented by a folding funnel as introduced by [34, 35] (Figure 1.2). In the unfolded state at the top of the funnel a protein samples many possible conformations and the free energy is dominated by entropic contributions. The native state is lowest in energy as it comprises the smallest conformational space, giving rise to the characteristic funnel shape

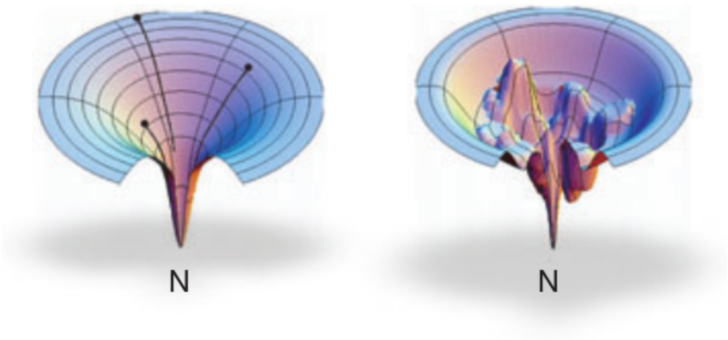


Figure 1.2: Schematic of a folding funnel depicting a smooth energy landscape of a fast folder (left) and a rough energy landscape with energy barriers and traps (right). The depth of the funnel is proportional to the free energy of the system, while the width is proportional to the conformational entropy. Adapted from Dill *et al.* [35].

seen in Figure 1.2. As the protein folds, the number of accessible conformations is reduced and contacts between individual residues are formed, thereby increasing the enthalpic contributions to the free energy. On their path from the unfolded to the native state, most proteins pass through the transition state, that comprises populations of molecules in which some contacts of the native structure are formed already but that are higher in energy than both the native and unfolded states [36]. If only the unfolded and folded states are significantly populated, the protein folds in a two-state mechanism. However, there are also accounts of proteins that never encounter a transition state barrier [37], so-called "downhill folders", and many others that fold through one or more stable intermediate that are separated by multiple transition state barriers and hence form a rugged energy landscape [38]. Furthermore, proteins can fold through sequential transition states that are separated by a high-energy intermediate [39] or through parallel pathways that involve two or more independent transition states [40]. Although these energy landscapes, or folding funnels, can correctly reconstitute a given folding pathway, they do not provide insight into the actual mechanisms that cause the heteropolymer to fold [41].

1.2.1 Driving forces of protein folding

It must be the physical interactions between amino acids and their side-chain characteristics that give rise to the folded, or native, structure. Through evolution a variety of sequences was selected during the folding of which such interactions do not interfere with each other but supportively and cooperatively lead to the structure with the lowest energy [42, 43]. A few factors have been identified, although their magnitude and hence importance tends to vary depending on the system under investigation.

With the explosion of the number of protein structures available for analysis, there is now considerable evidence that interactions between hydrophobic side chains are crucial to the folding process: In all soluble proteins (i.e. not membrane proteins) that adopt a definite three-dimensional structure hydrophobic residues can be found at the centre, sequestered from the solvent. This indicates that a major driving force is the demixing of polar and nonpolar amino acids, as well as the demixing of nonpolar amino acids and water [42, 44]. Electrostatic interactions, albeit unlikely to dominate folding in most cases, can stabilize the native state in some proteins and were also shown to make significant contributions to the denatured state of proteins, suggesting that their formation could both drive or inhibit the folding process in these cases [35, 45]. Both experiments and simulations show that hydrogen bond formation between carbonyl and amide groups of amino acids play a major role in the formation of secondary structure (leading to α-helices and β-sheets) [46, 47]. Van der Waals forces must also contribute to the tight side chain packing in the hydrophobic core and are an intrinsic parameter of many simulations [41, 48]. Due to a sum of short- and long-range interactions, some backbone and side-chain conformations are unfavoured or even inaccessible which limits the orientations the polypeptide chain can adopt [41]. Interactions with the solvent can influence any conformation of a polypeptide chain and hence, in aqueous solution, even the unfolded state will adopt some form of collapsed structure instead of a random coil, and this is thought to precede the folding process [31, 45, 49–51].

Finally, the characteristics of certain amino acids, such as proline and large aromatics, or the presence of many charged amino acids, can influence the folding pathway and the native structure [31]. Furthermore, disulfide bridges can not only affect the folding kinetics but also significantly stabilise the native state by restricting their conformation or the conformational space available to the folding polypeptide chain [52]. Lastly, post-translational modifications which allow precise biological control (e.g. phosphorylation), can alter stability of proteins significantly as well as shift their native state to another conformation or even induce folding [53, 54].

1.2.2 Experiments and simulations to elucidate protein folding mechanisms

Insights into folding mechanisms can be gained by the application of various techniques that probe the thermodynamics (or equilibrium) and the kinetics of the system. The conformation of a protein can be altered by changing the pH [55], adding a denaturant (e.g.guanidinium hydrochloride or urea) [56], or by raising the temperature of the sample [57]. At equilibrium, populations of folded and unfolded protein exist at the same time and the addition of a denaturant shifts this equilibrium towards the unfolded state [58]. A system is only at equilibrium if the denaturation process is reversible and hence temperature denaturation is rarely used as most proteins form aggregates when heated.

In kinetic, non-equilibrium, experiments the protein ensemble is subjected to a sudden change in environment using denaturant, pH or temperature jumps. By determining the kinetics of a folding process, one can gain insight into intermediates, if present, and the transition state(s) [58].

The conformational change as a function of denaturant or temperature can be monitored by a variety of experimental techniques [38]. Changes in secondary structure can be resolved using Far-UV circular dichroism (CD), whereas fluorescence of aromatic residues can detect changes in tertiary structure. Nuclear magnetic resonance (NMR) spectroscopy resolves structural changes in atomistic detail at equilibrium and in kinetic experiments, and it can give detailed information on inter-residue contacts in the unfolded, native and intermediate states [59, 60]. In hydrogen-deuterium exchange (HDX) experiments, solvent accessible hydrogens of amides are replaced by deuterium, or vice-versa. The rate of exchange will depend on the protection of the amide within a given conformation. HDX can be monitored using NMR, which is residue specific but provides only an average value of the population, or mass spectrometry, which is lower in resolution but resolves different populations [61]. Lastly, single-molecule Förster resonance energy transfer (FRET) can be used to probe the conformational state of a polymer or protein in a given condition. Although FRET cannot resolve structural detail, it provides temporal information on the end-to-end distance of molecules within a population or a mixture thereof [62].

Using protein engineering one can alter the thermodynamics and kinetics of the folding process. Conservative single point mutations can change the free energy difference between the native/unfolded and the transition states, and hence inform on the inter-residue contacts present in the transition state [49]. In contrast, more drastic single point mutations and truncations may alter a protein's folding pathway but they can give valuable information on the stability of individual domains [63].

Computational approaches can not only provide more detail and resolve distinct conformations along the folding pathway, they can also can give insight into the underlying physical principles [38]. Coarse-grained models are usually structure-based and rely on the theory of energy landscapes that predicts a more or less smooth path from the unfolded to the native state. From early, very simplified lattice models [64, 65], structure-based methods have evolved to include interaction potentials based on the information provided by the native state [66]. In contrast, atomistic simulations do not require pre-existing knowledge of the native state, but instead model folding based on a combination of different driving forces [67]. These models are computationally very expensive and therefore limited to short simulation times. However, recent advances made it possible that the complete thermodynamics and kinetics of protein folding can be modelled in cases where the computational and experimental time-scales overlap [68].

1.3 Folding mechanisms in the solenoidal repeat protein class

A number of studies have investigated the folding of tandem-repeat proteins. These vary in the type of repeat, the number of repeats and whether the repeat domains studies are naturally occurring or contain consensus sequences.

There is little reported on the unfolding of natural α-helical repeat proteins. One extreme of possible sizes constitutes an investigation of three 3-repeat TPR proteins, which unfold in a cooperative manner at equilibrium [69]. At the other extreme is a study on the 15-HEAT repeat protein, PR65, which folds through a hyperfluorescent intermediate at equilibrium [70]. Indeed, its energy landscape is so complex that that only rough boundaries between five individual subdomains could be identified. In the intermediate HEAT1-2 and HEAT11-13 are structured [71]. Data from kinetic studies suggests that there are parallel pathways leading to that intermediate. In the transition from the intermediate to the unfolded state, the remaining domains unfold sequentially. Transient unfolding under native conditions has been proposed to aid in the search for binding partners that associate with the N-terminal repeats [71].

The vast majority of the repeat protein folding field is focussed on ANK repeats. Indeed, the first folding study of a repeat protein to be reported was that of the tumour suppressor p16$^{\mathrm{INK4A}}$, containing four ANK repeat that unfold in one cooperative, two-state transition at equilibrium [72]. Subsequent computational and experimental studies involving point mutations and a deletion series could elucidate that its folding pathway was from the C- to the N-terminus [73–75]. Then followed many more studies of ANK proteins of increasing repeat array size, most of which fold in a two-state transition at equilibrium [76–79]. Myotrophin (4 ANKs) was found to fold from the C-terminus as well, however point mutations could reverse the folding pathway [76]. Although simulations predicted the *Drosophila* ankyrin-domain of Notch (7 ANKs) to fold starting at terminal repeats, equilibrium and kinetic studies of truncations and point mutants showed that folding is nucleated in ANK3-5 and only then propagates to the other repeats [75, 77, 78, 80]. Interestingly, this folding pathway can be rerouted by adding more stable consensus repeats to the C-terminus [81]. Recently, gankyrin (7 ANKs) was shown to unfold from the N-terminus to the C-terminus and to fold along a different pathway, from the N-terminus to the C-terminus [79]. In this protein, mutations affected the fractional population of the individual pathways [79].

Three ankyrin proteins were shown to fold through intermediates at equilibrium. The first two repeats of p19^{INK4d} (5 ANKs) and its archaeal homologue are unfolded in a stable kinetic and thermodynamic intermediate [82–84]. At body temperature, phosphorylation of two serine residues in these unfolded repeats leads to ubiquitination of a nearby lysine and hence downstream proteosomal degradation [85]. Simulations of IκBα (6 ANKs) predicted that the internal repeats would fold first and this was confirmed by experiment, although the exact nucleation sites were different in simulation and experiment [75, 86, 87]. Under native conditions FRET studies showed that ANK5-6 are disordered and only fold upon interaction with its binding partner, NF-κB [88, 89]. The intrinsic disorder of repeats 5-6 is crucial for the degradation of IκBα and therefore regulation of the transcription factor NF-κB [88].

As was shown by proteolytic digest and equilibrium studies, the C-terminal 12 repeats of ankyrin-R, called D34, unfold through an intermediate in which the last six repeats remain structured [90, 91]. This remainder can then unfold *via* two possible pathways [92]. Further investigations of the native state using single-molecule FRET showed that although the C-terminal half is more stable, it can exist in two states, one of which is fully folded while another is more extended [93]. The ability of D34 to form folded intermediates was proposed to provide more fine-tuned control in its biological role of linking the cell membrane to the actin-spectrin cytoskeleton [93].

The next most intensively studied repeat type within the natural repeat classes) is that of LRRs. All three proteins reported to date unfold in a two-state transition at equilibrium [94–96]. The N-terminal capping motif and first two repeats provide a folding nucleus for the C-terminal repeats in internalinB (7 LRRs) [97, 98]. YopM (15 LRRs) on the other hand is thought to unfold from the C-terminus [99]. However, in addition to the C-terminal β-strand, two internal repeats were found to have gate-keeper function for the folding of the weaker internal repeats and a point mutation could cause a switch to multi-state unfolding at equilibrium [99, 100]. In the study of PP32 (5 LRRs), [96] investigated the role of the N-terminal and C-terminal caps of LRRs and found that in this case, it is the C-terminal cap that initiates folding. However, when the C-terminus is destabilized by mutation the transition state moved towards the central repeats [101].

Very little is known about the folding of β-helical proteins. Pertactin, which contains 20 repeats, clearly folds through an equilibrium intermediate in which the C-terminal half remains structured [102]. Templating function of the C-terminal domain is thought to be relevant for efficient secretion of pertactin as it could prevent backsliding into the periplasm [102]. The 7-repeat pectate lyase C however, folds in an apparent two-state mechanism at equilibrium [103]. From a difference in differently determined spectroscopy parameters, the authors furthermore show that the folding behaviour deviates from two-state at high and low pH.

In summary, the examples detailed here highlight the multitude of folding pathways that are accessible to natural repeat proteins. Some fold from the inside out, others from the outside in. Some fold through intermediates that can be detected at equilibrium, others do not. However, one characteristic that most have in common is that they exhibit higher order kinetic folding behaviour that usually involves the nucleation of a few repeats that template the formation of the remainder. Only myotrophin and internalinB were shown to be true two-state folders both at equilibrium and kinetically [76, 97]. Furthermore, many exhibit slow refolding phases that can be attributed to proline isomerization. If the stability of repeats is distributed evenly across the array, repeat proteins tend to exhibit two-state behaviour at equilibrium, but if the stabilities vary significantly, the protein tends to (un)fold through intermediates. Mutations can shift the folding from one pathway to another, and can abolish or introduce an on-pathway kinetic intermediate. Due to their symmetry and structural simplicity, straightforward truncations can be used in addition to single point mutants which can facilitate the mapping of individual repeat protein folding pathways. Furthermore, although most repeat proteins have highly repetitive sequences, misfolding, as was reported for globular domains [104, 105], is rarely observed. In contrast to globular proteins, many repeat proteins fold much slower than predicted, the reason of which lies in their folding mechanism [63]. The folding nucleus usually contains elements of at least two repeats, which suggests that single repeats are unlikely to be found folded on their own and formation of an interface between two repeats is stabilising. This observation leads to the assumption that a transition state barrier within repeat proteins involves the folding of repeats without the formation of an interface [63]. However, due to the variability in stability, more detailed insight can only be gained from proteins in which all repeats are identical.

1.4 Working towards standard models: consensus repeat proteins

Consensus variants for multiple repeat types have been successfully designed based on the sequence of repeats within a family [19, 21, 22, 24, 106–109]. Most approaches use repeat alignments derived from a single protein, or a whole repeat protein family. However, these simple consensus approaches mostly optimize local energetics and may not necessarily reflect co-variance of residues which are involved in long-range interactions [63]. Therefore, more thorough analyses of such covariances can improve the consensus [106, 107, 110]. Further optimization can be achieved by removal of cysteines [19], balancing charges [106, 107] and overall amino acid composition [106], and improving packing the hydrophobic core [108].

Consensus ANK repeats were the first to be reported [24, 106], followed closely by TPRs [19] and LRRs [107]. All of them were thermally much more stable than natural proteins of the same repeat type and adopted the same structural conformation. These consensus repeats were originally designed with the aim to provide libraries of proteins that have randomized amino acids on their surface for binding capabilities. Depending on the target, binding function could then either be designed rationally or screened for. Indeed, soon TPR and various ANK repeat proteins were created that could bind their target specifically and with sufficient affinity [111, 112]. All consensus proteins were originally designed with capping motifs for one or both termini to shield the hydrophobic core and to increase solubility. Only recently, our group and others found that TPRs without a C-terminal capping helix remain soluble, monomeric and very stable [113, 114]. Early equilibrium denaturation data of small TPRs and ANKs suggested that they unfold in a two-state mechanism [24, 32, 115].

1.4.1 Folding of CTPR proteins and the re-discovery of the Ising model

A study by Kajander *et al.* [116] showed that equilibrium data of different length CTPR constructs can also be described using Ising models in which the total free energy of the protein is decomposed into the intrinsic energy of an individual repeat, ΔG_i, and the interfacial energy, ΔG_{ij}, that describes the coupling between two repeats. Ising models were originally developed to describe atomic dipole spins in ferromagnetic materials [117, 118] and were later applied to helix-coil transitions [119–121]. In an Ising model, there are only two energetic states accessible to the repeating unit: a spin of $+1$ or -1 (dipole spins), or folded and unfolded (helices, repeats). The total free energy of the system is a linear addition of the intrinsic energies of each repeat and each interaction potential between them, or coupling term. Arising from the underlying assumptions of this model, Kajander *et al.* [116] suggested that the microscopic picture of repeat protein folding at equilibrium should include partially unfolded species near the transition midpoint, and subsequently a low fraction of partially folded intermediates could be detected by NMR [122]. A calorimetric study of CTPRs containing 2-20 repeats provided evidence that a 2-repeat protein could be described by a two-state model, whereas a 3-repeat variant showed slight deviations from the fit that were only magnified as more repeats were added [123]. Similar deviations from two-state behaviour were also detected in kinetic studies [124]. As repeats are added, an on-pathway intermediate appears but the size of the two transition states (approximately 1.5 and 2.5 repeats, respectively) appeared to be independent of array length [124]. The refolding kinetics increase in speed whereas the unfolding kinetics slowed down with increasing repeat number, which was attributed to the difference in stability between the intermediate and the native state [32, 124].

A coarse-grained simulation showed that due to their one-dimensionality, TPR cooperativity and stability are directly related and depend on a balance between intrinsic and interfacial energies [125]. A protein becomes more stable and cooperative as the coupling is increased. However, if the intrinsic stability is increased, cooperativity decreases instead. Ferreiro *et al.* [125] could furthermore show that mutations can significantly affect these two parameters and thereby affect the degree of cooperativity in a repeat protein. Although, the simulations could reproduce intermediates, they did not resolve a folding pathway. Instead, their results suggested that if the folding correlation length of three repeats is exceeded, nucleation can occur anywhere as long as three repeats are available per nucleation site [125].

The stability and folding behaviour of CTPRs can be altered, either by mutation or by alteration of external factors such as ionic strength of the buffer. Historically, there are two CTPR consensus sequence that differ in the composition of the inter-repeat loop. The original consensus contained the "DPNN" sequence whereas most studies that required larger constructs contained the "DPRS" loop as a result of constructing the gene [19, 126, 127]. Ising analyses showed, the original consensus has significantly higher intrinsic and interfacial stabilities [113, 116]. Stabilizing mutations introduced into either the A or the B helix were found to only affect the intrinsic stability [127]. Furthermore, a reduction in salt content of the buffer can affect the stability of repeat proteins by decreasing the intrinsic repeat stability [128]. Most recently, our group was able to show that the CTPR scaffold could accommodate loops of up to 25 amino acids in length and that it primarily caused a decrease in the interfacial stability at the site of loop introduction [114].

The early Ising models took the repeating unit as a single helix due to the presence of the C-terminal capping helix [116]. It was assumed that the stability and interactions of the C-terminal helix were not significantly different from the internal helices and only recently has it been shown that the C-terminal cap is more stable than internal helices actual repeating unit is a whole repeat. Hence, Marold *et al.* [21] created a variant that contained full repeats but had hydrophobics at the N and C-terminal repeats substituted for polar amino acids. Results from our lab show that even these substitutions are not necessary and hence for the first time, we provided Ising model data based on truly repetitive units of whole repeats [114].

1.4.2 Folding of CANK repeats can also be described by Ising models

In contrast to the CTPRs, the capping motifs of consensus ANK (CANK) repeats are essential for solubility, and they even underwent multiple rounds of optimization to yield successively more stable constructs [129–131]. Due to these capping motifs, Ising models of ANK repeat proteins had to take into account different stabilities for the caps and the internal repeats [132]. CANKs are extremely stable both thermally and chemically and repeat arrays with more than four internal repeats and two capping repeats cannot be unfolded [132, 133]. Constructs containing one or two internal repeats showed two state behaviour at equilibrium and in kinetic experiments [132]. However, upon addition of the third repeat, a pre-transition was observed at equilibrium that corresponded to the unfolding of the C-cap [129]. While the refolding rates remained unaffected by the addition of more internal repeats, the unfolding was drastically slowed down [132]. Using the Ising model formalism, Wetzel *et al.* [132] could show that CANK proteins were likely to fray from the terminal repeats. In a later, study it was confirmed using HDX that indeed the core was more protected than the terminal caps and repeats [134].

Independently, Barrick and co-workers developed a CANK based on the Notch-Ankyrin domain with caps that only slightly varied in sequence from the internal repeats [133]. Using a deletion series of CANKs combined with both chemical and thermal denaturations, they were not only able to assign intrinsic and interfacial energies but they could also delineate the entropic and enthalpic contributions to these two energetic terms. While the intrinsic energy is enthalpy dominated and favours helix formation, the interfacial energy is entropy dominated due to the exclusion of water upon interface packing [133]. Further studies, involving global ising analyses of both equilibrium and kinetic data, could confirm an inside-out folding mechanism for CANKs [135]. Nucleation involving two repeats preferentially occurred at the centre, although it could also be detected at the N-terminus, and folding of the adjacent repeats would then proceed in parallel pathways. In contrast to the previous study by Wetzel *et al.* [132], the authors showed that it was due to these parallel pathways, that the unfolding rates decreased and the refolding rates increased with increasing number of repeats [135].

1.4.3 Further insights from other consensus repeat types

In addition to ANKs and TPRs, an ARM consensus was created to provide novel, designed proteins to bind peptide targets [108]. The consensus was designed using two ARM-repeat protein families, and its hydrophobic core was further optimised by computational sampling of side chain rotamers. The resulting constructs are thermostable and unfold reversibly. Using MD simulations and NMR, substitutions could be identified that reduced the dynamics of the C-cap and thereby increased the overall stability further [136].

It is difficult to determine the consensus of HEAT repeats as they can be long, irregular and vary strongly between subfamilies. However, based on a shorter, more homogeneous, HEAT-like motif, Urvoas *et al.* [109] developed a consensus that was thermostable but had a significant propensity to form dimers. Unfolding of CHEATs was found to be two-state even for longer constructs [109].

More recently, Barrick and co-workers reported the development of two new α-helical

Table 1.1: Ising models of different consensus repeats consist of the intrinsic energy, ΔG_i, its denaturant dependence, m_i, and the interfacial energy, ΔG_{ij}.

Protein	ΔG_i **[kcal mol^{-1}]**	ΔG_{ij} **[kcal mol^{-1}]**	m_i **[kcal mol^{-1} M^{-1}]**	**Ref**
CTPRn [a]	1.39 ± 0.04	-4.30 ± 0.07	-0.383 ± 0.005	[21]
CTPR_QKn [b]	-0.63 ± 0.04	-5.8 ± 0.1	1.03 ± 0.02	[114]
CTPR_QKa [b]	-1.07 ± 0.03	-4.08 ± 0.06	1.03 ± 0.02	[139]
42PR	2.01 ± 0.03	-4.63 ± 0.04	-0.572 ± 0.004	[21]
CANK	3.3 ± 0.2	-14.2 ± 0.7	1.1 ± 0.1	[132]
CANK[c]	5.2 ± 0.2	-12.5 ± 0.3	0.58 ± 0.03	[133]
CTALE(NS)[d]	5.89	-7.79	-0.5	[22]
CTALE(HD)[d]	3.49	-5.02	-0.5	[22]
DHR54[d]	-2.04	-6.76	-1.24	[138]
DHR71[d]	-1.41	-9.93	-1.57	[138]
DHR79 [d]	-3.48	-4.83	-1.12	[138]

[a] Based on the consensus reported by Main *et al.* [19] but repetition of a whole repeat.

[b] Contain the DE-QK substitutions reported by Cortajarena *et al.* [111] an no capping repeats.

[c] In addition to a denaturant dependence of ΔG_i, a denaturant dependence of ΔG_{ij}, $m_{ij} = 0.27 \pm 0.05$, was included.

[d] Assymetric errors obtained from Bootstrap analysis left out for ease of the reader.

consensus types. A 42 residue containing TPR-like repeat (42PR) surprisingly showed a reduction in overall stability compared to the usual 34 residue consensus although their helices are longer and hence were thought to be more intrinsically stable [21]. However, the unfolding of these 42RPs is more cooperative than that of the original TPR consensus. CTALEs with different DNA recognition motifs result in repeat types of different stability, which is thought to be relevant for DNA binding [22]. Both consensus repeat types can be described using Ising models and in the case of CTALEs an unfolding mechanism that involves end-fraying was suggested.

With the advancement of computational structure prediction many consensus repeats were re-designed [137], and even novel, totally unnatural designed helical repeats (DHRs) of variable repeat length were designed to create libraries of proteins with strikingly different geometrical shapes [7]. These DHRs were found to fold inside-out similar to ANKs, but due to their relatively high intrinsic stability of individual repeats the energy barrier of forming the first two repeats was absent [138].

1.4.4 The consensus of it all

To summarize, Ising model analysis of various consensus repeat types has shown that the conformational free energy of a repeat protein can be decomposed into intrinsic and interfacial energies. It is important to note that the local free energy of unfolding in the

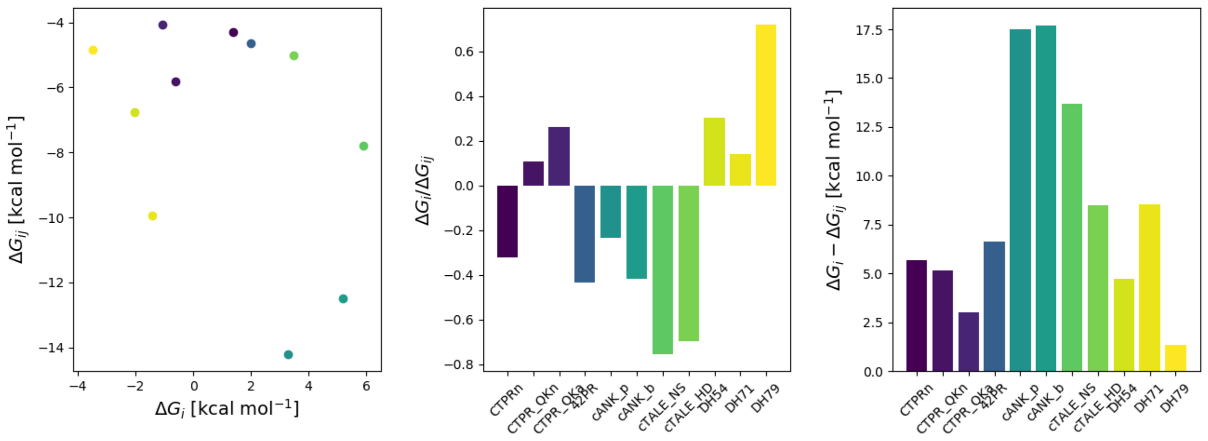


Figure 1.3: Graphical comparison of the intrinsic and interfacial energies (ΔG_i and ΔG_{ij}, respectively) between different consensus repeat types. Left - scatter of ΔG_{ij} versus ΔG_i (colours are the same as in bar plots); middle - ratio of ΔG_i and ΔG_{ij}, and right - difference between ΔG_i and ΔG_{ij} highlighting the energetic mismatch.

native state differs across repeats and is determined by the number of direct and indirect neighbours [21]. That is, the probability of unfolding for terminal repeats is higher than for central repeats which agrees with NMR and HDX data [32, 122, 134]. Energies for the different repeat types are summarized in Table 1.1 and Figure 1.3 to highlight the key differences and similarities. CANKs exhibit the highest interfacial coupling while the unnatural DHRs are the most intrinsically stable. While the CTALEs and the 42PR tend to be somewhere in between, the energetic mismatch or ratios of different TPRs can vary. This would suggest that stabilizing TPR mutations could potentially result in a similar folding mechanism to DHRs without the formation of the 2-3 repeat nucleus. For all repeat types, changes in consensus sequence can significantly alter the ratio or the energetic mismatch. In most cases, both intrinsic and interfacial energies are affected highlighting their tight interdependence and therefore their influence on cooperativity in repeat proteins [125]. As was predicted by Ferreiro *et al.* [125], natural repeat proteins will have evolved with a very fine balance between these two energetic contributions. Geiger-Schuller *et al.* [138] furthermore suggest that evolutionary pressure in natural repeat proteins will cause stabilization of long-range interactions, rather than an increase of the stabilities of individual repeats or subdomains. This observation supports the assumptions that if stable intermediates exist in natural repeat proteins they must be functionally relevant. Furthermore, considering that many repeat proteins are observed to fold in a two-state mechanism at equilibrium but have kinetic intermediates, raises the question of how exactly we define cooperativity. If true cooperativity relates to the absence of any thermodynamic and kinetic intermediates between the unfolded and native states, then only very few repeat proteins are truly cooperative on a global scale. That is, unlike in globular proteins, using repeat proteins one can define a degree of cooperativity which depends on the strength of the coupling between both ends of the repeat array.

1.5 Probing the mechanics of biomolecules using single-molecule force spectroscopy

Single-molecule force spectroscopy of proteins can provide amazing detail on various time and length scales. To date, the three commonly used methods by which to apply forces to biomolecules in a controlled manner are Atomic Force Microscopy (AFM), optical tweezers and magnetic tweezers.

1.5.1 Different methods for different purposes

In an AFM setup, the protein is attached to a surface and the AFM cantilever is lowered to pick up a molecule (Figure 1.4a). As the cantilever is moved away from the surface a force is exerted onto the molecule. At all times the position of the cantilever is monitored using a laser and unfolding events are detected by changes in the deflection of the cantilever. Usual AFMs have a millisecond and sub-nanometre resolution and can be used to apply forces of up to several nano-Newton which is particularly useful for very stable proteins [140]. However, the force resolution is often not sufficient for mechanically weak proteins [141].

The optical tweezers as we know them today are based on the discoveries of Arthur

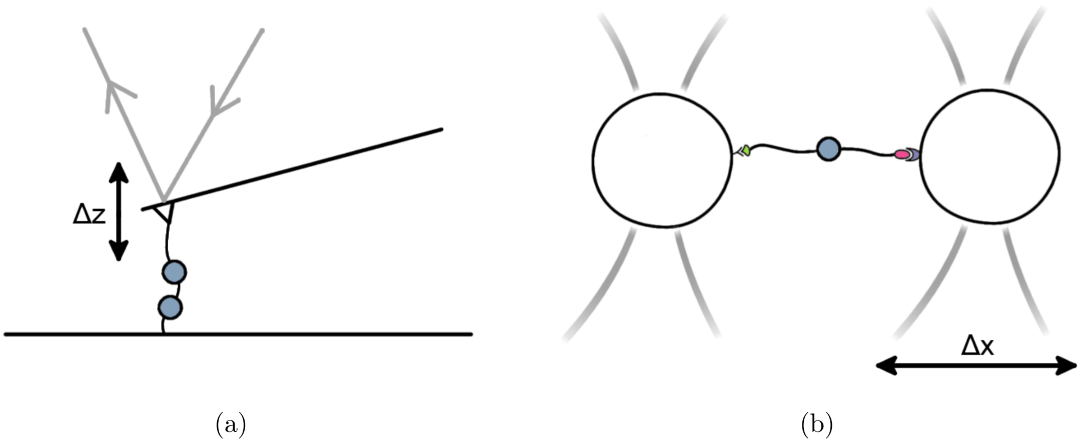


Figure 1.4: Two single-molecule methods that can be used to apply forces to proteins. (a) In AFM, the cantilever is moved up or down to apply a force to the molecule and to relax it again. (b) In a dumbbell optical tweezers assay two beads are trapped by an infra-red laser (see figure 1.5) and the force is applied by moving one trap away from the other.

Ashkin who found that micron to nanometre-sized dielectric particles could be trapped by a continuous laser beam (Figure 1.5) [142, 143]. Using chemically modified beads, biomolecules can be specifically coupled to such particles at one end while the other end is either attached to another bead (trapped in a second focus or held by a micropipette) or to the surface of the measurement chamber (Figure 1.4b). When the mechanical properties of smaller biomolecules are to be investigated, they are usually attached between two DNA molecules (so-called handles) to increase the distance between traps to reduce interference [144, 145]. As a force is applied to the biomolecule by moving a bead further away from the other (or with respect to the surface), it causes the bead to move out of the laser beam (Figure 1.5). Any (un)folding events are registered by monitoring a change in this deflection of the bead from the trap centre. Optical tweezers can resolve processes on the sub-nanometre and sub-millisecond scale but are limited to forces of ~65 pN by the transition of DNA from B- to S-structure and by to laser power which causes heating and oxidative stress of the sample [141].

AFM and optical tweezers can be used in the constant velocity and constant force

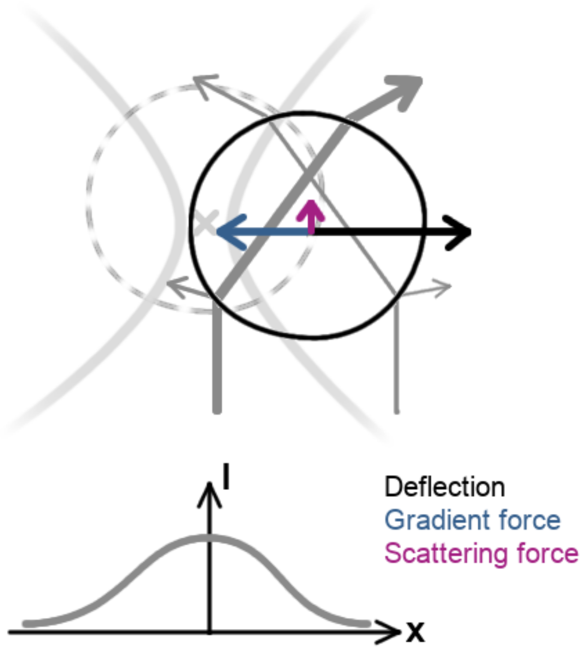


Figure 1.5: Trapping of a dielectric sphere using laser light. In a Gaussian beam light from the centre of the beam is of higher intensity than that from the edge. When a sphere is deflected, the part further away from the focus will refract less light than a part closer to the focus resulting in a net deflection of light away from the focus. Due to conservation refracted an equal and opposite force, the gradient force, acts on the sphere pushing it back into the focus. However, some of the light is also reflected at the surface of the sphere, and hence the scattering force pushes the bead away from the laser focus.

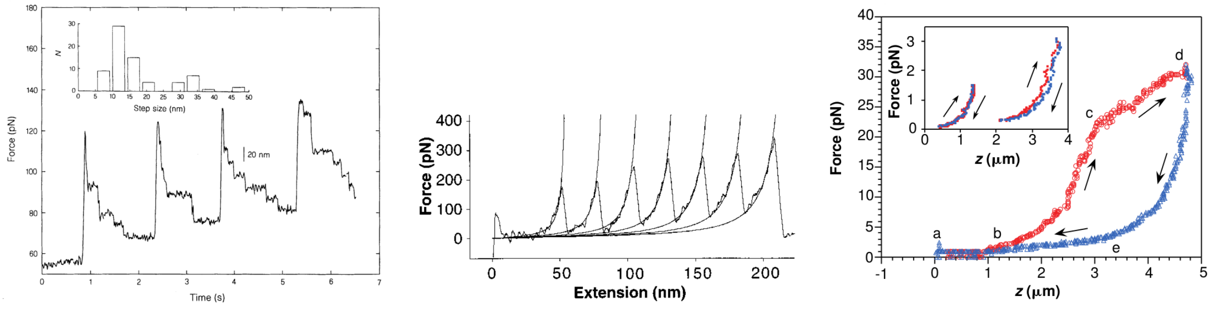
modes. When pulled at constant velocity, extension and force increase and can cause one or even successive unfolding events of a biomolecule. This method is used to assess the stability of a protein at a given speed and to detect possible intermediate [141]. In constant force mode, the biomolecule is kept at a constant force while unfolding and folding events are monitored at equilibrium. This method can be used to extract the kinetics of (un)folding [141].

While both AFM and optical tweezers apply forces linearly across a molecule, magnetic tweezers can induce torques [140]. In such experiments, the molecule is attached to the surface on one end and to a magnetic bead on the other. Either static or electromagnetic fields are then used to induce rotation of the bead. The bead position is monitored using bright-field imaging and any coiling or un-coiling events can be detected in a change of the bead height.

Current efforts in each of the respective fields focus on improving resolution both in time and distance, and on combining force and fluorescence spectroscopy [146].

1.5.2 Force-spectroscopy found application across all fields of biology

The first protein whose force-response was investigated was the muscle protein titin, which consists of multiple tandem-repeats of immunoglobulin and fibronectin III domains. Within days of each other, three groups published data on force extension of tintin spanning different force regimes (Figure 1.6). Using a titin molecule attached to a surface on one end and a trapped bead on the other, Tskhovrebova *et al.* [147] showed that when pulled to and held at a fixed distance, step-like unfolding events could be observed (Figure 1.6a). Rief *et al.* [148] reported a sawtooth-like pattern of unfolding events with some force peaks larger than 200 pN when titin was subjected to constant pulling velocities by AFM (Figure 1.6b). They were furthermore able to fit the the data with a Worm-like chain (WLC) model that describes the extension of a polymer under force [149] and hence suggested that each peak corresponded to the unfolding of a single imunoglobulin domain. The authors were furthermore able to show that when the force on the molecule was relaxed parts of it refolded and exhibited the same unfolding pattern upon renewed extension. In another set of optical tweezers experiments Kellermayer *et al.* [150] pulled from both ends of a titin molecule (Figure 1.6c). However, since their setup was fixed to apply forces to a maximum of 70 pN, they never observed discrete unfolding events. Instead they probed the polymer properties of the whole titin molecule and could showed that part of the extension exhibited WLC behaviour while another was linear. Relaxation was also observed but with considerable hysteresis [150]. After these initial investigations, further experiments focussed on engineered polyproteins con-



(a) (b) (c)

Figure 1.6: Early data on the force response of titin. (a) When held at constant force using, step-wise unfolding events could be observed. Adapted from [147]. (b) When subjected to constant pulling velocities, the molecule exhibited regular unfolding force peaks. Adapted from [148]. (c) In experiments that did not pull the titin molecule to full extension, a polymer stretching combined with a linear force response was observed. Adapted from [150].

taining multiple tandem-immunoglobulin domains with the aim to delineate its folding energy landscape and to establish a mechanism by which titin could function as a shock-absorber in muscle [151–153]. Using these constructs it was furthermore shown that the unfolding pathway and kinetics are the same when unfolding was induced by force or a chemical denaturant [154].

In the 20 years since titin unfolding was first reported, force spectroscopy was used to probe the mechanical properties of many proteins as well as DNA and RNA molecules. Due to the breadth of research, it is impossible to review all applications here. The early protein research focussed on molecules that were thought to be relevant in maintenance of the mechanical integrity of cells or in force-sensing, such as spectrins, polycystins and coiled coils, to name only a few [155–158]. Our understanding of how cellular motors move along cytoskeletal components was gained using optical tweezers experiments and in fact, the kinesin stepping along a microtubule was reported years before the unfolding of titin [159–162]. Others investigated the properties of DNA and RNA as well as RNA/DNA binding enzymes and complexes that alter DNA topology [163–165]. Force spectroscopy proved to be crucial for (i) mapping the energy landscapes of small globular proteins, including that of the fastest folding protein known and intrinsically disordered proteins, (ii) examining the physical basis of mechanical stability, and (iii) unfolding hierarchy of natural and engineered multi-domain proteins [152, 166–177]. Additionally, misfolding and protein degradation could be examined [178, 179]. Recently, even conformational changes of proline isomerization as well as catalytic reactions were resolved [180, 181].

Lastly, it became possible to study how protein and DNA interact with ions, co-factors and binding partners [182–189]. Mechanical disruption of a protein's conformation can affect not only the stability of a protein, but also its affinity for ligands and binding partners, or its catalytic efficiency [140]. Most interestingly, there is a large variation in how proteins respond to forces when bound to their interaction partners. In some cases, the binding of the ligand renders the protein more stable to mechanical stress [182, 190, 191]. In others, forces induce binding of a ligand [140], or simply open a buried binding site as seen with e.g. titin kinase [192] and the proline switch in filamin [180]. Furthermore, the application of a force to the catenin-cadherin complex bound to F-actin can increase the interaction strength and life-time considerably [193].

1.5.3 The nano-spring behaviour of repeat proteinss

The force-induced unfolding of repeat proteins contrasts with their chemical induced unfolding. Although most were shown to be highly cooperative or to possess only one or two kinetic intermediates, under force many repeat proteins unfold one or multiple repeats at a time.

In a CANK protein, each peaks corresponds to the unfolding of exactly one repeat (Figure 1.7) [194]. Gankyrin tended to unfold one repeat at a time but unfolding of two repeats at once could be detected in a small fraction of the molecules [195]. Larger ANK constructs such as ankyrin-B (24 repeats) and the ankyrin-R D34 domain (12 repeats) were instead often observed to unfold multiple repeats at once [196, 197]. The ribonuclease inhibitor (12 LRRs) showed similar unfolding behaviour to ANK proteins [197]. In contrast, unfolding events of fully α-helical proteins, such as the ARM repeat protein β-catenin and the HEAT repeat protein clathrin, could correspond to anything from a single

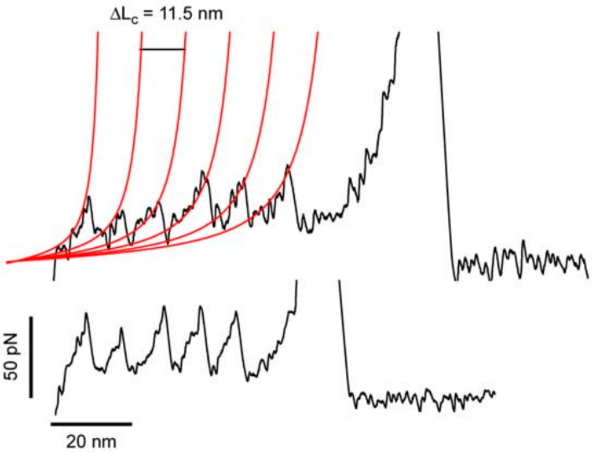


Figure 1.7: Force-induced unfolding of a consensus ankyrin. Shown are two traces, one where all six repeats of the protein are unfolded and one in which the protein was picked up at a random point internally and hence only five repeats were observed to unfold. The red lines in the upper panel correspond to WLC fits. Adapted from [194].

helix to multiple repeats at a time [197, 198]. However, unfolding steps of one repeat were the most frequent, followed by steps involving the unfolding of a single helix. Unfolding forces of each repeat ranged from 30 pN for α-helical proteins to 50-70 pN for ANKs. MD simulations of different ANK proteins and β-catenin suggest that although breaks can appear along the repeat array unfolding occurs from the termini inwards [191, 195, 198, 199]. In some cases, even stable intermediates could be detected [195, 198, 199]. However, more recent experimental evidence indicates that CANK proteins unfold from the C- to the N-terminus [173].

The real surprise of repeat protein folding lies within their refolding behaviour: all repeat proteins refold under force although hysteresis was present upon refolding and varied between repeat types. α-helical proteins and ANKs showed the least hysteresis, while LRRs exhibited the most due to the slow refolding of extensive β-structures [197]. Furthermore, hysteresis increased the more of the molecule was unfolded, suggesting that the folded remainder has a templating function [197]. Recently, the force-induced unfolding of the first β-helical protein was reported. It also unfolded one turn of the β-helix at a time, but at forces exceeding 100 pN, and refolded with minimal hysteresis [170]. Thus, repeat proteins are elastic and only dissipate minimal energy, even over multiple stretch-relax cycles of the same molecule [197].

Given the supramolecular shape of larger repeat proteins, they have also the potential to function as springs at the level of their tertiary structure, i.e. by changing shape without unfolding. Since all experiments to date were performed using AFM, the resolution was not sufficient to detect any such response of the tertiary structure apart from in experiments of the 24-repeat protein ankyrin-B [196]. This observation suggests that it may require very long arrays for a tertiary structure response to be resolved. In contrast to experiments, multiple steered MD simulations could visualize a stretching of the repeat array without loss of secondary structure. Importin-β (19 HEAT repeats) and a model of ankryin-R could extend to more than double their starting end-to-end distance [199, 200]. PR65 (15 HEAT repeats) was also shown to respond to large pulling and pushing forces by altering its shape [201]. In all three proteins, the repeat interfaces rearranged to accommodate the structural changes. Furthermore, all returned to conformations very close to the original structure after the force was released [199–201].

Some further evidence of the structural flexibility of repeat proteins comes from solutions experiments. Small-angle X-ray scattering data of Importin-β indicate that it is highly flexible and its conformations can fluctuate substantially [6, 202, 203]. Both the Notch ankyrin domain and the tumour suppressor p19^{INK4d} have one or two very dynamic repeats, the unfolding of which is required for post-translational modifications and hence protein homeostasis and signalling [204]. Furthermore, two FRET studies can provide some indirect evidence for both secondary and tertiary flexibility. At equilibrium, IκBα fluctuates between extended and fully compact states which can be related to the unfolding of repeats 5-6 [88]. For a CTPR3, a reduction in FRET signal was observed in a sub-population at low denaturant concentrations in which all secondary structure was still intact [205]. This lead the authors to assume that it could expand without actually unfolding.

Together, these studies either show directly or imply that repeat proteins have quite unique mechanical properties. In the cases of ankyrin-B/R and β-catenin these properties might be relevant to their biological function as they localise to force-gated ion channels and adherens junctions, respectively [198, 199]. Importin-β is likely to require a high degree of flexibility to be able to transfer a wide range of cargoes into the cell nucleus [203]. However, for most other repeat proteins, there is to date no understanding of how exactly the spring-like nature of the scaffold relates to its function [191].

1.6 PP2A and its scaffold subunit PR65

Cell development and homeostasis are dependent on the correct interplay between kinases and phosphatases, many of which have very broad substrate specificities [15]. Phosphorylation, as one of many post-translational modifications, can be the key to the activity of many a protein. Protein phosphatase 2A (PP2A) belongs to the PhosphoProtein Phosphatase (PPP) family of Ser/Thr phosphatases [206]. Depending on subunit combination and isoforms, there are more than 90 different PP2A variants that can account for up to 1% of all cellular protein, and each is likely to exert a non-redundant function [207, 208]. Together with PP1, PP2A is responsible for 90% of the Ser/Thr phosphatase activity in the cell [208]. Hence, it is not surprising that PP2A is involved in a multitude of processes that regulate metabolism, the cell cycle and transcription [206]. The loss of PP2A or altered PP2A function results in very pleiotropic effects. Different forms of cancer and Alzheimer's Disease (AD) have been associated with malfunctioning PP2A. Although some isoforms can be attributed to a distinct function, the general lack of isoform-specific antibodies and inhibitors hinders the elucidation of PP2A function [207].

1.6.1 Structure of PP2A

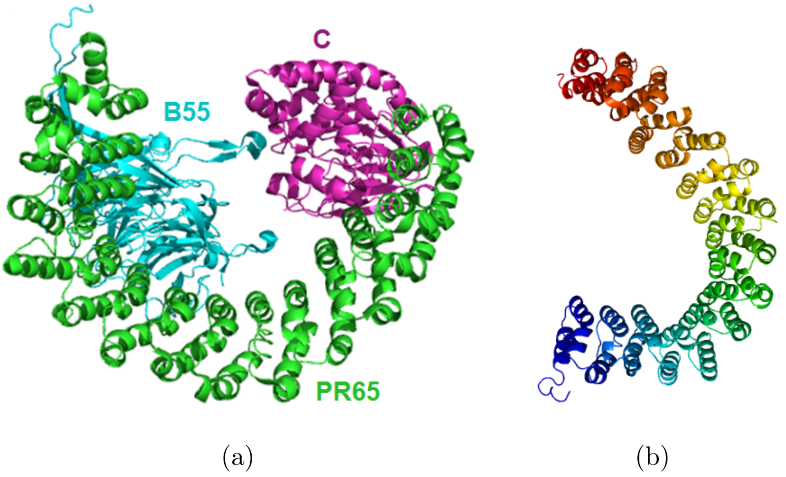
PP2A is a heterotrimeric enzyme that consists of a catalytic C-subunit (35 kDa), a structural A/PR65 subunit (65 kDa) and a variable B-subunit (Figure 1.8a) [15]. There are two C isoforms, yet only Cα knockouts in mice and yeast are lethal [208]. There are four subclasses of B-subunits, originating from 15 different genes and 26 splice forms. These subfamilies share no sequence similarity apart from the regions that interact with PR65 [208]. The two PR65 isoforms share 87% sequence similarity, yet seem to each have a non-redundant function [208]. PR65 is a non-globular repeat protein comprising 15 HEAT (huntingtin, EF3, A subunit of PP2A, Tor1) repeats (Figure 1.8b) whose sequence similarity is relatively low [15, 70]. HEAT repeats are roughly 37-50 residues long and are formed of two helices that arrange in a linear fashion by stacking upon each other [15]. Such solenoid structures lack sequence-distant contacts, but are instead stabilized by stacking interactions due to hydrophobic, aromatic and hydrogen bond forming residues [15]. The B-subunits bind to repeats H1-10 of PR65, while the C-subunits bind to the C-terminal repeats H11-15 [206].

Depending on the subunit composition, post-translational modifications of all three subunits, protein-protein interactions, inhibitors, activators and modulators, each PP2A heterotrimer has a distinct catalytic activity, substrate specificity, localization within cells and tissues, biogenesis and stability [207].

1.6.2 Localization and function of PP2A

Although PP2A has been associated with many cellular processes, only in some has the exact subunit composition been identified. Here, the focus will be on the following three heterotrimers, each representing one B-subunit family: PR65/C/B55 (ABC), PR65/C/B56 (AB′C) and PR65/C/PR70 (AB″C).

Holoenzymes containing B55 can bind and dephosphorylate Tau, and loss of B55 function results in Tau hyperphosphorylation and consequently AD phenotypes [207]. PP2A alone accounts for 71% of Tau dephosphorylation [210]. Mutations in the C-subunit can either cause defective PP2A assembly and B-subunit selectivity, or abolish catalytic activity [210]. pTau also sequesters normal Tau protein, which usually promotes assembly and



(a)

(b)

Figure 1.8: Crystal structures of (a) PP2A consisting of its A (PR65α), B (B55α) and C (catalytic) subunits; and (b) PR65, the scaffold subunit (colouring is blue-red from N to C). The PDB codes are 3dw8 [209] and 1b3u [16], respectively.

stabilization of microtubules (MTs), and therefore interferes with cytoskeletal dynamics [209]. B55 can bind directly to Tau at its MT binding region, and thus inhibition of B55 leads to Tau phosphorylation and consequently Alzheimer's disease [207]. ABC has also been observed to bind MTs directly and regulation of MTs likely is crucial to morphogenesis and tumorigenesis [211]. Yet, PP2A enzymes binding to MTs lose their catalytic activity [211].

The role of AB′C type PP2As appears to be very variable, possibly due to different B isoforms and splice variants. AB′C holoenzymes localize both to the centromere [212], the kinetochore [213] and to cell adhesion sites [211]. At the centromere, PP2A is crucial for protection of cohesion from cleavage in meiosis and mitosis [212, 214, 215]. In both mammalian and yeast systems, PP2A interacts with shugoshin proteins which are necessary for its localization. Shugoshins also interact with the PP2A inhibitor I2PP2A, which colocalizes with PP2A during meiosis II and whose relocalization seems to be tension dependent [214]. It has been suggested that centromeric protection itself is also tension dependent, however it is unclear whether it affects PP2A directly or simply dislocates its substrate [201, 216]. At the kinetochore AB′C is involved in a negative feedback loop that controls spindle assembly checkpoint (SAC) silencing by counteracting Mps1 and Aurora B kinases [213].

The PR70 members of the PP2A family have been shown to bind and phosphorylate Cdc6, whose regulation is crucial for G1/S phase transition in the cell cycle [217]. Cdc6 has a key role in DNA replication and its phosphorylation inhibits ubiquitination and subsequent degradation by the proteasome. PR70 targets PP2A to Cdc6 and its calcium binding EF-hands mediate an enhanced interaction between AB″C and Cdc6 [218].

Altogether, these three examples show how varied and yet specialized PP2A function can be. The core enzyme, consisting of the C-subunit and PR65, is constant throughout, suggesting that one regulatory mechanism of PP2A lies within the exchange of subunits.

1.6.3 PR65 flexibility may be crucial for PP2A function

Comparing crystal structures of different PP2A heterotrimers, the core enzyme and PR65 alone, it is apparent that PR65 can adopt conformations with varying geometries [16, 209, 217, 219, 220]. The end-to-end distance can differ by as much as 40 Å, indicating a considerable flexibility of the scaffold. The elasticity of PR65 has been proposed to influence the catalytic activity of the enzyme [201]. Since the B- and C-subunits bind to the N-terminal and C-terminal regions of PR65, respectively (Figure 1.8a), any change in curvature or twist of PR65 could potentially displace the subunits relative to each other. An MD simulation by Grinthal *et al.* [201] shows that changes in PR65 shape can occur at forces that are characteristic for cellular processes such as chromosome separation.

Furthermore, spontaneous fluctuations due to normal modes intrinsic to the scaffold may regulate PP2A activity [201]. Such fluctuations could open or close the B/C-subunit interface and thereby control substrate binding and release, and B-subunit exchange. Furthermore, our lab has extensively studied the folding mechanism of PR65, concluding that HEAT repeats 1-2 and 11-13 stabilize and guide the folding of the weaker, central domains [70, 71]. These results allow for a scenario in which unfolding of central HEAT repeats, while B- and C-subunits are bound to the terminal regions, is intrinsically linked to phosphatase activity.

1.7 General Motivation and Aims

Since the establishment of techniques that permit the manipulation of single molecules and cells, research has shown that forces are crucial to biological processes. They influence protein stability and their interaction with other proteins inside and outside of the cell. Tandem repeat proteins, possessing unique mechanical characteristics, have been shown to be important for various cellular processes. Using an interdisciplinary approach, I am interested in investigating the following overarching questions:

- How do structure and shape relate to mechanical and dynamic properties?
- How is repeat protein stability, as described by the Ising model, connected to their mechanical and dynamic properties?
- Are solenoidal repeat proteins a simple interaction platform for biological processes or are their mechanical properties functionally relevant?
- Is PR65 elasticity important for PP2A activity, and can spontaneous fluctuations within PR65 generate forces of sufficient amplitude to influence PP2A activity?

One of the original objectives was to first develop different PR65 variants which are designed to have altered mechanical characteristics and to test their biophysical properties and force responses experimentally before proceeding to phosphatase assays. Motivated by coarse grained models of repeat protein dynamics (Chapter 3), I attempted to extend the HEAT-repeat array and thereby alter its properties (Chapter 4). Additionally, point mutations and stapling of neighbouring repeats using the bi-arsenical dye FlAsH were envisaged to alter the mechanics. To overcome the resolution limit of AFM, I chose to investigate mechanics of PR65 using optical tweezers. However, I had to first establish DNA attachment methods for force-spectroscopy (Chapter 5) and examine their effects on protein stability (Chapter 6). After preliminary PR65 force-spectroscopy data was available, I decided to extend my studies to CTPR proteins (in lieu of consensus HEAT repeats) to aid the understanding of the PR65 force response (Chapter 7-8).

Chapter 2

General materials and methods

2.1 Materials

All chemicals were purchased from Sigma Aldrich, ThermoFisher, Merck or Asco Chemicals unless otherwise stated. 2x yeast tryptone (2xYT) and Lysogeny Broth (LB) Miller were purchased from Formedium. Unmodified DNA oligo nucleotides were purchased from Integrated DNA Technologies or Sigma Aldrich. *E. coli* strains and expression vectors are listed in Tables 2.1 and 2.2, respectively.

2.1.1 Preparation of chemically competent *E. coli*

E. coli cells from a glycerol stock were streaked onto LB Agar and incubated at 37°C over night, containing the appropriate antibiotic if Lemo21 or Rosetta cells were propagated. The next day, a single colony was picked and transferred to 5 ml 2xYT in a sterile 50 ml tube and again incubated over night while shaking (150-200 rpm) at 37°C (containing antibiotic if appropriate). The overnight culture was diluted 1:100 in 250 ml of fresh, sterile 2xYT (containing antibiotic if appropriate) and grown while shaking at 200 rpm at 37°C until the optical density at 600 nm reached 0.25-0.3. The cell suspension was filled into 50 ml tubes and cooled down on ice for 10 min to stop growth before centrifugation at 4000xg at 4°C for 5 min in a fixed angle rotor. The cell pellet in each tube was resuspended in 10 ml of Transformation Buffer I (30 mM KOAc pH 5.8, 100 mM RbCl, 10 mM $CaCl_2$, 50 mM $MnCl_2$, 3 mM Hexamine cobalt Cl, 15% (v/v) glycerol), combined into one sterile 50 ml tube and incubated on ice for 5 min before renewed centrifugation. The cell pellet was resuspended in 5 ml Transformation Buffer II (10 mM MOPS pH 6.5, 10 mM RbCl, 75 mM CaCl 2 , 15% (v/v) glycerol) and incubated on ice for 15 min. 50 µl aliquots of the cell suspension were dispensed directly into 2 ml microfuge which were pre-chilled at -80°C for at least 30 min. Competent cells were stored at -80°C.

Table 2.1: *E. coli* strains

Strain	Resistance	Description	Source
DH5α	-	K-12 strain derivative for routine molecular cloning applications; chemically competent cells for routine applications were propagated in-house; cells for high efficiency transformations were purchased	NEB, Bioline
C41 (DE3)	-	BL21 derived strain for high-level expression of recombinant protein; protein expression cassette under the T7 polymerase promoter	Kommander Lab, LMB
MDS42 ΔrecA	-	K-12 multiple deletion strain, MG1655, used for propagation of unstable genes and plasmids; used for molecular cloning and protein expression of amber suppression constructs	Scarab Genomics LLC [221]
Lemo21 (DE3)	Chloramphenicol	BL21 derived strain bearing the pLemo plasmid which codes for T7 lysozyme under a rhamnose inducible promoter for fine-tuning T7 RNA polymerase activity and consequently optimizing protein over-expression	NEB, [222]
Rosetta 2 (DE3)	Chloramphenicol	BL21 derived strain containing a plasmid supplying tRNAs for eukaryotic codons that are rarely used by *E. coli* under native promoters	Novagen

2.1.2 Preparation of electro-competent *E. coli*

E. coli cells from a glycerol stock were streaked onto LB Agar and incubated at 37°C over night. The next day, a single colony was picked and transferred to 5 ml 2xYT in a sterile 50 ml tube and again incubated over night while shaking (150-200 rpm) at 37°C. The overnight culture was diluted 1:100 in 500 ml of fresh, sterile 2xYT and grown while shaking at 200 rpm at 37°C until the optical density reached 0.35-0.4. All hardware (50 ml and 2 ml tubes) were pre-chilled on ice before use. Water and glycerol dilutions were autoclaved and pre-chilled on ice as well. The cell suspension was filled into 50 ml

Table 2.2: Expression vectors

Vector	Resistance	Description	Source
pET28a	Kanamycin	Low copy number expression plasmid for expression of a C-/N-terminally H_6-tagged gene under a T7 promoter	Novagen
pRSETa	Ampicillin	High copy number expression plasmid for expression of an N-terminally H_6-tagged gene under a T7 promoter. The tag is separated from the gene of interest by an enterokinase cleavage site, which was modified to a thrombin cleavage site in our laboratory	Invitrogen
pGST	Ampicillin	pRSET derivative in which the H_6-tag was exchanged to a glutathione S-transferase.	Itzhaki group
pRSET_H_6_ybbR, pRSET_GST_ybbR	Ampicillin	pRSET derivative for the construction of ybbR-tagged proteins contain N-terminal H_6- or GST-tags, followed by a poly-N sequence, a TEV protease cleavage site, the N-terminal ybbR-tag and a BamHI restriction site. The C-terminal ybbR-tag is located between a HindIII restriction site and a stop codon. In-frame insertion using BamHI and HindII restriction sites will result in fusion proteins where N- and C-terminal ybbR tags are separated from the protein by the amino acids GS and KL, respectively.	M. Synakewicz
pKW1	Spectinomycin	pCDF derivative with a strongly re-designed backbone containing an *Methanosarcina barkeri* pyrrolysine tRNA and corresponding amino-acyl synthetase	[223]
pRSF_oRibo-Q1	Kanamycin	pRSF derivative containing an orthogonal 16S subunit under a T7 promoter and the gene of interest preceded by an orthogonal ribosome binding sequence under an constitutively active promoter. Proteins of interest have to be tagged with an N-terminal GST- and a C-terminal H_6-tag and inserted into pRSF_oRibo-Q1 using the SwaI (within GST) and SpeI restriction sites.	[224]

tubes and cooled on ice for 20-30 min to stop growth before centrifugation at 1000xg at 4°C for 20 min. The cell pellet in each tube was resuspended in 20 ml of ice-cold, autoclaved MilliQ-$H_2$0, combined into 5 sterile 50 ml tube before renewed centrifugation. Each cell pellet was resuspended again in 20 ml ice-cold MilliQ-$H_2$0, combined into two 50 ml tubes and centrifuged again. Each cell pellet was then resuspended in 20 ml ice cold, autoclaved 10 % (v/v) glycerol and harvested by centrifugation. Finally, both pellets were resuspended 1 ml 10% glycerol and 50 µl aliquots of the cell suspension were dispensed directly into 2 ml microfuge and snap-frozen in liquid N_2. Competent cells were stored at -80°C.

2.2 Molecular biology

2.2.1 Cloning methods

Restriction digests and ligation

Unless otherwise stated DNA was digested using FastDigest enzymes (Thermo Scientific), which reduced the incubation times to 15-30 min. To create a particular construct, insert (1-10 µg of vector or purified PCR product) and 1-10 µg vector were digested in 30 µl using the appropriate restriction enzymes according to the manufacturer's protocol. The vector was dephosphorylated using alkaline phosphatase and the digested product of the correct size was identified by agarose gel electrophoresis, extracted and purified according to the QIAquick gel extraction protocol. An insert derived from a vector was purified in the same manner as the vector. However, if PCR product was used it was DpnI digested during restriction digest and then purified using the QIAquick PCR purification protocol. All DNA was eluted into MilliQ-H_2O and stored at -20°C if not used immediately. Usual yields of purified product were $>$30 ng µl^{-1} for the vector and 10-20 ng µl^{-1} for the insert (depending on the concentration before digest)

For standard BamHI-HindIII cloning, 1 µl of digested vector and 2 µl of insert (irrespective of the DNA concentrations) were added to 1 µl Anza^{TM} T4 DNA Ligase Master Mix (ThermoFischer), incubated for 10-20 min at room temperature and the whole mixture was transformed into in-house produced, chemically-competent DH5α *E. coli* cells. Specific constructs with potential for recombination, such as TPRs, were ligated with QuickStick Ligase (Bioline) according to the manufacturer's protocol using a 1:3 molar ratio of vector to insert. A maximum of 3 µl of the reaction mixture was transformed into high efficiency chemically competent cells such as NEB® 5-alpha Competent *E. coli* (High Efficiency) (NEB) or Alpha-Select Gold Efficiency *E. Coli* (Bioline). Transformed cells were plated onto LB agar containing the appropriate antibiotic and incubated at 37°C over night. Usually, 2-4 colonies were selected and propagated in suspension culture for plasmid standard purification. Incorporation of the gene into the plasmid was first confirmed by renewed digest using the same enzymes and subsequent agarose gel electrophoresis, before its sequence was verified by Sanger sequencing (Eurofins).

Ligation-independent cloning: FastClone

If the introduction of a gene, or part of a gene, using restriction enzymes is not desired as it introduced additional amino acids, ligation-independent methods can be used. Among various commercial products (NEB-Builder, In-Fusion Cloning), FastClone can provide an easy, cheap alternative [225]. One set of primers are designed to linearise the plasmid at the desired insertion site (Figure 2.1a). The other set of primers is designed to amplify the gene construct to be inserted. Both sets of primers have 15-17 bp of overlap, that is the vector primers contain 15 bp of the insert and the insert primer contain 15 bp of the vector. After PCR amplification, DpnI is added to both reactions after which both samples are mixed incubated at 37°C for at least 1 hr to digest all template DNA. 3-5 µl of the mixture were then transformed as usual into chemically competent *E. coli.*

Site-directed mutagenesis: Traditional and Round-the-Horn methods

Single point mutations can be inserted using site directed mutagenesis (SDM). Two overlapping primers are designed which contain the mutation (Figure 2.1b). After a standard polymerase chain reaction (PCR) with these primers and an appropriate DNA polymerase (e.g. Phusion High-Fidelity, NEB) the template strand is digested using DpnI and 3-5 µl are directly transformed into chemically-competent DH5α *E. coli* cells. However, this

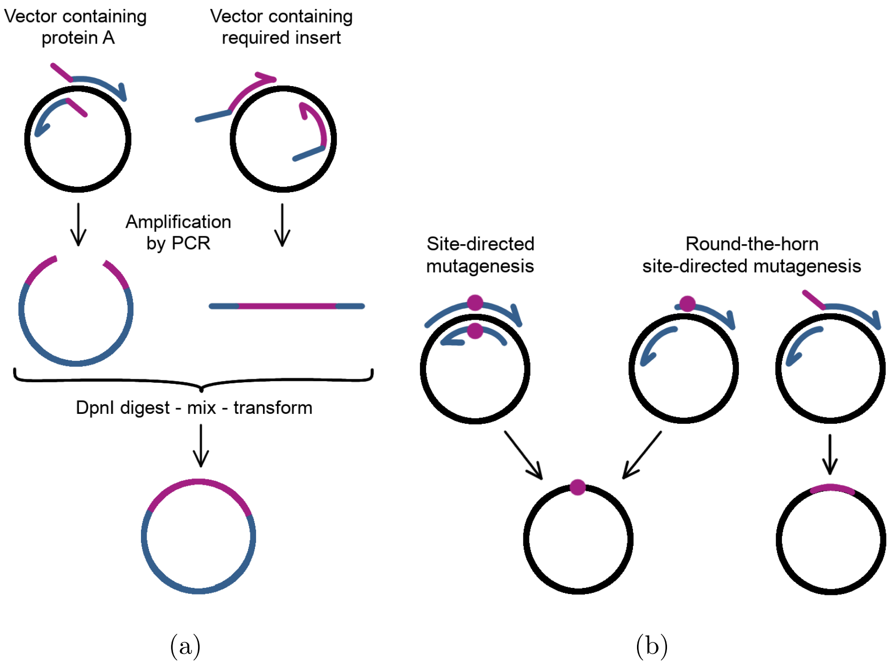

Vector containing protein A
Vector containing required insert
Amplification by PCR
DpnI digest - mix - transform
Site-directed mutagenesis
Round-the-horn site-directed mutagenesis
(a)
(b)

Figure 2.1: Molecular biology techniques. (a) Ligation-independent cloning using the FastClone method [225]. (b) Site-directed mutagenesis

method can only result in a linear amplification of the template and hence it is very inefficient in most cases. Instead, one can design non-overlapping primers to amplify the gene and vector in an exponential PCR [226, 227]. The mutation is only incorporated in one primer, usually at the 5'-end, and even longer insertions can be introduced this way (Figure 2.1b). For Round-the-Horn site-directed mutagenesis (RTH-SDM), 100 µM primers were phosphorylated using polynucleotide kinase (ThermoFischer) and 2-3 mM ATP according to the manufacturers protocol. The enzyme was heat-inactivated at 85°C for 10-15min. Phosphorylated primers were stored at -20°C until used. PCRs were performed using these primers and an appropriate DNA polymerase (e.g. Phusion High-Fidelity, NEB). PCR products were gel-purified and 50-100 µg of DNA material was added to 1 µl Anza$^{\mathrm{TM}}$T4 DNA Ligase Master Mix (ThermoFischer) in a total volume of 4 µl, incubated for 10-20 min at room temperature and transformed into in-house produced, chemically-competent DH5α *E. coli* cells. As before successful transformants were selected and propagated for plasmid purification.

2.2.2 Genetic constructs

PR65

PR65 was previously expressed in a pET28a vector and purified using an N-terminal H_6-tag. This plasmid was a kind gift from D. Barford. However, to increase protein yields and to facilitate purification, a C-terminally H_6-tagged PR65 gene was produced by RTH-SDM and transferred to a pGST vector (pRSETa backbone, N-terminal GST-tag cleavable by thrombin) resulting in a GST-PR65-H6 fusion gene. The construct was verified by Sanger sequencing using T7-Forward, pGEX5', T7-terminator and an internal sequencing primer.

CTPR constructs

DNA constructs of CTPR_RV proteins in a pRSET backbone were build sequentially from from single/double repeat modules using BamHI/BglII cloning as previously described by Kajander *et al.* [116]. A single repeat is preceded by a BamHI restriction site and followed by a BglII restriction site, double stop codon and HindIII restriction site. The sequences of constructs used here as well as sequences of previously published constructs are listed in Table 2.3. To create a construct containing N repeats, a vector containing M repeats is digested using BglII and HindIII and a vector containing a single repeat is amplified by PCR using Phusion DNA polymerase and T7-forward and -terminator primers at an annealing temperature of 56°C and elongation temperature of 72°C. Both the vector restriction digest and the PCR product are gel purified. The PCR product is subsequently digested using BamHI and HindIII and the enzymes are heat-inactivated at 80°C for 10

Table 2.3: CTPR consensus sequences, mutations and their applications. Since all CTPR constructs used here are based on the QK mutant, CTPR_QKa and CTPR_QKRVa are abbreviated to CTPRa and CTPR_RV, respectively. Abbreviations: ENM - elastic network model, SMFS - single molecule force spectroscopy.

Protein	**Sequence**	**Reference**	**Application**
CTPRn	AEAWYNLGNAYYEQGDY DEAIEYYQKALELDP**NN**	19	original consensus
CTPRa	AEAWYNLGNAYYEQGDY DEAIEYYQKALELDP**RS**	126	ENM, geometry
CTPR_QKn	AEAWYNLGNAYYEQGDY **QK**AIEYYQKALELDP**NN**	111	charge optimization
CTPR_QKa	AEAWYNLGNAYYEQGDY **QK**AIEYYQKALELDP**RS**	139	Ising, SMFS
CTPR_QKRVa	AEA**LN**NLGN**VYR**EQGDY **QK**AIEYYQKALELDP**RS**	139	Ising analysis, SMFS
CTPR_QKRVn	AEA**LN**NLGN**VYR**EQGDY **QK**AIEYYQKALELDP**NN**	139	ENM, geometry

min. Since BamHI and BglII produce the same 5'-overhangs the single repeat can be ligated into the vector using QuickStick ligase. The ligation product is transformed into *E. coli* and plasmid purified, hopefully resulting in a construct containing $M + 1$ repeats. The whole procedure is repeated until an N-repeat construct has been obtained.

2.2.3 Construction of pRSET_H$_6$_ybbR/pRSET_GST_ybbR

The pRSET/pGST vectors were modified into pRSET_H$_6$_ybbR/pRSET_GST_ybbR in sequential steps of RTH-SDM. First, a TEV cleavage site was introduced instead of the thrombin cleavage site, in such a manner that the BamHI restriction site remained unaffected. Second, a poly-N spacer containing 10 asparagines was inserted between the H$_6$- or the GST-tag. Third, the N-terminal ybbR-tag was inserted between the TEV cleavage site and the BamHI restriction site. Finally, a C-terminal ybbR-tag and stop codon were inserted directly downstream of a HindIII restriction site.

2.3 Protein expression and purification

Expression and purification protocols of each protein are detailed below. Unless otherwise specified, all constructs were transformed into chemically competent C41 *E. coli*. Suspension cultures were grown at 37°C shaking at 200 rpm until an OD$_{600}$ of 0.6 to 0.8 was reached. Protein expression was induced with isopropyl β-D-1-thiogalactopyranoside

(IPTG) for varying amounts of time and temperatures. Cells were harvested by centrifugation at 4000xg for 10 min at 4°C, before re-suspending in the appropriate lysis buffer supplemented with EDTA-free protease inhibitor cocktail (Sigma Aldrich), and, in some cases, DNase I (Sigma Aldrich) and Lysozyme (Sigma Aldrich). The cells were lysed by passing the suspension twice to three times through an Emulsiflex-C5 (AVESTIN) at pressures between 10000 and 15000 psi. Soluble protein was separated from cell debris and other insoluble fractions by centrifugation at 35000xg for 35 min at 4°C and each protein was purified by the respective affinity chromatography method detailed in the following sections. Buffers for the purification of each protein are listed in Table 2.4. If not used immediately, all proteins were flash-frozen in liquid N_2 and stored at -80°C.

2.3.1 PR65 WT, mutants and PR65-Impβ chimeras

All PR65 variants containing either an N-terminal H_6-tag or an N-terminal GST-tag were expressed in C41 *E. coli* as described above. Protein expression was induced using 250 µM IPTG at 25°C overnight. Cell pellets were resuspended in the lysis buffer appropriate for the respective affinity tag and lysed as usual.

Variants containing an N-terminal H_6-tag

The soluble protein fraction of N-terminally H_6-tagged variants was applied to 1 ml of Ni-NTA resin per litre of culture (Amintra Affinity Resins, Expedeon) or a 5 ml HisTrap Excel column (GE Healthcare) equilibrated in lysis buffer. Resin was incubated at 4°C for 1-2 hours with rotation followed by three 50 ml washes and three to four 10 ml elutions using the respective wash and elution buffers listed in Table 2.4. Column-bound protein was washed with 20 column volumes and eluted in 2 ml fractions using the respective buffers. After analysis of the elution fractions obtained from either method by SDS-PAGE, the proteins were further purified, first by anion exchange and second by size exclusion gel filtration chromatography. The Ni-NTA/HisTrap fractions were pooled, concentrated and diluted with wash buffer to reduce the salt content to < 50 mM before application to a Mono Q 10/100 GL (GE Healthcare). After washing the column, the protein was subsequently eluted using a 20 column volume salt gradient from 0 to 1 M NaCl. MonoQ fractions containing the protein were concentrated if necessary before application to a HiLoad 26/600 Superdex 200 pg (GE Healthcare) equilibrated in PBS (force-spectroscopy constructs) or MES (used for denaturations) storage buffers.

Variants containing an N-terminal GST-tag

The soluble protein fraction of N-terminally GST-tagged variants was applied to glutathion resin (Amintra Affinity Resins, Expedeon) equilibrated in lysis buffer. Resin was incubated at 4°C for 1-2 hours with rotation and washed using three times 50 ml of wash buffer, followed by on-matrix cleavage of the protein in low salt buffer using 50 units of either human or bovine thrombin (Sigma) per litre of culture at 4°C overnight. Cleaved protein was removed from the resin using three 10 ml washes of elution buffer. The ratio of the absorbance at 260 and 280 nm was measured using a Nano-spectrophotometer (ThermoFisher) and if DNA contamination was present, the elution fractions were combined and purified by anion exchange before gel filtration as described above. If DNA contamination was not present the elution fractions containing cleaved protein were concentrated and purified by gel filtration directly.

2.3.2 CTPR proteins

N-terminally H_6-tagged CTPR proteins were transformed in C41 *E. coli* and plated on LB Agar containing 100 µg ml^{-1} Ampicillin. All colonies were used to inoculate 0.5 l of 2xYT. Protein expression was induced using 0.5 mM IPTG over 3-5 hours at 37°C. Longer constructs were induced for less time than shorter CTPRs to minimize recombination. After lysis the cell suspension was heated to 70-80°C in a water bath before centrifugation. The soluble protein was then filtered through a 0.22 µm PES membrane and applied to a 5 ml HisTrap Excel column equilibrated in wash buffer. The column was washed using 20 column volumes of wash buffer before proteins were eluted in 2 ml fractions. All fractions containing protein were pooled, and if necessary, concentrated to <15 ml total volume using a Vivaspin® centrifugal concentrator. The protein was then further purified by size exclusion chromatography using a HiLoad 26/600 Superdex 75 pg (GE Healthcare) equilibrated in either Tris (used for force-spectroscopy) or sodium phosphate (used for equilibrium denaturation) buffer. The fractions were analysed using SDS-PAGE and those containing the least amount of recombined protein were pooled and concentrated. Constructs with 10 repeats or more exhibited significant recombination resulting in proteins that had a decreasing number of repeats. Due to the elution profile of CTPRs, it was not possible to remove such recombination products entirely, which meant that only the first few fractions of the elution peak could be pooled for concentration, while >60% of the fractions had to be discarded.

2.3.3 Sfp-synthase

The pCK plasmid containing Sfp-synthase-H_6 was a kind gift from the Gaub laboratory (Ludwig Maximilian Universität, München, Germany). For expression, the plasmid was transformed into chemically competent C41 *E. coli* and plated onto LB Agar containing Ampicillin. All colonies from the plate were used to inoculate 1 l of 2xYT. Protein expression was induced at $OD_{600} \sim 0.6$ by the addition of 1 mM IPTG and followed by

Table 2.4: Buffer systems used for purification of the individual proteins. PR65-Impβ chimeras were purified in the usual PR65 buffers. If different buffers were used for lysis and washing of the affinity matrix, lysis buffers contained in higher concentrations of salt, SigmaFast^{TM}EDTA-free protease inhibitor cocktail (PIC, Sigma Aldrich), and in some cases DnaseI (Sigma Aldrich) and Lysozyme (Sigma Aldrich). These variations are shown in parenthesis.

Protein	**Method**	**Wash (lysis)**	**Elution**
GST-PR65-H6	glutathione	50 mM Tris-HCl pH 7.5, 150 mM (500 mM) NaCl, 1 mM DTT (+ PIC)	50 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM DTT
H6-PR65/PR65-H6	Ni-NTA or HisTrap Excel	50 mM Tris-HCl pH 7.5, 500 mM NaCl, 20 mM imidazole, 1 mM DTT (+ PIC)	100 mM Tris-HCl pH 8.0, 2 mM EDTA, 300 mM imidazole, 1 mM DTT
	MonoQ	100 mM Tris-HCl pH 8.0, 2 mM EDTA, 0.5 g/l EGTA, 1 mM DTT	100 mM Tris-HCl pH 8.0, 2 mM EDTA, 0.5g/l EGTA, 1 mM DTT, 1M NaCl
	size exclusion and storage		PBS pH 7.4, 2 mM DTT or MES pH 6.5, 2 mM DTT
GST-CaM-H6	glutathione	PBS pH 7.4 (+ PIC)	PBS pH 7.4
CaM-H6	Ni-NTA	PBS pH 7.4, 10 mM imidazole	PBS pH 7.4, 250 mM imidazole
Sfp synthase	HisTrap Excel	20 mM Tris-HCl pH 7.5, 500 mM NaCl, 5 mM imidazole (+ PIC, DnaseI)	20 mM Tris-HCl pH 8.0, 300 mM imidazole, 300 mM NaCl, 2 mM EDTA
	storage		10 mM Tris-HCl pH 7.5 or PBS pH 7.4, 1 mM EDTA, 10% (v/v) glycerol
CTPR	HisTrap Excel	50 mM Tris-HCl pH 7.5, 500 mM NaCl, 20 mM imidazole (+ PIC, DnaseI, Lysozyme)	50 mM Tris-HCl pH 7.5, 150 mM NaCl, 300 mM imidazole
	size exclusion and storage		50 mM Tris-HCl pH 7.5 or 50 mM sodium phosphate pH 6.8, 150 mM NaCl
TEV protease	HisTrap Excel	50 mM Tris-HCl pH 8.0, 300 mM NaCl, 20 mM imidazole, 1 mM DTT (+ PIC, DnaseI, Lysozyme)	50 mM Tris-HCl pH 8.0, 300 mM NaCl, 300 mM imidazole, 1 mM DTT
	storage		50 mM Tris-HCl pH 8.0, 150 mM NaCl 50% (v/v) glycerol, 0.01% (v/v) 1-thioglycerol

further incubation at 25°C overnight. After lysis, the soluble protein fraction was filtered through a 0.22 µm PES membrane and applied to a 5 ml HisTrap Excel column (GE Healthcare) equilibrated in wash buffer. After washing with 20 column volumes of wash buffer, Sfp was eluted in one step. The elution fractions were analysed by SDS-PAGE and those containing protein at >90% purity were pooled. This solution was then split in half and dialysed twice against either PBS- or Tris-based storage buffers. After concentration using a Vivaspin®centrifugal concentrator before freezing.

2.3.4 TEV protease

C41 glycerol stocks containing H_6-TEV_{S219V} are stored in liquid N_2. A sample of this stock was used to inoculate 2x 10 ml of 2xYT containing 100 µg ml^{-1} Ampicillin, which was then grown over night at 37°C. 5 ml of starter culture were used to inoculate 0.5 l 2xYT supplemented with 100 µg ml^{-1} Ampicillin. Protein expression was induced with 0.2 mM IPTG at 20°C over night. The cells were harvested by centrifugation for 10 min at 7000xg and resuspended in lysis buffer. After lysis, the soluble protein fraction was filtered using a 0.22 µm membrane and applied to a 5 ml HisTrap Excel column (GE Healthcare) equilibrated with wash buffer (Table 2.4). The column was washed using 20 column volumes of wash buffer, before the protein was eluted in 1 ml fractions into a 96-well block already containing 1 ml wash buffer to immediately dilute the imidazole. Fractions containing the protein were pooled and concentrated to approximately 250 µM before 100% glycerol was added to produce a TEV protease stock containing a final concentration of 50% glycerol.

2.3.5 Determining protein concentration and purity, and verification of molecular weight

A rough estimate of purity was provided by SDS-PAGE. Molecular weights (Table 2.5) were verified by MALDI mass spectrometry (PNAC Facility, Department of Biochemistry), or ESI mass spectrometry. Protein concentrations were measured by absorbance at 280 nm using a nano-spectrophotometer. Absorbances were converted to molar concentration using the Beer-Lambert law and the appropriate theoretical extinction coefficient obtained from the primary amino acid sequence using the Expasy ProtParam tool (Table 2.5) [228].

Table 2.5: Molecular weights and theoretical extinction coefficient at 280 nm.

Protein	Molecular weight [Da]	ϵ [M^{-1} cm^{-1}]
PR65	66131/66276[a]	42400
PR65-ImpβH19	71429/71353[b]	47900
PR65-ImpβH18-19	76275/76199[b]	47900
CTPR_RV2	9824	11920
CTPR_RV4	17708	23840
CTPR_RV8	33478	47680
yCTPR_RV3y	16117	17880
yCTPR_RV5y	24001	29800
yCTPR_RV10y	43713	59600
yCTPR_RV20y	83136	119200
yCTPRa5y	26214	73690
yCTPRa9y	42383	131450
TEV	~27	33460
Sfp-synthase	~26	29130

[a] N-terminal H_6-tag/thrombin cleaved GST-PR65-H_6

[b] PR65 WT/interface mutations

2.4 Force-spectroscopy

2.4.1 Protein-DNA conjugation and purification

Conjugation of DNA oligonucleotides (Table 2.6) to proteins and their purification is the subject of Chapter 5 and described in detail therein.

Table 2.6: Functionalised DNA oligonucleotides (5' to 3').

Purpose	Sequence	Modification (*)
DNA-protein	GGCAGGGCTGACGTTCAACCAGACCAGCGAGTCG*	variable
DNA-handles	*GGCGA*CTGG*CGTTGATTTG	biotin/digoxigenin
	CGACTCGCTGGTCTGGTTGAACGTCAGCCCTGCC	abasic site
	*CCTGCCCGGCTCTGGACAGG	

2.4.2 Production of DNA handles for force-spectroscopy

Functionalised DNA handles of approximately 600 bp length were amplified from λ-DNA (e.g. Jena Bioscience) using the a triple biotinylated primer, a triple digoxigenin modified primer and a primer with a stable abasic site (Table 2.6, Metabion). A standard PCR was performed using *Taq* DNA polymerase in ThermoPol buffer (NEB) at an annealing temperature of 60°C and elongation temperature of 68°C. This PCR produces 5'-overhangs complementary to the oligo used for protein-DNA attachments. The final reaction was cleaned according to QIAquick PCR Purification Protocol with home made reagents and redissolved in water. The concentration was measured at 260 nm wavelength using a nano-photospectrometer (ThermoFischer).

2.4.3 Functionalised silica beads

Carboxyl-functionalised 1 µm silica beads (Bangs Laboratories) were modified in-house with anti-digoxigenin and tetramethylrhodamine-BSA (Sigma) [186]. Streptavidin coated beads were either produced in a similar manner or purchased (Bangs Laboratories). Anti-digoxigenin and streptavidin bead stocks were vortexed rigorously and diluted 1:20 and 1:130 in the appropriate buffer, respectively.

2.4.4 Sample preparation

In general, 4-10 µl of purified DNA-protein construct were incubated with 100-200 ng of functionalised DNA handles for 0.5-1 hr at room temperature. Of that mixture, 0.5-3 µl were then incubated with 1 µl diluted anti-digoxigening beads in 10 µl sample buffer for no longer than 5 min. Finally, 0.5-0.7 µl of this mixture were added to 50 µl buffer containing 0.5-0.6 µl streptavidin beads, an oxygen scavenger system consisting of 0.65% (w/v) glucose (Sigma), 13 U ml^{-1} glucose oxidase (Sigma) and 8500 U ml^{-1} catalase (Calbiochem), and, if appropriate, 1-2 mM DTT or 1-5 mM TCEP.

2.4.5 Chamber preparation

Two parafilm strips were fixed to a microscope slide and covered by a cover slip (Figure 2.2). The combination is then heated to 80°C to melt the parafilm and thereby seal the chamber sides. The chamber is first blocked with 10 mg/ml BSA (Sigma) for 5 min and washed twice using the appropriate buffer before the sample is introduced. The edges are sealed with vacuum grease directly afterwards.

2.4.6 Setup

Experiments were conducted on a custom-built, dual-beam set up with back-focal plane detection [229]. All measurements presented in this work were performed in constant velocity mode, in which the mobile trap is moved away from the fixed trap at constant velocities ranging from 10 nm s^{-1} to 50 µm s^{-1} to obtain force-extension traces. Trap

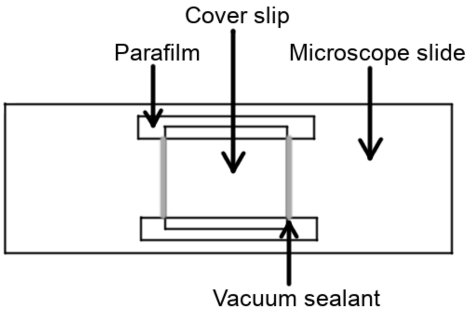


Figure 2.2: Sample chamber used for optical tweezers experiments

stiffness ranged from 0.24 pN (commercial streptavidin beads) to 0.35 pN (in-house functionalised beads). Data were acquired at sampling rates of 10-30 kHz.

2.4.7 Data analysis

Force-extension data was analysed using the Igor software (WaveMetrics). Traces were fitted with worm-like chain (WLC) polymer models which describe the extension of DNA and a protein amino acid chain under force [149, 164]. The DNA force response can be described by the Modified Marko-Siggia WLC model [164]:

$$F_{eWLC} = \frac{k_B T}{p_{DNA}} \left(\frac{1}{4\left(1 - \frac{\xi}{L_{DNA}}\right)^2} - \frac{1}{4} + \frac{\xi}{L_{DNA}} - \frac{F_{eWLC}}{K} \right), \tag{2.1}$$

where ξ is the extension, k_B is the Boltzmann constant, T the temperature, p_{DNA} the persistence length of DNA, L_{DNA} the contour-length of the DNA and K its elastic stretch modulus. The protein force response can be modelled using the original Marko-Siggia WLC [149]:

$$F_{WLC} = \frac{k_B T}{p_{protein}} \left(\frac{1}{4\left(1 - \frac{\xi}{L_{protein}}\right)^2} - \frac{1}{4} + \frac{\xi}{L_{protein}} \right), \tag{2.2}$$

where $p_{protein} = 0.7$ is the persistence length of the protein and $L_{protein}$ is the contour-length of the protein. The final extension of the protein-DNA construct is an addition of the stretching of both protein and DNA:

$$\xi_{construct} = \xi_{eWLC}(F) + \xi_{WLC}(F). \tag{2.3}$$

Here, DNA contour and persistence lengths, as well as protein contour lengths were processed for further analysis. Unless otherwise stated, a data set of DNA or protein contour lengths was represented as a histogram, the bin width of which was automatically chosen using the Freedman-Diaconis rule. The mean contour-length was determined by fitting a Gaussian distribution to the data:

$$P(L) = ae^{\frac{1}{2}\left(\frac{L-\mu}{\sigma}\right)^2}, \tag{2.4}$$

where a is the scaling factor, μ the mean and σ its standard deviation. If two peaks were present, a sum of two Gaussians was fitted to the histogram. Experimental contour-lengths were compared to the expected contour lengths which were estimated assuming 0.36 nm per amino acid.

2.5 Software

2.5.1 Molecular graphics

Pymol 1.7.2.1 and 1.8.4.0 were used to produce graphical representations of protein structures [230]. VMD 1.9.1 together with the NMWIZ plugin was used to produce graphical representations of normal mode vectors [231, 232].

2.5.2 Python

Python 2.7.12 in combination with the libraries NumPy 1.14.2 and SciPy 1.0.1 was used for all data analysis, calculations and modelling, except for the analysis and transformation of force-extension data, and elastic network models [233–236]. Any graphical outputs were generated using Matplotlib 1.5.3 [237].

2.5.3 ΔΔPT

ΔΔPT [238], which is freely available from *sourceforge.net*, was used to generate elastic network models and to perform any calculations described in Sections 3.2.2 and 3.2.3.

2.5.4 IgorPro

IgorPro 7.0 (WaveMetrics) under a multi-user license obtained by the Rief group was used to perform primary analyses on force-extension data. The procedures used to load force-extension data, fit WLCs to force-extension curves, perform contour length transformations, calculate standard deviations, and to extract non-equilibrium energies were written by former members of the Rief group.

2.6 Error propagation

Error propagation was performed as described in Hughes and Hase [239] using the following equations or a combination thereof:

- Sum of two parameters

$$Z = A \pm B \implies \alpha_Z = \sqrt{(\alpha_A)^2 + (\alpha_B)^2} \tag{2.5}$$

- Product of two parameters

$$\left.\begin{aligned} Z &= AB \\ Z &= \frac{A}{B} \end{aligned}\right\} \implies \alpha_Z = Z\sqrt{(\frac{\alpha_A}{A})^2 + (\frac{\alpha_B}{B})^2} \tag{2.6}$$

- Product of a parameter and a constant

$$Z = kA \implies \alpha_Z = |k|\alpha_A \tag{2.7}$$

- Parameter to the power of a constant

$$Z = A^k \implies \alpha_Z = \left|k\frac{Z}{A}\right|\alpha_A \tag{2.8}$$

- Inverse of a parameter

$$Z = \frac{1}{A} \implies \alpha_Z = Z^2\alpha_A \tag{2.9}$$

2.7 Statistics

To test whether there is a significant different difference between data sets, the appropriate test was selected as shown in Figures 2.3 and 2.4. In each case the Null-hypothesis was rejected if the test statistic produced values of $p < 0.05$ and accepted if it produced results with $p > 0.05$. If one (or more) out of a given number of data sets to be compared was not normally distributed, non-parametric tests were used.

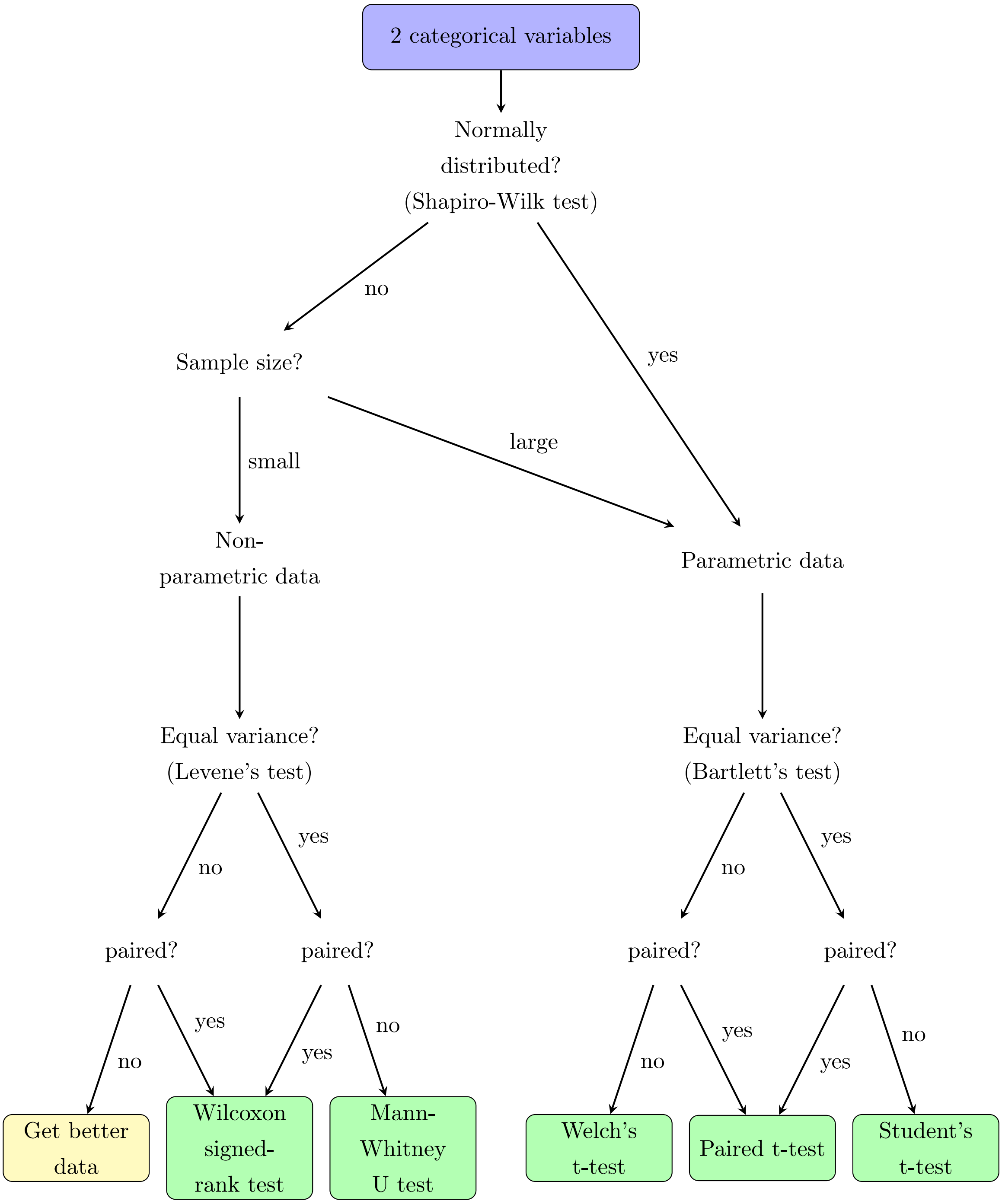


Figure 2.3: Determining whether there is a significant difference between 2 categorical variables.

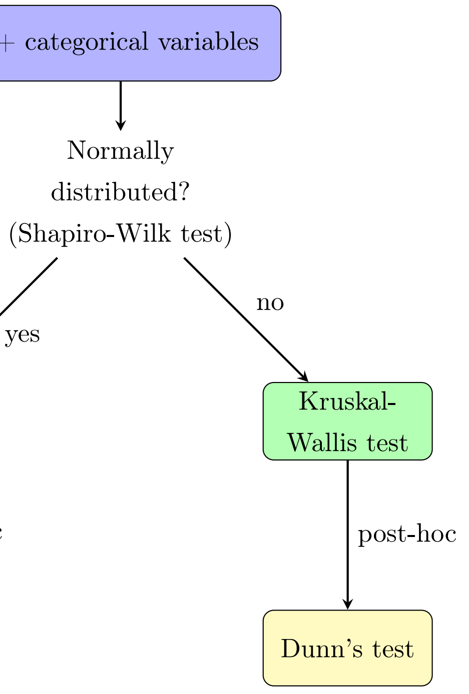


Figure 2.4: Determining whether there is a significant difference between more than 2 categorical variables

2.8 Nomenclature of protein names

Abbreviations used for all proteins with and without various tags are listed in Table 2.7.

Table 2.7: Nomenclature of all proteins used in this study

Protein	Variations	Description
PR65	H_6-PR65	Wild-type, human amino acid sequence of PR65 with an N-terminal hexahistidine-tag.
	GS-PR65-H_6	Wild-type PR65 with an N-terminal GS after cleavage of the GST-tag, and a C-terminal hexahistidine-tag.
	yPR65y	Wild-type PR65 with N- and C-terminal ybbR-tags without spacing amino acids.
	yPR65-GSy	Wild-type PR65 with N- and C-terminal ybbR-tags, containing a seven residue GS-spacer between the final residue of the protein and the C-terminal ybbR tag.
	alkPR65alk	D5 and L588 substituted with alkyne derivatives of pyrrolysine.
	azPR65az	D5 and L588 substituted with azide derivatives of pyrrolysine.
	cycPR65cyc	D5 and L588 substituted with cyclopropene derivatives of pyrrolysine.
	yPR65-277az	PR65 with an N-terminal ybbR-tag and E277 substituted to an azide derivative of pyrrolysine.
	yPR65-514az	PR65 with an N-terminal ybbR-tag and Q514 substituted to an azide derivative of pyrrolysine.
Chimeras	c1WT	PR65 with repeat HEAT19 of Importin-β added C-terminally
	c1int	PR65 with its C-cap mutated to an interface and repeat HEAT19 of Importin-β added C-terminally
	c2WT	PR65 with repeats HEAT18-19 of Importin-β added C-terminally
	c2int	PR65 with its C-cap mutated to an interface and repeats HEAT18-19 of Importin-β added C-terminally
	c3WT	PR65 with repeats HEAT17-19 of Importin-β added C-terminally
	c3int	PR65 containing C-cap-to-interface mutations and repeats HEAT17-19 of Importin-β added C-terminally
	c4WT	PR65 with repeats HEAT16-19 of Importin-β added C-terminally
	c4int	PR65 with its C-cap mutated to an interface and repeats HEAT16-19 of Importin-β added C-terminally
CTPRaN	yCTPRa5y	CTPR_QKa (as defined in Table 2.3) with 5 repeats and N- and C-terminal ybbR-tags that are separated from the protein by residues GS and KL, respectively.
	yCTPRa9y	Same as yCTPRa5y, but with 9 repeats.
CTPR_RVN	$N = 2, 4, 8$	CTPR_QKRVa with N repeats, as defined in Table 2.3.
	yCTPR_RV5y	CTPR_RV5 with N- and C-terminal ybbR-tags
	yCTPR_RV10y	CTPR_RV10 with N- and C-terminal ybbR-tags
	yCTPR_RV20y	CTPR_RV20 with N- and C-terminal ybbR-tags

Chapter 3

Gaining insight to repeat protein function using Elastic Network Models

3.1 Introduction

Traditionally, proteins were thought of as more or less rigid bodies, but over the past few decades research has progressed significantly showing that the dynamic motion of a protein is important for its structure, function and even evolution [240–249]. Within a protein, motion can be distinguished on different levels: the movement of single atoms, the movement of small groups of atoms (e.g. side chains) and the movement of whole structures and domains. Due to the tight association of atoms into larger domains, their movement will be correlated [240]. These global motions are mostly dependent on inter-residue contacts and are thought to be quite insensitive to atomistic details [241]. However, dynamics can also be more localized, e.g. to side chains of a protein, and these are proposed to be equally important, as they could potentially couple sub-global motions to the global movement representative of a whole domain or protein [250]. Various techniques techniques have been used to probe protein dynamics. NMR probes the interaction of atomic dipoles with its local environment and when all of these are combined it can provide a model of the global dynamics of a protein on various time scales [251]. HDX can provide insight to dynamics on a much longer time time scale [252]. SAXS on the other hand can give coarse grained information on changes in shape and conformation of proteins in solution [253].

Considering the complexity of protein structure it is not surprising that their dynamics are similarly complicated. Yet, their motion can be decomposed into normal modes: those of higher frequency are associated with local motions of single atoms and side chains and global modes usually have lower frequencies [250, 254]. Very early on, in a study comparing open and closed conformations of different proteins, Tama and Sanejouand [255] showed that the observed structural changes are well represented by one or a small number of modes. There are now many more examples of structural changes that correlate strongly with modelled vibrational motions [1, 246, 247, 256–259].

The dynamics of large molecular systems like proteins can be described using normal mode analysis (NMA) or essential dynamics analysis (EDA) [260]. In NMA, the dynamics of a protein are decomposed into its normal modes, the lowest of which are thought to be functionally relevant. NMA is based on the assumption that small fluctuations about the energy minimum of a conformation can be approximated by a quadratic potential [260]. However, these assumptions usually break down at physiological temperatures (and outside a vacuum). EDA does not rely on harmonic approximations and indeed can capture anharmonic fluctuations under physiological conditions. Essentially, it is a principal component analysis (PCA) of simulation trajectories and it captures the collective degrees of freedom, or the principal components (PCs), of the observed dynamics [260]. In EDA, the largest PCs are used to describe the functional dynamics and these can even represent the slowest normal modes [259].

Both standard NMA and EDA are usually derived from molecular dynamics (MD) simulations. While MD simulations can capture detailed movements at atomic resolution, they are computationally expensive, and time scales are often too short to describe whole-protein motions [258]. Contrastingly, coarse-graining allows for sampling of larger time scales and conformational changes [255, 258, 261].

3.1.1 Elastic Network Models

One method of coarse graining is an elastic network model (ENM), where a protein is reduced to a network of harmonic springs with its α-carbon atoms being subject to the potential fluctuations of their nearest neighbours [262, extensively reviewed in 245, 263]. Another advantage of the ENM is that energy minimization is not necessary and therefore, the model is based on the actual crystal structure [245, 255]. The ENM can be used to perform NMA of a protein and it has been shown that ENMs predict global modes equally well as full atomistic simulations [258, 260, 264]. Furthermore, a study on MD, X-ray crystallography and NMR data sets of HIV-1 protease has shown that normal modes, while not exactly the same as the dynamics in these data sets, represent the essential motions well in both simulation and experiment [259]. Early work by Bahar *et al.* [265] could show that ENM fluctuations did not only correlate well with structural B-factors, but also with HDX data. More recently, using anisotropy terahertz microscopy and inelastic neutron scattering of protein crystals, Niessen *et al.* [248] found that the direction of the intramolecular vibrations significantly changed when a protein was bound to an inhibitor.

To date, basic ENMs have been developed further (a) to include more detail such as side-chain chemistries [266] or models of membranes [256], (b) to shed light onto transition pathways both in combination and without MD simulations [257, 258] and (c) to aid structural refinement for cryo-EM [267]. Due to their robustness and the fact that optimization of model parameters is not required in most cases [261, 268], ENMs have become very popular as they enable even a lay user to obtain quite accurate information on the dynamics of their particular system using an easy web-interface [232]. Given the simplicity of ENMs (with and without adaptations), it is astounding in how many instances it can identify and make predictions of dynamic motions and hinges of subdomains, allosteric mechanisms, transition states and drug-able sites [246, 247, 255–258]. Nevertheless, most studies on protein dynamics examine globular proteins and very little is known about the dynamics of repeat proteins and how they relate to their biological function.

3.1.2 The protein systems under investigation

Of the multitude of possible repeat protein systems I chose to investigate the 15-HEAT repeat protein PR65 and a family of 7-TPR proteins involved in bacterial quorum sensing. Multiple crystal structure of significantly different conformations are available for both examples, and conformational changes have been implied to be of functional (allosteric) relevance [11, 201].

PR65

The structure and function of PR65 as a PP2A subunit have been introduced in detail in Section 1.6. Here, I perform a quantitative analysis of the structural geometries and

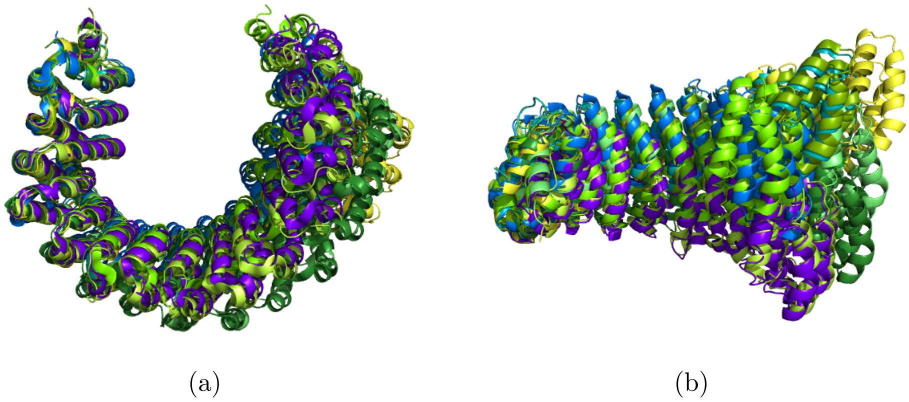


Figure 3.1: Top (a) and side view (b) of an N-terminal alignment of different PR65 structures (PDBids: 1b3u, 2iae, 2ie4, 2nym, 3dw8, 3k7w, 4i5l, 5w0w) using two N-terminal HEAT repeats [8, 16, 209, 217, 220, 269–271]. N-termini are to the left and C-termini are to the right of each image.

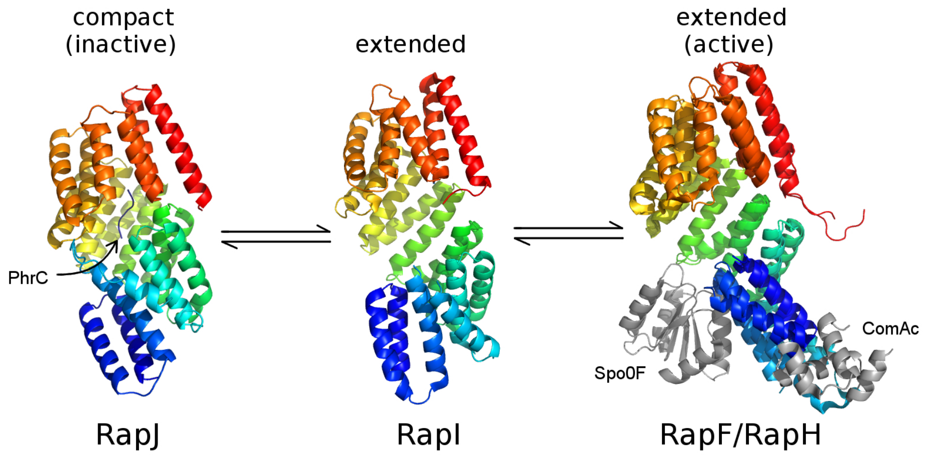


Figure 3.2: Structures of different Rap proteins (C-terminus in red) depicting a possible mode of action (PDBids, from left to right: 4gyo, 4i1a, 3q15, 3ulq [11, 278, 279]). When the TPR domain binds to a signalling peptide, it causes the Rap protein to adopt a compact conformation. Upon binding an interaction partner, however, conformational changes in the TPR domain are minimal, whereas the N-terminal three-helix bundle flips by approximately 180°.

vibrational dynamics of PR65 from different crystal structures, in which it was crystallized (a) on its own [16], (b) in complex with regulatory and catalytic subunits [8, 209, 217, 269], (c) in complex with the catalytic subunit bound to toxins [220, 270], or (d) bound to the catalytic subunit and an inhibitor protein [271].

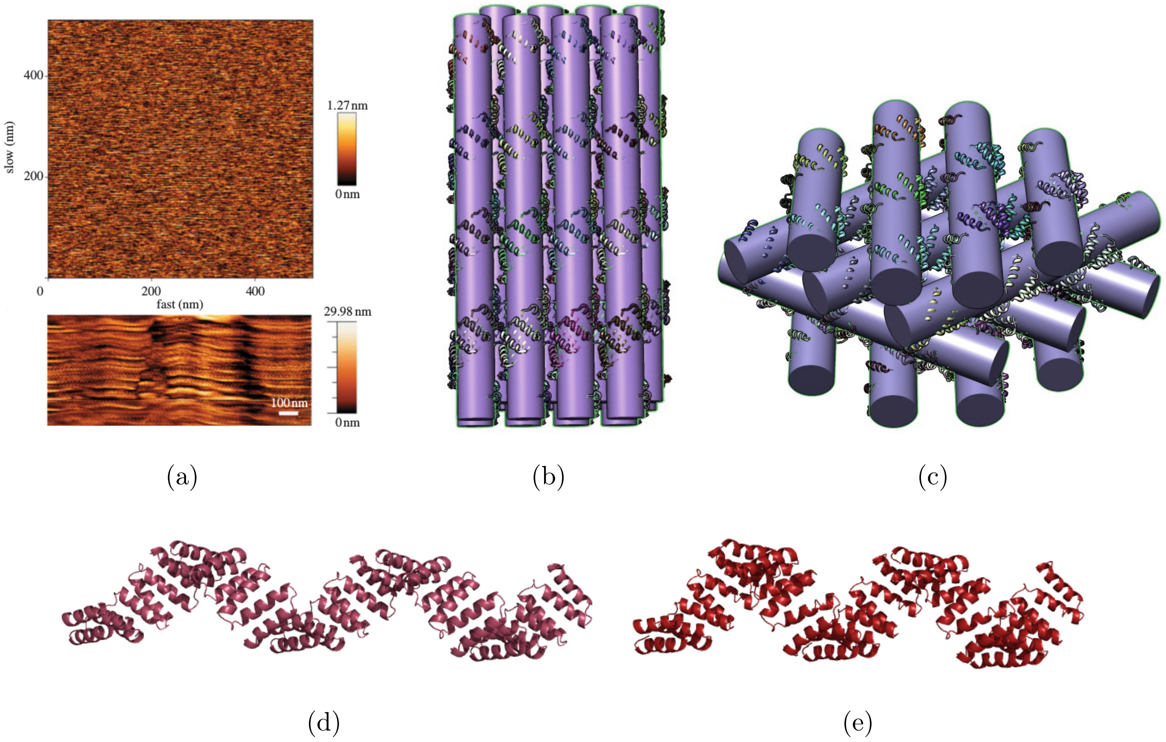
The Rap protein family

Rap phosphatases and their peptide activators were originally described in *B. subtilis* by Perego and co-workers [272–277]. They belong to the bacterial RRNPP family, members of which are inhibited by short quorum-sensing peptides [20]. For example, RapH acts as phosphatase of Spo0F and prevents sporulation, while RapF binds and inhibits gene regulators such as ComA [20]. Co-crystallization of RapH and RapF with Spo0F and ComAc, respectively, revealed the the same overall conformation when bound to their partner molecule (Figure 3.2) [278]. In another study, the crystal structure of RapI was compared with that of RapJ in complex with the PhrC peptide. The solenoid structure of the RapJ-PhrC complex showed a higher degree of compaction relative to the RapI (Figure 3.2) [11].

All four Rap proteins exhibit the same domain organization: an N-terminal three-helix bundle, a flexible helical linker and a C-terminal TPR domain. Notably, the N-terminal domain and the helical linker can form a four-helix bundle that resembles a pair of TPRs. Peptide binding to the C-terminal TPR domain causes a conformational change that propagates to the N-terminal domain [11]. As they lacked a complete set of crystal structures of the same Rap homologue in three conformational states, Parashar and coworkers used homologous structures to propose a mechanism of action for signal transduction and concluded that quorum-sensing peptides inhibit Rap function via an allosteric mechanism.

TPRs of different shape

To complement the studies on natural repeat proteins, I extended my analyses to CTPR proteins, which have been observed to exhibit ordered end-to-end packing, both *in crystallo* and on surfaces (Fig. 3.3) [126, 280]. As part of a project involving the re-shaping of CTPRs, a former member of our group, Albert Perez-Riba, has designed TPR proteins whose interfaces differ in packing and hydrophobicity [139]. One variant,CTPR_RV, was thermodynamically less stable than the consensus, and exhibited a different super-helical geometry and crystal packing (Figure 3.3).



(a) (b) (c)

(d) (e)

Figure 3.3: TPR proteins arrange in a systematic fashion and can form supramolecular structures. (a) A film of CTPR18 as visualized by AFM (adapted from [280]). (b,e) CTPR8/20 and (c,f) CTPR_RV4 arrange into helices *in crystallo* with different packing mechanisms [139].

3.2 Methods

3.2.1 Principal Component Analysis

Principal component analysis (PCA) is widely used to determine the axes of variation across a data set. The first principal component (PC1) of a given data set is the axis that exhibits the largest variance, the second principal component (PC2) is the axis with the second largest variance, etc. The PCs are calculated by orthogonal distance regression, where the sum of squared orthogonal distances from data points to the regression line or plane are minimized. PCA can therefore be used to fit a plane to a 3-dimensional data set. First, the error matrix, X, is determined for N data points using

$$\mathbf{X} = \begin{pmatrix} x_1 - \bar{x} & y_1 - \bar{y} & z_1 - \bar{z} \\ x_2 - \bar{x} & y_2 - \bar{y} & z_2 - \bar{z} \\ \vdots & \vdots & \vdots \\ x_N - \bar{x} & y_N - \bar{y} & z_N - \bar{z} \end{pmatrix}, \tag{3.1}$$

where $\bar{x}$, $\bar{y}$ and $\bar{z}$ are the averages of x, y and z respectively. Second, the covariance matrix is calculated:

$$\mathbf{C}(x,y,z) = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix} = \frac{1}{N-1}\mathbf{X^T X}, \tag{3.2}$$

where for example

$$cov(x,y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N-1}. \tag{3.3}$$

Lastly, the eigenvalues and eigenvectors of the covariance matrix are computed and sorted: the eigenvector associated with the largest eigenvalue is PC1, the eigenvector with the second-largest eigenvalue is PC2, and the eigenvector with smallest eigenvalue is PC3.

Using PCA to calculate angles between repeats

The geometry of any repeat protein can be described by relative angles between repeat planes, which are: curvature, twist and lateral bending. Methods for calculations of angles were adapted from Forwood *et al.* [6] (Fig. 3.4). In brief: a PCA is performed on the C_α-atom coordinates of each repeat to determine PC1, PC2 and PC3. Then, curvature between repeat N and $N+1$ is the angle between respective PC2s projected onto the plane formed by PC2 and PC3 of repeat $N+1$; twist is the angle between respective PC1s projected onto the plane formed by PC1 and PC2 of repeat $N+1$; and lateral bending is the angle between PC3s projected onto the plane formed by PC1 and PC3 of repeat $N+1$. To ensure correct assignment of angles (positive or negative), the conventions

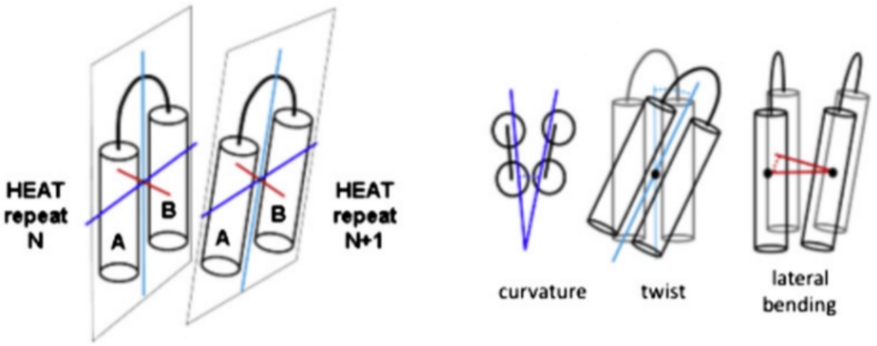


Figure 3.4: Calculation of repeat protein angles using PCA: light blue - PC1, blue - PC2, red - PC3. PC3 is also the normal vector of the repeat plane. Adapted from [6].

were introduced that PC1 has the same orientation as the super-helical axis, and PC3 has the same orientation as a vector pointing from the centroid of repeat N to the centroid of repeat $N+1$. The super-helical axis is defined by the right-hand-rule after having determined the handedness of the super-helix using the N- to C-terminal direction of the polypeptide chain [15]. The correct orientation of PC2 then arises from the cross-product of PC3 and PC1. All calculations were performed in Python [233, 234] using the extension modules NumPy [235] and Matplotlib [237].

PCA on a conformational ensemble

The coordinates from M number of structures with N number of C_α atoms each are placed into a $M \times N$ coordinate matrix [259, 281]:

$$\mathbf{X} = \begin{pmatrix} x_{11} & y_{11} & z_{11} & x_{12} & y_{12} & z_{12} & \cdots & x_{1N} & y_{1N} & z_{1N} \\ x_{21} & y_{21} & z_{21} & x_{22} & y_{22} & z_{22} & \cdots & x_{2N} & y_{2N} & z_{2N} \\ \vdots & & & & & & \ddots & & & \vdots \\ x_{M1} & y_{M1} & z_{M1} & x_{M2} & y_{M2} & z_{M2} & \cdots & x_{MN} & y_{MN} & z_{MN} \end{pmatrix}. \tag{3.4}$$

Using $\mathbf{X}$ one can obtain the elements for the error matrix, $\mathbf{Y}$,

$$y_{m,n} = x_{m,n} - \frac{1}{M}\sum_{m=1}^{M} x_{m,n}, \tag{3.5}$$

which can be used to calculated the $3N \times 3N$ covariance matrix for the ensemble

$$\mathbf{C} = \frac{1}{M-1}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}. \tag{3.6}$$

The covariance matrix is decomposed into a matrix containing the principal components (PCs), $\mathbf{P}$, and a eigenvalue matrix, Λ,

$$\mathbf{C} = \mathbf{P}\Lambda\mathbf{P}^{\mathrm{T}}. \tag{3.7}$$

The eigenvectors are sorted according to their eigenvalue in descending order. Since $M < 3N$, only $M-1$ non-zero eigenvalues with corresponding eigenvectors are obtained [260].

3.2.2 Elastic Network Models

The Elastic Network Models (ENMs) were generated using the open-access ΔΔPT toolbox described in detail by Rodgers *et al.* [238]. Some PDB structures had atoms with multiple possible positions between different crystal unit cells, e.g. due to different side chain rotamers. This is indicated by the occupancy, n. If $n \neq 1$, the atom position with the highest occupancy was kept and set to $n = 1$, while all others were deleted. If the occupancies for both positions equal to 0.50, the first position was kept, while the second one was deleted. Atoms with occupancies with $n \neq 1$ but only one given coordinate were kept and set to $n = 1$. Next, PDB structures were reduced to C_α-atoms only and springs set between atoms i and j with an equilibrium distance, R, separated by a distance, r, and within a cut-off radius, R_c. The corresponding potentials are

$$V_{ij} = \begin{cases} \frac{k_{ij}}{2}(r_{ij} - R_{ij})^2 & R_{ij}^2 \leq R_c^2 \\ 0 & R_{ij}^2 > R_c^2 \end{cases}, \tag{3.8}$$

where k_{ij} is the spring constant of the potential. Unless otherwise stated, spring constants were set $k = 1$ kcal mol^{-1}Å^{-2} for all atom pairs, and the cut-off radius was set to $R_c = 12$. The calculated potential is used to construct a mass-weighted Hessian matrix, $\mathbf{D}$, with elements

$$\mathbf{D}_{i\alpha,j\beta} = \left. \frac{\partial^2 V}{\partial r_{i\alpha}\sqrt{m_i}\partial r_{j\beta}\sqrt{m_j}} \right|_R, \tag{3.9}$$

where m is the mass of the respective atom and α and β refer to the direction of motion. $\mathbf{D}$ is then diagonalized

$$\boldsymbol{e}^{-1}\mathbf{D}\boldsymbol{e} = \begin{pmatrix} \omega_1^2 & 0 & \cdots & 0 \\ 0 & \omega_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_n^2 \end{pmatrix} \tag{3.10}$$

and the eigenvectors of this matrix, $\boldsymbol{e}$, are the normal modes, ν, while the eigenvalues are the squares of the associated frequencies, ω_ν.

The frequencies from the eigenvalues can be used to calculate the free energy and entropy of each mode using

$$G_\nu = -k_B T \ln \left(\frac{1}{1 - \exp(-\frac{\hbar\omega_\nu}{k_B T})} \right) \tag{3.11}$$

and

$$S_\nu = k_B \left(\frac{\frac{\hbar\omega_\nu}{k_B T}}{\exp(\frac{\hbar\omega_\nu}{k_B T}) - 1} - \ln\left(1 - \exp\left(-\frac{\hbar\omega_\nu}{k_B T}\right)\right)\right), \tag{3.12}$$

where k_B is the Boltzmann constant and T the temperature [238].

Using the root mean square deviations (RMSDs) of the lowest 25 vibrational modes (ignoring the lowest 6 modes that account for rotation and translation) of an atom i, one can obtain a similar quantity to the crystallographic B-factor using

$$B_i = \frac{8 k_B T \pi^2}{3 m_i} \sum_\nu \frac{|\boldsymbol{e}_i^2(\nu)|}{\omega_\nu^2}. \tag{3.13}$$

Here, $k_B T$ only functions as a scaling factor, where $T = 298K$. B_i gives a measure of the atom fluctuation about its equilibrium position, but since damping due to solvent or water does not exist, these fluctuations are very large. Therefore, they are scaled using average experimental and ENM B-factors of the whole structure and their respective root mean square displacements. The Pearson's Correlation Coefficient, r, is used as a measure of the goodness of fit of the model to the experimental data, and are usually found to be > 0.5 [238, 282].

3.2.3 Qualitative comparison of protein dynamics

The ΔΔPT toolbox can also be used to calculate the collectivity and cross-correlation of atoms within one protein [238]. Together they permit a qualitative description atom motion, either with respect to each other or with respect to a normal mode.

The collectivity, κ, of a given mode, can be described using [249]:

$$\kappa_\nu = \frac{1}{N} \exp\left(-\sum_i^N \alpha |\boldsymbol{e}_i^2(\nu)| log(\alpha |\boldsymbol{e}_i^2(\nu)|)\right), \tag{3.14}$$

where N is the total number of atoms and α is the collectivity constant defined by $\sum_i^N \alpha |\boldsymbol{e}_i^2(\nu)| = 1$. Using κ_ν, it is possible to determine the fraction of atoms most affected by a mode ν. The lowest frequency modes tend to have $\kappa_\nu > 0.4$ [238].

The cross-correlation, C_{ij}, of atoms i and j over the lowest modes indicates how much they move into the same direction, and can be calculated using [283]:

$$C_{ij} = \sum_\nu \left(\frac{\boldsymbol{e}_i(\nu) \cdot \boldsymbol{e}_j(\nu)}{(|\boldsymbol{e}_i(\nu)|^2 \; |\boldsymbol{e}_j(\nu)|^2)^{0.5}} \right). \tag{3.15}$$

For perfect correlation or anti-correlation $C_{ij} = 1$ or $C_{ij} = -1$, respectively. A value in between can arise from motion that is less correlated in terms of phase and/or period, or motions of atoms that are not (anti)parallel. If $C_{ij} = 0$, the atoms move with the same period and phase but their motions are orthogonal.

3.2.4 Quantitative comparison between ENM dynamics and conformational changes

The ENM eigenvectors of a given structure can be compared to (i) the eigenvectors of individual structures, (ii) the conformational changes between structures, $\Delta\boldsymbol{R}$, and (iii) conformational variation seen across X-ray, NMR or MD simulation ensembles. To calculate conformational changes, the backbones structures were aligned and an RMSD was calculated using PyMol [230]. A displacement vector, $\Delta\boldsymbol{R}_{ij}$, was obtained by vector subtraction $\Delta\boldsymbol{R}_{ij} = \boldsymbol{r}_j - \boldsymbol{r}_i$ [255], where $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$ are the crystal coordinates of two different structures.

Comparisons between data sets (that is between (i) two eigenvectors $\boldsymbol{e}_i(\nu)$ and $\boldsymbol{e}_j(\nu)$, (ii) an eigenvector, $\boldsymbol{e}_i(\nu)$ or $\boldsymbol{e}_j(\nu)$, and a conformational change, $\Delta\boldsymbol{R}_{ij}$, or (iii) an eigenvector, $\boldsymbol{e}_i(\nu)$ or $\boldsymbol{e}_j(\nu)$, and a principal component) were made using the overlap, O, which is a measure of alignment between two vectors, as defined by:

$$O = \frac{|M \cdot N|}{||M||\ ||N||}, \tag{3.16}$$

where M and N are placeholders for any of the vectors described by $\boldsymbol{e}_i(\nu)$, $\boldsymbol{e}_j(\nu)$, $\Delta\boldsymbol{R}_{ij}$, or a PC [255, 259]. If the directions align either in an antiparallel or parallel fashion, the overlap is 1; if they are exactly orthogonal, the overlap is 0. The cumulative overlap, CO, as defined by Yang *et al.* [259] is

$$CO(k) = \left(\sum_{j=1}^{k} O^2\right)^{\frac{1}{2}} \tag{3.17}$$

which gives a measure of how well k normal modes overlap with a single vector, such as $\Delta\boldsymbol{R}_{ij}$.

3.2.5 ENMs of PR65 and PP2A

Structures of PR65 and PP2A complexes from different crystals were reduced to a common set of residues, to (a) account for missing residues in various structures, and (b) avoid abnormal eigenvectors of the loose N-terminal residues (see Table 3.1 for all residue boundaries). The structures were aligned to 1b3u chain A and RMSDs of all structures compared to 1b3u were calculated using the pymol *align* command with the *cutoff* set to 10 to include all residues [230]. The conformational change vectors, $\Delta\vec{r}$, from 1b3u to 2iae, 3dw8 and 4i5l coordinates were calculated as described in Section 3.2.4. Conformational variation across all 8 PR65 structures were represented using PCA (for detail, see Section 3.2.1).

ENMs with $R_c = 8$Å were produced for a subset of PR65 structures and PP2A complexes (1b3u, 2iae, 3dw8, 4i5l), which represent conformational changes seen over the

Table 3.1: The construct specific residues used to compute ENMs. The numbering is based on Uniprot sequences.

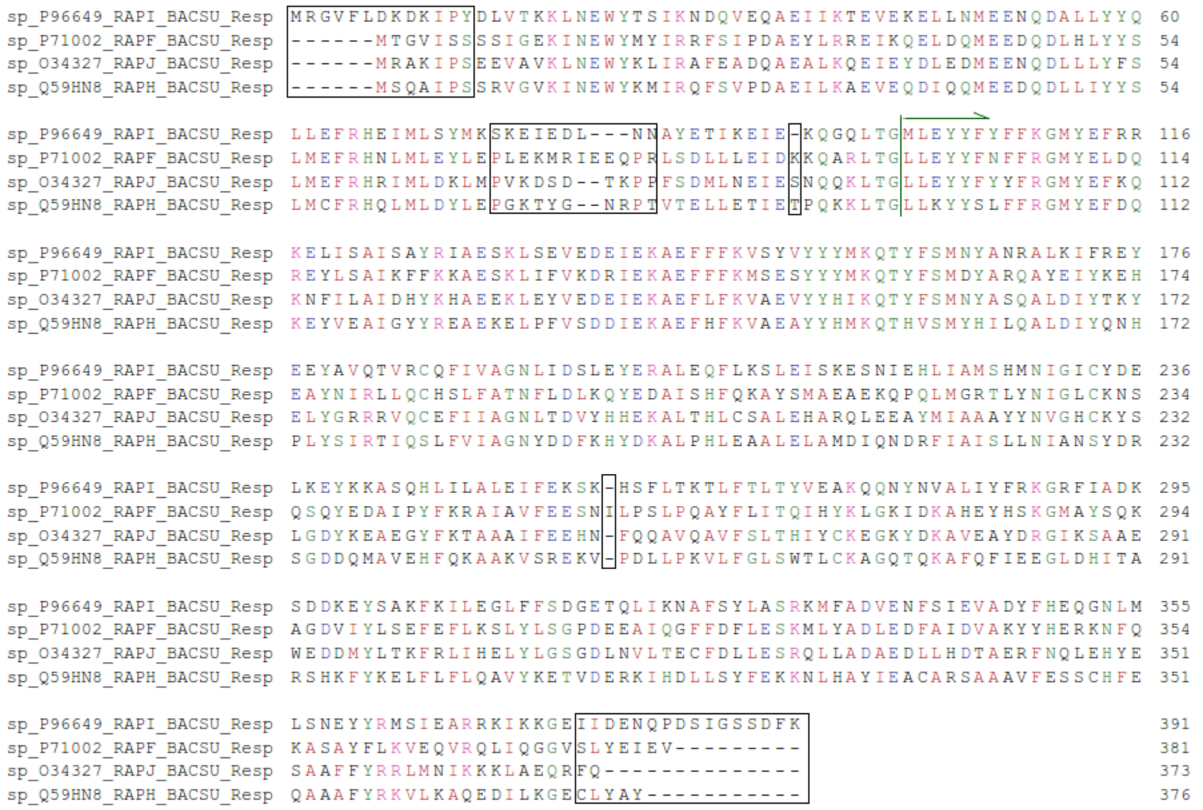
PDB ID	Reference	Complex	Chain	Residues	Purpose
1b3u	16	A	-	10-588	Structural analysis and ENMs
			-	1-588	Comparison to MD simulations
2iae	8	A/B'/C	D	11-589	Structural analysis and ENMs
			E	30-403	ENMs
			F	6-292	ENMs
2ie4	220	A/C + OA	A	11-589	Structural analysis
2nym	269	A/C/B'	A	11-589	Structural analysis
3dw8	209	A/B/C	A	11-589	Structural analysis and ENMs
			B	10-446	ENMs
			C	6-292	ENMs
3k7w	270	A/C + DTX1/2	A	11-589	Structural analysis
4i5l	217	A/B"/C	A	11-589	Structural analysis and ENMs
			B	141-478	ENMs
			C	6-292	ENMs
5w0w	271	A/C + TIPRL	A	11-589	Structural analysis

whole ensemble (for details, see Section 3.2.2). Normal modes were displayed using the NMWIZ plugin of VMD [231, 232].

Using the overlap (Equation 3.16), the normal modes of a given structure were compared to (i) conformational changes between the apo- and afore mentioned holoenzymes, (ii) PCs of the conformational ensemble (iii) PCs derived from four traces of 100 ns equilibrium MD simulations based on a PR65 mutant containing a split tetra-cysteine motif between H1 and H2 [284]. The MD data were kindly provided by Giovanni Settanni (Johannes Gutenberg Universität, Mainz).

3.2.6 ENMs of Rap proteins

If not specified otherwise, the binding partners and the peptide were removed to construct ENMs of RapH, RapF and RapJ, respectively. The structures had to be cut to their final size, which depended largely on which residues were shared among different structures of one protein. A particular challenge is imposed by downstream calculations such as overlap analyses, which require the ENMs to have the same number of atoms. The four Rap proteins share a large structural homology, particularly within the TPR domain [11]. Therefore, all four structures were aligned using PROMALS3D using both sequence and structural information (see Fig. 3.5) [285]. Missing residues that are conserved but not present in the structure were modelled using MODELLER [286]. Based on the



```
sp_P96649_RAPI_BACSU_Resp MRGVFLDKDKIPYDLVTKKLNEWYTSIKNDQVEQAEIIKTEVEKELLNMEENQDALLYYQ 60
sp_P71002_RAPF_BACSU_Resp ------MTGVISSSSIGEKINEWYMYIRRFSIPDAEYLRREIKQELDQMEEDQDLHLYYS 54
sp_O34327_RAPJ_BACSU_Resp ------MRAKIPSEEVAVKLNEWYKLIRAFEADQAEALKQEIEYDLEDMEENQDLLLYFS 54
sp_Q59HN8_RAPH_BACSU_Resp ------MSQAIPSSRVGVKINEWYKMIRQFSVPDAEILKAEVEQDIQQMEEDQDLLIYYS 54

sp_P96649_RAPI_BACSU_Resp LLEFRHEIMLSYMKSKEIEDL---NNAYETIKEIE-KQGQLTGMLEYYFYFFKGMYEFRR 116
sp_P71002_RAPF_BACSU_Resp LMEFRHNLMLEYLEPLEKMRIEEQPRLSDLLLEIDKKQARLTGLLEYYFNFFRGMYELDQ 114
sp_O34327_RAPJ_BACSU_Resp LMEFRHRIMLDKLMPVKDSD--TKPPFSDMLNEIESNQQKLTGLLEYYFYYFRGMYEFKQ 112
sp_Q59HN8_RAPH_BACSU_Resp LMCFRHQLMLDYLEPGKTYG--NRPTVTELLETIETPQKKLTGLLKYYSLFFRGMYEFDQ 112

sp_P96649_RAPI_BACSU_Resp KELISAISAYRIAESKLSEVEDEIEKAEFFFKVSYVYYYMKQTYFSMNYANRALKIFREY 176
sp_P71002_RAPF_BACSU_Resp REYLSAIKFFKKAESKLIFVKDRIEKAEFFFKMSESYYYMKQTYFSMDYARQAYEIYKEH 174
sp_O34327_RAPJ_BACSU_Resp KNFILAIDHYKHAEEKLEYVEDEIEKAEFLFKVAEVYYHIKQTYFSMNYASQALDIYTKY 172
sp_Q59HN8_RAPH_BACSU_Resp KEYVEAIGYYREAEKELPFVSDDIEKAEFHFKVAEAYYHMKQTHVSMYHILQALDIYQNH 172

sp_P96649_RAPI_BACSU_Resp EEYAVQTVRCQFIVAGNLIDSLEYERALEQFLKSLEISKESNIEHLIAMSHMNIGICYDE 236
sp_P71002_RAPF_BACSU_Resp EAYNIRLLQCHSLFATNFLDLKQYEDAISHFQKAYSMAEAEKQPQLMGRTLYNIGLCKNS 234
sp_O34327_RAPJ_BACSU_Resp ELYGRRRVQCEFIIAGNLTDVYHHEKALTHLCSALEHARQLEEAYMIAAAYYNVGHCKYS 232
sp_Q59HN8_RAPH_BACSU_Resp PLYSIRTIQSLFVIAGNYDDFKHYDKALPHLEAALELAMDIQNDRFIAISLLNIANSYDR 232

sp_P96649_RAPI_BACSU_Resp LKEYKKASQHLILALEIFEKSK-HSFLTKTLFTLTYVEAKQQNYNVALIYFRKGRFIADK 295
sp_P71002_RAPF_BACSU_Resp QSQYEDAIPYFKRAIAVFEESNILPSLPQAYFLITQIHYKLGKIDKAHEYHSKGMAYSQK 294
sp_O34327_RAPJ_BACSU_Resp LGDYKEAEGYFKTAAAIFEEHN-FQQAVQAVFSLTHIYCKEGKYDKAVEAYDRGIKSAAE 291
sp_Q59HN8_RAPH_BACSU_Resp SGDDQMAVEHFQKAAKVSREKV-PDLLPKVLFGLSWTLCKAGQTQKAFQFIEEGLDHITA 291

sp_P96649_RAPI_BACSU_Resp SDDKEYSAKFKILEGLFFSDGETQLIKNAFSYLASRKMFADVENFSIEVADYFHEQGNLM 355
sp_P71002_RAPF_BACSU_Resp AGDVIYLSEFEFLKSLYLSGPDEEAIQGFFDFLESKMLYADLEDFAIDVAKYYHERKNFQ 354
sp_O34327_RAPJ_BACSU_Resp WEDDMYLTKFRLIHELYLGSGDLNVLTECFDLLESRQLLADAEDLLHDTAERFNQLEHYE 351
sp_Q59HN8_RAPH_BACSU_Resp RSHKFYKELFLFLQAVYKETVDERKIHDLLSYFEKKNLHAYIEACARSAAAVFESSCHFE 351

sp_P96649_RAPI_BACSU_Resp LSNEYYRMSIEARRKIKKGEIIDENQPDSIGSSDFK                         391
sp_P71002_RAPF_BACSU_Resp KASAYFLKVEQVRQLIQGGVSLYEIEV---------                         381
sp_O34327_RAPJ_BACSU_Resp SAAFFYRRLMNIKKKLAEQRFQ--------------                         373
sp_Q59HN8_RAPH_BACSU_Resp QAAAFYRKVLKAQEDILKGECLYAY-----------                         376
```

Figure 3.5: PROMALS3D structure-based sequence alignment of all four Rap proteins. The black boxes indicate residues that were removed before construction of ENMs, while the green line and arrow indicate the start of the TPR repeats.

Table 3.2: The construct-specific residues used to compute ENMs. See Fig. 3.5 for the corresponding structural alignment. The numbering is based on Uniprot sequences.

Protein	Chain	PDB ID	Reference	Residues removed	TPR domain
RapI	B	4i1a	11	1-13, 75-82, 375-391	100-391
RapJ	B	4gyo	11	1-7, 69-78, 88	96-373
RapH	A	3q15	279	1-7, 69-78, 88, 372-376	96-376
RapF	A	3ulq	278	1-7, 69-80, 90, 257, 375-381	98-381

PROMALS3D alignment residues were removed that are flexible and cannot be modelled meaningfully using ENMs, not conserved and not represented in the crystal structure, or present in some structures but not others. The final boundaries are displayed in Figure 3.5 and listed Table 3.2. After removal residues, the structures were aligned again using PyMOL [230]. ENMs were also built for the TPR repeats only (Tab. 3.2).

Optimal values for k and R_c were explored in the intervals of $0.5 < k < 2.0$ and $5 < R_c < 15$, respectively, by maximising the correlation between crystallographic and predicted B-factors. The spring constant was found to not have a measurable effect and correlation values were maximal for $R_c > 11$Å for all proteins but RapJ. Correlations for the more extended conformations are consistently lower than 0.5, indicating that the experimental B-factors do not exhibit the large global flexibility seen in the network model, possibly due to crystallographic packing [245]. Therefore, final ENMs for all structures were built using default cut-off of $R_c = 12\text{Å}$ because it was an overall optimal value.

3.2.7 CTPR structural and ENM models

Structures of CTPRa and CTPR_RVn series were used for ENMs and geometric calculations (see Section 3.2.1, Table 2.3). CTPRa was crystallized both as 8- and 20-repeat proteins with very little difference in structure and crystal packing [126]. However, the asymmetric crystal unit contained only four repeats, and models of larger repeat number had to be re-constructed using crystal symmetry and unit cell translation [126]. Here, CTPRa4 represents the asymmetric unit and CTPRa20 the re-constructed 20mer. CTPR_RVn was crystallized as a 4 repeat protein, and both the original structure (CTPR_RV4) and a similarly reconstructed model for CTPR_RV20 were used here [139]. Initial ENMs of the 20mers were distorted by abnormally large eigenvectors of the discontinued C-termini of each 4-repeat unit cell. Therefore, the inter-repeat loops after each fourth repeat were modelled with the MODELLER software [286] using the 20mer structures as fixed templates. The full structures could then be used for ENM without any further problems.

3.3 PR65 as a putative modulator of PP2A dynamics

3.3.1 Vibrational Dynamics of PR65 and full PP2A complexes

The lowest vibrational mode, mode 7, is a bending mode that changes the curvature of PR65 thereby moving the N- and C-termini into opposite directions (Fig. 3.6a). Mode 8 describes a screw-like twisting of the N- and C-terminal repeats about an axis through the protein, while mode 9 is a combination of twisting and wagging of both N-terminal and C-terminal domains (Fig. 3.6b,c). In higher order normal modes, the movement becomes more localised, which is reflected in the collectivity: The lowest 6 modes are all associated with collectivities of > 0.5, indicating that a significant proportion of the molecule is affected by that mode. Collectivities of the higher modes vary strongly between 0.1 and 0.7.

In ENMs of holoenzymes, the dominant low modes of PR65 are translated into the complex. However, the direction of each holoenzyme mode strongly depends on the position of the other subunits relative to each other and PR65:

1. If the B- and C-subunits were bound tightly to either end of PR65, the B-subunit and C-subunits moved with the N- and C-terminal repeats of PR65, respectively.

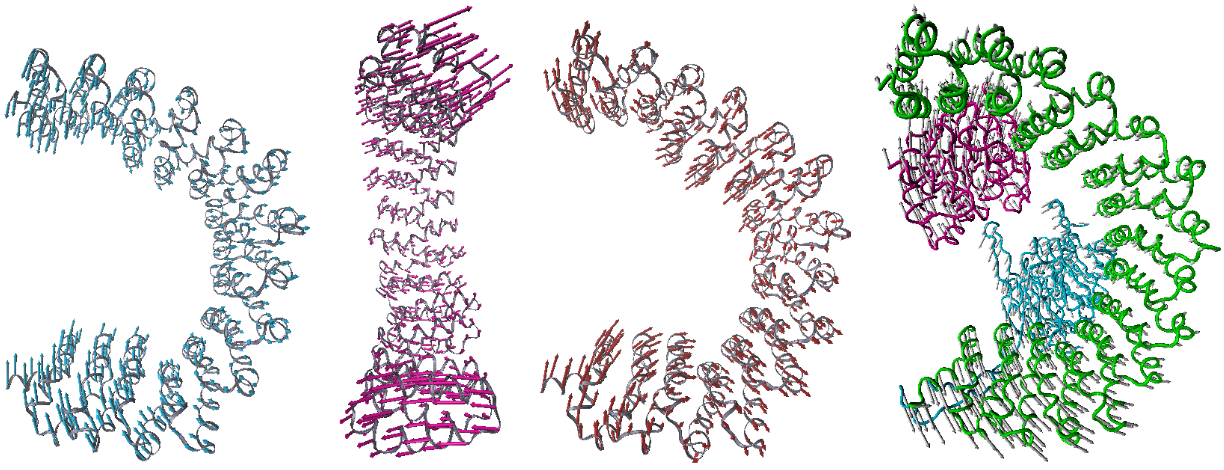

(a) Mode 7 (b) Mode 8 (c) Mode 9 (d) Holoenzyme

Figure 3.6: Vibrational modes of PR65. All structures are shown with the C-terminal at the top of the image.

That is, although the vibrational frequency was much lower due to the increase in size, the original movements of PR65 were preserved which modulated the distance between the subunits.

2. If the B- and C-subunits were close enough such that springs are set in between them, forming a closed ring, the subsequent motions describe an overall twisting and wagging of the whole complex about this connection.

3. If PR65 did not form extensive contacts to the surface of a B-subunit, they moved independently about this hinge.

When any of the PR65 subunits of these complexes is modelled without the subunits, the motion is qualitatively very similar to unbound PR65 (1b3u) and frequencies of the lowest modes are of the same magnitude. Overlaps between the individual normal modes are high and are dominated by either the first mode only (e.g. $O = 0.74$ between 1b3u and 2iae, and $O = 0.89$ between 1b3u and 3dw8), or they are spread between all three lowest normal modes (e.g. 1b3u and 4i5l).

3.3.2 Analysis of PR65 conformational changes between crystal structures

An N-terminal alignment of PR65 molecules from different *apo*-, *holo*- and inhibitor/regulator-bound conformations reveals large changes in the backbone (Figure 3.1). The protein- or inhibitor-bound structures differ from the *apo*-crystal by varying degrees with RMSDs ranging from 5 Å to more than 9 Å. The repeat angles of all structures are quantified in

Table 3.3: Average and cumulative angles of different PR65 structures

PDB ID	**Curvature [°]**		**Twist [°]**		**Bending [°]**	
	$\bar{x}$	$\sum x$	$\bar{x}$	$\sum x$	$\bar{x}$	$\sum x$
1b3u	9 ± 3	130	-9 ± 4	-120	-4 ± 2	-60
2iae	11 ± 3	165	-7 ± 3	-104	-3 ± 1	-36
2nym	16 ± 3	235	0 ± 3	-5	-5 ± 2	-69
3dw8	11 ± 2	158	-5 ± 3	-71	-4 ± 1	-50
3k7w	12 ± 3	164	-7 ± 3	-104	-3 ± 1	-37
4i5l	13 ± 3	186	-6 ± 3	-82	-4 ± 1	-48
5w0w	16 ± 3	234	-2 ± 3	-26	-4 ± 2	-55

Table 3.3. Although on average curving and twisting angles are the same within error, their variation between structures is much larger than that observed for bending. The large standard errors arise from discontinuities in packing. Repeat angles may be similar when averaged, but once they are added over the whole array they reflect the changes seen in the crystal structures: curvature and twist exhibit variations of over 100°.

Tama and Sanejouand [255] found that the conformational changes between open and closed forms of 20 proteins often correlated with a single mode. Here, the same methods were applied to PR65. The RMSDs of the conformational changes ($\Delta\vec{r}$) and the overlaps of $\Delta\vec{r}$ with the vibrational modes of the *apo* protein are presented in Table 3.4. For every $\Delta\vec{r}$, one of the lowest three modes scores the highest overlap. Changes from 1b3u to 2iae and 4i5l are clearly dominated by mode 7, whereas in 3dw8 the conformational change overlaps with all three lowest modes almost equally. The lowest three modes together describe more than 80% of the conformational changes, and all 19 modes examined account for >90% of the variation.

Next, this analysis was extended to a whole PR65 structural ensemble including 8 different crystal structures (Table 3.1). Using PCA, one can calculate the directional

Table 3.4: Overlaps between conformational changes and ENM modes of PR65. $\Delta\vec{r}$ RMSDs were calculated between residues 10-587 (*apo*) and 11-588 (*holo*) of PR65. Bold numbers signify the largest overlap of all 19 modes.

$\Delta\vec{r}$	1b3u → 2iae	1b3u → 3dw8	1b3u → 4i5l
RMSD [Å]	8.4	5.7	9.2
Mode 7	**0.76**	0.45	**0.87**
Mode 8	0.23	0.45	0.02
Mode 9	0.40	**0.53**	0.36
CO [7-9]	0.87	0.83	0.94
CO [all]	0.92	0.93	0.97

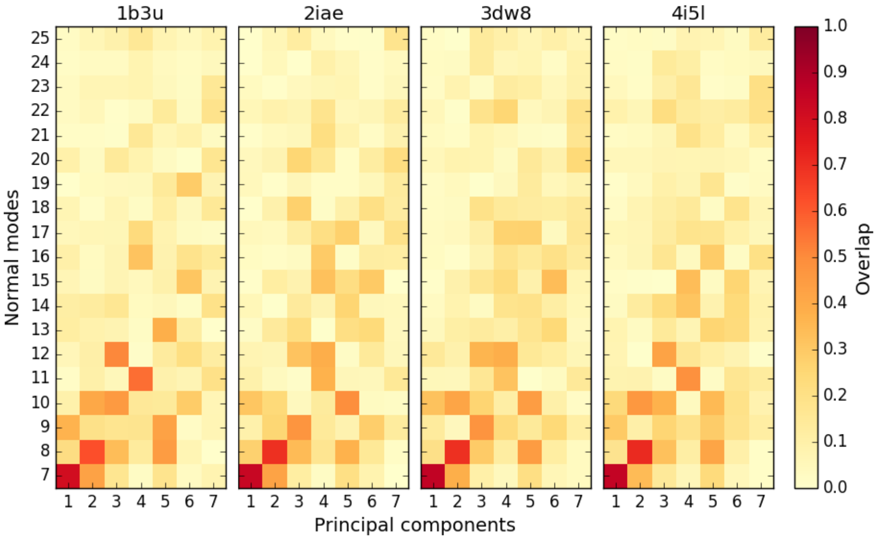


Figure 3.7: Normal mode overlaps with conformational changes across different PR65 crystals. The first six normal modes corresponding to rotational and translational motion have been omitted from this analysis.

variation of backbone coordinates between all crystal structures. Figure 3.7 shows the overlaps between the resulting principal components and normal modes of the four PR65 conformations examined previously. This overlap matrix shows that the first two normal modes, bending and twisting, overlap strongly with the first and second PC, indicating that on average, the majority of structural changes seen between crystals are accessible from any of the four shown starting structures *via* their respective normal modes.

3.3.3 Comparison of ENMs with full atomistic simulations

A comparison between ENM results and full atomistic simulations will allow the verification of whether the ENM is able to capture the global dynamics of the protein. Grinthal *et al.* [201] published results of a normal mode analysis (NMA) based on MD simulation data. For the lowest two modes they quote wave numbers of 0.37 cm^{-1} and 0.81 cm^{-1}, which are close to the ENM values of 0.38 cm^{-1} and 0.76 cm^{-1}. It is known in the literature that NMA and ENMs tend to differ only slightly [260].

To test whether the harmonicity of ENMs could still capture more realistic, anisotropic motions, the 19 lowest ENM modes were compared to the ten largest PCs from an essential dynamics analysis (EDA). The highest scoring overlaps, cumulative overlaps and subspace overlaps are shown in Tables 3.5a-c. Mode 7 and PC1 score the highest overlap. PC2 does not overlap well with mode 8 but instead scores the overall second highest overlap with mode 9. Cumulative overlaps calculated over more modes describe the motion of a given PC better, although the first three modes can account for most of the variation seen in PC1 and PC2 (Table 3.5b). The motion spaces overlap well as long as only the

Table 3.5: Comparing normal modes to MD simulation data. (a) Overlaps between the three lowest modes and the three largest PCs. (b) Cumulative overlaps between the three largest PCs and different sets of normal modes. (c) Subspace overlaps (RMSIP) between PCs and normal modes.

(a) Overlaps

	PC1	**PC2**	**PC3**
Mode 7	**0.66**	0.21	0.01
Mode 8	0.09	0.27	0.22
Mode 9	0.27	**0.59**	0.38
Mode 10	0.28	0.24	**0.41**

(b) Cumulative overlaps

	PC1	**PC2**	**PC3**
3 modes	0.72	0.68	0.44
6 modes	0.80	0.79	0.76
all	0.86	0.84	0.84

(c) Subspace overlaps

	3 PCs	**6 PCs**
3 modes	0.63	0.52
6 modes	0.79	0.71

lowest three PCs are included (Table 3.5c).

3.3.4 ENMs can capture PR65 flexibility observed *in crystallo* and *in silico*

The vibrational dynamics of PR65 are dominated by very simple motions: bending, twisting and wagging. These motions become more complicated once the bound B/C subunits are modelled and if these are in close proximity to each other. When they are close enough to form springs that result in an overall ring-like structure, frequencies and motions of the lowest modes change considerably. Currently, it is unknown whether physical contacts between the B- and C-subunits are present in solution or whether these arise from crystal packing.

Conformational changes between the PR65 *apo*-crystal and different holoenzymes or inhibitor/regulator-bound crystals are well represented by a single mode or a combination of the lowest three modes, which agrees well with previously published results [263] and suggests that conformational changes within all repeat proteins can occur along the direction of their lowest normal modes. It is possible that binding of interaction partners is energetically more favourable when conformational changes happen along low-frequency modes, although this information would be difficult to extract experimentally.

The comparison between MD simulations and ENM shows that the ENM captures the global motions of repeat proteins well even though it is only a harmonic approximation. More importantly, the overlaps between modes and PCs from EDA indicate that most of the motion seen along the largest PCs is captured by the lowest modes. Comparing these results to a study on HIV-1 protease by Yang *et al.* [259], PR65 obtains very similar scores. Non-unity overlaps are most likely due to the remaining anharmonicity inherent to each individual system.

Altogether, these data show that PR65 has the intrinsic ability to be a very flexible linker that can adapt to form the multitude of PP2A complexes that are present in the cell. This flexibility can be captured using simple, coarse-grained ENMs. Biophysical analysis has shown that binding of the catalytic subunit to the C-terminal repeats of PR65 increases the affinity of the N-terminal repeats of PR65 for an inhibitor (SV40 small t antigen) by an order of magnitude [219]. However, there are no obvious direct contacts between the small t antigen and the catalytic subunit, which suggests a mechanism by which PR65 functions as an allosteric transmitter of catalytic-subunit binding. One could speculate that binding of the catalytic subunit to the C-terminus of PR65 will influence the dynamics in such a way as to facilitate binding of the regulatory subunit, e.g. by slowing down the vibrational motions and trapping PR65 in a different conformational ensemble. It could also be possible that the intrinsic dynamics of PR65 could aid subunit exchange and consecutive dephosphorylation cycles [201].

3.4 The dynamics underlying quorum sensing

3.4.1 Dynamics of the Rap protein family

Given the functional relevance of the conformational changes seen for the Rap proteins crystallographically, I investigated whether they might arise from the intrinsic dynamics of each protein. The lowest three normal modes of RapI are displayed in Figure 3.8. The

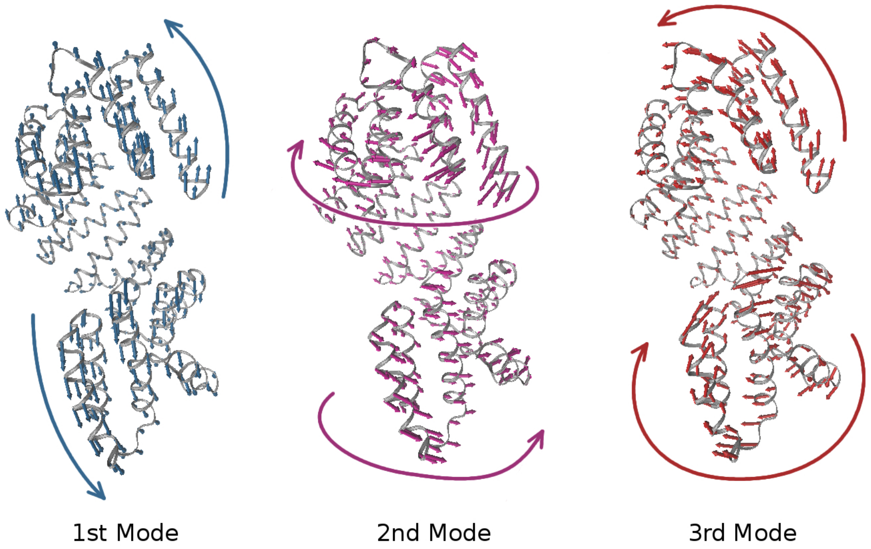


Figure 3.8: The three lowest normal modes of RapI. The first mode (teal) describes a bending motion that alters the distance between the N- and C-termini; the second mode (magenta) tightens the superhelix in a screw-like motion; the third mode (red) twists the N-terminal three-helix bundle and the C-terminal TPR relative to the repeat array superhelix.

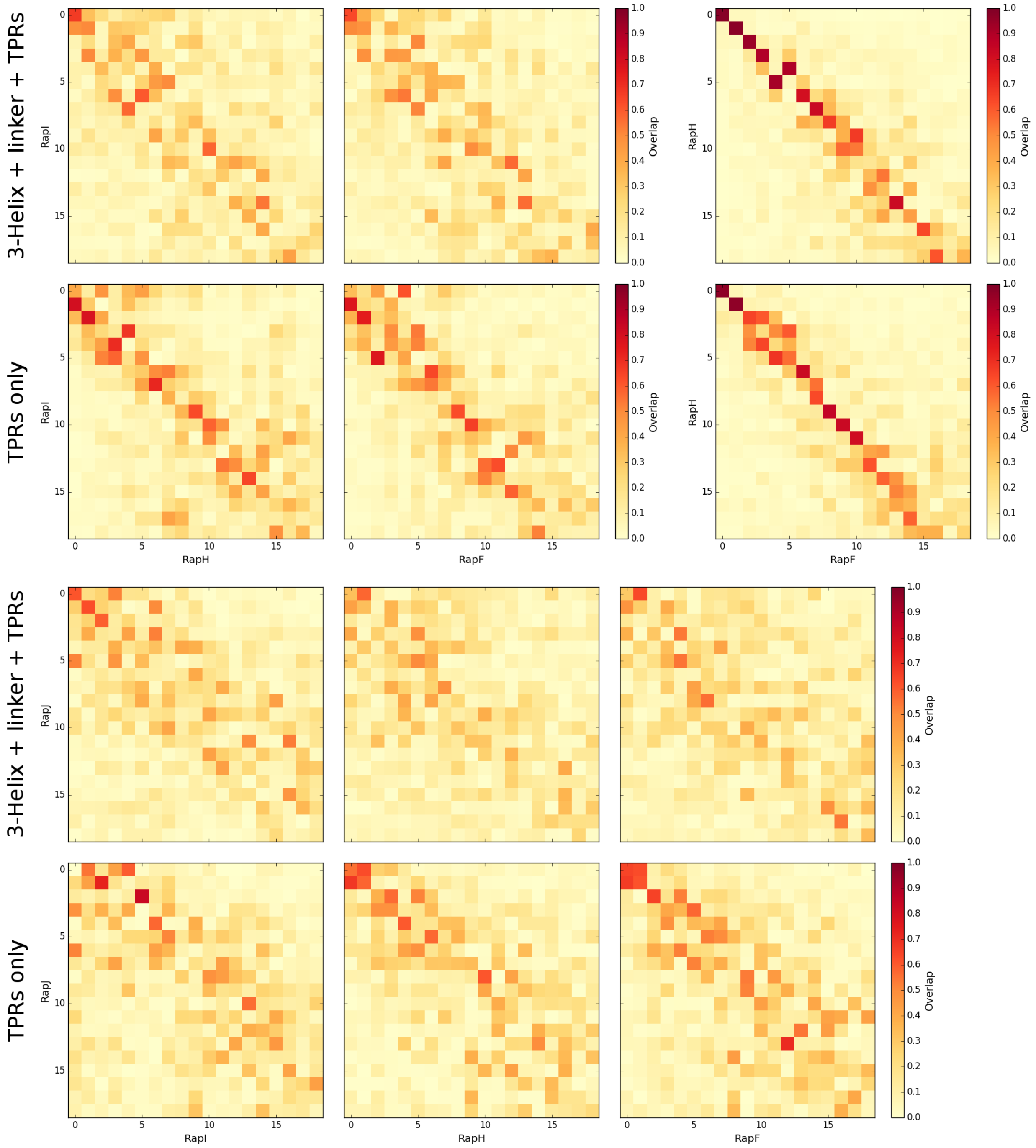


Figure 3.9: Overlaps between ENM normal modes of Rap proteins, for the whole proteins and TPR repeats only. If the normal modes of two proteins are nearly identical, the overlap matrix will have diagonal elements with $O \to 1$ and off-diagonal elements with $O \to 0$ (e.g. RapH vs RapF). If normal modes are close in frequency their order can switch, giving rise to off-diagonal elements with a large overlap (e.g. RapI vs RapF/H). The first six normal modes corresponding to rotational and translational motion have been omitted from this analysis.

motion is dominated by the bending of the super-helix, followed by a screw-like twist of repeats along the super-helical axis and finally, more localized motions of the N- and

C-terminal three-helix bundle and TPR repeats, respectively. The general dynamics are very similar in all four proteins, where the lowest normal modes involve the collective motion of approximately 40-80% of the structure. Due to the high structural similarity between RapH and RapF (the original crystal structures align with an RMSD of 1.65 Å), both proteins explore a nearly identical vibrational space (Figure 3.9, top right). Since the orientation of the N-terminal helix bundle is the only major difference between the active configurations and Rap I, the overlaps of RapI dynamics with RapF and RapH normal modes are high and concentrate in the diagonal, especially when the TPR domains are modelled independently (Figure 3.9, top left). The dynamics change dramatically upon compression of the superhelix in RapJ, and only the lowest normal modes overlap to a significant degree with the other structures ($O > 0.5$, Fiugre 3.9). High overlaps along the diagonal are lost, indicating that a clear correlation between motional spaces is lost. When comparing ENMs of the TPR repeats only, similarities in motion are only slightly higher.

It is possible to compare the dynamics of peptide-bound and un-bound conformations of RapJ. This is properly done by calculating an ENM of the sub-system RapJ within the system of the whole complex [246]. However, the peptide C_α atoms represent only a small fraction ($\sim 1.4\%$) of the whole RapJ+PhrC complex. Therefore, overlaps between RapJ modelled on its own and modelled in complex with the PhrC peptide (with PhrC normal mode vectors removed from the ENM) are a good enough approximation. The presence of the peptide does not significantly influence the motion of the lowest 7 normal modes, most of which score overlaps of > 0.99, and only moderately affects the dynamics

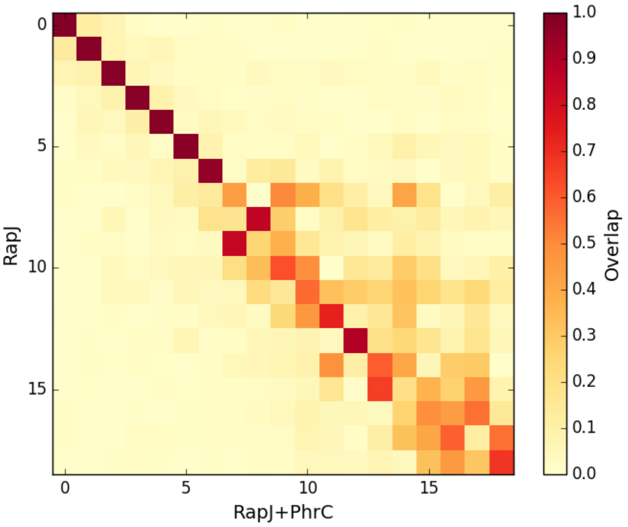


Figure 3.10: Approximation of overlaps between normal modes of ENMs of RapJ bound to the PhrC peptide and RapJ only, where the atomic vectors corresponding to the peptide have been removed.

of the higher modes (Figure 3.10). Since most comparisons of global dynamics involve the lowest modes, the effects of the peptide are negligible.

3.4.2 Relating ENMs to the conformational changes in Rap proteins

The models were examined to determine whether the motions observed in the ENMs can account for the conformational changes between two structures by measuring how well the lowest five normal modes of one protein overlap with the conformational transition to another protein (Figure 3.11). The transition between extended and compact conformations is very well described by the normal modes of either RapI or RapJ. The first five normal modes of RapI can account for 0.87 of the conformational change between RapI and RapJ, while the first five normal modes of RapJ only describe 0.75 of the change. Overlaps between normal modes and transitions of either compact or extended conformation to the active forms (RapH and RapF) range mostly between 0.3 and 0.4 (Figure 3.11). If modelled without the N-terminal three-helix bundle, overlaps between conformational transitions in the TPR domain and respective normal modes are again generally > 0.7. This observation is not entirely unexpected, as more localized motions, such as the motion of the helix bundle are only captured by normal modes of higher order.

When examining the entropic contributions of each normal mode, the extended confor-

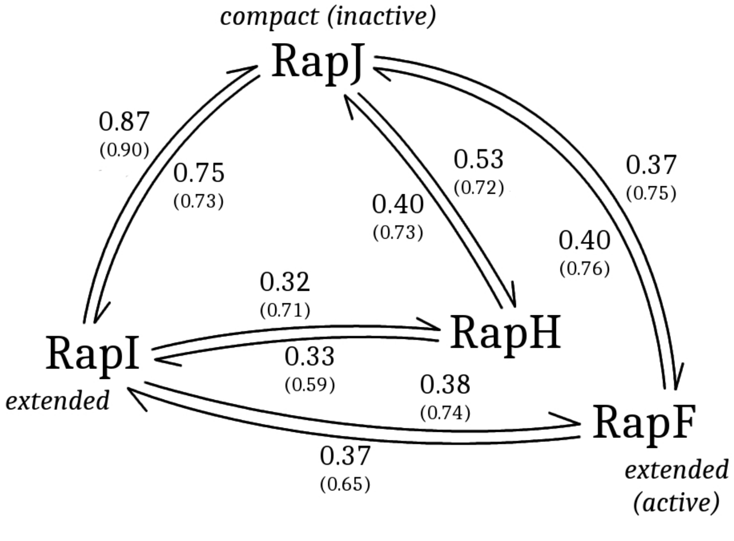
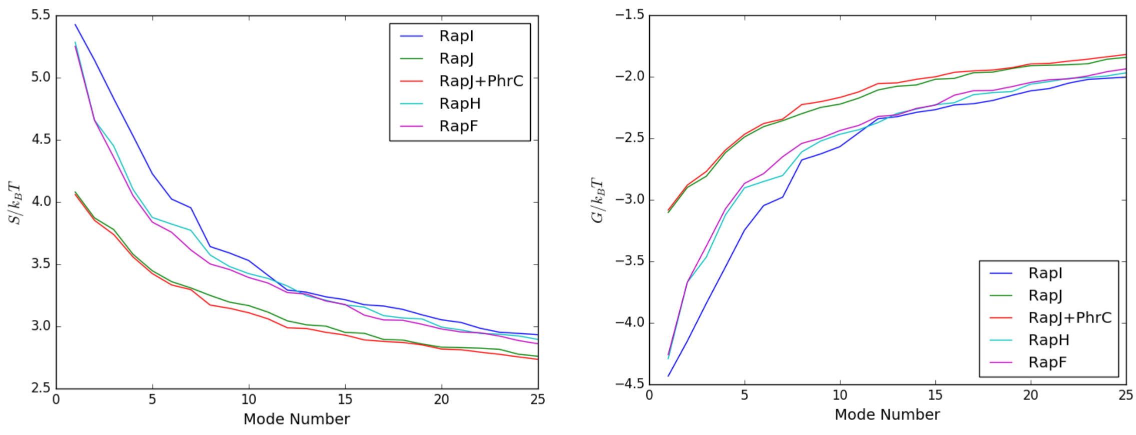


Figure 3.11: Quantitative comparison between the lowest five normal modes and conformational changes between different Rap proteins. The arrows represent the conformational change vector, and the values equal the corresponding cumulative overlap between the vector and the ENM of the starting structure. Numbers in brackets correspond to the cumulative overlaps between the dynamics of truncated and independently modelled TPR domains and their respective conformational changes.

mation is entropically, and thereby also energetically, the most favourable (Figure 3.12). The entropic contributions of the lower modes of RapI are higher than those of RapF or RapH. The compact conformation of RapJ comes along with a considerable entropic cost, which is largely independent of the presence of the peptide. Since differences between ENMs with and without peptide are small, the effect of peptide binding on the entropy of the system is negligible compared with the entropic cost of the conformational change.



(a)

(b)

Figure 3.12: Entropic (a) and the resulting energetic (b) contributions of each normal mode to the total motion. The closed conformation was modelled both with and without the PhrC peptide.

Lastly, by analysing the correlation of motion between different residues (Figure 3.13), one can obtain insights into why the three-helix bundle in the compact conformation has very little potential to rotate and to bind partner proteins. The TPRs exhibit correlated motion only with their nearest neighbours, giving rise to a pattern of that resembles consensus TPR contact maps [63]. Movements of the rotated N-terminal three-helix bundle, linker domain and first TPR (Figure 3.13a, blue box) are strongly correlated in the active conformations and are not TPR-like, suggesting that they form a subdomain relative to the rest of the TPRs. Some nearest-neighbour correlations in that region are reduced in the extended conformation of RapI, whereas they are either further reduced or even reversed in RapJ (Figure 3.13a, arrows), and the separate behaviour of the N-terminal-domain is lost. The binding of the peptide marginally increases nearest-neighbour correlation at the centre of the TPR domain (Figure 3.13a, purple box). In cross-correlation maps based on fewer normal modes, the reversal of correlations upon structural change becomes more pronounced (Figure 3.13b,c). In particular, the cross-correlations calculated from the lowest three modes exhibit a pattern in which the protein is divided into different subdomains, the borders of which shift depending on the conformational state. Most notably, the open conformation of RapI displays nearly symmetric correlation that is mirrored about the centre of the protein.

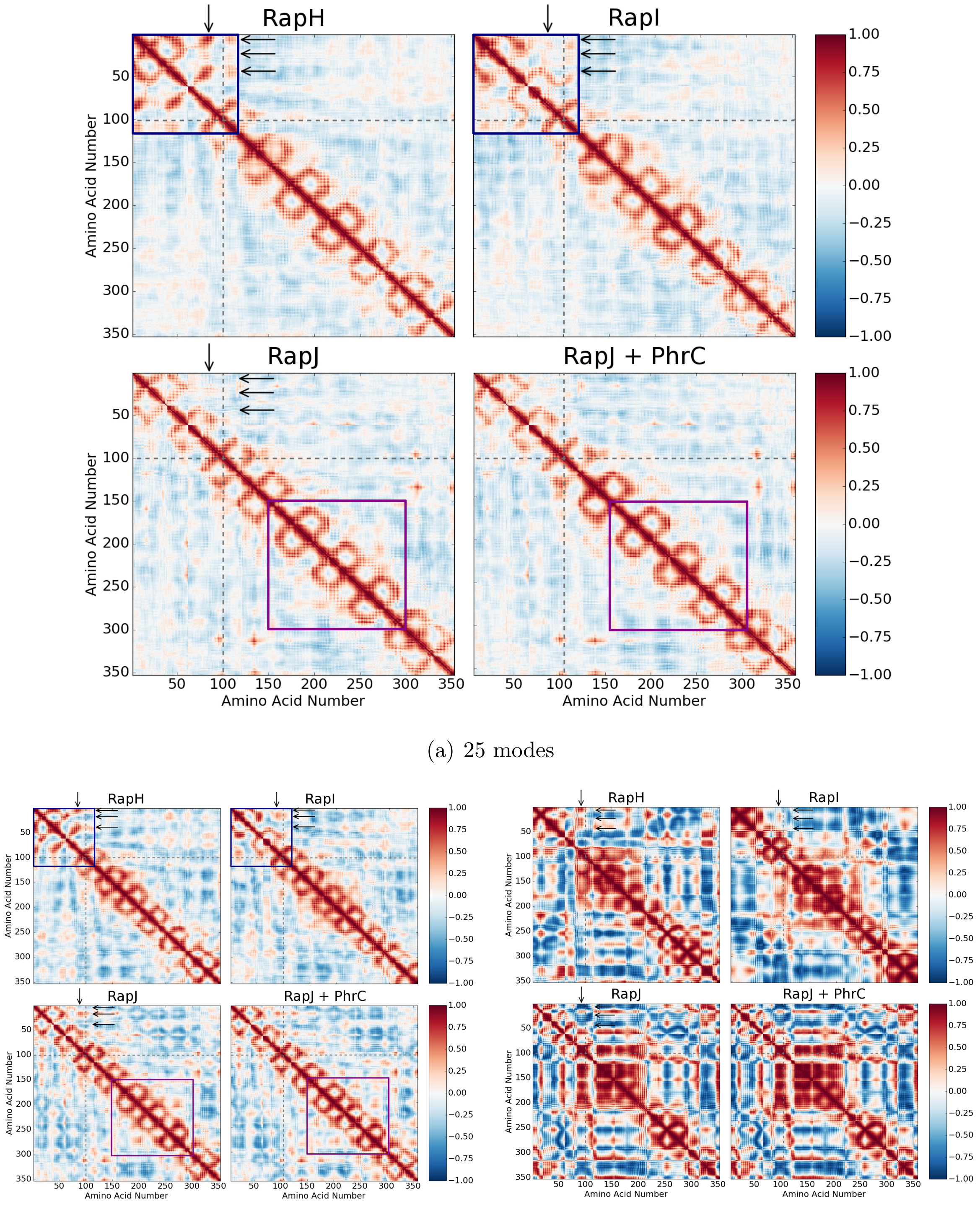
3.4.3 The intrinsic dynamics constitute a possible allosteric mechanism

The four different Rap structures presented here differ in their arrangement of the TPR arrays. Comparisons of ENMs of each protein show that although they differ in structure, their dynamic behaviour, as characterised by the lowest normal modes, is quite similar. The motion of each can be more or less sub-divided into the that of the TPR domain and that of the N-terminal helix bundle, which forms a sub-domain with the first TPR in the extended and active conformations. As expected, the similarity between motions correlates with overall structural similarity, i.e. extension of the superhelix.

It was further shown that changes (excluding rotation of the N-terminal bundle) observed between any of the conformational states can be captured by a few of the lowest normal modes. The highest overlaps were scored by the RapI-RapJ transition, indicating that moving between extended and inactive conformations are the easiest to access *via* normal modes. The fact that overlaps from an extended to open conformation are higher than for the reverse direction, may simply arise from a larger number of springs set in the closed conformation [266, 267]. Since formation of the compact state entails a considerable entropic cost, enthalpic contributions of multiple contacts between peptide and TPRs are required to achieve an overall energy minimization [11].

The second highest overlaps are scored by the RapJ-RapF/H transitions. The most plausible reason for this is that the active conformations are more compact than the un-bound state, and therefore slightly resemble the inactive conformation. This is also reflected in a slight reduction of entropy of the lowest normal modes. However, a possible transition between inactive and active conformation is unlikely due to the correlated motions of the N-terminal helix bundle, which are still somewhat present in the extended conformation, while they are totally absent in the compact conformation.

Considering these comparisons, a mechanism emerges by which Rap proteins on their own could potentially explore the different conformations observed, including rotation of the N-terminal domain. Binding of the peptide, substrate or transcription factor could then simply trap the protein energetically in a given conformation. It is notable that the presence of the peptide does not directly influence the correlated motion of the N-terminal domain, indicating that locking of the three-helix bundle to the TPRs is entirely due to the conformational change. However, the peptide could induce indirect or allosteric effects by stabilizing the TPR domain in the compact conformation. These effects could then be



(a) 25 modes

(b) 10 modes

(c) 3 modes

Figure 3.13: Representative cross-correlation maps for the partner-bound, open and peptide-bound conformations summed over the lowest (a) 25, (b) 10 and (c) 3 normal modes. The cross-correlation between residues is a measure of how much these move in the same direction. The N-terminal helix bundle and TPR repeats are divided by grey dashed lines. Purple and blue boxes highlight the peptide-binding TPR repeats and the N-terminal sub-domain, respectively. Arrows point to sign-reversals in correlation.

transmitted through the array *via* the interaction potential between repeats. That is, an alteration in cooperativity between individual repeats in the C-terminal domain might be translated into a rearrangement of the helix bundle.

In summary, the compact and extended conformations of Rap proteins have different supramolecular geometries, arising from differences in the inter-repeat packing. Consequently, they must have different values of the interfacial repeat stability. The ENMs showed that all conformations are easily accessible through the motions of the TPR domain, albeit that an extended conformation of the array may be preferred owing to the entropic cost of the compact state. Ultimately, the intrinsic flexibility of the TPR array may allow for the existence of two functionally different conformations that can be locked by their respective binding partners.

3.5 Insights from consensus repeat proteins

To gain more information on how the overall shape of a super-helix relates to its vibrational dynamics, the geometries and ENMs of two different CTPR variants are analysed in this section. I compare the two variants on both a small and a large scale using the asymmetric 4-repeat crystal unit and the re-constructed 20-repeat models, respectively (see Section 3.2.7).

3.5.1 Geometries of designed CTPR proteins

When comparing the 4-repeat units the structural differences are very small. The backbone RMSD between CTPR4 and CTPR_RV4 is only 1.4 Å.[1] The average curving and bending angles are similar between the two variants, only the twist is clearly larger in the CTPR4 (Table 3.6). The cumulative angles, however, already predict the trend seen for

Table 3.6: Average angles as calculated. Averages are quoted with standard errors of the mean.

Type	Number	Curvature [°]		Twist [°]		Bending [°]	
		$\bar{x}$	$\sum x$	$\bar{x}$	$\sum x$	$\bar{x}$	$\sum x$
CTPRa	4	28 ± 1	83	13.07 ± 0.03	39	22.7 ± 0.4	68
	20	26.2 ± 0.8	497	13.5 ± 0.2	256	23.4 ± 0.3	444
CTPR_RV	4	30 ± 2	91	10.3 ± 0.2	31	21 ± 1	62
	20	31.5 ± 0.7	599	11.1 ± 0.4	211	19.4 ± 0.7	369

[1] For comparison, the backbone RMSDs between different crystal forms of CTPRa arrays [126] are as follows: the structures with PDB ID 2hyz (used here) and 2avp differ by 0.507 Å; the structures with PDB IDs 2fo7 and 2avp differ by 0.461 Å; and 2hyz and 2fo7 differ by 0.641 Å.

larger repeat arrays. In the 20-repeat arrays, the backbone RMSD is > 9 Å and differences in average repeat angles becomes more prominent (Table 3.6). The four interface mutations result in an increase of the curvature by 5°, a decrease in twist by approximately 2°, and a decrease in lateral bending of 4°. Cumulatively, these relatively small changes translate into total differences ranging from 45° in twist to more than 100° in curvature. Effectively, the increase in curvature is buffered by decreases in both twist and bend, resulting in a super-helix of larger diameter and consequently shorter en-to-end distance.

3.5.2 Dynamics of designed CTPR proteins

The normal modes of ENMs based on 4 repeats resemble those described for horse-shoe-like structure such as PR65. However, in 20-repeat arrays, the two lowest normal modes are bending motions, while the third and fourth modes twists and stretch the super-helix, respectively (Figure 3.14). The overlaps between modes 7 to 10 of both 4- and 20-repeat proteins are larger than 0.8 for all but one, indicating large similarities in dynamics (Table 3.7). Apart from mode 7, overlaps between 20-repeat arrays are larger than those containing four repeats, indicating that motion differs more on the small than on the large scale.

Larger repeat arrays have significantly higher entropic contributions compared to smaller arrays, and therefore lower free energies. In both small and large repeat arrays,

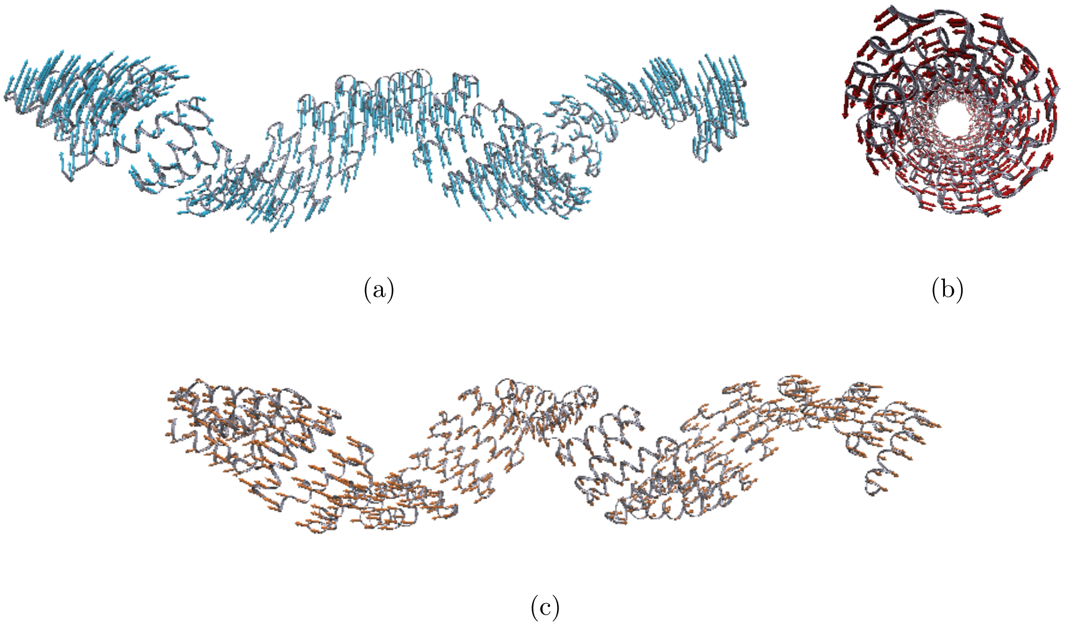
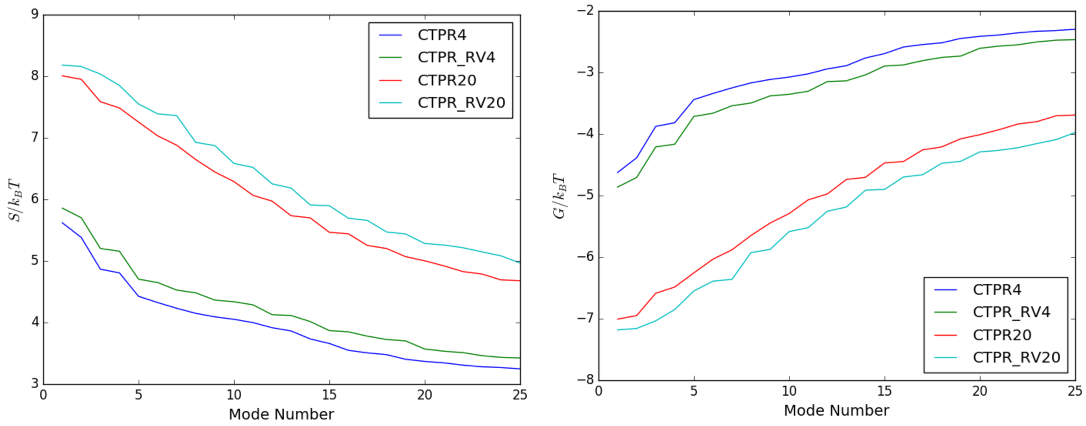


Figure 3.14: Dynamics of the CTPRa20 super-helix: (a) mode 7 and 8 describe a bending motion, (b) mode 9 is a screw-like twist, and (c) mode 10 is a stretching mode. The dynamics of the CTPR_RV20 super-helix are qualitatively very similar.



(a) (b)

Figure 3.15: Entropic (a) and the resulting energetic (b) contributions of each normal mode to the total motion.

normal modes of CTPR_RV are associated with higher entropies and hence lower free energies (Figure 3.15). That is the shape of CTPR_RV is energetically more favourable than that of CTPR, at least from a vibrational perspective.

3.5.3 Raps and CTPRs may not be directly comparable

In contrast to Rap proteins, a designed TPR system exhibits motions that resemble a simple physical spring. The substitution of a small number of amino acids in each repeat results in different interfacial geometries and vibrational motions. The change in shape is nearly undetectable in 4-repeat arrays but very prominent in those containing 20 repeats. However, 4-repeat arrays show larger differences in motion than larger repeat arrays. It remains to be investigated whether differences in motion can be quantitatively connected to alterations in shape.

In both 4- and 20-repeat arrays, the CTPR_RV variant is entropically more favoured than the slimmer more extended CTPRs. The data from these designed TPRs therefore suggest that a compact conformation is entropically more favourable, which stands in

Table 3.7: Overlaps of ENM modes 7-9 of CTPRa4 and CTPR_RV4, and their 20-repeat helices.

	4 repeats	**20 repeats**
Mode 7	0.85	0.82
Mode 8	0.76	0.84
Mode 9	0.91	0.95
Mode 10	0.76	0.97

direct contrast to the conclusions obtained from the analyses of Rap proteins. However, in CTPRs the arrangement of network nodes is more regular and hence the two systems may not be comparable. Furthermore, it is likely that in an even narrower super-helix, more springs are set between neighbouring repeats and therefore cause the helix to be less flexible. Experimentally, this decrease in flexibility could be caused by higher intrinsic repeat and interface stabilities, which would correlate with the fact that CTPRs are more stable both thermally and chemically than CTPR_RVs.

3.6 Final discussion

Bending, twisting and wagging motions of linear repeat proteins are quite similar to those described to be universal to bilobate structures [287], even though repeat proteins possess no hinges, and these motions are independent of repeat type. Instead of moving about a hinge, repeat proteins move about the apparent centre of mass. It would be interesting to see how the geometric parameters of repeat orientation (curvature, twist and lateral bending) translate into the supra-molecular shapes of different repeat proteins [7]. One could then design repeats with different repeat angles on a small scale in a targeted manner, similar to how CTPR_RV was designed [139], and predict their cumulative changes in shape. ENMs rely heavily on the overall shape of a protein and therefore it would be interesting to see how differences in repeat arrangement ultimately give rise to differences in dynamics.

The origin of repeat array flexibility and how it depends on the repeat types and overall shape of a protein is currently unknown. I previously built ENMs for armadillo,

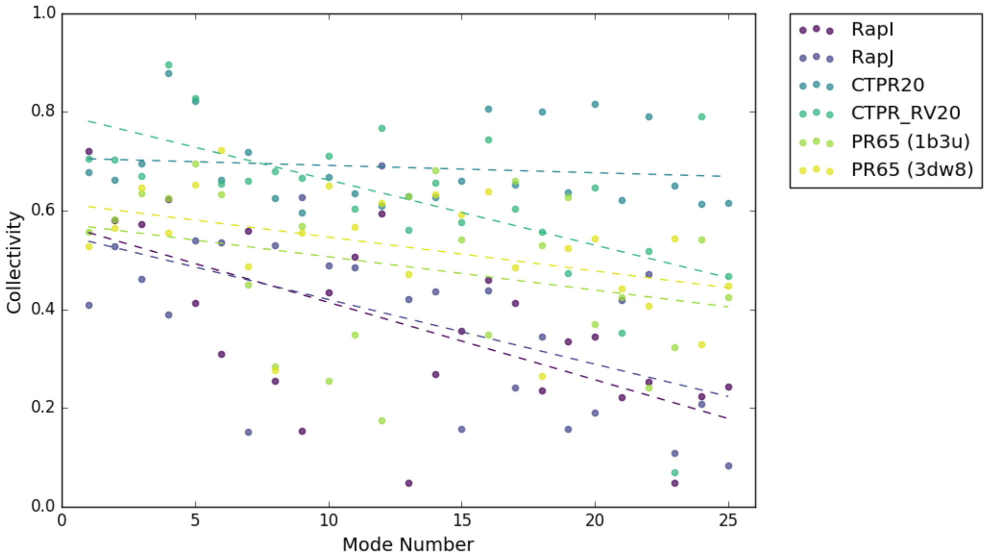


Figure 3.16: Comparison of collectivity of protein motion

ankyrin, WD40 and LRR repeats (data not shown). However, differences in the respective crystallographic data sets were extremely small and are within backbone uncertainty. Although this may suggest that such repeats are less flexible than HEAT repeats and TPRs, it cannot be excluded that the lack of structural variation is due to crystal packing. Furthermore, it remains to be examined whether repeat types with different packing interactions, interfacial energies and cooperativities will exhibit correspondingly different dynamics and macromolecular flexibility. Since these parameters result from side-chain characteristics, a major amount of detail is lost in C_α-atom based ENMs. However, interfacial energies and cooperatitivities are likely to depend on structural parameters such as the interface packing, which in turn will influence repeat angles locally and thereby determine supra-molecular shape globally. Using more advanced ENMs, that incorporate chemical characteristics of side chains or are combined with MD simulations [258, 266], it would be interesting to perform a quantitative comparison of how collectivities of the modes vary with interfacial surface area or packing density. For example, Figure 3.16 shows the collectivities of two examples from each system examined here using basic ENMs. The variations seen for collectivities within a protein are large. Generally however, collectivities tend to decrease from lower to higher modes. Interestingly, this is not the case for CTPR20, which is experimentally the most stable protein of them all.

Furthermore, it is unclear how the global dynamics are affected by local changes, such as mutations, that do not affect the supra-molecular shape but clearly alter protein stability. Within the Ising model formalism, mutations can cause changes in cooperativity [125], which will not only affect that repeat, but will influence neighbouring repeats as well. Some time ago, different groups suggested that local changes can couple to global dynamics and this can be used by allosteric mechanisms [250, 288, 289], yet there is no direct concrete experimental evidence of how this may happen. The relative simplicity of repeat proteins compared to globular proteins may aid addressing this question, and the PP2A and Rap protein systems presented here could be good first candidates for future investigation.

Chapter 4

Building PR65-Impβ chimeras

4.1 Introduction

Both normal mode analysis and principal component analysis of ENMs and MD simulations have shown that as a result of its intrinsic shape, PR65 has a very dominant bending mode that allows for alternating increase and decrease in curvature [201, Itzhaki and Settanni, unpublished]. If PR65 motion drives the interaction between B- and C-subunits, then altering the frequency of that motion should affect the rate of phosphorylation. We hypothesise that an extension of PR65 by a few repeats will slow down vibrational frequencies. The N-terminus of PR65 interacts with one of several B-subunits, which each contact different repeats in the region HEAT1-10, whereas the C-subunit binds to the C-terminal repeats of PR65, HEAT12-15 [8, 209, 217]. Therefore, the C-terminus appears to be a more suitable place of the array to add one to four HEAT repeats than the N-terminus (Fig. 4.1). Doug Barrick and colleagues showed using the Notch ankyrin repeat-domain that repeat proteins can tolerate a variety of repeat insertions and deletions without disrupting their structure [290, 291]. By using repeats from Importin-β, we exploit the advantages of adding the same repeat type without introducing additional

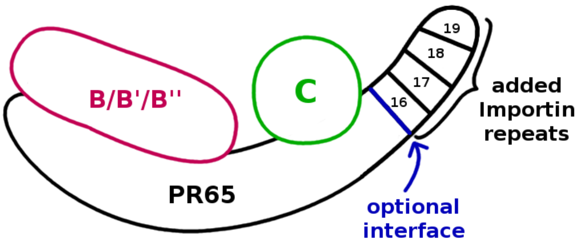


Figure 4.1: The design of PR65-Importinβ chimeras: C-terminal Impβ repeats are added to the C-terminus of PR65 without interrupting the binding regions of either the B-subunits or the C-subunits of PP2A.

binding possibilities for either C- or B-subunits.

One objective of this work is to examine how publicly available, server-based computational methods can be used to aid the design process of repeat protein constructs. First, I use a docking software to investigate the formation of a stable interface between the final PR65 repeat and individual Importin-β repeats. Second, I employ structure prediction methods to test whether these chimeras would be able to fold into stable structures. The final objective, was to produce the chimeras and characterize their biophysical properties prior to testing their effects on function in phosphatase assays.

4.1.1 Docking using HADDOCK

Docking describes the process of modelling multi-molecular assemblies from individual starting structures. That is, given two (or more) starting structures a docking algorithm will find a binding mode in which those structures are interacting [292, 293]. Current *ab initio* docking approaches only take into account the starting structures, while HADDOCK (High Ambiguity Driven protein-protein DOCKing) allows for data-driven docking where experimental data or user-defined predictions guide the process [292]. This additional input, consisting of the definition of both "active" and "passive" residues, forms Ambiguous Interaction Restraints (AIRs) that drive the docking. That means, every active residues has to form contacts with another active or passive residue, otherwise it will incur an energy penalty. Docking proceeds in three stages: first, energy of the complexed system is minimized using rigid-body rotations; second, the complex undergoes semi-flexible refinement where only those parts of the complex that are in contact are minimized in energy; third, the complex is refined in explicit solvent. At every stage, a HADDOCK score, a weighted sum of different energies (see Section 4.2.2), is calculated for each model which is used for ranking. After the last refinement, the final set of models are clustered according to structural similarity. This can be done either according to the backbone RMSDs of the interface and ligands (i-RMSD and l-RMSD) [294] or according to the fraction of common contacts (FCCs) [295]. While RMSD-based clustering performs well with small proteins, FCC-based clustering is better for larger systems such as supra-molecular assemblies and the study of interactomes [294, 295]. The incorporation of structural flexibility upon complex formation distinguishes HADDOCK from other docking algorithms that focus on *ab initio* modelling [292]. The web-server performs well in international competitions that test the accuracy of algorithms using unpublished crystal structures or computational models [293, 296].

4.1.2 Compuational structure prediction

The ability to predict protein structure from primary sequence computationally can be a powerful tool, especially where experimental structure determination is difficult and time-intensive. Traditionally, these methods are divided into three groups [297]:

- Comparative modelling based on known structures with high homology,
- Threading methods, that identify evolutionary unrelated templates that have a similar structure, and
- *Ab initio* modelling which starts from first principles.

However, the boundaries between these three independent methods continue to blur as composite approaches have been shown to be highly advantageous in recent studies [297]. The quality of models generated by any method, largely depends on the difficulty of the target, i.e. on the degree of homology to known structures and the size of the target itself [41]. Here two web servers were tested to predict the structure of the proposed chimeric proteins: Robetta and I-TASSER. While Rosetta, the underlying software used by the Robetta server, is mainly based on a comparative modelling algorithm, which has recently been improved to include co-evolution data for unknown structures [298–301], I-TASSER (Iterative Threading ASSEmbly Refinement) uses threading methods [297, 302, 303]. Both of them also have *de novo*/*ab initio* prediction methods, which are applied to loop regions and are otherwise limited to small proteins of around 120 residues or smaller.

Robetta

Robetta is a fully automated webserver based on the Rosetta software. It splits the target sequence into domains and models them using either comparative or *de novo* modelling, depending on the degree of homology to known structures. Domains are predicted using the Ginzu protocol which uses a hierarchical screening procedure to identify domains with homologues in the PDB [299]. Domains with high homology are modelled using the comparative modelling, RosettaCM [298], while un-assigned regions of the template, e.g. loops and termini, are modelled using the Rosetta *de novo* protocol [300]. RosettaCM uses sequence alignments and structural constraints from homologues to optimize a physically realistic all-atom energy function by sampling secondary structure elements from templates combined with local sampling of fragments using a Monte Carlo method-based protocol [298, 304]. After modelling of each domain is complete, they are iteratively assembled starting from the N-terminus followed by full-atom side-chain and backbone refinement using the Rosetta energy function [298, 299, 304]. The Rosetta software has been used to create a series of *de novo* consensus repeat proteins based on different repeat families [137] and also to design helical-repeat types that are unrelated to any known sequence and can be used to precisely alter super-molecular geometries [7].

I-TASSER

Queries submitted to the I-TASSER webserver are first threaded through a non-redundant structure library to identify possible templates, that is the query is cut into 5 residue fragments which are then placed into any given structure to detect common folds [302]. This alignment process is conducted using 8 different programs, the results from which are then scored and ranked based on the goodness of alignment. The Z-score, an energy-based score that is relative to the average of all alignments [297], can be used to identify whether the target is easy (homologues exist) or hard (no homologous structures). The sequence is then divided into template aligned and unaligned regions (which are modelled *ab initio*), and reduced to C_α atoms and side-chain centre of masses [297, 302]. The model is assembled from the threading alignments using a Monte Carlo method that is guided by statistical terms derived from the PDB, spatial restraints from templates and sequence-based contact predictions [297]. The lowest energy conformations are identified by clustering. The cluster centroids are re-submitted to the assembly algorithm to remove steric clashes and to refine the global topology [297, 302]. Models from this step are clustered again and the lowest energy structures are built to full-atom models by optimizing the hydrogen bonding network [297]. The model quality is assessed using the C-score, which is derived from the quality of threading alignments and the convergence of the simulation, i.e. the clustering density [297, 302]. The C-score correlates strongly with widely used benchmarking measures for the comparison of computer derived models to actual structures, thereby providing the lay modeller with an easy evaluation measure [303]. By definition of its force-field, I-TASSER is biased towards single-domain globular proteins, but multi-domain proteins can be modelled independently followed by docking [297].

4.2 Methods

4.2.1 Geometry calculations

Curvature, twist and lateral bending were calculated for PR65, using chain A from 3dw8 [209], and Importin-β, using chain A from 1qgk [305], as described in Section 3.2.1. In both cases the flexible termini and inter-repeat loops were excluded from the calculations to ensure the correct assignment of principal components that represent the repeat orientation [7]. Exact residues are given in Appendix A.2.

Table 4.1: Nomenclature of PR65-Importin-β chimeras, where PR65 is full length and Importinβ residues refer to the respective PDB structures.

Protein	Repeats	Residues		WT C-cap	Designed
	added	3nd2 [6]	1qgk [305]		interface
PR65-ImpβHEAT19	1	674-861	831-876	c1WT	c1int
PR65-ImpβHEAT18-19	2	715-861	786-876	c2WT	c2int
PR65-ImpβHEAT17-19	3	768-861	732-876	c3WT	c3int
PR65-ImpβHEAT16-19	4	818-861	686-876	c4WT	c4int

Table 4.2: The designated active residues specified for HADDOCK. PR65 H7 and H8 refer to a control run where PR65 on its own was split in half.

Protein/Fragment	Hydrophobic core	R-D ladder
PR65 C-cap	K561L, E565L, T568L, V573, Q580A	K576
ImpβHEAT19	I818, V827, I831, A848, A851	-
ImpβHEAT18	L772, I779, F782, A801, L805, M812	D789
ImpβHEAT17	I724, L727, A731, A753, A757, I761	E737
ImpβHEAT16	M681, L685, M688, L698, V705, I709	E697
PR65 HEAT7	M245, L248, A252, A264, F267	R260
PR65 HEAT8	I278, L283, V298, F309	D293

4.2.2 HADDOCK

The structures were submitted to the "Easy Interface" (`https://haddock.science.uu.nl/services/HADDOCK2.2/haddockserver-easy.html`). For PR65 two conformations were used, the apo crystal (1b3u, residues A2-A589 [16]) and chain A from a PP2A complex (3dw8, residues M8-A589 [209]). The C-cap interface mutations were introduced manually using Coot [306]. Up to four Importin-β repeats (see Table 4.1) were taken from the yeast Kap95 structure 3nd2 and docked to PR65 with and without C-cap interface mutations. In both structures, those residues that clearly point towards the hydrophobic core were chosen to be active, as well as charged residues that (may) continue the R-D-ladder (see Table 4.2) [16]. Passive residues were set automatically, which thereby include any residues within 6.5 Å of an active residue [292].

HADDOCK scores are calculated after each of the three docking stages (rigid body (it0), semi-flexible refinement (it1) and explicit solvent (water)):

$$HS_{it0} = 0.01E_{vdw} + 1.0E_{elec} + 1.0E_{desol} + 0.01E_{air} - 0.01BSA, \tag{4.1}$$

$$HS_{it1} = 1.0E_{vdw} + 1.0E_{elec} + 1.0E_{desol} + 0.1E_{air} - 0.01BSA, \tag{4.2}$$

$$HS_{water} = 1.0E_{vdw} + 0.2E_{elec} + 1.0E_{desol} + 0.1E_{air}, \tag{4.3}$$

where E_{vdw} corresponds to the van der Waals intermolecular energy, E_{elec} is the electrostatic intermolecular energy, E_{desol} the desolvation energy, E_{air} the distance restraints energy (for unambiguous and ambiguous interaction restraints) and BSA the buried surface area [292]. Within the "Easy Interface" the top 200 docking results (or models) are submitted for automated clustering based on l-i-RMSD, named according to cluster size and ranked according to the average HADDOCK score of the four best scoring cluster members [292].

The l-RMSD describes the backbone variation of the docked "ligands" after the docking results have been aligned using the "protein", accepted values lie between 5 and 10 Å [293]. The i-RMSD describes the backbone variation of interface residues after their superposition and here acceptable values lie between 2 and 4 Å [293]. The i-l-RMSD is similar to the i- and l-RMSD but is calculated from only those backbone atoms that are at the interface, resulting in values somewhere between i- and l-RMSDs [294, 295].

4.2.3 Robetta

The primary sequence containing PR65 residues with/without interface mutations and the respective Importin-β residues were submitted in FASTA format to the Robetta server online (`http://robetta.bakerlab.org/submit.jsp`). Ginzu domain predictions were accepted and submitted to the server. Final models in PDB formats were downloaded and used for subsequent analysis. Backbone RMSDs were calculated using the pymol align command without a cut-off (to include all backbone atoms). The alignment score to the parent structure was recorded as a measure of homology.

4.2.4 I-TASSER

The primary sequence containing PR65 residues with/without interface mutations and the respective Importin-β residues were submitted in FASTA format to the I-TASSER server online (`https://zhanglab.ccmb.med.umich.edu/I-TASSER/`). No other options (e.g. additional restraints, templates, etc.) were selected or provided. The results were downloaded from the website after completion. Normalised Z-scores of the best 10 LOMETS threading samples were averaged and C-scores of the best, folded structures were recorded. Normalised Z-scores >1 identify a good alignment, the higher the Z-score, the better the alignment and the easier the target. The C-score is an estimate for the quality of the predicted models and is based on the alignment and convergence of the simulations, they usually vary between -5 and 2 where larger scores (e.g. -1.5) correspond to models with high confidence.

Table 4.3: Cloning details of the PR65-Importin-β chimeras using the human Importin-β (hImpβ) sequence. PR65 is full length. Predicted molecular weights (MW) were estimated using the Expasy ProtParam Tool. Nomenclature as defined in Table 4.1.

Chimera	hImpβ residues	H_6-chimera [kDa]	GST-chimera-H6 [kDa]
c1WT/c1int	831-876	71	98
c2WT/c2int	786-830	76	102
c3WT/c3int	732-785	82	108
c4WT/c4int	685-731	87	113

4.2.5 Experimental procedures

The solvent-exposed surface of the last PR65 repeat was converted to an interface by introducing the following mutations using RTH-SDM: K561L, E565L, T568L and Q580A. The H6-Importin-β (Impβ) gene in a pET28 vector was a kind gift from Alan Lowe (Birkbeck/UCL, UK). It was used to create the PR65-Impβ chimeras detailed in Table 4.3 using the FastClone method [225]. C-terminal Impβ repeats were added to the C-terminus of PR65 (with and without interface mutations) in frame, maintaining the original N- to C-terminal repeat order of both proteins. The protocol described by Li *et al.* [225] was followed closely, using Phusion High-Fidelity DNA polymerase (ThermoFisher or New England Biolabs). Both N-terminally H_6-tagged chimeras (pET28a) and GST fusion proteins with C-terminal H_6-tags (pRSETa) were created.

Small scale expression tests

Transformed cells were used to inoculate 10 ml of 2xYT media with the appropriate antibiotics. Cultures were grown and induced in the same way as large scale cultures (see Chapter 2, Section 2.3), and 1 ml of final culture was used for further analysis. Cells were pelleted in a microfuge tube at maximum speed, re-suspended in 300 µl BugBuster® Master Mix (Merck Chemicals), and incubated at room temperature for 30 min. The soluble and insoluble fractions were separated by centrifugation at maximum speed, soluble fractions were isolated and the insoluble pellet was washed with 1 ml of 10 % BugBuster® before re-suspending it in 100 % BugBuster®. Total cell protein, soluble and insoluble protein fractions were analysed by SDS-PAGE.

Far-UV Circular Dichroism spectroscopy

Far-UV spectra were taken from 300 µl samples in a 1 mm quartz cuvette using a Circular Dichroism (CD) spectrophotometer (Applied Photophysics). PR65 and chimeras were diluted to concentrations between 0.5 and 1 µM using 25 mM MES pH 6.5, 1 mM DTE.

The mean residue molar ellipticity, $[\theta]_{MRW}$, can then be calculated using

$$[\theta_{MRW}] = \frac{\theta}{Nlc}, \tag{4.4}$$

where θ is the ellipticity in millidegrees, l the path length, c the concentration of the sample and N the number of residues.

Temperature denaturation monitored by circular dichroism

N-terminal H6-tagged PR65 WT and c1 chimeras, with (c1int) and without (c1WT) C-cap to interface conversion, were buffer exchanged into 25 mM MES, 1 mM DTE and diluted to 1 µM final concentration. Far-UV CD spectra were collected between 204 nm and 265 nm in a 0.1 cm path length cuvette using a Chirascan$^{\mathrm{TM}}$ spectrometer (Applied Photophysics). Samples were subjected to a temperature gradient from 25°C to 86°C. Since the heat capacity of PR65 is unknown, the normalised ellipticity at 222 nm, θ_{norm}, with respect to temperature, T, was fitted using a Sloppy Boltzmann to extract the apparent temperature mid-point, T_m [307]:

$$\theta_{norm} = \theta_N + \frac{\theta_U - \theta_N}{1 + \exp\left(\frac{T_m - T}{n}\right)}, \tag{4.5}$$

where θ_N and F_U are the ellipticity of the native and denatured state, respectively, which can be described by

$$\theta_N = \alpha_N + \beta_N T \tag{4.6}$$

$$\theta_U = \alpha_U + \beta_U T, \tag{4.7}$$

and n is the slope of the transition.

Urea induced denaturation

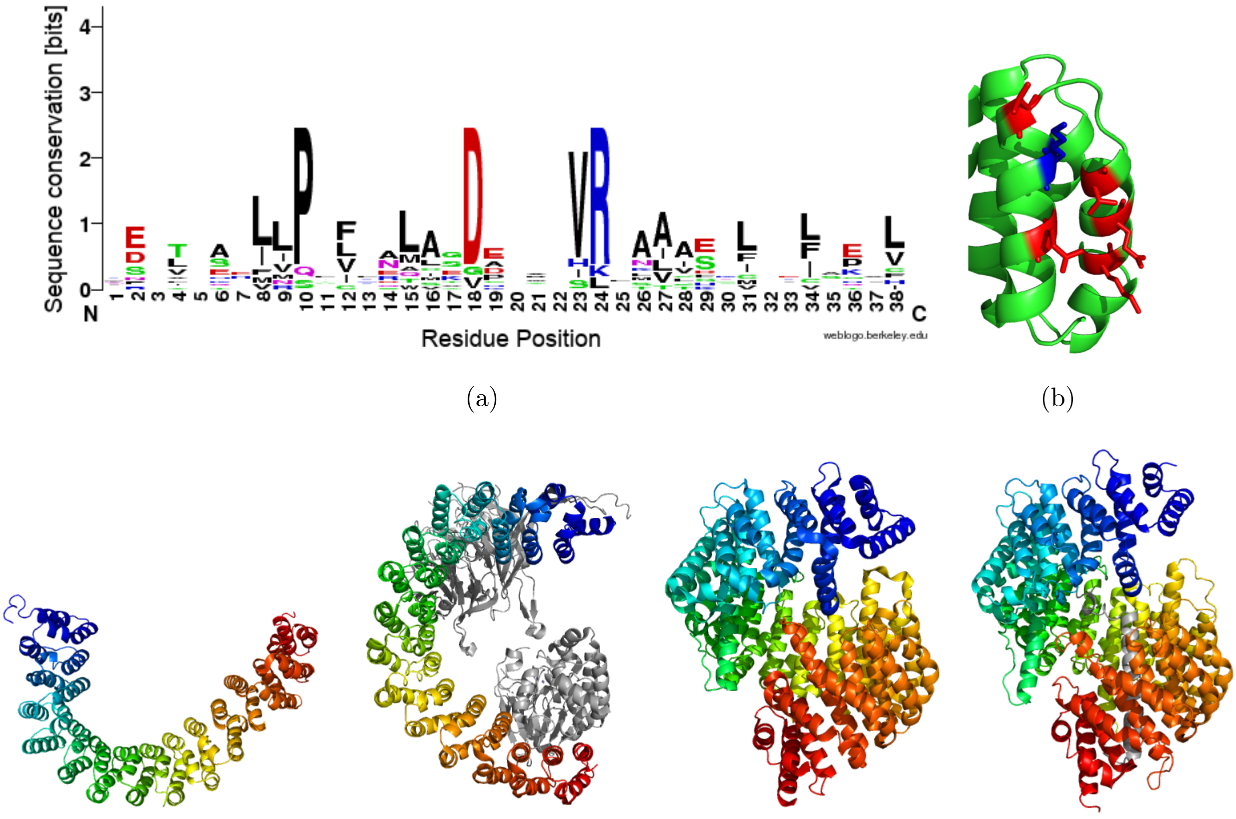
Samples of a total volume of 150 µl were prepared in a 96-well format, with urea gradients of 0-8 M in approximately 0.1 M intervals. The final protein concentrations were 1 µM, 0.25 µM and 0.45 µM for PR65 WT, c1WT and c1int, respectively. Samples were incubated at 25°C for 2h and the fluorescence at 360 ± 10 nm was measured using a CLARIOStar microplate reader (BMG Labtech).

4.3 Results

4.3.1 Structure prediction of the PR65-Importinβ chimeras

Sequence analysis of an alignment of internal PR65 repeats (i.e. excluding N- and C-capping repeats) reveals a distinct conservation of leucine and alanine residues (Fig. 4.2a) which form the hydrophobic core [17]. Those residues that are conserved in both internal and capping repeats, mostly correspond to the interface between repeat H14 and H15 in the PR65 C-cap. Polar and charged residues on the outer face of the C-cap were substituted to the hydrophobic consensus (K561L, E565L, T568L, Q580A, see Fig. 4.2b) to create a new interface between the C-cap and future adjacent repeats. Chimeras with these interface mutations and without (i.e. original C-cap) were carried forward in both simulations and experiments, to provide an intrinsic control, as the original C-cap sequence is more likely to hinder folding than that containing interface mutations.

Since server-based modelling programs as well as the physical generation of chimeras can take considerable time, most experimental and computational methods proceeded in



(a)

(b)

(c) 1b3u

(d) 3dw8

(e) 3dn2

(f) 1qgk

Figure 4.2: Designing PR65-Importin-β chimeras. (a) Sequence alignment of internal PR65 repeats, to obtain the consensus residues. Generated using the WebLogo server [308, 309] (b) View onto the C-terminal capping repeat of PR65, where residues in red and blue are the respective hydrophobic and charged amino acids which are set as active residues. Bottom row: Structural representations of PR65 (c) [16], PR65 bound to B- and C-subunits (d) [209], Kap95/yeastImportin-β (yImportin-β, e) [6] and human Importin-β (hImportin-β) bound to the IBB of Importin-α (f) [305]. All repeat proteins are coloured from blue (N-terminus) to red (C-terminus), while other subunits and binding partners are coloured in grey.

Table 4.4: Average and cumulative angles between HEAT repeats in Importin-β and PR65 as described in Chapter 3, Section 3.2.1. Values for Importin-β are similar to previously published results by Forwood *et al.* [6] and only differ in the orientation of bending. The cumulative angles of Importin-β were normalised to 15 HEAT repeats to facilitate comparison with PR65.

	Importin-β		**PR65**	
	$\bar{x}$	$\frac{15}{N}\sum x$	$\bar{x}$	$\sum x$
Curvature [°]	19 ± 5	267	12 ± 2	171
Twist [°]	10 ± 7	135	-5 ± 2	-66
Bending [°]	-7 ± 2	-98	-1 ± 2	-19

parallel. Important to note is, that HADDOCK modelling was completed with the yeast Importin-β (PDB id 3nd2) before any experimental work was carried out, as it is the only structure available that is not bound to another protein. When we obtained a plasmid containing human Importin-β all subsequent modelling which did not require a structure (Robetta and I-TASSER) was performed using the human sequence. The docking was not repeated using a human structure, because

a. 3nd2 aligns with the hImp-β structure 1qgk (bound to IBB domain of Importin-α) to 2.49 Å which is close to the resolution of either structure, being 2.4 Å and 2.5 Å, respectively,

b. sequence identity between both is >30% and sequence similarity is >50%,

c. and the secondary structure content agrees to 80-90% between both proteins.

The alignments were performed using the EMBL-EBI server PDBeFold [310] and RCS-Balign using the jCE algorithm [311].

As can be seen from Figure 4.2c-f, overall geometries of PR65 and Importin-β are quite different: while PR65 is a planar molecule, Importin-β forms a supramolecular helix. To examine how much they differed locally, angles between the repeats were calculated (Table 4.4). Although the total curvature of Importin-β is 1.5 times than that of PR65, their average curvatures are within error. However, in both twist and lateral bending they differ significantly: (a) Importin-β has a significant negative lateral bend, whereas PR65 bends only very little, and (b) the twist of Importin-β is positive, while PR65 repeats twist in the opposite direction.

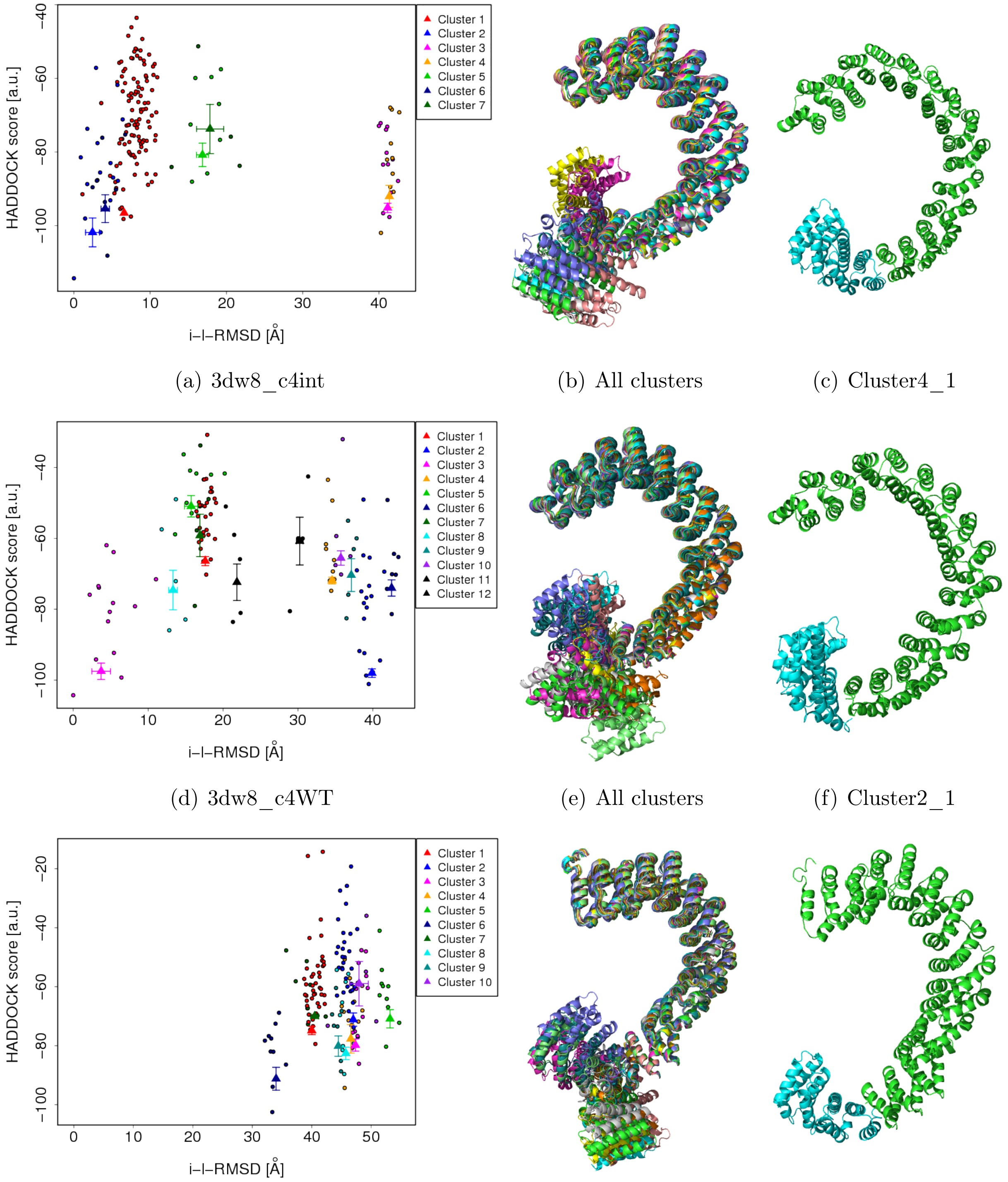
Modelling of the chimeric interface using HADDOCK

Here, PR65 (with and without interface mutations) was taken as the "protein", while respective Importin-β repeats were considered the "ligand" (Table 4.1). Residues that could be involved in hydrophobic or electrostatic interactions were chosen as active residues (Table 4.2). Due to the break in the HEAT repeat array of the PR65 apo-structure (1b3u), I chose to use the PR65 chain from a PP2A structure (3dw8) as my main reference. The variation between the C-terminal repeats of each structure lies between 0.62 Å (backbone only) and 1.038 Å (all residues). However, as an additional control 1b3u with interface mutations was submitted as well, leading to a set of three PR65 structures submitted for docking:

- 3dw8 with interface mutations (3dw8_int)
- 3dw8 without interface mutations (3dw8_WT)
- 1b3u with interface mutations (1b3u_int)

The output from the HADDOCK web server was analysed both qualitatively by classifying the protein-ligand orientations, and quantitatively using HADDOCK scores. An example of the HADDOCK output is given in Figures 4.3 for the docking of Importin-β repeats HEAT16-19 to all three PR65 templates. The web server produces plots of HADDOCK score versus i-l-RMSD (Figure 4.3a,d,g) which were used for clustering and ranking. When Importin-β HEAT16-19 was docked onto 3dw8_int, the majority of models from multiple clusters fall within the accepted range for interface and ligand RMSDs (Figure 4.3a). In the absence of the interface mutations, clusters are spread across a range of RMSDs (Figure 4.3d). In the case of 1b3u_c4int (Importin-β HEAT16-19 docked onto 1b3u_int, for definition of nomenclature see Table 4.1), all clusters are spread across RMSDs beyond those considered acceptable (Figure 4.3g).

Figure 4.3 also shows the best models of each cluster aligned using the PR65 domain, as well as one docking result in which PR65 and Importin-β adopt a conformation appropriate for HEAT repeats. As can be seen from the multiple alignments of all clusters (Figure 4.3), there is a large variety of orientations when the Importin-β repeats are docked onto any PR65 template. In some models, the final PR65 repeat and first Importin-β repeat are aligned "correctly", that is the C-terminus of PR65 and the N-terminus of Importin-β are on the same side of the repeat array and could adopt this conformation in an actual fusion protein. In other models, these repeats were "inverted", i.e. PR65 C-terminus and Importin-β N-terminus are on opposite sides of the repeat array. Sometimes the first Importin-β repeat was found at "right angles" with the last PR65 repeat (i.e. the α-helices stacked at angles of 90° instead of in a parallel or antiparallel fashion), while other Importin-β fragments did not dock at the interface at all but instead docked to the side fo the C-cap. In the 3dw8_c4int run models of the 4th cluster adopt a correct conformation (Figure 4.3c). The large variation in RMSDs seen for 3dw8_c4WT is clearly represented in the structural alignment (4.3d,e) and even a model

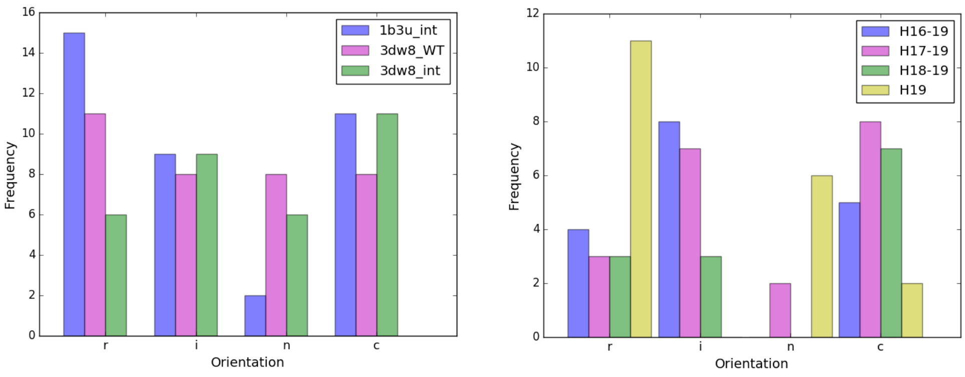


(a) 3dw8_c4int (b) All clusters (c) Cluster4_1

(d) 3dw8_c4WT (e) All clusters (f) Cluster2_1

(g) 1b3u_c4int (h) All clusters (i) Cluster9_1

Figure 4.3: Overview of HADDOCK output of c4 chimeras based on different PR65 templates: 3dw8_c4int, 3dw8_c4WT and 1b3u_c4int (from top to bottom). The first panel shows the i-l-RMSDs of all 200 models which are used by HADDOCK for clustering. In the middle panel, structural representations of the fist model from each cluster are aligned using the PR65 template. The right panel shows models that exhibited the right docking orientation with the lowest score.



(a)

(b)

Figure 4.4: Analysis of all HADDOCK results. Models from all docking runs were sorted according to their docking orientation: "r" - right angle, "i" - inverted, "n" - not at interface, "c" - correct. (a) Bar graphs of these data grouped according to PR65 template and (b) the same data grouped according to Importin-β repeat.

from the top scoring cluster with the correct orientation does not completely form an interface (4.3f). This variation is slightly less for 1b3u_c4int, but the interface is formed by the two repeats adopting an abnormal angle (4.3h-i). In all cases, the clusters in which the interface adopts a HEAT-repeat geometry and which score as high as possible, have some of the worst RMSDs. Furthermore, HADDOCK scores of the best five (3dw8_c4int and 1b3u_c4int) or 2 (3dw8_c4WT) clusters are within one standard deviation of each other and contain a mixture of orientations.

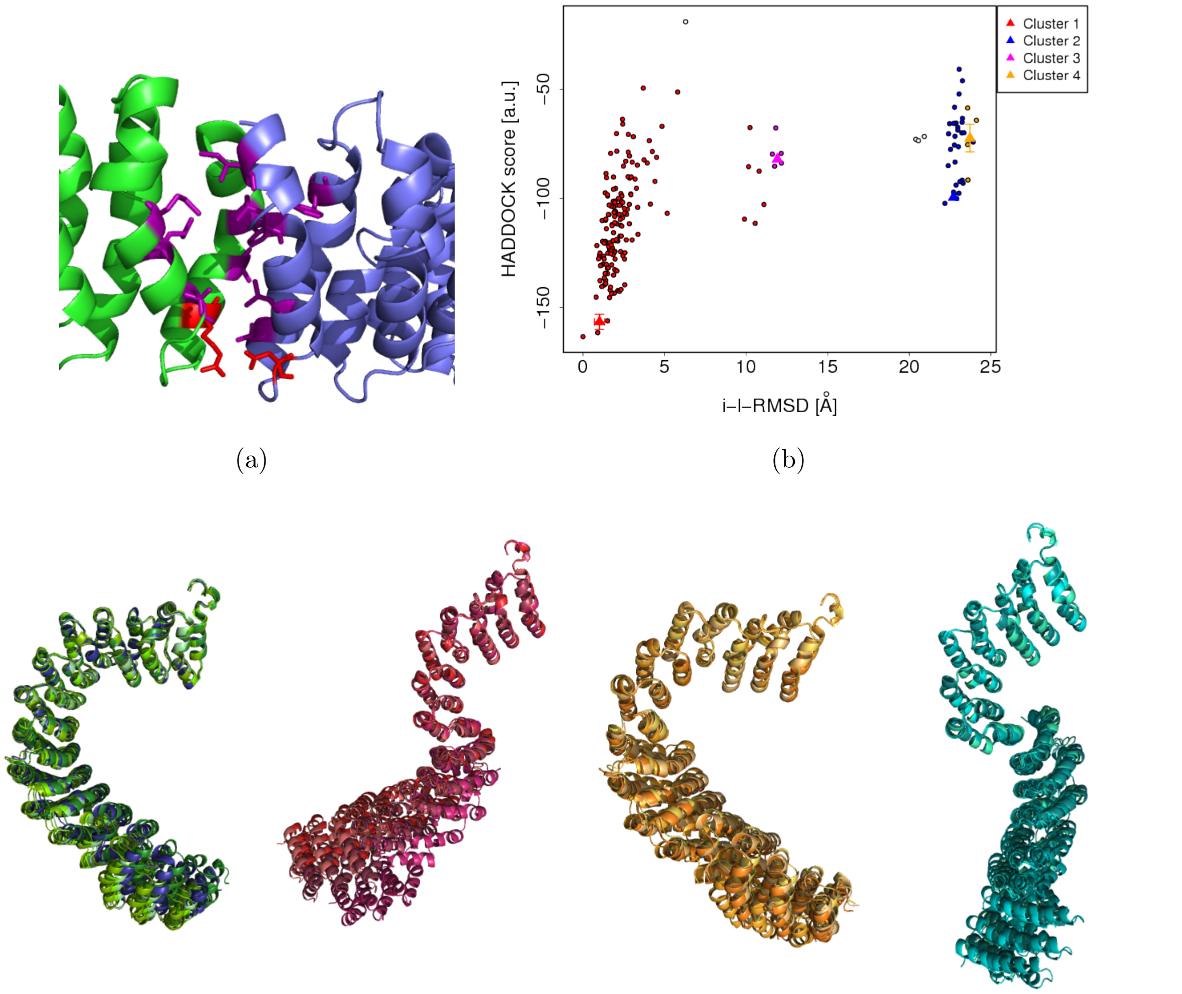
Statistical tests were performed to examine whether the HADDOCK score correlates with a particular orientation, template or Importin-β HEAT repeat. First, the data were grouped according to PR65 template to examine whether there is a preference of orientation depending on the template (Figure 4.4a). Both inverted and correct orientations were slightly favoured more by templates with interface mutations as opposed to 3dw8_c4WT. Most models based on 1b3u_int docked at right angles, but otherwise no clear preference can be seen. The fraction of correct orientation for each template are 30%, 23% and 34% for 1b3u_int, 3dw8_WT and 3dw8_int, respectively. Second, the data was grouped according to which Importin-β fragment was used for docking (Figure 4.4b). HEAT17 and HEAT19 were the only ones found to ever dock away from the specified interface. HEAT19 preferentially aligns at right angles with PR65, while models of HEAT18 docked onto PR65 were mostly correct. The fraction of correct orientations for each fragment were 29%, 40%, 54% and 11% for HEAT16-19, HEAT17-19, HEAT18-19 and HEAT19, respectively. Next, it was tested whether there was any correlation between HADDOCK score and orientation within a specific grouping. Across all runs, the HADDOCK scores lie between -20 and -140, and average to around -90. Statistical tests were performed within each grouping (PR65 and Importin-β templates) which indicate that there is no significant difference in HADDOCK score between any orientation in either data set.

Since the HADDOCK score is energy-based, it must somehow relate to the prediction of a physically probable interaction between two structures. However, none of the above analyses could determine whether the score in conjunction with RMSD is a good predictor for these chimeras. Furthermore, the assignment of the number of active residues is somewhat arbitrary and relies on the correct identification of residues that are involved in interface formation. As a first control, residues at the side of PR65 were chosen as active residues (R527, K519, A480, A488, P483). The resulting docking models were not significantly different in score and RMSD compared to models where interface residues were set as active. As a second control, PR65 was split into two halves between HEAT7 and HEAT8. The active residues on both HEAT7 and HEAT8 were selected using the same rationale as before (see Table 4.2, Figure 4.5a), but were also verified as hydrophobic core and R-D-ladder residues using published data [16]. Figures 4.5b shows the i-l-RMSD for this control run, which display very successful clustering in comparison to the docking of chimeras. All clusters are clearly separated from each other and, apart from the lowest scoring cluster, their standard deviations do not overlap. Furthermore, the majority of models are in the first cluster which has an i-RMSD of < 2 Å and an l-RMSD of < 10 Å and, most importantly, these models adopt the correct orientation which only differ from the starting structure (1b3u) by 1.1 Å (Figure 4.5c,g). In contrast, the 2nd and 4th clusters are inversions of the C-terminal half, while the C-terminal half is only slightly rotated in the 3rd cluster. These orientations are reflected in the large RMSDs from the best structure in the first cluster, and thereby the starting structure (Figure 4.5g). Yet, clusters 2 to 4 contain only 22% of all docking runs. Within the top four structures of the first cluster, the best scoring models have a rotation of one of the active residues on either the C-terminal (I278) or the N-terminal half (M245). Otherwise, the orientation of interface residues is quite similar to the starting structure. That is in 78% of the models HADDOCK could reproduce the starting structure.

Homology model based structure prediction from primary sequences using I-Tasser and Robetta

Given the HADDOCK results and considering that docking is still somewhat limited in modelling structural changes that affect the whole complex, the sequences of all chimeras were submitted to Robetta and I-TASSER web servers to predict their tertiary structures.

The output from the Robetta webserver is limited to PSI-BLAST alignments with their associated homology score and five 3D models of the predicted structure (Table 4.5 and Figure 4.6). Unfortunately, there is no quantitative assessment of model quality, as models with their respective Rosetta Energy Functions are not accessible via the web-



(a)

(b)

(c) Cluster 1

(d) Cluster 2

(e) Cluster 3

(f) Cluster 4

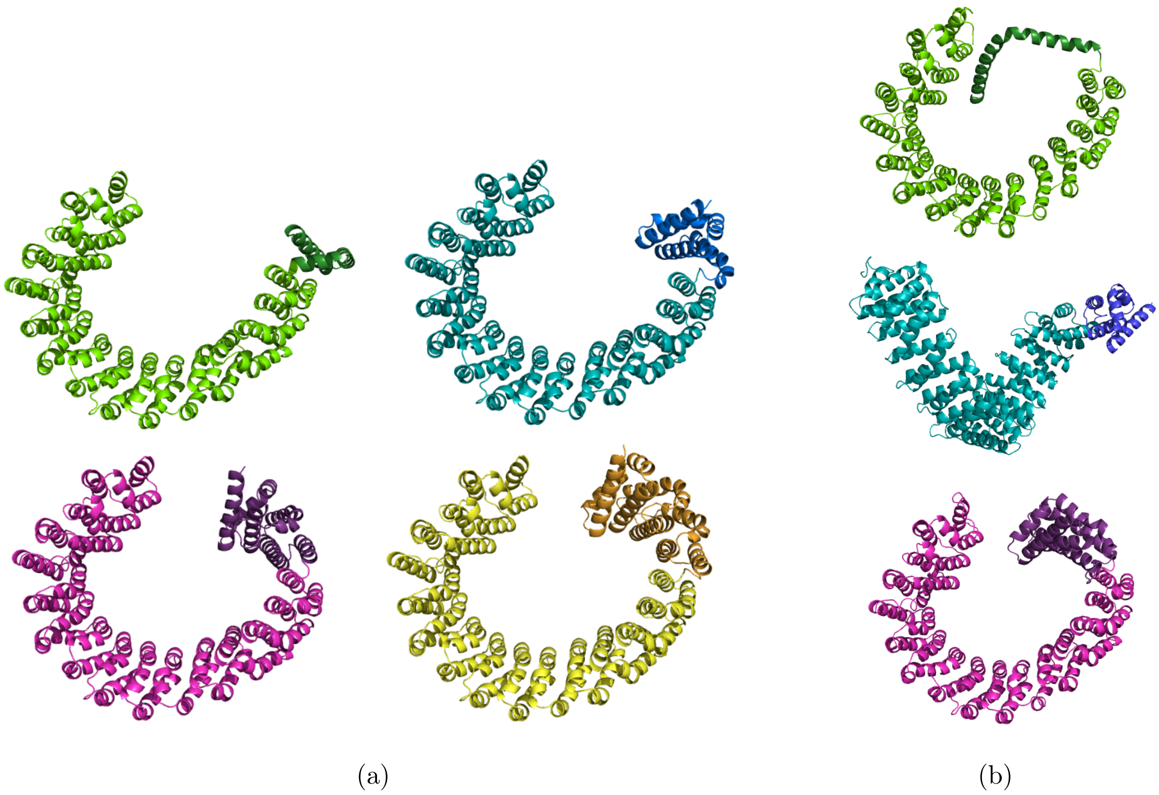
Cluster	Size	Score	RMSD [Å]
1	156	-157 ± 7	1.1 ± 0.8
2	31	-100 ± 2	21.1 ± 0.9
3	5	-82 ± 3	6.5 ± 0.2
4	4	-72 ± 13	22.8 ± 0.6

Complex	Score	RMSD [Å]	Rotation
1	-163.32	1.82	I278
2	-161.52	2.01	M245
3	-156.05	1.59	-
4	-145.40	1.11	-

Figure 4.5: Design and HADDOCK results for the PR65 control split between H7 and H8, that is artificially cut between residues 275 and 276. (a) Structural detail of the H7 (green) - H8 (blue) interface. Hydrophobics are coloured in purple while RD-ladder residues are coloured in red. (b) i-l-RMSDs and clustering of the PR65 control run. (c)-(f) Structural representations of the 4 clusters where N- and C-termini of PR65 are arranged top to bottom. Left table: Haddock scores and RMSDs to the overall lowest energy structure of all clusters. Right table: Haddock scores and RMSDs to PR65 (blue in (c)) of the top 4 structures in cluster 1, also indicating the observed side chain rotations in the models compared to the starting structure.

Table 4.5: Robetta alignment output and model information. The best folded models were aligned to the parent 2iae (resolution 3.5 Å).

Chimera	Parent structure	Score	Folded models Any(PR65-like)/Total	Best model	RMSD [Å]
c1WT	3w3v	0.8	1(0)/5	-	-
c1int	2iae/2qna	0.76/0.76	3(3)/5	m1	3.598
c2WT	2iae/2qna	0.76/0.99	2(0)/5	-	-
c2int	2iae/2qna	0.76/0.99	5(3)/5	m1	3.293
c3WT	2iae/1qgr	0.75/0.99	1(0)/5	-	-
c3int	2iae/1qgr	0.73/0.76	4(1)/5	m4	3.412
c4WT	2bpt	0.31	5(4)/5	-	-
c4int	3w5k	0.58	5(4)/5	m1	2.799
PR65	5ve8	0.79	5(5)/5	m1	1.856



(a)

(b)

Figure 4.6: Robetta models of chimeras: (a) Fully folded, PR65-like models, (b) most common misfolded or Impβ models. Colouring is as follows: green - c1 chimeras, blue - c2 chimeras, magenta - c3 chimeras, yellow - c4 chimeras. Structural parts in lighter shades correspond to PR65 repeats, while darker shades are the Importin-β repeats.

server. Almost all homologues identified by Ginzu were either existing PR65 (e.g. 2iae a crystal structure of a PP2A complex) or Importin-β structures (e.g. in complex with a binding partner as in 2qna, 1qgr, 3w5k). The c1WT chimera (for nomenclature see Table 4.1) and PR65 WT were the only structures that were modelled based on entirely different protein: Kap121 and Kap123, two yeast karyopherins containing HEAT repeats. In all but PR65, c1WT and both the c4WT and c4int chimeras, Ginzu predicted two different domains and hence PR65 and Importin-β residues were first modelled as independent domains before final energy minimization in the complex. Homology scores were usually > 0.7, signifying a good alignment. However, the c4 alignments only scored 0.31 and 0.58 due to them being aligned as a single domain, in which the PR65 portion will dominate the alignment to the Importin-β parent. When the WT C-cap residues were submitted, only 45% of models were folded on average, as compared to 85% when the models contained the interface mutations. Examples of "misfolding" are shown in Figure 4.6b, where either the Importin-β repeat did not fold or the interface failed to form correctly. Folded models adopted the shapes of both PR65 like structures, that is planar HEAT repeat arrays (Figure 4.6a), or Importin-β-like structures, i.e. super-helical HEAT repeat arrays (Figure 4.6b). In the top folded and PR65-like models (Figure 4.6a), the Importin-β repeats form an extension of the PR65 HEAT repeat array. Their PR65 domains differ from the parent 2iae by 2.8 - 3.6 Å, which is close to the original resolution of the parent (3.5 Å). Although the WT control of PR56 was modelled against an unrelated parent, all five models were fully folded and PR65-like, the best of which had an RMSD variation from 2iae of just under 2Å.

Table 4.6: Overview of I-TASSER results. Detailed are the Z-score (which gives information on the difficulty of the target), threading templates (where Imp-β can be be from any organism), number of folded models and C-scores (which are a measure of model quality).

Chimera	Z-score[a]	Parent Structure (PR65/Imp-β/others)	Folded Models Any(PR65-like)/Total	C-scores[b]
c1WT	4.252	10/0/0	1(1)/5	0.0
c1int	4.333	10/0/0	4(4)/5	0.42
c2WT	2.833	7/1/2	0/5	-
c2int	2.787	7/0/3	0/5	-
c3WT	2.613	6/4/0	1(0)/5	-0.95
c3int	2.728	6/3/1	2(1)/5	0.17
c4WT	4.11	8/2/0	1(0)/5	0.17
c4int	4.082	8/2/0	1(0)/5	0.16
PR65	4.418	10/0/0	2/2	1.3

[a] Average Z-score of top 10 threading templates

[b] Best C-score among fully folded models

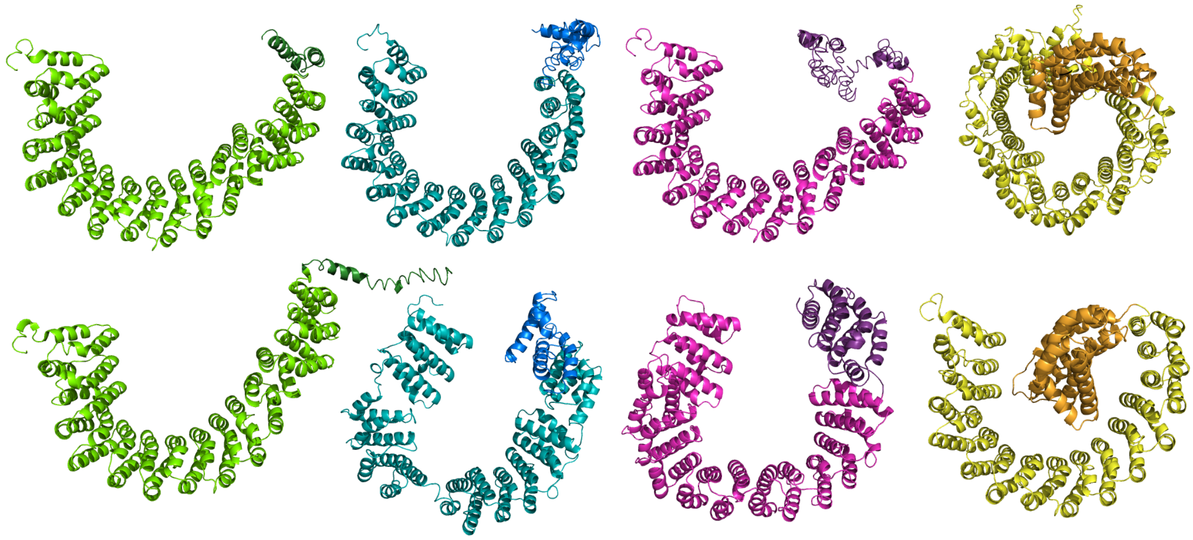


Figure 4.7: Structural representations of I-TASSER models giving examples of the variety of folding and misfolding events. The top row are models of WT chimeras, while the bottom row are models of interface chimeras. The colour scheme is the same as in Figure 4.6.

Of the predictions provided by the I-Tasser web server only 15% of WT-cap models and 35% of chimeras with mutated C-cap were folded on average (Table 4.6). In most cases the PR65 domain formed independently of the Importin-β repeats (Figure 4.7). These were either (a) totally unfolded, (b) independently folded with Importin-β-like geometries, or (c) in a state similar to a collapsed coil. If collapsed, they could be connected to PR65 by a flexible linker or have some contacts with the interface. In those cases where the models were fully folded, 40% were Importin-β-like. If a model resembled the shape of PR65, it usually exhibited some sort of odd packing topology. For example, in folded c1 models the Importin-β repeat was not tightly packed against the PR65 domain. In the example of c3int, the first helix of the 17th Importin-β repeat was stacked into the interface.

Folding did not correlate with Z-score, that is quality of the multiple sequence alignment to prospective parents. The PR65 control scored $Z > 4$ but so did some of the chimeras where only one model was folded (Table 4.6). However, in comparison to PR65, models of all chimeras received C-scores smaller than 1, most of them being either close to or below zero. C1int, which scored closer to 0.5 resulted in 4 folded models. For the PR65 control, which scored $C = 1.3$, the server only returned two models, a good indication that the simulations converged well, which in turn results in models of higher quality.

4.3.2 ENM of chimeras

To confirm that an extension of PR65 resulted in a reduction of frequency, ENMs were performed. The models ranked first by Robetta, which were folded and PR65-like, were

Table 4.7: ENM mode frequencies of chimeras.

	PR65	**c1int**	**c2int**	**c3int**	**c4int**
Mode 7 [cm^{-1}]	0.38	0.26	0.23	0.19	0.18
Mode 8 [cm^{-1}]	0.76	0.46	0.36	0.28	0.27
Mode 9 [cm^{-1}]	0.79	0.66	0.56	0.35	0.38

taken forward for alignment calculations and ENMs. Since these conditions excluded all but one chimera with WT C-cap, only chimeras with interface mutations were carried forward. The frequencies of normal modes were reduced significantly: an extension by 4 repeats resulted in a reduction of more than 50% for modes 7-9 (Table 4.7). The reduction of frequency with increasing number of repeats appears to be non-linear.

4.3.3 Cloning and expression of chimeras

The chimeric clones (with and without C-cap-to-interface mutations) were designed with N-terminal H_6-tagged constructs and also with N-terminal GST-fusions and C-terminal H_6-tag. There are two reasons for testing constructs with different tags:

1. Switching WT PR65 from an N-terminal H_6-tag to N-terminal GST-tag increased the yield of soluble protein per litre of culture by nearly 5 times.

2. All GST-fusion proteins will be close to or larger than 100 kDa. Expression of large proteins in *E. coli* usually results in low yields.

Soluble expression of all GST-chimeras was unsuccessful. The following conditions were screened for soluble GST-chimera expression:

- Cell type: C41, Rosetta, Lemo 21 and MDS42
- IPTG: 0.25 mM, 0.5 mM and 1.0 mM
- Induction temperature: 20°C, 25°C and 37°C,

The results of these expression tests were difficult to reproduce and hence it was not possible to detect clear trends. For example, only in some tests a significant insoluble fraction was present, and in other experiments GST was present after cleavage, but the protein was not. A pattern or dependency on a specific condition could not be found. Originally, the chimeras were designed to have no linking residues between PR65 and Importin-β repeats, as the flexible inter-repeat loops from Importin-β were thought to be sufficient to connect both domains without disturbing the α-helices. To test whether a linker extension by 2 residues would aid expression and solubility, a glycine and serine were introduced by mutagenesis. Expression of GST-tagged constructs with this alteration was repeated in both small and large scale, but remained unsuccessful.

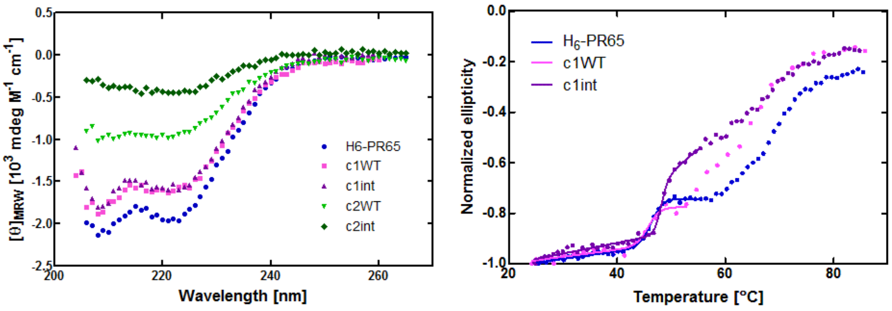
The H_6-tagged chimeras showed sufficient but reduced expression in C41 cells in comparison to the WT. For example, c1WT and c1int expressed nearly at similar levels as H_6-PR65, while c4WT and c4int expressed at considerably lower levels. In most cases, clones with interface mutations were expressed more than clones with a WT C-cap. All chimeras were observed to be more susceptible to degradation during the course of the three step purification process (IMAC, anion exchange, size exclusion). After purification, only c1 and c2 chimeras were obtained with sufficient yield for Far-UV Circular Dichroism Spectroscopy. The yields of c1 chimeras were high enough to verify their molecular weight by mass spectrometry and to subject them to equilibrium denaturation monitored by fluorescence in a 96-well plate format.

4.3.4 Biophysical characterisation of c1 and c2 chimeras

If the Importin-β repeats folded into their correct secondary structure, one would expect the α-helical content of chimeras to increase when compared to PR65. Yet, such a change on its own cannot probe whether or not the Importin-β repeats stack against PR65. The addition of repeats to one end of PR65 is expected to increase its stability, and therefore, denaturation studies can provide insight into whether a new interface has been formed between the added repeats and the PR65 C-cap.

The Far-UV spectra of the c1 chimeras show the characteristic minima of α-helical structures at 208 and 222 nm, indicating that both are folded (Figure 4.8a). However, the signal of both chimeras is considerably less than that from PR65. The c2 chimeras appear to have even less α-helical content: c2WT still shows some α-helical characteristics, but c2int lost the 208 nm minimum. Thermal denaturation was performed to assess whether the overall structure was grossly disrupted or indeed whether stability had been increased upon adding one or two repeats. Denaturation curves of PR65 and the c1 chimeras, and the corresponding temperature mid-points are shown in Fig. 4.8b. The significant loss in α-helicity in c2 chimeras suggests that they are only partially folded and hence I did not perform denaturation studies on these construct. The PR65 WT melting curve shows two transitions, where the first is attributed to the melting of protein monomers while the second one is more likely to arise from protein aggregates [312]. All but c1int have a very similar melting temperature of the first transition compared to the WT. The second transition differs between constructs and could not be fit meaningfully.

At the time these experiments were done, another member of the Itzhaki group had adapted urea-induced denaturations to a 96-well plate format as they required much less protein than experiments done in a 1 ml fluorimeter cuvette [313]. I tested this format



Protein	T_m [°C]
PR65	46.2 ± 0.1
c1WT	45.3 ± 0.7
c1int	48.6 ± 0.1

(a) (b) (c)

Figure 4.8: Circular dichroism spectroscopy of chimeras. (a) Far-UV wavelength spectra and (b) temperature induced denaturations of PR65 and c1 and c2 chimeric proteins monitored at 222nm. The data are representative of at least two independent replicates. The first transition of the thermal melts were fitted using a Sloppy Boltzman to estimate the apparent melting temperature (c).

using the PR65 WT, which unfolds in three states via a hyperfluorescent intermediate [71], and the c1 chimeras. Data from 96-well, half-area plates (150 μl well volumes) did not resolve the first transition well and exhibited poorer signal to noise ratios than seen previously [284].

4.4 Discussion

Our understanding of the protein folding problem has continually improved and so have computational algorithms for structure prediction [41]. There are multiple structural representations of both PR65, Importin-β, and closely related proteins in the PDB that could guide computational design. Importin-β was my first protein of choice as its structure and conformational flexibility upon interaction with other proteins is well characterised (e.g. see references [6, 203, 305, 314–316]). Although overall Importin-β geometry and repeat stacking differ significantly from those observed in PR65, it was unclear how these differences would impact chimeric structures. HEAT repeats show a large structural and sequence variety and have previously been classed into at least three different sub-families [15, 17]. In this classification, PR65 and Importin-β fall into closely related but different HEAT repeat families, which nevertheless have very similar sequence conservation at the hydrophobic core [17].

All three computational methods employed here produced models which predict both properly folded chimeras as well as models in which folding was incomplete. Using HADDOCK it was possible to produce some models of how an interface between PR65 and Importin-β repeats could form. Clustering of models, independent of whether or not they could represent a folded structure, is better with PR65 templates that contain the interface mutations. In the presence of an interface, slightly more models adopted orientations appropriate for the stacking of two HEAT repeats, although no clear correlation to a better HADDOCK score could be detected. The global topology of PR65 from the 3dw8 crystal performs better than the 1b3u structure in both clustering and orientation of docked models, even though the C-cap in both is not significantly different. This could simply be due to the altered topology and different exposure of surface residues which could influence the scoring and clustering at the end of the docking run [292]. There is no clear conclusion as to which Importin-β repeat preferentially results in a physically correct stacking of HEAT repeats. HEAT17 and HEAT18 perform slightly better in clustering and docking orientation, which could be due to more consensus-like residues being present at the interface-forming side of these repeats [305]. Docking of H19 is particularly unsuccessful but this may be because (a) it resembles an Armadillo repeat and (b) it has charged and polar residues that have to be accommodated in the interface. Even though it was possible to draw these few conclusions, their reliability must be questioned considering the variability observed across all docking runs and especially in light of the control docking run. None of the chimeras cluster or score as well as the control. Therefore, the small trends observed for a given PR65 template or Importin-β repeat are likely insignificant in context of the overall performance of docking runs.

In the majority of models predicted by I-Tasser folding was either absent or incomplete, while only few were returned with a properly folded topology. However, the inclusion of interface mutations in the C-cap increased the average number of folded models, of which 60% were similar to PR65 in shape even though some were packed oddly. The small number of fully folded predictions returned by I-Tasser could be attributed to its bias towards globular proteins, but this is unlikely to be the only factor. The ratio of folded to unfolded models returned by Robetta was altogether larger: on average, in 60% of all folded structures the Importin-β repeats form a continuous HEAT-repeat array with a PR65 domain that was structurally close to other observed PR65 conformations. Appropriately folded topologies also clearly required the use of interface mutations to optimize the new interface. Yet, misfolding or incomplete folding did occur in multiple cases and sometimes final models exhibited an Importin-β-like topology.

The overall lack of success in all three types of simulations to return appropriately folded models is consistent with the difficulty of obtaining solubly expressed and stable protein constructs *in vitro*. In fact, the loss of α-helicity observed upon adding HEAT repeats indicates that a fraction of residues is not completely folded. Although the melting transitions of chimeras differed from the PR65 at high temperatures (possibly due to different aggregation behaviour), the first transitions are almost identical. Together, these CD spectroscopy results indicate that PR65 is probably folded independently, as was observed in folding simulations, while the Importin-β repeats adopt individual substructures that are less well folded and are unlikely to stack against the final repeat of PR65.

In conclusion, if these proteins could be produced to adopt conformations similar to those observed in Robetta models, their dynamic properties would indeed change significantly. However, at all stages, starting from docking, over folding simulations to actual *in vitro* experiments, evidence accumulated that the problems observed are most likely due to the initial design, by trying to combine repeats from two different HEAT-repeat families. It remains to be seen, whether a different combination of HEAT repeat proteins will obtain better results. Given the correlation between simulations and experiments observed here, it should be possible in the future to obtain computational data of different proteins to identify a good candidate before proceeding to renewed rounds of experimentation.

Chapter 5

Developing click-able DNA-protein chimeras for force spectroscopy

5.1 Introduction

In a dumbbell optical tweezers set up the protein of interest is attached to DNA handles that are bound to two trapped polystyrene beads (Figure 5.1). Functionalisation of the beads and introduction of compatible DNA modifications is relatively straightforward, e.g. using streptavidin-biotin or digoxigenin-antidigoxigenin interactions. The current bottleneck in single-molecule force-spectroscopy (SMFS) of proteins is the site-specific attachment of DNA (optical tweezers) or protein/peptide handles (AFM). Various methods for the bio-conjugation of specific functionalities in proteins have been established [317], but only a subset are stable enough to find application in SMFS. The current choices are: (i) endogenous cysteines for thiol-/maleimide-based attachments [144], (ii) fusions of protein tags (HaloTag, SNAP-tag, SpyTag/Catcher), or (iii) N- and/or C-terminal introduction of small peptide tags [318]. Protein handles can additionally function as fingerprints and thereby aid the refinement of data-sets [318]. For example, the HaloTag has been successfully used to attach substrate for the ClpX protease to micro-beads [319]. In a dumbbell optical tweezers set up, the DNA handles, by which the protein is attached to beads on either side, have specific characteristics that can function as a fingerprint. Furthermore, the introduction of large protein-tags can pose a problem for expression of the protein of interest. Small peptide tags, such as the ybbR- [320] or sortase-tags [321] can be beneficial in these cases as they enable linking of protein to handles post-translationally. Using the 4'-phosphopantetheinyl transferase (Sfp) enzyme, coenzyme A(CoA)-bearing moieties can be covalently cross-linked to a serine in the ybbR-tag (DSLEFIASKLA) [322, 323]. For example, the ybbR-tag has been used to link proteins to DNA oligonucleotides for both optical tweezers [179] and AFM [320, 324]. Nevertheless, the use of such peptides is limited to the N- and C-termini, as they can interfere with protein folding and/or stability

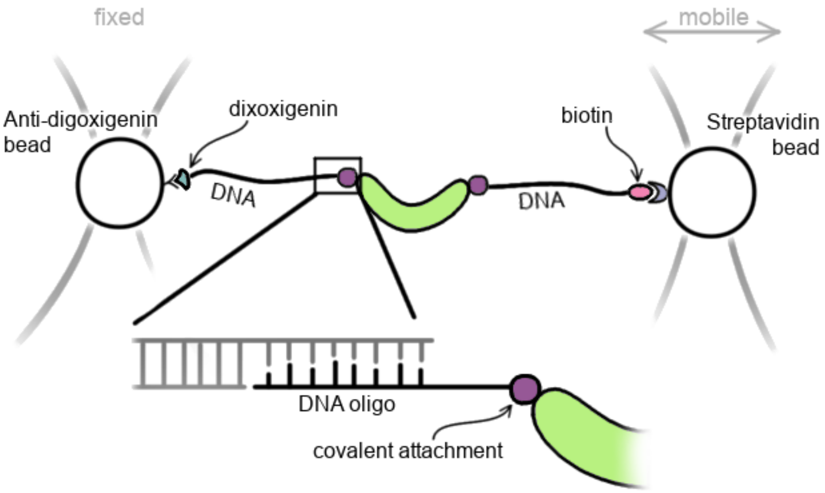


Figure 5.1: Site-specific attachment of DNA handles to a protein is necessary for force-spectroscopy in a double-dumbbell optical tweezers set up. A small oligonucleotide, which is covalently linked to the protein, hybridizes with single-stranded overhangs in DNA handles. The other ends of the DNA handles bear functionalities such as biotin and digoxigenin which can bind to streptavidin and anti-digoxigenin coated polystyrene beads, respectively.

when introduced at internal sites [318]. Moreover, sortase-mediated attachment is limited to one terminus at a time due to the possible formation of circular protein products when termini of multiple molecules are conjugated [325].

The current DNA-protein cross-linking protocol for optical tweezers experiments uses site-specific introduction of cysteines at either termini or internally of the protein of interest. After thiol-pyridine activation of cysteine side chains, proteins can then be conjugated to thiol-DNA oligomers [144]. However, unwanted by-products of this reaction can be poly-protein constructs and oligo-oligo dimers, both of which may not be easily separated from the protein-DNA chimera in purification steps following the conjugation reaction. Cysteines can also be reacted directly to maleimide-modified DNA oligos, but maleimide has to be supplied in large excess and the potential of maleimide dimerization remains [326, 327]. To improve reaction speeds and minimize the excess use of expensive components, Mukhortava and Schlierf [327] developed a two-step protocol, in which cysteines are first functionalised with DBCO-maleimide followed by subsequent conjugation to azide-modified DNA oligos. Yet, these improvements are limited to proteins where cysteines are not present in the WT or where un-wanted cysteines can be removed without severely affecting protein stability.

Due to the presence of a large number of cysteine residues in PR65, it would be very difficult to achieve specific attachment. Previous experiments in the Itzhaki group have shown that only a few of these can be substituted before protein stability is seriously affected. Therefore, I developed a novel, cysteine-independent method using different bioorthogonal chemistries. These methods require the site-specific introduction of unnatural amino acids (UAAs) bearing desired chemical functionalities to react selectively with corresponding modified DNA oligo. In parallel, I designed protein constructs containing ybbR-tags and show that the two attachment methods can be easily combined.

5.1.1 Bioorthogonal chemistries

A bioorthogonal probe has to react selectively under physiological conditions, result in stable linkages, and its reactants have to be kinetically, thermodynamically and metabolically stable without being toxic to living systems [328]. A variety of bioorthogonal chemistries have been developed for post-translational modification of bio-molecules [reviewed extensively in 329, 330].

Copper(I)-catalysed azide-alkyne cycloaddition (CuAAC, Fig. 5.2), as was first reviewed by Huisgen in 1963 [331], demonstrates such bioorthogonality, as it does not engage any of the functional groups present in amino acids [332]. Although the reaction is highly favourable thermodynamically, it requires high pressures and temperatures to obtain reasonable yields [328]. The use of copper(I) as a catalyst both eliminates the need for high temperatures and increases the regioselectivity for 1,4 disubstituted 1,2,3-triazoles [333, 334]. The main disadvantage of CuAAC is that Cu(I) or Cu(I)-mediated generation of reactive oxygen species can be toxic to cells and can hydrolyse proteins [328, 332]. These side-effects can be reduced by providing a copper ligand, such as THPTA, to the reaction mix. CuAAC has now been widely used to label azide-functionalized molecules in a variety of *in vivo* and *in vitro* systems [328], to single and double label proteins combined with other chemoselective reactions including thiol chemistry and other click chemistry systems [335, 336], and even to staple proteins and peptides [337, 338].

Metal dependence can be eliminated by placing the alkyne under ring-strain, such as in cyclooctynes, which can then undergo strain-promoted azide-alkyne cycloadditions with azides (Figure 5.2). Probes used in early studies were slow in their reactivity, but after years of iterative modifications they have been improved significantly [329, 330]. For example, by fusing benzene rings to a cyclooctyne, reactions to azides became much faster. Such dibenzocyclooctynes (DIBO/DBCO) are now widely used to tag proteins, both *in vivo* and *in vitro.*

Similar to SPAAC, in inverse-electron-demand Diels-Alder (IED-DA) cycloadditions, an electron-rich dienophile (strained alkene), reacts with electron-poor azadienes (tetrazines) to form diazanorcaradienes without requiring a metal catalyst (Figure 5.2) [329]. The fastest IED-DA reported to date is that between *trans*-cyclooctyne (TCO) and tetrazine probes [330]. However, large groups, such as that of TCO, are not as bio-compatible as smaller ones. Cyclopropenes, although not as fast in their reaction kinetics as TCO,

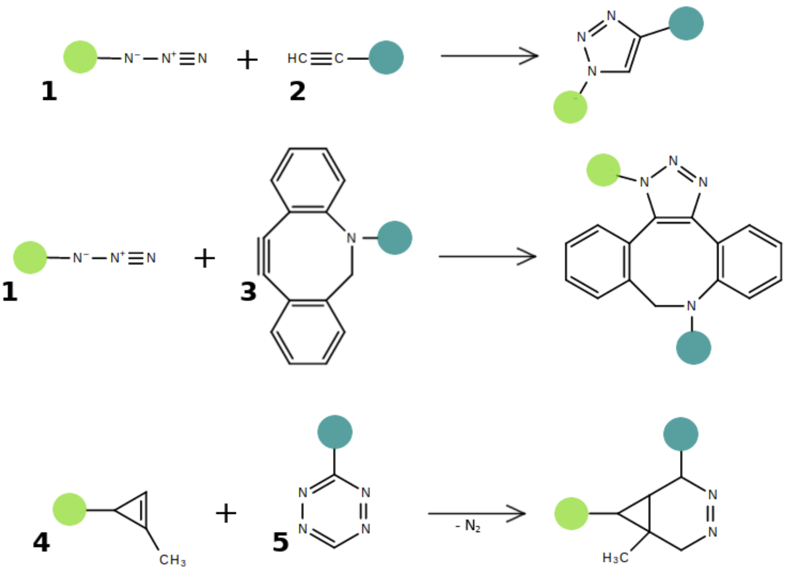
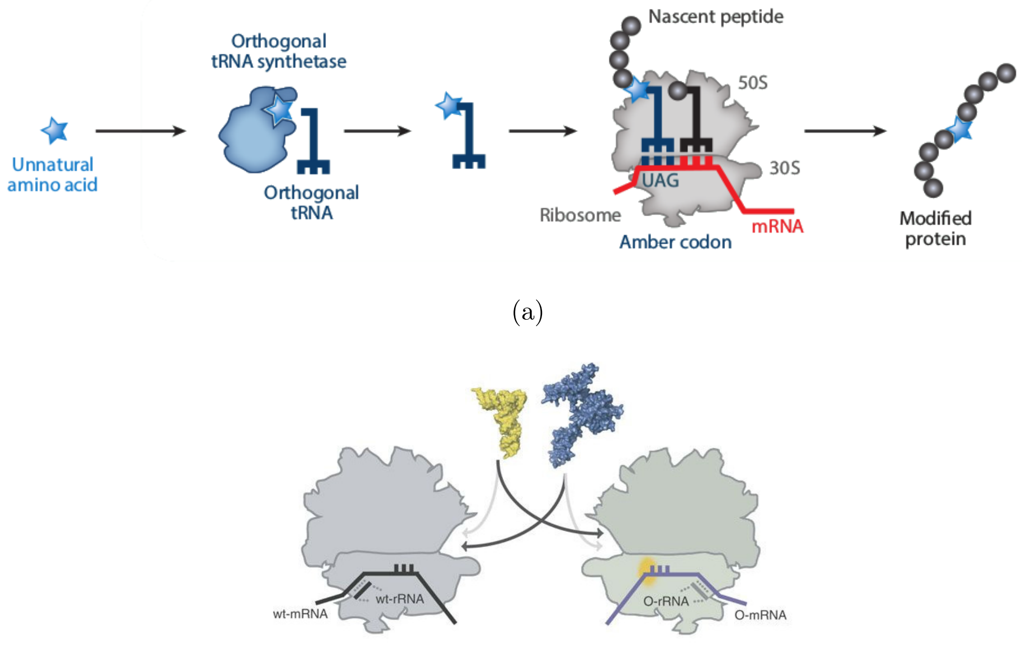


Figure 5.2: Bio-orthogonal reactions. Terminal azides (**1**) and alkynes (**2**) can form 1,4-disubstituted 1,2,3-triazoles in a 1,2-dipolar cycloaddition catalysed by copper. They can also react with strained alkynes such as DIBO/DBCO (**3**) in strain-promoted azide-alkyne cycloadditions that do not require a catalyst. Cyclopropenes (**4**) react with tetrazines (**5**) in inverse-electron-demand Diels-Alder cycloadditions to diazanorcaradienes.

provide a good alternative that is compatible with many applications, including living systems [339–341]

5.1.2 Incorporation of unnatural amino acids

Chemoselective reactions require distinct chemical functionalities, which do not occur naturally in proteins. Such unnatural amino acids (UAAs) can be incorporated either in a residue or a site-specific manner [328]. In the first scenario, an amino acid analogue is supplied to compete for natural aminoacyl tRNA synthetases (aaRSs), resulting in proteome-wide and often only partial incorporation of UAAs [328]. Site-specific incorporation of UAAs requires the introduction of aaRS/tRNA pairs that insert the UAA in response to a specific codon, leading to a homogeneous population of modified protein (Fig. 5.3a) [328]. Often the amber stop codon (UAG) is used as it is the least frequent stop codon in *E. coli* and mammals [342]. Moreover, aaRS/tRNA pairs have also been adapted to incorporate UAAs instead of natural amino acids in response to sense codons [339]. A variety of different aaRS/tRNA pairs from heterologous hosts available, the most commonly used being those from *Methanocaldococcus jannashii* and *Methanosarcina barkeri*, which allowed the expansion of the genetic code *via* the amber stop codons in bacteria, yeast, mammalian and plant cells [328, 342–346].



(a)

(b)

Figure 5.3: Amber suppression using (a) an orthogonal aaRS/tRNA pair, and (b) an orthogonal ribosome that decodes an orthogonal mRNA. Orthogonality is achieved by complementary mutations in the mRNA Shine-Dalgarno sequence and the 16S rRNA [347], thereby eliminating RF-1 (blue) competition with the orthogonal tRNA (yellow) for amber codons in the orthogonal mRNA. Adapted from [328] and [348], respectively.

Various aaRS/tRNA pairs have been evolved to incorporate a vast range of amino acids with and without specific functional groups [328, 339, 343, 345, 349–351]. Despite the successes of amber suppression, the decoding of a stop codon is still less efficient than that of a sense codon [343]. First attempts to increase protein yields were to increase the copy number of tRNAs and aaRSs, or to create Release Factor 1 (RF1) knock-out bacterial strains [343]. The biggest increase in protein production was achieved by introducing a second 16S ribosomal subunit that is orthogonal to the endogenous translation machinery and only decodes amber stop codons using the respective orthogonal tRNA (Fig. 5.3b) [347]. As the orthogonal ribosome is not responsible for the general proteome of the cell, its characteristics can be altered by directed evolution to incorporate UAAs in response to the amber stop codon as well as quadruplet codons [224, 346, 348]. Here, I use the *M. barkeri* PylRS/tRNA$_{\mathrm{CUA}}$ pair combined with the orthogonal ribosome to introduce alkyne, azide and cyclopropene derivatives of pyrrolysine site-specifically into PR65.

5.2 Methods

5.2.1 Genetic constructs and mutagenesis

pRSF-oRibo-Q1-oGST-$\mathrm{CaM}_{1\mathrm{TAG}}$ [224], containing an orthogonal ribosome under an IPTG-inducible promoter and the protein of interest under a constitutively active promoter, and pKW1 [223], containing the orthogonal aaRS and tRNA, for amber suppression were a kind gift from the Chin Lab (MRC LMB, Cambridge, UK). The PR65 template was a thrombin cleavable GST-PR65-H_6 fusion protein in a pRSETa backbone. CTPR_RV templates of varying repeat numbers were available in a pRSETa backbone with an N-terminal H_6-tag [139].

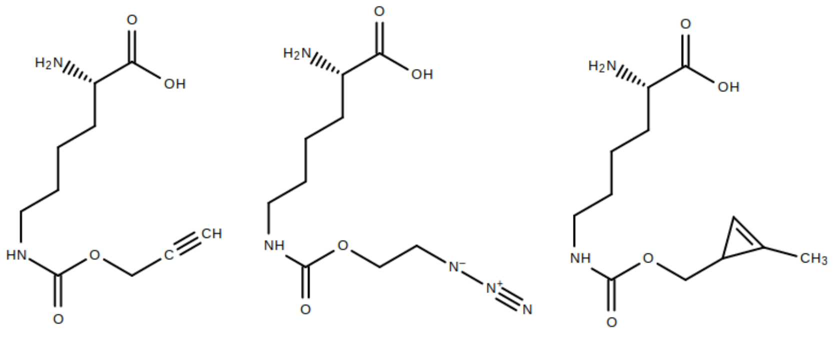
Constructs for amber suppression

Amber suppression constructs were created by first introducing the TAG codon at the positions of M1/A589, D5/L588, E277 and Q514 into the GST-PR65-H6 fusion protein using RTH-SDM (Section 2.2.1). The constructs were then transferred into pRSF-oRibo-Q1-oGST by (a) using the GST- internal SwaI and post-H_6 SpeI restriction sites and either QuickStick Ligase (Bioline) or Anza^{TM}T4 DNA ligase, (b) FastClone [225] or (c) In-Fusion Cloning (Takara Bio). For In-Fusion cloning, pRSF-oRibo-Q1-oGST was digested using the SwaI and SpeI restrictions sites and the PR65 insert was obtained by PCR with primers that had 15bp overlap with these restriction sites and the vector backbone. Both vector and insert were gel purified and 1 µl of each was mixed with 0.5 µl 5X In-Fusion HD Enzyme Premix on ice. The reaction was incubated for 15 min at 50°C in a pre-heated thermal cycler and placed back on ice immediately after. 2-4 µl of the ligation reaction were transformed into high efficiency DH5α *E. coli* as usual.

Introduction of ybbR-tags

N- and C-terminal ybbR-tags (DSLEFIASKLA) were introduced between thrombin cleavage site and M1, and A589 and the stop codon using RTH SDM with primers bearing the tag sequence in the 5' overhang. A variant containing a GS-linker (SGSGSGS) between A589 and C-terminal ybbR-tag was created in the same manner. These two constructs are named yPR65y and yPR65-GSy.

Initially, ybbR-tagged CTPR_RV proteins were created by adding the tag N- or C-terminally to a single repeat using RTH-SDM. CTPR_RVs of N repeats were then built up using BamHI/BglII cloning (Section 2.2.2), starting with yCTPR_RV, followed by N-2 times CTPR_RV and ending with CTPR_RVy, giving rise to yCTPR_RVNy.



(a) (b) (c)

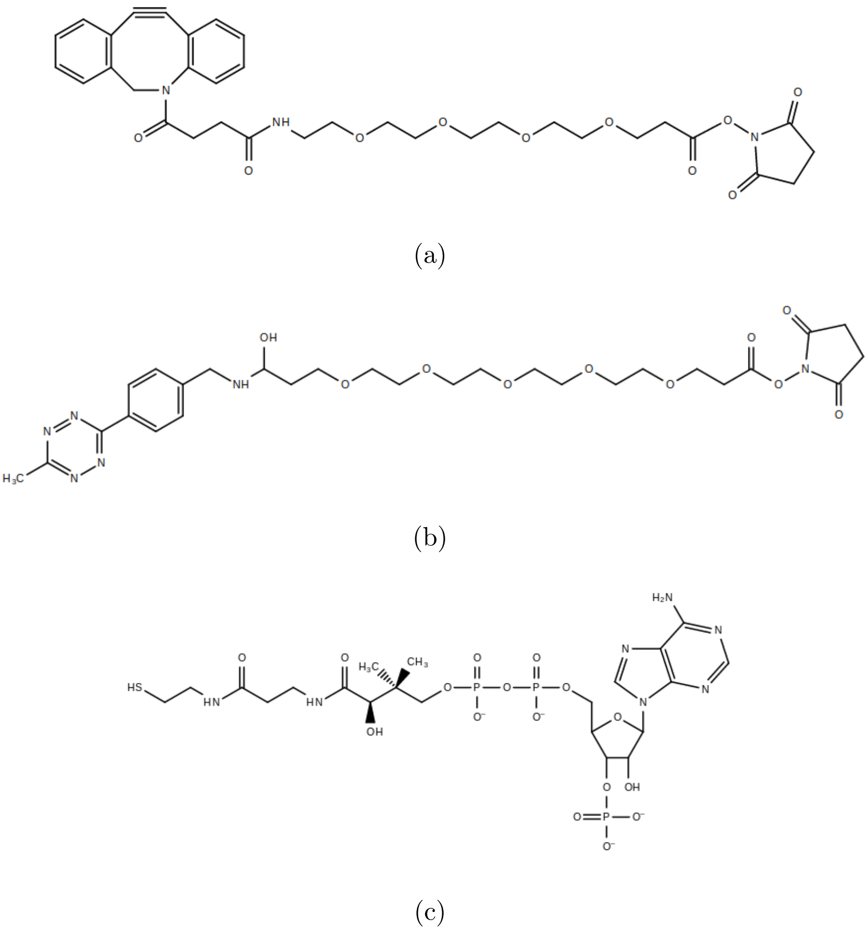
Figure 5.4: Pyrrolysine derivatives bearing carbamate-linked (a) alkyne, (b) azido or (c) cyclopropene functional groups. Structures were drawn using the ChemSpider online tool.

5.2.2 Protein expression and purification

Expression and purification of constructs containing ybbR-tags was performed as detailed in Section 2.3. Expression and purification of amber suppression constructs was adapted from a protocol described by Sachdeva *et al.* [336]. GST-fusion proteins of CaM$_{1\text{TAG}}$, PR65$_{5/588\text{TAG}}$ and PR65$_{1/589\text{TAG}}$ were expressed in electro-competent MDS42 ΔrecA *E. coli* cells, grown at 37°C in 2xYT containing 25 µg/ml Kanamycin and 37.5 µg/ml Spectinomycin. Expression was induced at 37°C for 5 hrs when OD$_{600}$ = 0.5-0.6, using 1 mM IPTG and 1-2 mM of either N-ϵ-(Prop-2-ynyloxycarbonyl)-L-lysine (Iris Biotech GmbH), N-ϵ-((2-Azidoethoxy)carbonyl)-L-lysine (Iris Biotech GmbH) or N-ϵ-[[(2-methyl-2-cyclopropene-1-yl) methoxy] carbonyl]-L-lysine (Sirius Fine Chemicals) (Figure 5.4), which were dissolved in 0.2 M NaOH, diluted 1:3 using 1M HEPES pH 7.4 and adjusted to the pH of the cell culture. Proteins containing azides and cyclopropene derivatives were purified in buffers without reducing agent. All proteins were first purified by glutathione pull-down and thrombin cleavage at 4°C, followed by IMAC to select for full-length product. The eluents were buffer exchanged into PBS or MES using Zeba Spin Desalting Columns (ThermoFisher Scientific).

5.2.3 Chemical modification of DNA oligomers

The sequence of DNA oligos complementary to the DNA handles was provided by the Rief Lab (TUM, Germany).



(a)

(b)

(c)

Figure 5.5: Bifunctional molecules used for Oligo modifications: (a) DBCO-(PEG)$_4$-NHS-ester, (b) 6-methyl-tetrazine-PEG$_5$-NHS-ester, (c) co-enzyme A. Structures were drawn using the ChemSpider online tool.

DBCO- and tetrazine-linked DNA oligomers

A protocol for chemical modification of oligos was adapted from Nojima *et al.* [324]. DBCO-PEG$_4$-NHS-ester (Sigma, Figure 5.5a) and 6-methyl-tetrazine-PEG$_5$-NHS-ester (Jena Bioscience, Figure 5.5b) were conjugated to 3'-amino modified DNA oligos (Integrated DNA Technologies) in 50 µl Bicine-KOH pH 8.0 containing 100 µM amine and 5 mM NHS-ester. Due to its hydrophobicity, DBCO containing reactions were performed in 25% DMSO (Sigma) to ensure full solubility. Conjugations of 6-methyl-tetrazine-PEG$_5$-NHS-ester to amino-oligo were also performed in CHES pH 9.0 and CAPS pH 10.0.

The reaction was incubated at 37°C for 2-3 hours on an orbital shaker and loaded onto an anion-exchange colum (1ml DEAE FF, GE Healthcare) equilibrated in 50 mM Tris-HCl pH 7.4. Bound oligo was eluted in one step using 50 mM Tris-HCl pH 7.4, 1M NaCl.

CoA-linked DNA oligomers

DNA oligos linked to co-enzyme A (Figure 5.5c) were either acquired ready-made from Biomers or produced in-house. CoA stocks were stored in 100 mM Sodium Acetate pH 5.0 at -20°C. 100 μM maleimide oligos (Biomers) were reacted with 5mM CoA in PBS for 1-2 hours at room temperature. The reaction mixture was purified by size-exclusion chromatography using a 10-300 Superdex 200 column (GE Healthcare), equilibrated in either PBS or Sodium Acetate pH 5.0.

Ethanol precipitation of DNA oligomers

The DEAE or S200 fractions containing the relevant oligo were isolated, split into 500 μl aliquots, combined with 1 ml ice cold, absolute ethanol and incubated at -80°C for at least 1 hour. The precipitate was pelleted for 30 min by centrifugation at 0°C, 20,000xg. The supernatant was carefully aspirated and discarded. The pellets were washed using 1 ml of room temperature, 95% (v/v) ethanol and collected by renewed centrifugation for 10 min at 4°C, 20,000xg. The supernatant was discarded and the pellet was dried. Oligos were resuspended in MilliQ H_2O and stored at -20°C.

5.2.4 Production of DNA handles for troubleshooting and optimization

The protocol described in Section 2.4.2 was adapted to produce unlabelled DNA handles. The reaction was performed using Phusion High Fidelity DNA Polymerase in HF Buffer with an annealing temperature of 60°C.

5.2.5 Conjugating DNA and protein

CuAAC

For optimization purposes chemical reactions were performed in 20 μl volumes of PBS or MES with 5 μM of alkyne bearing protein which were reacted to 100 μM azide using a range of catalyst concentrations. Copper sulfate ($CuSO_4$), sodium ascorbate (NaAsc) and THPTA were pre-mixed into a "click mix" (CM). The 100X CM as defined by Sachdeva *et al.* [336] contains 10 mM $CuSO_4$, 25 mM NaAsc, 50 mM THPTA and was used in a final concentration of 1X. A 100X click mix with ten times the amount of NaAsc (CM-A,[352]) contains 10 mM $CuSO_4$, 250 mM NaAsc and 50 mM THPTA. The samples were incubated at 25°C for 0.5-2 hrs, or overnight. To stop the reaction the sample was either buffer exchanged or mixed directly with SDS-PAGE sample buffer. For proof of concept experiments and optimization, alkyne-bearing proteins were reacted with 5-FAM-azide

(Lumiprobe). To produce protein-DNA chimeras, azide-functionalised DNA oligomers (Integrated DNA Technologies) were reacted with alkPR65alk.

To test the functionality of the purchased azide-oligo, 20 µM of oligo was labelled with 20 µM 5-FAM alkyne dye (Lumiprobe) and increasing amounts of click mix, sampling $CuSO_4$ concentration (20 µM to 1 mM). The optimized conditions were then used to react 20 µM of protein with 100 µM azide oligo in a 20 µl volume using 10X CM-A.

SPAAC and IED-DA

Trial reactions were performed in 10 and 20 µl volumes of PBS containing 5 µM proteins and 10-20 µM modified oligo respectively. Control reactions for azide-bearing proteins were performed using TAMRA-DBCO (Jena Bioscience). Reaction mixtures were incubated for varying durations (0.5 hours to over night) at room temperature or 37°C in an orbital shaker. Large scale reactions for force-spectroscopy were performed in 50 µl volumes of PBS containing 10 µM protein and 20-40 µM oligo at room temperature over night.

Sfp-synthase mediated DNA-protein conjugation

Originally reactions were performed as previously described in 100 µl of 50 mM HEPES pH 7.5, 10 mM $MgCl_2$, using 5 µM ybbR-tagged protein, 5 µM biotin-CoA and 0.1 µM Sfp-enzyme [323]. The reaction was incubated for 30 min at room temperature. Here, optimization tests were performed in 10 µl volumes with 5 µM protein and 20 µM CoA-oligo or CoA-5-FAM. Buffer, pH, Sfp concentration, $MgCl_2$ concentration, incubation time and temperature were varied. Large-scale reactions were carried out with 10 µM protein, 20-40 µM CoA-oligo in a 50 µl volume at room temperature over night.

Sfp-SPAAC combination

On a small-scale, 5 µM of PR65 constructs containing one ybbR-tag and one azide functionality were reacted to 10 µM of each CoA-oligo and DBCO-oligo in 10 µl of 50 mM $NaPO_4$ pH 6.5, 150 mM NaCl, 50 mM $MgCl_2$. Control reactions were performed by omitting one oligo at a time. Large-scale reactions were carried out with 10 µM protein, 20 µM of each CoA-oligo and DBCO-oligo in a 50 µl volume at room temperature over night.

Analysis and purification of DNA-protein chimeras

Reaction products of oligo-dye couplings were analysed by electrophoresis using 1% unstained agarose gels. Protein-dye and protein-oligo reactions were analysed by SDS-

PAGE. Before polyacrylamide gels were stained with Coomassie Blue, fluorescent bands were imaged under UV using a trans-illuminator (UVP, LLC).

If oligo reactions were successful, the mixture was purified by size exclusion chromatography using either a Superdex 200 10/300 GL (GE Healthcare) or a YMC-Pack Diol-300 (Yamamura Chemical Research). Fractions were analysed by SDS-PAGE and those containing the majority of protein conjugated to two DNA oligos were hybridized to DNA handles and analysed by agarose gel electrophoresis.

5.2.6 AFM microscopy

66 ng of DNA handles were added to 10 µl of an S200 fraction and incubated for 0.5 hours at room temperature. Protein-DNA samples were diluted to 1:100 before 45 µl were deposited onto freshly cleaved mica and left to adsorb for 5 min. Samples were then washed eight times with 0.5 ml BPC-grade water (Sigma) and dried under a stream of nitrogen. Dry sample imaging was conducted on a Bruker Dimension FastScan atomic force microscope using a FastScan AFM Scanner and silicon probes (FASTSCAN-A, Bruker) with stiffness of 18 Nm^{-1}that tuned near to their resonant frequency of 1.4 MHz. Drive amplitude and amplitude set-point were optimized and images were captured at scan rates of 18-20Hz with 512 scan lines per area (e.g. 2µm x 2µm).

5.3 Results

5.3.1 Incorporation of Pyrrolysine derivatives into PR65

Amino acid incorporation can be dependent on the positioning of the codon within the sequence and click reaction efficiencies can depend on the local environment, e.g. hydrophobicity (Kaihang Wang, Chin lab, personal communication). Therefore, two different combinations for end-to-end attachment were chosen: 5 and 588 (Fig 5.6a, purple), and 1 and 589 (Fig 5.6a). For internal attachments, UAAs were substituted for solvent exposed residues at boundaries between "domains" of PR65 which are thought to unfold in separate steps, that is between H7 and H8 and between H13 and H14 (Figure 5.6a) [71].

Cloning into pRSF_oRibo-Q1_oGST

Cloning of amber suppression constructs into pRSF_oRibo-Q1_oGST was a major challenge. Efficiencies of standard restriction digest-ligation reactions were poor and resulted in a variety of recombined constructs (Fig. 5.6b). Different ligases, vector:insert ratios and chemically competent cell lines were tested. A correct clone of PR65$_{5/588TAG}$ was

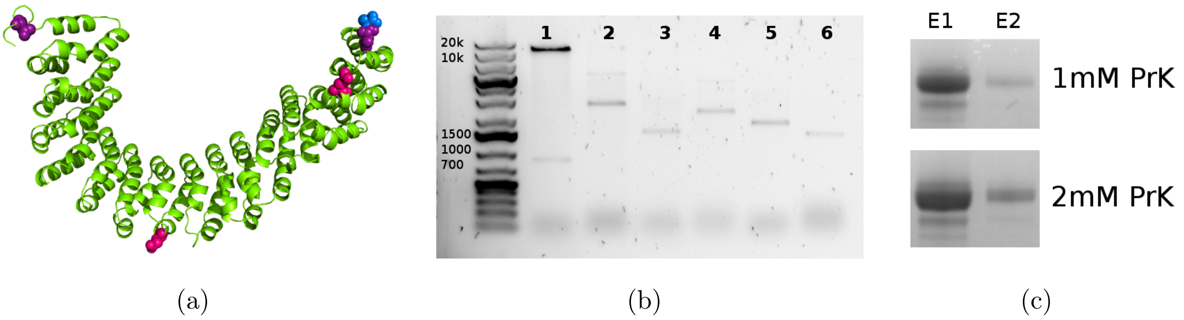


Figure 5.6: Cloning and expression of PR65 constructs for incorporating UAAs. (a) Incorporation sites of UAAs into PR65: Purple - D5, L588; blue - M1 (missing), A589; magenta - E277, Q514. (b) BamHI restriction digest of different colonies from ligation of PR65$_{1/589TAG}$ into oRibo. Correct constructs are digested into 900 bp and 12 kb fragments (lane 1), whereas the majority exhibited variable recombination (lane 2-6). (c) First and second elution of thrombin-cleaved PR65$_{5/588TAG}$ after affinity purification, induced with different amounts of alkPyl.

obtained this way. The only positive clone of PR65$_{1/589TAG}$ had acquired the G149D mutation at the outer face of helix B in HEAT4. At the time, the mutant was carried forward to be used in proof-of-principle experiments.

Since SwaI is a blunt end cutter, it was attempted to delete the second BamHI site in the vector backbone, to enable digest and ligation using BamHI and SpeI, which both leave sticky ends. Unfortunately, both standard SDM and RTH-SDM performed with a variety of PCR conditions were unsuccessful. Next, it was attempted to introduce a gene of interest using the restriction enzyme-independent FastClone method [225], but it was impossible to obtain a correct clone due to recombination of the vector. Lastly, I employed the commercial, ligation-independent In-Fusion technique, with which no recombination was observed at all. This technique is highly efficient and hence it was possible minimize the reaction from 25 to 2.5 µl and the volume of high-efficiency competent cells from 50 µl to as low as 10 µl. Correct constructs for yPR65-277az and yPR65-514az were obtained this way.

Expression of amber suppression clones

Our PR65 expression protocol was compared with the method described by Sachdeva *et al.* [336], but a noticeable difference in soluble protein yield could not be found. An expression test was performed, in which cultures were induced with 1 mM IPTG and either 1 mM or 2 mM alkPyl. When comparing the protein yield after glutathione affinity purification, increasing alkPyl concentration leads to a slight, but detectable increase in yield (Fig. 5.6c). Both PR65 constructs incorporating alkPyl purified to final yields of

<1 mg l^{-1}. Yields of yPR65$_{277\mathrm{TAG}}$ and yPR65$_{514\mathrm{TAG}}$ incorporating azPyl were 0.87 mg l^{-1} and 0.31 mg l^{-1}, respectively. In contrast, yields of PR65$_{5/588TAG}$ incorporating azPyl or cycPyl were ~1.6 and ~1.1, respectively, when induced with only 1 mM of UAA.

Table 5.1: Molecular weights of PR65$_{5/588\mathrm{TAG}}$ and PR65$_{1/589\mathrm{TAG,G149D}}$ incorporating alkyne-Pyl. Theoretical molecular weights (MW) were calculated using the Expasy ProtParam Tool. Measured values represent averages ± standard deviations of different peaks of mass/charge ratios.

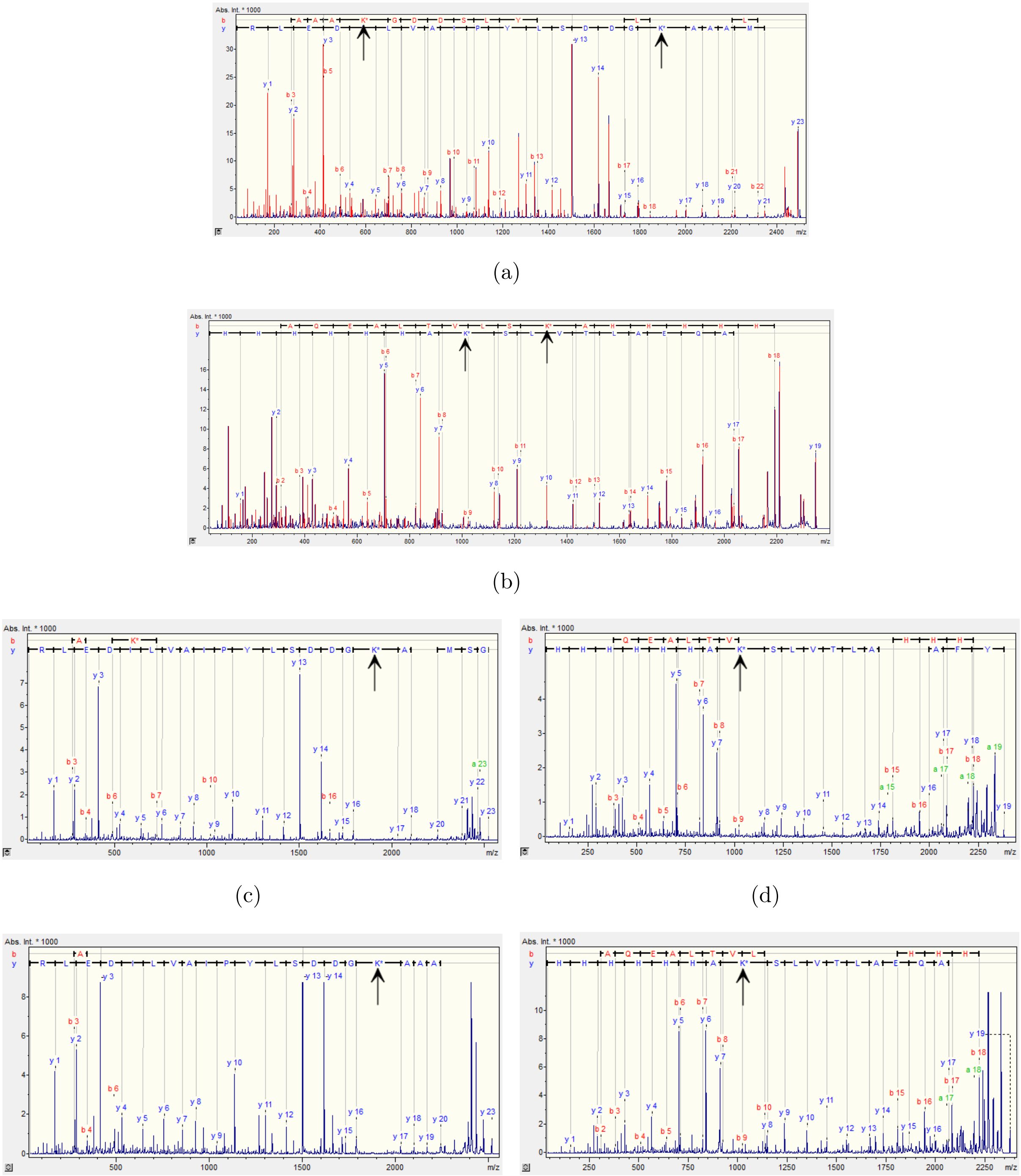
	alkPR65(5/588)alk	**alkPR65$_{\mathbf{G149D}}$(1/589)alk**
Theoretical [Da]	66467.2	66551.8
MALDI-MS [Da]	66484.5 ± 7.3	66561.9 ± 0.4
Difference [Da]	17.3 ± 7.3	10.1 ± 0.4

Theoretical masses and masses measured by mass spectrometry of both alkPR65alk variants are shown in Table 5.1. Both proteins are larger than expected. Since mass spectrometry of larger proteins can be difficult and requires high protein concentrations, tandem mass spectrometry of trypsin digested PR65 peptides proved more reliable to validate UAA incorporation at the N- and C-termini. Figure 5.7 shows tandem mass spectrometry results of N- and C-terminal peptides of PR65$_{5/588TAG}$ incorporating alkPyl (Figure 5.7a,b), azPyl (Figure 5.7c,d) and cycPyl (Figure 5.7e,f). In spectra of azPR65az peaks of azido-peptides showed additional peaks of similar size corresponding to peptides in which the azide had been reduced to an amine. Furthermore, it was possible to detect masses of C-terminal peptides with azide and lysine.

Verification of azPyl incorporation at the positions of E277 and Q514 was not attempted by mass spectrometry because the assignment of internal peptides can vary between enzymatic digests. Instead, azPyl incorporation was verified by labelling with TAMRA-DBCO (see Figure 5.13a)

5.3.2 Modification of DNA oligonucleotides

Amine-modified DNA oligomers were reacted to either NHS-ester conjugated DBCO or tetrazine molecules, and mass spectrometry of purified products was performed to identify whether the reaction was complete. As Figure 5.8 shows, the reaction between DBCO-PEG$_4$-NHS and amine oligo was complete, whereas the reaction of tetrazine-PEG$_5$-NHS to amino oligo was incomplete. The tetrazine conjugation was repeated at higher pH and by incubating on ice, none of which increased the reaction efficiency. Purification by anion-exchange removes DMSO and uncoupled bifunctional molecules, although modified and un-reacted oligomers could not be separated. HPLC was attempted, but the results



(a)

(b)

(c)

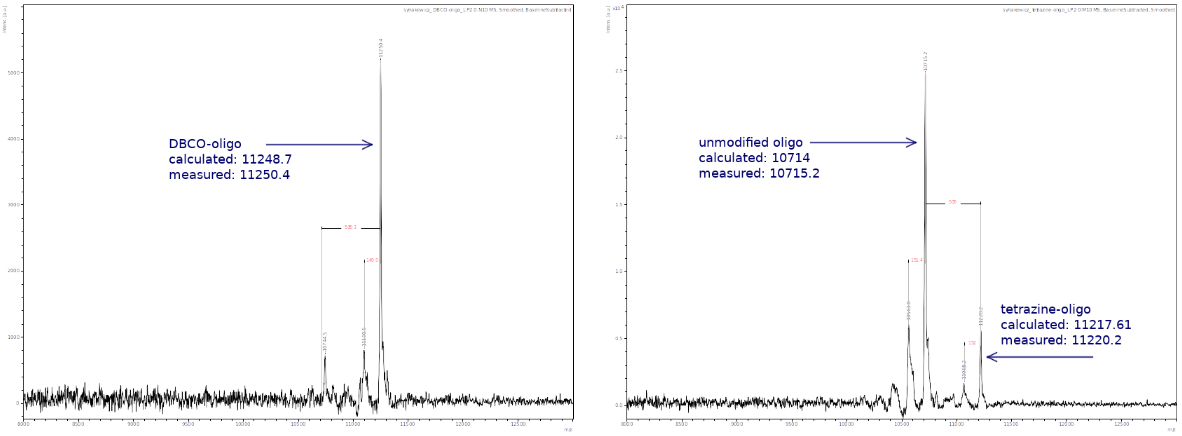
(d)

(e)

(f)

Figure 5.7: Tandem mass-spectrometry of PR65 peptides with UAAs incorporated at positions 5 and 588 (arrows). Shown are the N-/C-terminal peptides of alkPR65 (a,b), azPR65 (c,d), and cycPR65 (e,f). Data was provided by the PNAC Facility of the Biochemistry Department, University of Cambridge.

indicated that extensive optimization of the method is required to separate the two oligo species.



(a)

(b)

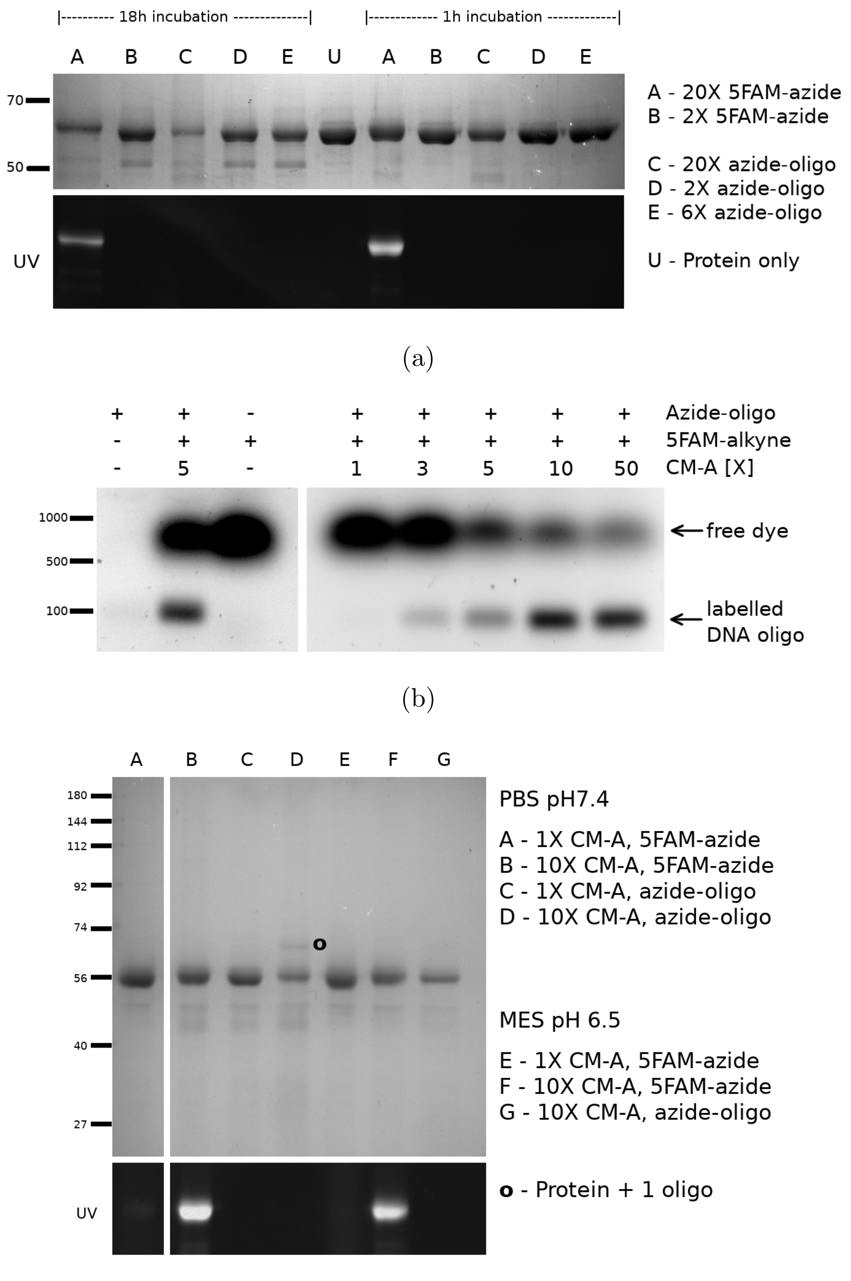
Figure 5.8: MALDI mass-spectrometry of modified oligonucleotides (a) DBCO-oligo (b) tetrazine oligo. Data was provided by the PNAC Facility of the Biochemistry Department, University of Cambridge.

5.3.3 Attaching fluorophores and DNA oligos to PR65 using bioorthogonal chemistries

CuAAC

Considering the expense of azide-functionalized DNA oligos, 5-FAM azide was used for proof-of-concept experiments, with the advantage that successful click reactions can be identified easily by SDS-PAGE under UV light. alkPyl containing PR65 variants were reacted with a 20X molar excess of 5-FAM azide according to a previously published protocol [336]. After continued failures, Calmodulin (CaM_{1TAG}) was expressed and successfully labelled with 5-FAM azide in PBS. CaM_{1TAG}-labelling was largely independent of the buffer system used (PBS pH 7.4 or MES pH 6.5). However, when 1 mM DTT was added, CaM failed to label. The reducing agents TCEP, DTE, β-mercaptoethanol and 1-thioglycerol were subsequently tested and all were found to inhibit the reaction. After PR65 variants were buffer exchanged into PBS without DTT, proteins labelled successfully (Fig. 5.9a).

Since DNA oligos are usually reacted with protein in a 1:1 stoichiometric ratio (personal communication, Daniela Bauer, Rief lab), 5-FAM-labelling was attempted using a 2X excess but was unsuccessful (2X, Figure 5.9a). Furthermore, it was investigated whether a long incubation would result in higher fluorescence but protein degradation/aggregation results in reduced fluorescence. PR65 variants were also incubated with 2X, 6X and 20X molar excess of azide oligo. If a reaction is successful, additional bands corresponding to ~75 kDa (one oligo attached) and ~85 kDa (two oligos attached) can be observed by SDS-PAGE [145]. However, no such bands could be visualized (Figure 5.9a).



(a)

(b)

(c)

Figure 5.9: Optimizing CuAAC reactions between alkPR65alk and 5-FAM-azide and azide-oligo. (a) Example of click reactions using different azide to alkyne ratios of 5-FAM azide or 3'-azide DNA oligo and alkPR65(5/588)alk. Concentrations of azide are given in molar excess. Reactions were performed in PBS with standard click-mix. (b) Click reactions of azide-oligo and 5-FAM-alkyne analysed with increasing concentrations of CM-A (containing 250 mM NaAsc) on an unstained 1% agarose gel. Click-mix concentrations are given in X CM-A. (c) alkPR65alk conjugation to 5-FAM- and oligo-azide in different buffer and click-mix conditions. Protein and azide concentration are the same as those described in (a).

The reaction was optimized further by increasing the ratio of reducing agent to copper [352]. To test, whether electrostatic interactions between DNA and protein decrease the efficiency, reactions were also performed in 1 M NaCl, but the high salt concentration was found to decrease the fluorescence signal in 5-FAM reactions and did not affect the outcome of protein-DNA reactions.

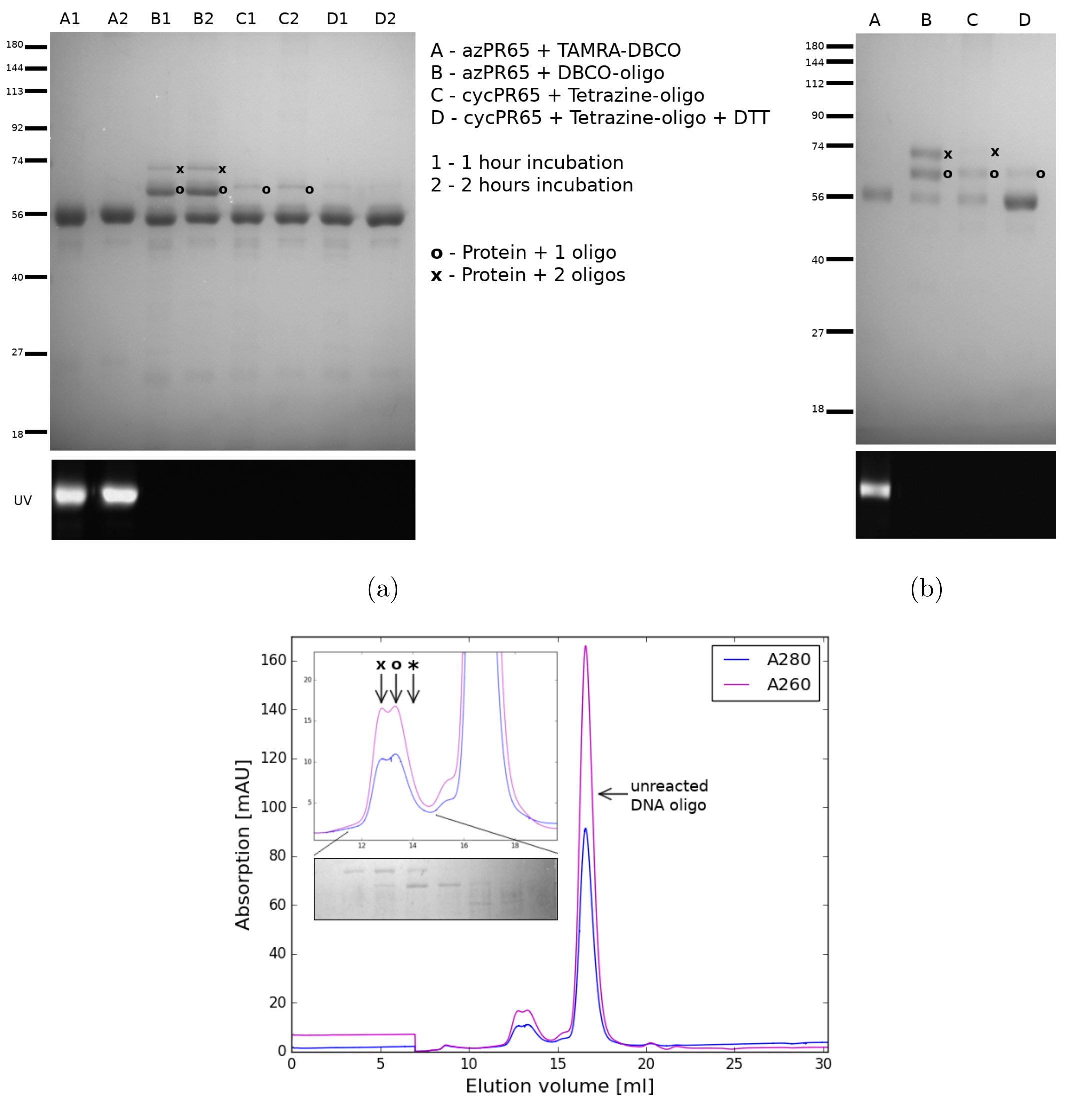
The azide functionality of the commercial oligo was confirmed using a 5-FAM alkyne dye and visualized by agarose gel electrophoresis (Figure 5.9b, left). Whereas the oligo alone runs as a barely noticeable band at the expected molecular weight in an unstained gel, the signal increases significantly once labelled with 5-FAM (Figure 5.9b, left). When increasing the amount of CM-A, it was found that maximum labelling efficiency was obtained with a 10X excess or higher (Figure 5.9b, right). When reacting alkPR65alk to 5-FAM-azide using different concentrations of CM-A, this increase in labelling efficiency could reproduced (Figure 5.9c). Lastly, reactions were conducted in two different buffers and it was found that labelling was slightly more successful in PBS than in MES (Figure 5.9c).

Performing the reaction in PBS and increasing NaAsc:the $CuSO_4$ ratio as well as the overall click mix content resulted in a detectable higher molecular weight band after SDS-PAGE, indicating the presence of PR65 conjugated to one oligonucleotide (Figure 5.9c). Yet, no band corresponding to alkPR65alk attached to two oligos was detected.

Copper-independent chemistries

After the limited success of conjugating DNA oligonucleotides to protein using CuAAC, I turned to orthogonal chemistries that were independent of copper and also possessed faster reaction kinetics [329, 330]. After just one hour of incubation at 37°C, azPR65az was labelled almost to completion with TAMRA-DBCO, which increased only slightly after 2 hours (Figure 5.10a, sample A), and was reduced after overnight incubation, presumably due to aggregation of protein in the absence of reducing agent, or degradation (Figure 5.10b, sample A). For the first time, it was also possible to visualize protein attached to both one and two oligonucleotides in reactions containing DBCO-oligo (Figure 5.10a, sample B). After 1 hour of incubation, a small amount of protein attached to two oligos can already be detected, which increases after 2 hours. When leaving the reactions over night, some protein is lost and some un-reacted protein can still be visualised. Yet, the dominating species are those with one or two oligomers attached.

Reactions between cyclopropenes and tetrazines were significantly less successful as it was impossible to estimate how much tetrazine-modified oligo was actually present. However, some attachment of one oligo could be observed after 1 and 2 hours (Figure 5.10a, sample C) and very small amounts of protein attached to two oligos appear to be present after over night incubation (Figure 5.10b). It was tested whether IED-DA was



(a)

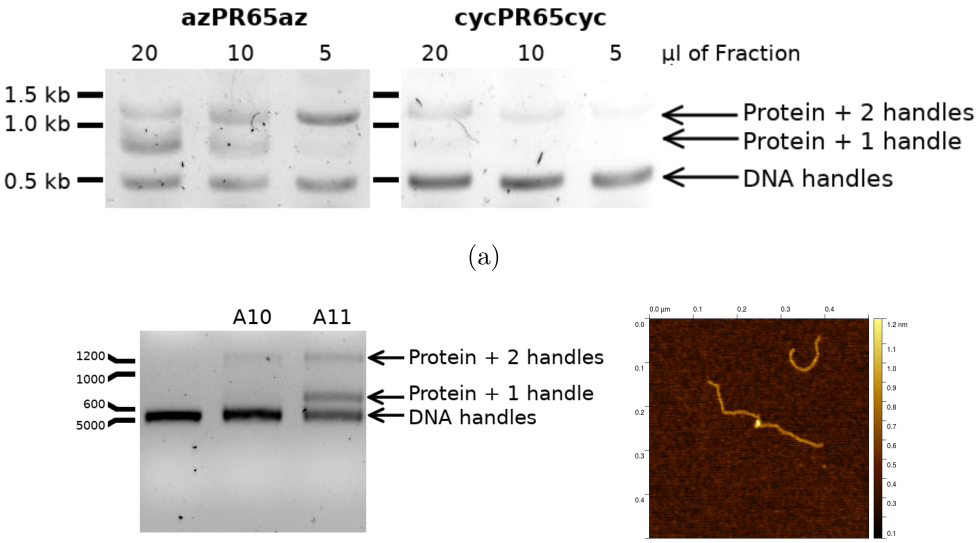
(b)

(c)

Figure 5.10: SPAAC and IED-DA between azide or cyclopropene containing PR65$_{5/588\mathrm{TAG}}$ and TAMRA or oligo. (a) Example reactions using 5 µM protein and 20 µM TAMRA or oligo (b) Same reactions as in (a), incubated over night. (c) Chromatogram of a large scale reaction of 10 µM azPR65az and 40 µM DBCO-oligo incubated over night at 25°C, separated on a S200 10/300 GL. The smaller peak doublet preceding the oligo peak corresponds protein conjugated to two oligos (**x**) and one oligo (**o**). Un-reacted protein can be detected by SDS-PAGE even though a clear A280 peak is not visible (*****, inset).

inhibited by reducing agent, and as can be seen in sample D in Figure 5.10b, 1 mM DTT already reduces the reaction efficiency. However, at this point it is unclear why exactly this occurs and hence any subsequent purification of cycPR65cyc were performed using buffers without reducing agents.

Using size exclusion chromatography it was possible to separate the different species



(a)

(b)

(c)

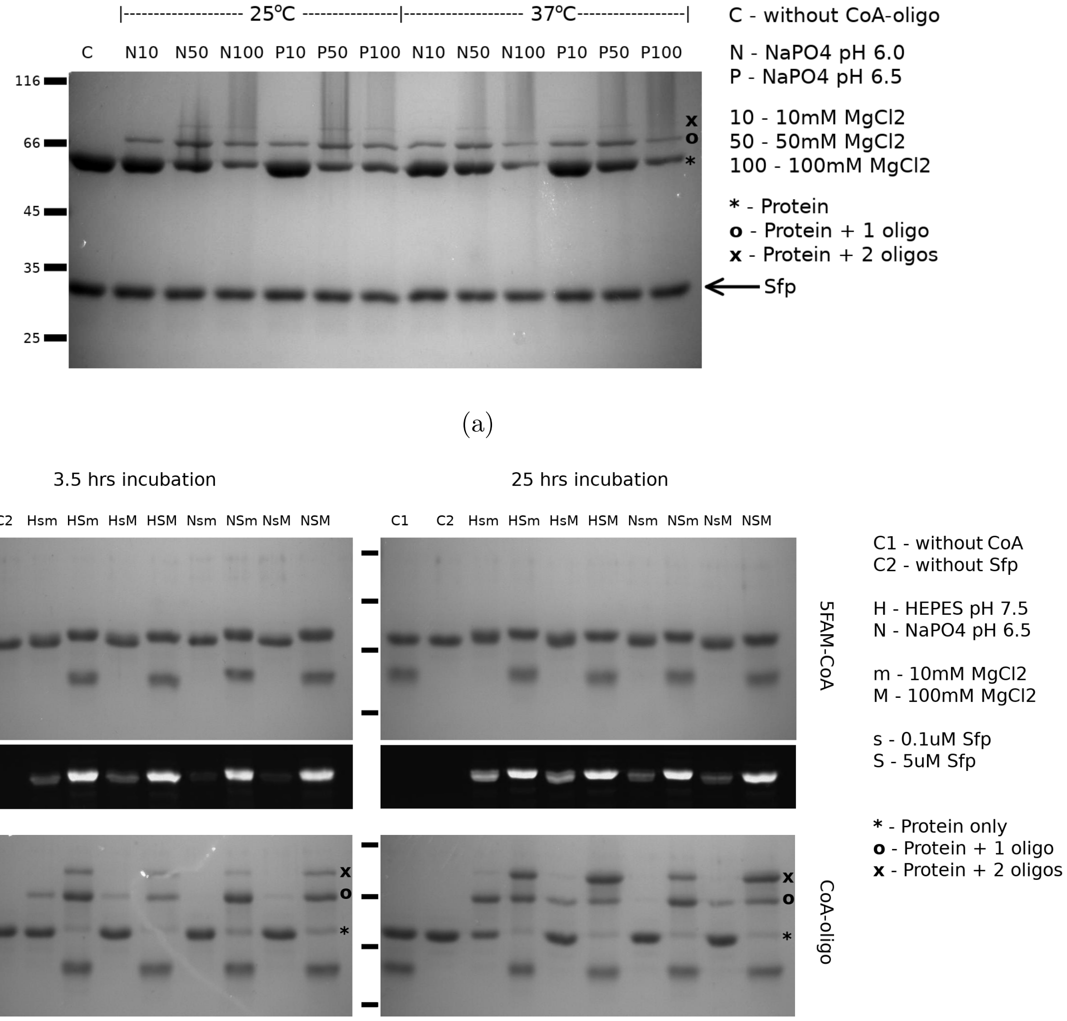
Figure 5.11: DNA handle hybridization to Fraction A11 of SPAAC and IED-DA conjugations purified using an S200 10/300 GL column and detected by agarose gel electrophoresis. The DNA amount is kept constant at 200 ng/µl and incubated with increasing volumes of Fraction A11. (b) Hybridization to 10 µl of Fractions A10 and A11 of a azPR65az S200 purification. (c) AFM micrograph of a protein attached to two DNA handles.

from large-scale reactions to some extend (Figures 5.10c), but complete separation could not be achieved with neither a Superdex 200 10/300 GL nor an HPLC column. When incubated with DNA handles, bands corresponding to protein-DNA chimeras hybridised to one ∼500 bp DNA handle and to two DNA handles can be observed (Figure 5.11a,b). The equilibrium between these two hybridization events depends on a combination of (i) the ratio between protein conjugated to one and two oligonucleotides which differs in neighbouring S200 fractions (Figure 5.11b), and (ii) the ratio of protein-oligo to DNA handles (Figure 5.11a), in which case higher protein-oligo concentrations result in more protein with one handle attached. It was possible to visualize successful attachment of DBCO-oligos to azPR65az using atomic force microscopy (AFM) (Figure 5.11c). Although it was not possible to visualize cycPR65cyc species after purification by SDS-PAGE, hybridization to two DNA handles and detection by agarose gel electrophoresis proves the presence of protein conjugated to two tetrazine-oligomers (Figure 5.11a). Those amounts were sufficient for force spectroscopy. Force extension curves of both azPR65az and cycPR65cyc can be seen in Figure 5.15a,b.

5.3.4 Optimization of Sfp-synthase method

In parallel with the bio-orthogonal chemistries, I created ybbR-tagged PR65 and CTPR constructs that can be conjugated to CoA-modified oligonucleotides and dyes [322]. Yin *et al.* [323] originally reported complete ybbR-CoA conjugation within 30 min using only 0.1 μM Sfp-synthase. These conditions resulted in very low labelling efficiencies with any of the ybbR-tagged proteins tested here. However, when using at least 5 μM Sfp-synthase [321], both yPR65y and yCTPR_RV10y constructs could be conjugated to CoA-oligos (Figure 5.12).

It was tested whether additional $MgCl_2$ could increase the reaction efficiency. When 100 mM $MgCl_2$ was added to a reaction mixture containing home-made CoA-oligo, a white precipitate was observed. At first, it was thought that this could be due to Mg^{2+} forming



(a)

(b)

Figure 5.12: Sfp-mediated CoA-ybbR conjugation of (a) yPR65y and (b) yCTPR_RV10y in various reaction conditions. Reactions were conducted using home-made (a) and commercially produced CoA-oligo (b).

insoluble $Mg(OH)_2$ in buffers with $pH > 7$. Yet mixing $MgCl_2$ with buffers at the concentrations used in the reaction did not cause precipitation. Constantinides and Steim [353] reported a precipitate of palmitoyl-CoA when mixed with as little as 5mM Mg^{2+}. Indeed, when $MgCl_2$ was added to home-made CoA-oligo in PBS or HEPES, it precipitated. To test whether precipitation was pH dependent, reactions were carried out in sodium phosphate buffers of pH 6.0 and 6.5 (Figure 5.12a). No precipitation occurred in these reaction mixtures. Higher concentration of $MgCl_2$ increased the conjugation efficiency significantly, such that after 2 hrs protein conjugated to two DNA oligos was visible by SDS-PAGE. However, higher amounts of $MgCl_2$ correlated with increased smearing on the gels and an apparent reduction in protein. The reactions between yPR65y and CoA-oligo appeared to be slightly better at room temperature or pH 6.5 than at 37°C or pH 6.0, possibly due to higher stability of both PR65 and Sfp at lower temperatures and higher pH. Similar trends were observed for conjugations of yCTPR_RV10y to home-made CoA-oligos.

Precipitation never occurred in reactions containing commercially made CoA-oligo and none of the above observations could be reproduced (Figure 5.12b). Indeed, using commercial 5-FAM-CoA and CoA-oligo, some labelling could be observed at Sfp concentrations of 0.1 μM, albeit it was far from complete after 3.5 hours or even overnight incubation. Furthermore, conjugations were better at pH 7.5 than at pH 6.5, as can be seen from the fluorescence intensity (Figure 5.12b, UV images) and oligo conjugations (Figure 5.12b, bottom) in reactions with low Sfp concentrations. This agrees with previous publications on Sfp activity in buffers of different pH [322]. $MgCl_2$ concentration only appears to increase yields in ybbR conjugations to CoA-oligo, not 5-FAM-CoA, and is apparent only in sodium phosphate buffer. After over-night incubation, a slight increase of protein attached to two oligos is also observed for HEPES reactions with high Sfp concentrations, while reactions with low amounts of Sfp show the opposite trend.

Eventually, it did not matter which set of reaction conditions were used. Although complete conjugation was not observed even after overnight incubation, if >5 μM Sfp were added, sufficient amount of protein attached to two DNA oligos was obtained. Using HPLC, it was possible to separate different DNA-oligo species of smaller CTPR proteins, while separation of CTPR proteins containing a large number of repeats was similarly incomplete as observed with PR65. Force extension curves for yPR65y, yPR65-GSy and yCTPR_RV10y can be seen in Figures 5.15c-e.

5.3.5 Combining SPAAC and Sfp-mediated DNA-protein conjugation

The presence of functional azPyl in both yPR65-277az and yPR65-514az was confirmed by labelling with TAMRA-DBCO (Figure 5.13a). Using azPR65az and TAMRA-DBCO the Sfp-synthase reaction conditions were screened to examine whether any components would interfere with SPAAC. TAMRA-labelling was found to be independent of buffer and $MgCl_2$ concentration. At the time, when yPR65-277az and yPR65-514az reactions were initially prepared for force-spectroscopy, the commercial CoA-oligo had not yet been available and the results described in Section 5.3.4 had not yet been obtained. Therefore, first attachments were performed in sodium phosphate pH 6.5 containing 50 mM $MgCl_2$.

To determine attachment efficiencies, proteins were reacted with one oligo at a time, as well as to both oligomers at the same time (Figure 5.13b, left). Attaching the N-terminal CoA-oligo caused the expected retention by approximately 10 kDa in SDS-PAGE in both proteins. However, attachment of the internal oligomer caused the construct to migrate at larger molecular weights: PR65 conjugated to an oligo at residue 277 caused a shift of the band to nearly 100 kDa, whereas PR65 conjugated to oligos at residue 514 caused a shift that was barely distinguishable from the N-terminal attachment (Figure 5.13b, left). After incubation overnight, the protein amounts were severely reduced but conjugation was deemed successful enough to proceed with force-spectroscopy (Figure 5.13b, right). When commercial CoA oligos were available, these reactions were conducted in the original Sfp reaction buffer containing HEPES and 10 mM $MgCl_2$. Reaction efficiencies and loss of protein during long incubations were comparable. Generally, yPR65-514az was more affected by long incubations suggesting that the protein is less stable.

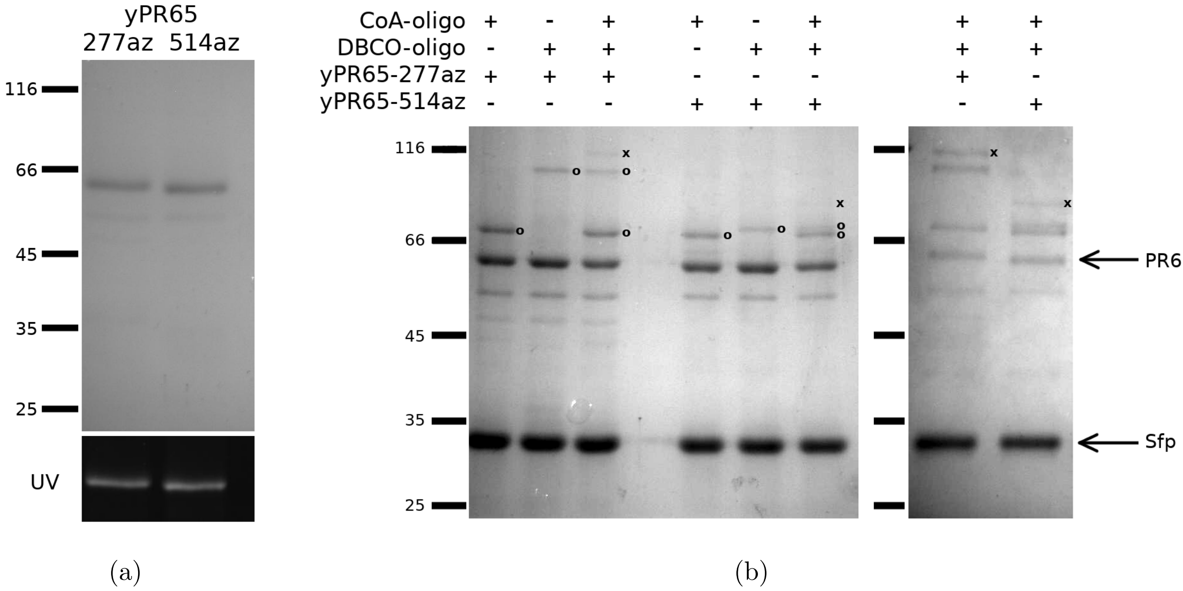

yPR65
277az 514az
116
66
45
35
25
UV
CoA-oligo + - + + - + + +
DBCO-oligo - + + - + + + +
yPR65-277az + + + - - - + -
yPR65-514az - - - + + + - +
116
66
45
35
25
x
o
PR65
Sfp
(a)
(b)

Figure 5.13: Combining CoA-ybbR and SPAAC conjugations. (a) Labelling yPR65-277az and yPR65-514az with TAMRA-DBCO to confirm azPyl incorporation. (b) Conjugating both yPR65-277az and yPR65-514az to DBCO- and CoA oligonucleotides after 2 hours (left) and over-night incubation (right).

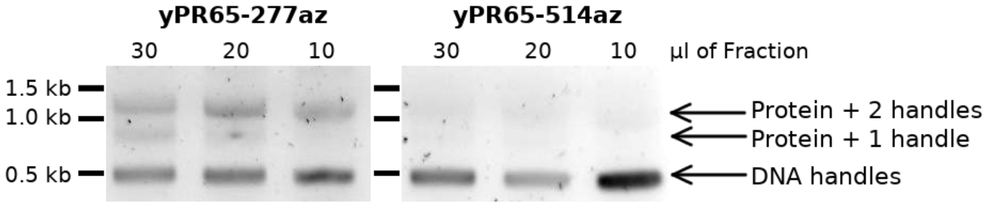


Figure 5.14: DNA handle hybridization to combined SPAAC and Sfp-mediated protein-oligo conjugations purified using an S200 10/300 GL column and detected by agarose gel electrophoresis. The DNA amount is kept constant at 200 ng/µl and incubated with increasing volumes of Fraction A11 of the purification. These reactions are not the same as in Figure 5.13.

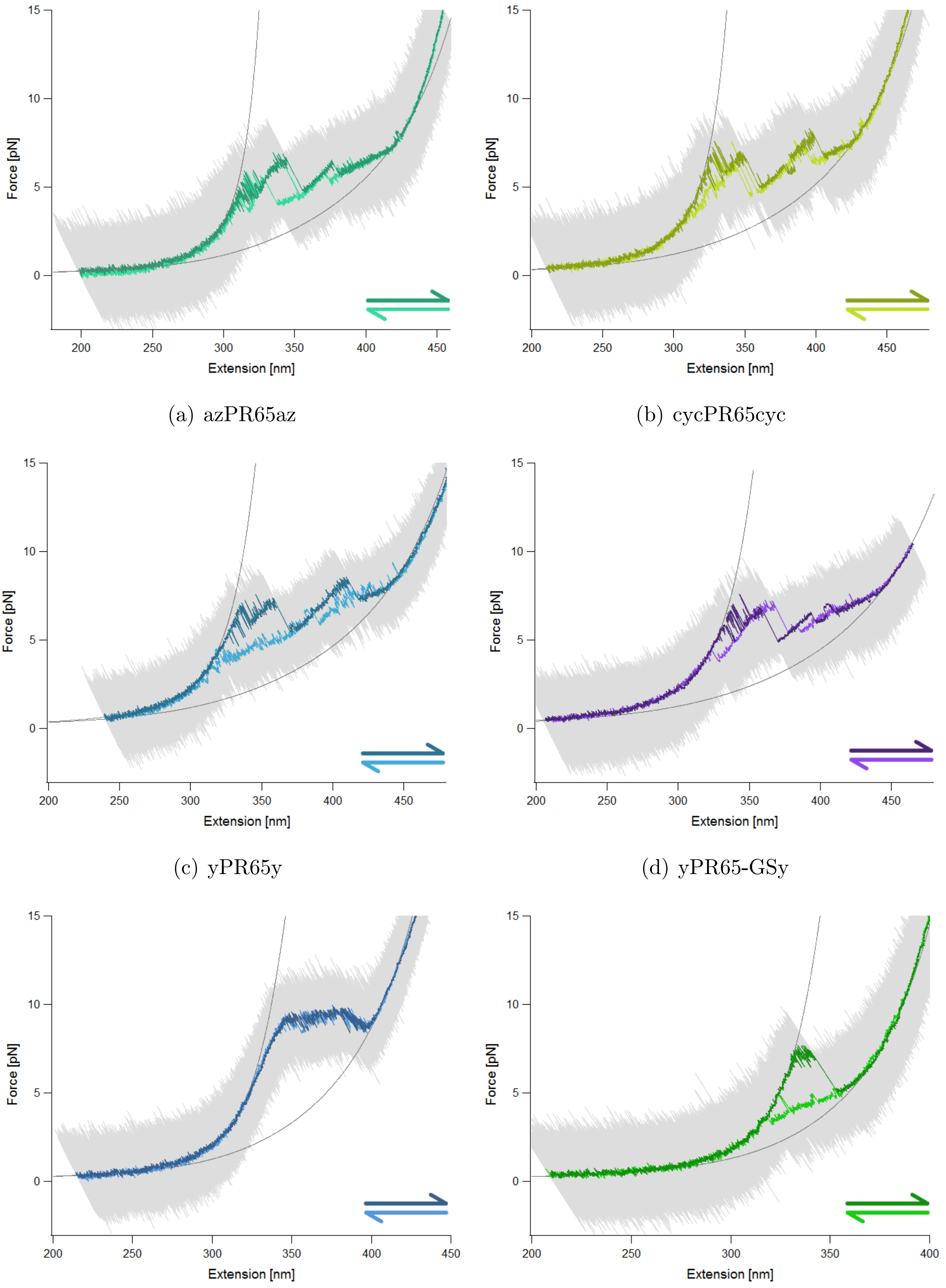
Separation of the different oligo-protein species proceeded as described above and a hybridization to DNA handles is shown in Figure 5.14. In this particular data set, only purified yPR65-277az reactions show successful conjugation, while too little amount of yPR65-514az was left after purification to determine conjugation success by any method. Force spectroscopy data of yPR65-277az reacted both in sodium phosphate and HEPES were obtained and a representative force-extension curve can bee seen in Figure 5.15f. No force spectroscopy data could be obtained for yPR65-514az, although measurements of multiple conjugations were attempted.

5.4 Discussion

The amber suppression system was implemented successfully to produce PR65 with pyrrolysine derivatives carrying alkyne, azide and cyclopropene functional groups which could then be conjugated to the respectively modified DNA oligos and fluorescent dyes. Additionally, a method based on the enzymatic modification of ybbR-tags was tested alongside the chemical approaches. To summarize the findings of this chapter an overview of all four methods is given in Table 5.2, the content of which will be discussed below.

The problems reported concerning cloning with pRSF-oRibo-Q1 were also encountered by other groups and even members of the Chin group themselves (Nicholas Huguenin-Dezot, Chin group, personal communication). The source of the high recombination rate is likely to arise from its size (>10 kb) as well as the duplication of backbone sequences [224, 348]. Different cloning strategies were explored, and it was found that recombination was absent using a minimized protocol of the commercially available In-Fusion cloning method.

Yields of full-length PR65 protein containing two UAAs were less than expected when alkPyl was incorporated. Nguyen *et al.* [350] originally reported higher incorporation efficiencies of alkPyl than azPyl, which is in contrast to what is observed here. Furthermore,



(a) azPR65az

(b) cycPR65cyc

(c) yPR65y

(d) yPR65-GSy

(e) yCTPR_RV10y

(f) yPR65-277az

Figure 5.15: Example force extension curves demonstrating successful attachment. Raw and smoothed data are shown in grey and colours, respectively. Arrows indicate the pulling direction of the corresponding trace.

Table 5.2: Overview of handle attachment approaches tried.

Method	**Yield**	**Stability of modification**	**Oligo modification**	**Attachment species**	**Attachment efficiency**[a]
CuAAC	<1 mg l^{-1}	good	-	single	0%
SPAAC	1-2 mg l^{-1}	azide reduction	complete	single, double	30-40%
IE-DA	$\sim$1 mg l^{-1}	good	incomplete	single, double	$<5\%$
ybbR	>10 mg l^{-1}	-	complete or not known[b]	single, double	30-40%

[a] Proportion of protein conjugated to two DNA oligos after overnight incubation.

[b] Commercial or home-made, respectivley.

Wang *et al.* [348] reported 20-60 % of WT yields when UAAs were incorporated in proteins containing a single amber stop codon, whereas only slightly more than 20% of WT yields were obtained with two amber stop codons. If native PR65 is expressed at 5-10 mg l^{-1} then the expected yields would be 1-2 mg l^{-1}. However, yields of PR65 incorporating two UAAs were almost always higher than incorporation of a single UAA internally. The exact reason for this is currently unknown, but may be due to the variability observed between different proteins and positions therein (Kaihang Wang, Chin group, personal communications).

Using tandem mass-spectrometry of trypsin digested protein, it was possible to prove the incorporation at the N- and C-terminal positions, albeit not for internal residues. No alterations or aberrant read through of the amber codon was observed in samples of PR65 containing either alkyne or cyclopropene. However, although azPyl containing PR65 was purified without DTT, a significant reduction from azide to amine could be detected using mass-spectrometry. This could be due to potential side reactions with endogenous thiols or reducing agents during expression in *E. coli* [329] and highlights the required absence of reducing agents from any reactions containing azides. Furthermore, a C-terminal peptide containing lysine instead of azPyl was observed, indicating that there was some read through of the stop codon.

In the future, the use of alkyne-bearing pyrrolysine will be limited to labelling reactions, which were as fast as previously reported [224, 336]. In a control reaction of 5-FAM-alkyne to azide oligo it was shown that the oligo itself inhibits CuAAC in some manner, perhaps due to either chelation of Cu^{2+} ions by the phosphate backbone or association with amines in the nucleotides. After adding catalyst in large excess, increased reactivity was observed, although only protein attached to one DNA oligo was obtained in a detectable quantity.

Currently, the use of cyclopropene modifications for force-spectroscopy is limited by the production of sufficient tetrazine-functionalised DNA oligo. Mixtures of unconjugated and tetrazine-modified oligo were reacted to PR65 containing cyclopropene with some success, albeit the low yield subsequently affected the number of successful trapping events during force spectroscopy. Hydrolysis of the NHS-ester competes with the coupling to primary amines, but the conjugation efficiency remained unaltered when reactions were carried out at different pHs. Hydrolysis is slower at 0°C, but performing the reaction at this temperature did not result in any conjugation at all. It is possible that the NHS-ester hydrolysed over time during storage at -20°C and hence, it may be necessary to verify its current state. Furthermore, it will be necessary to develop an HPLC purification protocol that can separate modified from un-modified DNA oligo.

The Sfp-synthase-mediated conjugation of CoA to a ybbR-tag was explored as an alternative to bio-orthogonal chemistries. Since only 11 residue were added, proteins can be expressed normally and purified to the usual yields. However, in-house production of the CoA-oligo revealed to be unreliable as the final product formed precipitates with Mg^{2+}. The precipitate was reproducible and independent of buffer, pH and protein. Since these results were never observed using the commercially modified oligo, this indicates that the protocols for in-house production and/or purification may require alteration or optimization. Labelling of ybbR-tagged protein using the commercial oligo was more successful and no dependence on $MgCl_2$ concentration was observed in HEPES buffer at pH 7.5. However, in sodium phosphate, an increase in $MgCl_2$ could compensate to some extent for the decrease in activity of Sfp. This decrease in activity may be due to (a) pH [322], or (b) sequestration of Mg^{2+} by phosphate ions.

Both SPAAC and Sfp-reactions are similarly fast as long as Sfp-synthase is used in μM concentrations. Bio-orthogonal chemistry allows site-specific modification of proteins without requiring the deletion of un-wanted cysteines or introduction of protein or peptide tags. Various chemistries are possible [328], only a few of which have been explored here. One disadvantage of using UAAs is that they can be expensive depending on the functionality, and protein yields are relatively low. However, yields from one litre of bacterial culture are more than sufficient for multiple tests and large-scale conjugations for force-spectroscopy. In contrast, Sfp-mediated site-specific modifications require the introduction of a small 11-residue tag which was not observed to interfere with protein expression. Sfp-synthase itself can be produced easily and cheaply to very high yields and reducing agents do not interfere with the reaction. In this conjugation method, the CoA oligos are the expensive component and prices are of similar magnitude as UAAs. Due to its helical propensity [322] the introduction of the ybbR-tag can significantly alter the stability of the protein (see Chapter 7). Internal introductions of loops can already destabilize proteins, and in a repeat protein their effects will be position dependent [114]. How a helical insertion affects proteins remains to be investigated.

In conclusion, using both SPAAC, IED-DA and Sfp-mediated reactions produced successful attachments between protein and DNA oligonucleotides. They therefore extend the toolbox available to scientists that want to interrogate single protein molecules by force. In the process of optimizing the Sfp-synthase reaction conditions, I was able to show that sufficient amounts of final product could be obtained in various buffers as long as high amounts of Sfp were used, which could be useful for proteins with a theoretical pI close to 7.5. Finally, the reaction conditions explored were compatible with SPAAC leading to protein-DNA chimeras that had one oligo attached N-terminally using a ybbR-tag and one oligo attached internally using SPAAC.

Chapter 6

The effect of ybbR-tags on repeat protein stability

6.1 Introduction

It is of general interest to compare the folding behaviour of proteins both using ensemble and single molecule techniques to ensure that each is an accurate representation of the system. Furthermore, any modification of proteins for attachment purposes could potentially affect their stability and folding pathways (personal communications, Rief (TUM) and Gaub (LMU) labs). Therefore, one needs to examine how the incorporation of UAAs, ybbR-tags as well as the conjugation to DNA oligonucleotides affects the individual protein system and subsequently their force- or chemical-induced unfolding response. Proteins with un-natural functional groups were too low in yield to allow ensemble measurements to be made. Furthermore, the protein-oligo attachments could not be carried out on a scale sufficient for biophysical measurements due to the expense and availability of functionalised DNA oligos. Therefore, only un-conjugated ybbR-tagged constructs were subjected to further biophysical studies.

In the case of PR65, I have force spectroscopy data of different attachments (SPAAC, IED-DA, ybbR-CoA) which should help distinguish whether they have different effects on protein stability, if any effect at all (Chapter 8). For the CTPRs, however, all of the data were obtained using ybbR-CoA attachments, and therefore the effects of the ybbR-tag were explored in more detail in this Chapter. All ybbR-tagged proteins were subjected to both circular dichroism (CD) spectroscopy and chemical-induced equilibrium denaturations (monitored by fluorescence). Sample availability was limited in some cases, especially that of PR65 variants, which impeded the generation of independent triplicates. Therefore, the data and discussion presented here are limited to preliminary, single measurements and technical duplicates/triplicates.

6.2 Methods

6.2.1 Far-UV Circular Dichroism spectroscopy

Far-UV spectra were taken using 600 µl samples in a 2 mm quartz cuvette using a Circular Dichroism (CD) spectrophotometer (Applied Photophysics). CTPR_RV samples were prepared at 10 µM (TPR2 to TPR5) and 5 µM (TPR8 and TPR10) in 50 mM sodium phosphate pH 6.8, 150 mM NaCl. PR65 variants were diluted to concentrations between 0.5 and 1 µM using 25 mM MES pH 6.5, 1 mM DTE. Pure samples of both CTPR_RV and PR65 variants exhibit A_{260}/A_{280} ratios of ~0.4 and ~0.5, respectively, that is lower than the expected value of pure protein (0.67). Therefore, the concentration, c, was corrected using

$$c^{*} = 1.55c - 0.76c\frac{A_{260}}{A_{280}}. \tag{6.1}$$

The molar ellipticity, $[\theta]$, can then be calculated using

$$[\theta] = \frac{\theta}{lc^{*}}, \tag{6.2}$$

where θ is the ellipticity in millidegrees and l the path length of the cuvette.

6.2.2 Equilibrium denaturations

Denaturing TPR proteins using GdHCl

Samples of a total volume of 150 µl were prepared in a 96-well format (Greiner, medium-binding), in 50 mM sodium phosphate pH 6.8, 150 mM NaCl with guanidinium hydrochloride (GdHCl) gradients of 0 to 4.5 M (CTPR_RV2 and yCTPR_RV3y) or 0 to 6 M (all other proteins). The final protein concentrations ranged from 0.3 µM for TPR10 to 11.3 µM for TPR2 and yTPR3y. That is, concentrations were adjusted according to protein size and the amount available for the experiment. Samples were incubated on an orbital shaker at 25°C for 2h. Tryptophan residues were excited at 295 ± 10 nm and fluorescence was monitored at 360 ± 10 nm using a CLARIOStar microplate reader (BMG Labtech). Due to the deletion of tryptophan residues from the CTPR_RV variant, tyrosine residues were excited at 280 ± 10 nm and their fluorescence measured at 330 ± 10 nm. The data from 9 reads were averaged, then normalised and fitted to a two-state equation:

$$F_{norm} = \frac{\alpha_N + \beta_N D + (\alpha_U + \beta_U D)\exp(\frac{m(D-D_{50\%})}{RT})}{1 + \exp(\frac{m(D-D_{50\%})}{RT})}, \tag{6.3}$$

where D is the denaturant concentration, α_N and α_U the fluorecence signal of the native and unfolded state at 0.0 M Urea, β_N and β_U are their respective rates of change with increasing denaturant, $D_{50\%}$ is the midpoint of the unfolding transition, and m is the constant of proportionality related to the change in solvent accessible surface area upon unfolding. Equation 6.3 is based on the assumption that the free energy of unfolding is linearly dependent on on the denaturant concentration

$$\Delta G_{U-N} = -RT \ln\left(\frac{[\mathrm{U}]}{[\mathrm{N}]}\right) = \Delta G_{U-N}^{H_2O} - m_{U-N}D, \tag{6.4}$$

where $\Delta G_{U-N}^{H_2O}$ is the free energy of unfolding in water, and [U] and [N] are the concentrations of the unfolded and native (folded) states, respectively. At the transition midpoint, $D_{50\%}$,

$$[\mathrm{U}] = [\mathrm{N}] \tag{6.5}$$

resulting in

$$\Delta G_{U-N} = -RT \ln(1) = \Delta G_{U-N}^{H_2O} - m_{U-N}D_{50\%} = 0 \tag{6.6}$$

$$\Delta G_{U-N}^{H_2O} = m_{U-N}D_{50\%}, \tag{6.7}$$

which allows us to calculate the free energy of unfolding in water from a two-state fit.

Assuming that all protein is folded at zero denaturant and fully unfolded at high denaturant concentrations, the fluorescence can be converted into fraction folded, θ, or fraction unfolded, $1 - \theta$, using

$$F = (\alpha_N + \beta_N D)\theta + (\alpha_U + \beta_U D)(1 - \theta), \tag{6.8}$$

which can be rearranged to give

$$\theta = \frac{F - \alpha_U - \beta_U D}{\alpha_N - \alpha_U + (\beta_N - \beta_U)D}, \tag{6.9}$$

or

$$1 - \theta = 1 - \frac{F - \alpha_U - \beta_U D}{\alpha_N - \alpha_U + (\beta_N - \beta_U)D} = \frac{-F + \alpha_N + \beta_N D}{\alpha_N - \alpha_U + (\beta_N - \beta_U)D} \tag{6.10}$$

Denaturation curves converted to the form of $\theta(F)$ are required for Ising model fitting.

Denaturing PR65 using urea

All PR65 variants were buffer exchanged into 50 mM MES pH 6.5, 1 mM DTE/DTT. If protein samples were used for both CD and equilibrium denaturations monitored by fluorescence DTE was used. If protein samples were used for equilibrium denaturations only, DTT was used. Samples of a total volume of 90 µl were prepared in a black 384-well plate format (Perkin Elmer, OptiPlate-384F HB) with urea gradients of 0 to 7-8 M. The final protein concentrations ranged from 0.5 to 1.5 µM. Samples were incubated on an orbital shaker at 25°C for 2h. Tryptophans were excited at 295± nm and emission was monitored at 340 ± 10 nm using a CLARIOStar microplate reader (BMG Labtech). The data from 4 reads were averaged, then normalised and fitted to a three-state equation:

$$F_{norm} = \frac{F_N + \exp\left(\frac{m_1(D - D_{50\%,1})}{RT}\right)\left[F_I + F_U \exp\left(\frac{m_2(D - D_{50\%,2})}{RT}\right)\right]}{1 + \exp\left(\frac{m_1(D - D_{50\%,1})}{RT}\right)\left[1 + \exp\left(\frac{m_2(D - D_{50\%,2})}{RT}\right)\right]}, \tag{6.11}$$

where F_N and F_U are the fluorescence of the folded and denatured states, respectively, and can be described by

$$F_N = \alpha_N + \beta_N D \tag{6.12}$$

$$F_U = \alpha_U + \beta_U D, \tag{6.13}$$

and F_I, the fluorescence of the intermediates, is assumed to be constant. Subscripts of 1 and 2 refer to the first and second transitions, respectively.

6.2.3 Ising model formalism

Data from different CTPR_RV denaturations were fitted using homozipper and heteropolymer Ising models from the PyFolding suite [354], which are based on the formalism developed by Aksel and Barrick [355]. Any Ising model is based on the assumptions that the free energy of (un)folding can be decomposed into the intrinsic energies of each repeat, ΔG_i, and the energy of interfaces between two repeats, ΔG_{ij}, where $j = i+1$ and hence ΔG_{ij} describes the coupling between the i^{th} and the $(i+1)^{th}$ repeat. These can be combined linearly, which, in the simplest case, gives the free energy of an N-repeat system of

$$\Delta G = N\Delta G_i + (N-1)\Delta G_{ij} = -RT\ln(\kappa^N \tau^{N-1}), \tag{6.14}$$

where R is the gas constant and T the temperature. κ and τ are the respective equilibrium constants (or statistical weights) of intrinsic folding and interfacial interactions, and are defined as

$$\kappa = e^{-\Delta G_i/RT} = e^{-(\Delta G_{i,H_2O} - m_i D)/RT} \tag{6.15}$$

$$\tau = e^{-\Delta G_{ij}/RT} = e^{-(\Delta G_{ij,H_2O})/RT}. \tag{6.16}$$

Depending on the underlying properties and hence the partition function of the equilibrium ensemble, one can define the following cases:

1. **The homozipper approximation** can be applied to protein systems that have identical repeats and large mismatches between ΔG_i and ΔG_{ij} such that the state of any given repeat is highly coupled to its neighbours. This means that any intermediate folding states where terminal repeats are folded while central repeats are unfolded is highly unlikely and can be excluded. For example, in a protein of $N = 8$ repeats, repeats 1, 2, 7 and 8 cannot be folded while repeats 3-6 are unfolded.

2. **The homopolymer model** can be applied to protein systems with identical repeats independently of the coupling strength between neighbouring repeats. Intermediates with gaps, i.e. in which folded repeats are separated by unfolded repeats, have a finite probability and hence have to be included. For example, in a protein of $N = 8$

repeats, microstates in which either repeats 1, 2, 7 and 8, or 2, 3, 4 and 7,8 are folded, must now be accounted for.

3. **The heteropolymer model** describes the case where repeats are not identical, and hence ΔG_i and ΔG_{ij} can take different values within one protein. Common examples are: a protein of $N = 3$ with N- and C-terminal capping repeats added for increased solubility, or a protein of $N = 3$ where one repeat has a point mutation that significantly alters the overall stability.

These three systems only differ in their expression of θ, the fraction of folded repeats, which is derived from the partition function, $q(N)$, of the each system. In the following sections, derivations of these parameters are highly abridged; for derivations in full detail see Aksel and Barrick [355].

Homozipper approximation

Considering the number of ways, $\Omega_{i,g}$ that i out of N folded repeats can be arranged with g gaps (unfolded repeats), the partition function is

$$q = 1 + \sum_{i=1}^{N} \sum_{g=0}^{i-1} \Omega_{i,g} \kappa^i \tau^{i-1-g}. \tag{6.17}$$

However, in the homozipper approximation gaps between folded repeats are absent, and hence $g = 0$. In this case $\Omega_{i,g=0} = (N - i + 1)$ and the partition function becomes

$$q = 1 + \sum_{i=1}^{N} (N - i + 1) \kappa^i \tau^{i-1-g}, \tag{6.18}$$

which can be simplified to

$$q = 1 + \frac{\kappa[(\kappa\tau)^{N+1} - (N+1)\kappa\tau + N]}{(\kappa\tau - 1)^2}. \tag{6.19}$$

Given a fractional population p_i of the i^{th} partly folded macrostate, the fraction of repeats that are folded can be calculated using

$$\theta = \frac{1}{N} \sum_{i=0}^{N} i p_i = \frac{\kappa}{Nq} \frac{\partial q}{\partial \kappa}, \tag{6.20}$$

which becomes

$$\theta = \frac{\kappa}{N(\kappa\tau - 1)} \frac{N(\kappa\tau)^{N+2} - (N+2)(\kappa\tau)^{N+1} + (N+2)\kappa\tau - N}{(\kappa\tau - 1)^2 + \kappa[(\kappa\tau)^{N+1} - (N+1)\kappa\tau + N]}. \tag{6.21}$$

Homopolymer model

The inclusion of gaps in Equation 6.17 makes the system highly degenerate and increases the complexity especially for proteins of large N. Instead the partition function, $q(N)$, is taken as the probability of each of the possible confirmation, and is defined recursively by progressively including proteins with fewer repeats (e.g. $q(N-1)$). This allows q to be defined as sum of two partition functions that describe folded states, $q_f(N)$, and unfolded states, $q_u(N)$, of individual repeats, respectively:

$$q(N) = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} q_f(N) \\ q_u(N) \end{bmatrix} \tag{6.22}$$

$$= \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} \kappa\tau & 1 \\ \kappa & 1 \end{bmatrix}^N \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{6.23}$$

$$= \begin{bmatrix} 0 & 1 \end{bmatrix} W^N \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \tag{6.24}$$

where W is the statistical weight matrix. Treating W as an eigenvalue problem gives

$$W = TDT^{-1}, \tag{6.25}$$

where D is a diagonal matrix containing λ_1 and λ_2, the eigenvalues of W, and T consists of the corresponding eigenvectors of W. Hence, the partition function becomes

$$q(N) = \begin{bmatrix} 0 & 1 \end{bmatrix} T \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}^N T^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \tag{6.26}$$

which after the calculation of eigenvalues and eigenvectors yields

$$q(N) = \frac{\kappa(1-\tau)(\lambda_1^N - \lambda_2^N) + \lambda_1^{N+1} - \lambda_2^{N+1}}{\lambda_1 - \lambda_2}, \tag{6.27}$$

with

$$\lambda_1, 2 = \frac{1}{2}\left(\kappa\tau + 1 \pm \sqrt{(\kappa\tau - 1)^2 + 4\kappa}\right) \tag{6.28}$$

Finally, differentiating with respect to κ gives the fraction of folded repeats

$$\begin{aligned} \theta =& \frac{\kappa}{N}\left[\frac{\frac{\partial\lambda_2}{\partial\kappa} - \frac{\partial\lambda_1}{\partial\kappa}}{\lambda_1 - \lambda_2}\right. \\ &\left. + \frac{(1-\tau)[\lambda_1^N - \lambda_2^N + \kappa N(\lambda_1^{N-1}\frac{\partial\lambda_1}{\partial\kappa} - \lambda_2^{N-1}\frac{\partial\lambda_2}{\partial\kappa})] + (n+1)(\lambda_1^N\frac{\partial\lambda_1}{\partial\kappa} - \lambda_2^N\frac{\partial\lambda_2}{\partial\kappa})}{\kappa(1-\tau)(\lambda_1^N - \lambda_2^N) + \lambda_1^{N+1} - \lambda_2^{N+1}}\right]. \end{aligned} \tag{6.29}$$

Heteropolymer model

When a protein is composed of repeats of varying ΔG_i and ΔG_{ij}, the partition function needs to include a unique statistical weight for each repeat, i,

$$q(N) = \begin{bmatrix} 0 & 1 \end{bmatrix} W_1 W_2 \cdots W_N \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \tag{6.30}$$

where

$$
\begin{aligned}
W_i &= \begin{bmatrix} \kappa_i \tau_{i-1,i} & 1 \\ \kappa_i & 1 \end{bmatrix} \\
\kappa_i &= e^{-\Delta G_i/RT} = e^{-(\Delta G_{i,H_2O} - m_i D)/RT} \\
\tau_{i-1,i} &= e^{-\Delta G_{i-1,i}/RT} = e^{-(\Delta G_{ij,H_2O})/RT}.
\end{aligned}
\tag{6.31}
$$

However, since there is no unified value of κ, differentiation in the manner applied above is not possible. Instead, the fraction of folded repeats is given by the average probability that each of the repeats is folded:

$$
\theta = \frac{1}{N} \sum_{i=1}^{N} \theta_i, \tag{6.32}
$$

where θ_i describes the probability of the i^{th} repeat being folded. θ_i in turn, can be related to the partition function through a sub-partition function

$$
q_i = \begin{bmatrix} 0 & 1 \end{bmatrix} W_1 W_2 \cdots W_{i-1} \begin{bmatrix} \kappa_i \tau_{i-1} & 0 \\ \kappa_i & 0 \end{bmatrix} W_{i+1} \cdots W_N \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \tag{6.33}
$$

which sums over all states with the i^{th} repeat folded and eliminates all conformations where the i^{th} repeat is unfolded, giving

$$
\theta_i = \frac{q_i}{q(N)}. \tag{6.34}
$$

Hence the fraction folded in Equation 6.32 becomes

$$
\theta = \frac{1}{Nq(N)} \sum_{i=0}^{N} q_i. \tag{6.35}
$$

6.3 Results

6.3.1 Far-UV circular dichroism of ybbR-tagged proteins

Far-UV CD spectra of three ybbR-tagged CTPR_RV proteins ($N = 3, 5, 10$) and three un-tagged proteins ($N = 2, 4, 8$) are shown in Figure 6.1. Originally, these proteins were chosen because reliable Ising model fitting requires $\geq$ 3 proteins of different length for accurate minimization. All CTPR_RV proteins exhibit a CD spectrum characteristic of TPRs, in which the double minimum typical of α-helical proteins shows a reduced signal at 208 nm relative to the minimum at 222 nm [124, 139]. For different TPR proteins it was furthermore shown that the molar ellipticity scales linearly with the number of repeats between $N = 3$ and $N = 6$ [124, 139]. Therefore, the molar ellipticity at 222 nm was used to back-calculate the number of repeats. A helical ybbR-tag could be expected to decrease the ellipticity of a protein further, since two tags at either end of the array would be equivalent to adding 0.65 repeats. A linear trend of the molar ellipticity with repeat number can indeed be observed (Figure 6.1). However, due to the lack of replicates, an uncertainty of the readings cannot be calculated and hence it is impossible to determine whether or not the addition of ybbR-tags can be detected using CD.

CD spectra of PR65 wild-type, yPR65y and yPR65-GSy are shown in Figure 6.2. In the context of PR65, a set of terminal ybbR-tags would be expected to decrease the ellipticity by only 4%. However, when the signal at 222 nm is normalised relative to PR65, both variants exhibit a much larger variation from the WT (roughly 30% and 10% for yPR65y and yPR65-GSy, respectively).

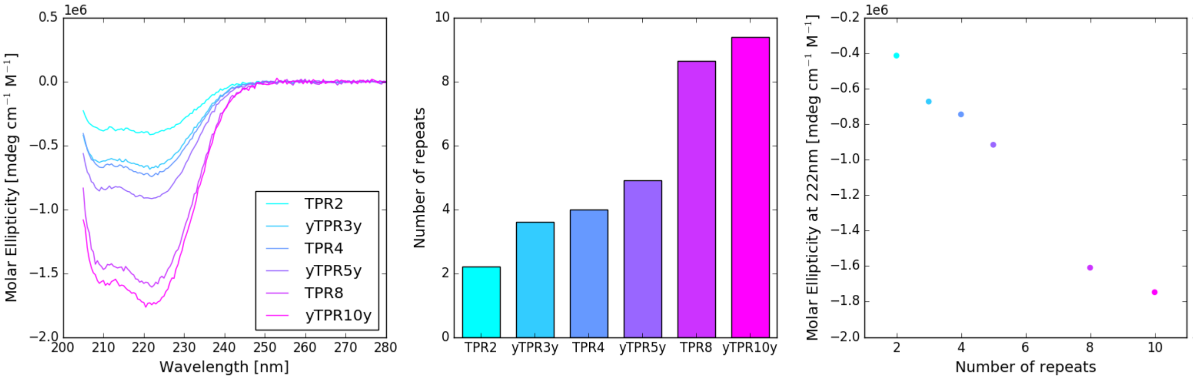


Figure 6.1: CD of CTPR_RV proteins. Left: Far-UV CD spectra of the individual proteins. Middle: Conversion of the molar ellipticity at 222 nm to ellipticity per repeat, and hence number of, repeats by normalisation using the signal from CTPR_RV4. For ease of the reader, the label "CTPR_RV" was abbreviated to "TPR" in the figure legend.

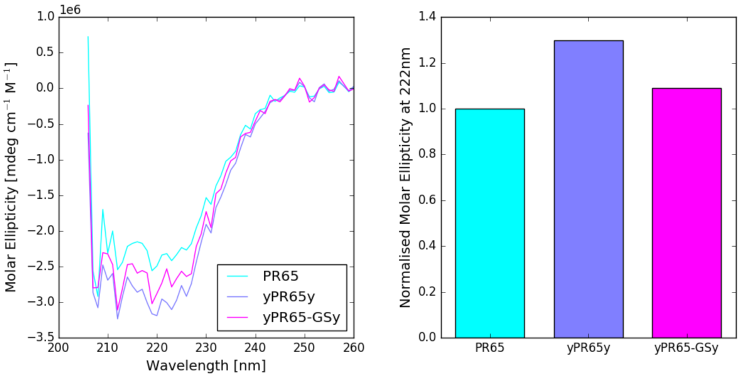
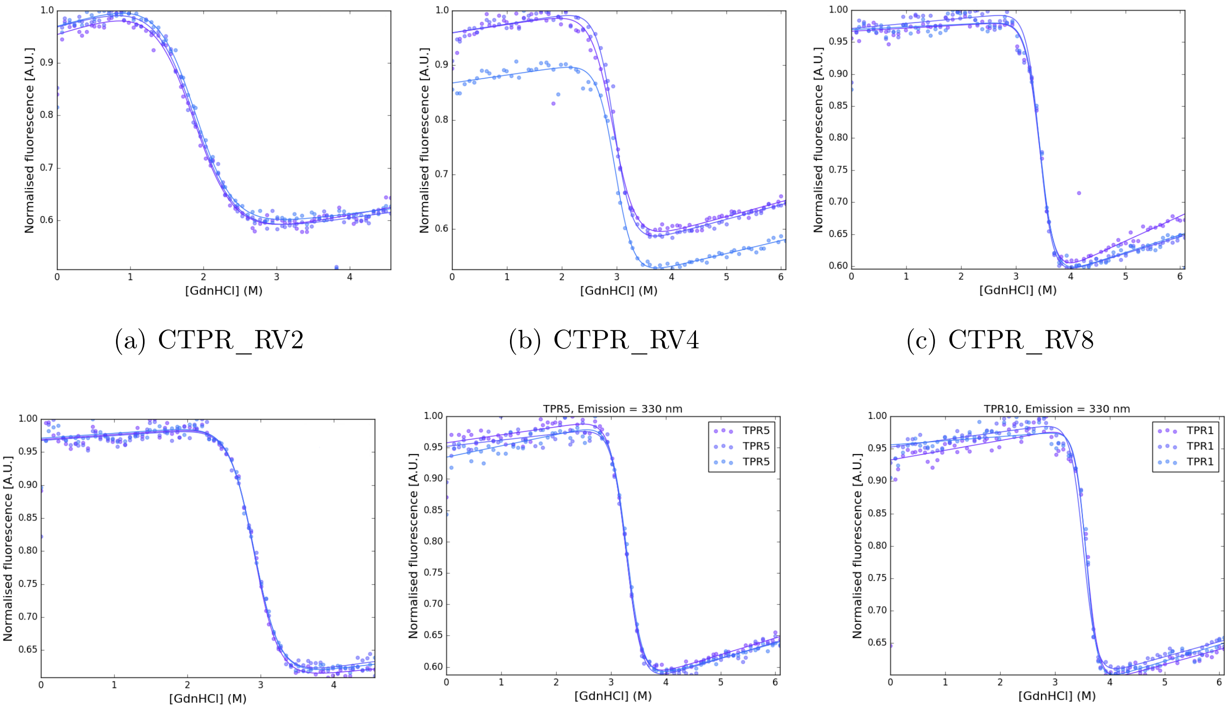


Figure 6.2: CD of ybbR-tagged PR65 variants. Left: Far-UV spectra, and right: normalized molar ellipticity at 222 nm.

6.3.2 Chemical stability of ybbR-tagged proteins

The fluorescence monitored equilibrium denaturation curves of all CTPR_RV constructs discussed in the previous section are shown in Figure 6.3. Two-state models were used to extract parameters for the native and unfolded baselines, the transition mid-point ($D_{50\%}$) and the m-value (denaturant sensitivity of the free energy). Results of the fits and the corresponding free energies of unfolding in water are listed in Table 6.1. Compared to the original CTPR series, the mutations of the RV variant resulted in changes in stability of -2.25 and -2.4 kcal mol^{-1} for constructs with 2 and 4 repeats respectively. The increase in m-value for both tagged and un-tagged proteins is consistent with previously published data on TPR proteins, the unfolding of which is populated by intermediate states [113, 116, 124, 127, 128, 133]. Surprisingly, adding ybbR-tags significantly stabilises the repeat arrays (Table 6.1). The denaturation curve of yCTPR_RV3y gives the same midpoint and m-value, and hence the same free energy of unfolding as CTPR_RV4. Thereby, the energetic gain from adding the tag can compensate for the loss from the mutations, since the stability of yCTPR_RV3y and CTPR3 are nearly within error. However, the yCTPR_RV5y construct is as stable as a CTPR6, suggesting that the



(a) CTPR_RV2

(b) CTPR_RV4

(c) CTPR_RV8

(d) yCTPR_RV3y

(e) yCTPR_RV5y

(f) yCTPR_RV10y

Figure 6.3: Equilibrium denaturation using guanidinium hydrochloride of untagged CTPR_RV and ybbR-tagged CTPR_RV proteins. The curves represent three technical replicates for each experiment that were fitted independently using a two-state unfolding model.

exact effect of the ybbR-tag depends on repeat number. Although the free energies are similar to CTPR constructs, they differ in their transition midpoints and m-values. Unfortunately, more quantitative comparisons are not possible, as they would require data of more constructs with and without ybbR-tag.

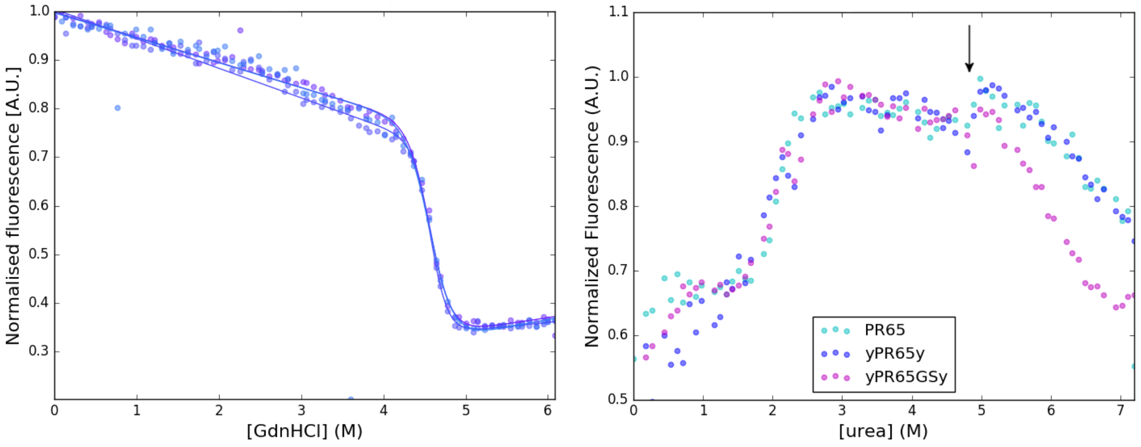
Table 6.1: Equlibrium denaturation fit parameters of CTPR_RV and CTPR proteins. The parameters represent averages ± s.e.m. of three technical replicates.

Protein	**D$_{50\%}$ [M]**	***m*-value [kcal mol^{-1} M^{-1}]**	**$\Delta G^{H_2O}_{U-N}$ [kcal mol^{-1}]**
CTPR_RV2	1.868 ± 0.004	2.17 ± 0.03	4.05 ± 0.07
CTPR_RV4	2.958 ± 0.009	3.4 ± 0.2	10.1 ± 0.6
CTPR_RV8	3.441 ± 0.003	4.69 ± 0.09	16.1 ± 0.3
yCTPR_RV3y	2.916 ± 0.001	3.47 ± 0.03	10.12 ± 0.09
yCTPR_RV5y	3.285 ± 0.005	4.20 ± 0.05	13.8 ± 0.2
yCTPR_RV10y	3.56 ± 0.01	5.4 ± 0.2	19.2 ± 0.8
yCTPR9y	4.577 ± 0.002	4.8 ± 0.4	22 ± 2
CTPR2[a]	2.98 ± 0.03	2.12 ± 0.04	6.3 ± 0.1
CTPR3[a]	3.66 ± 0.04	2.93 ± 0.07	10.7 ± 0.3
CTPR4[a]	3.90 ± 0.01	3.2 ± 0.1	12.5 ± 0.4
CTPR6[a]	4.23 ± 0.02	3.3 ± 0.2	14.0 ± 0.8

[a] As previously reported by Perez-Riba and Itzhaki [313].

[b] Errors were propagated using Equation 2.6.

I had originally chosen to use proteins of the RV series in force spectroscopy experiments because it was known from the denaturation of a 4-repeat protein that these proteins are less stable than the CTPRs [139]. Not knowing at which forces CTPRs unfold, especially considering their high thermal and chemical stability, I proceed with this less stable TPR variant to avoid the regime of non-linearity of the trap at higher forces. However, after preliminary data was available, I designed and produced ybbR-tagged CTPR proteins with 5 and 9 repeats. The latter was meant to contain 10 repeats to enable direct comparison to the CTPR_RV series. However, during gene construction, recombination by the *E. coli* (a common occurence with repetitive sequences) resulted in an array exactly one repeat shorter. At the time, it was necessary to proceed with this construct because a visit to the Rief lab to measure these constructs had already been arranged. Both constructs were purified in parallel and due to equipment problems, yCTPR5y was contaminated with yCTPR9y. TPR proteins of these sizes cannot be separated by size exclusion chromatography due to their particular elution profile. CTPRs elute with highly similar retardation factors, irrespective of repeat numbers and with a significant tail suggesting an interaction with the column matrix. For force-spectroscopy, this contamination did not matter as constructs with different number of repeats can easily be distinguished. However, the protein was not pure enough to obtain reliable chemical denaturation data. Therefore, equilibrium denaturation data were only obtained for yCTPR9y (Figure 6.4a), the two-state fitting results of which are included in Table 6.1. Again, when compared to an un-tagged CTPR, the free energy of unfolding of yCTPR9y is comparable to that previously published for an untagged CTPR10 containing two amino acid substitutions and a C-terminal capping helix [124].



(a)

(b)

Figure 6.4: Equilibrium denaturations of ybbR-tagged CTPR9a (a) and PR65 variants (b). Trypophan fluorescence (excited at 295 nm) was monitored at wavelengths of 360 (a) or 340 nm (b). Data in (a) represent three technical replicates in a 96-well plate format, while data in (b) are averages of two technical replicates of denaturations performed in 384-well plate format.

Denaturation curves of PR65 wild-type, yPR65y, and yPR65-GSy are shown in Figure 6.4b. Previous work from our group has shown that PR65 unfolds in two transitions, located at approximately 2 and 4.5 M urea, through a hyperfluorescent intermediate [70, 71, 284]. It has 5 tryptophans, one in HEAT4 and HEAT7 and one in each of HEAT11-13, which can report on the local unfolding [71]. Repeats HEAT3-10 and HEAT14-15 unfold in the first transition, while the more stable repeats HEAT1,2 and HEAT11-13 unfold in the second transition [71]. If ybbR-tags affect the stability of the array in a similar manner to that observed for TPRs, one would expect the N-terminal and C-terminal tags to primarily shift the second and first transitions to higher urea concentrations, respectively.

In the 384-well plate format applied here, the second transition is broader than seen with data obtained from a 1 ml sample using a fluorimeter. This meant that the unfolded baseline was not reached at 6 M urea and hence no meaningful three-state fits could be obtained. Furthermore, artefacts such as the discontinuity in fluorescence intensity seen at 5 M (see arrow in Fig. 6.4b) can be accredited to the edge wells of the plate, suggesting that significant evaporation must have taken place during repeat measurements of the plate.

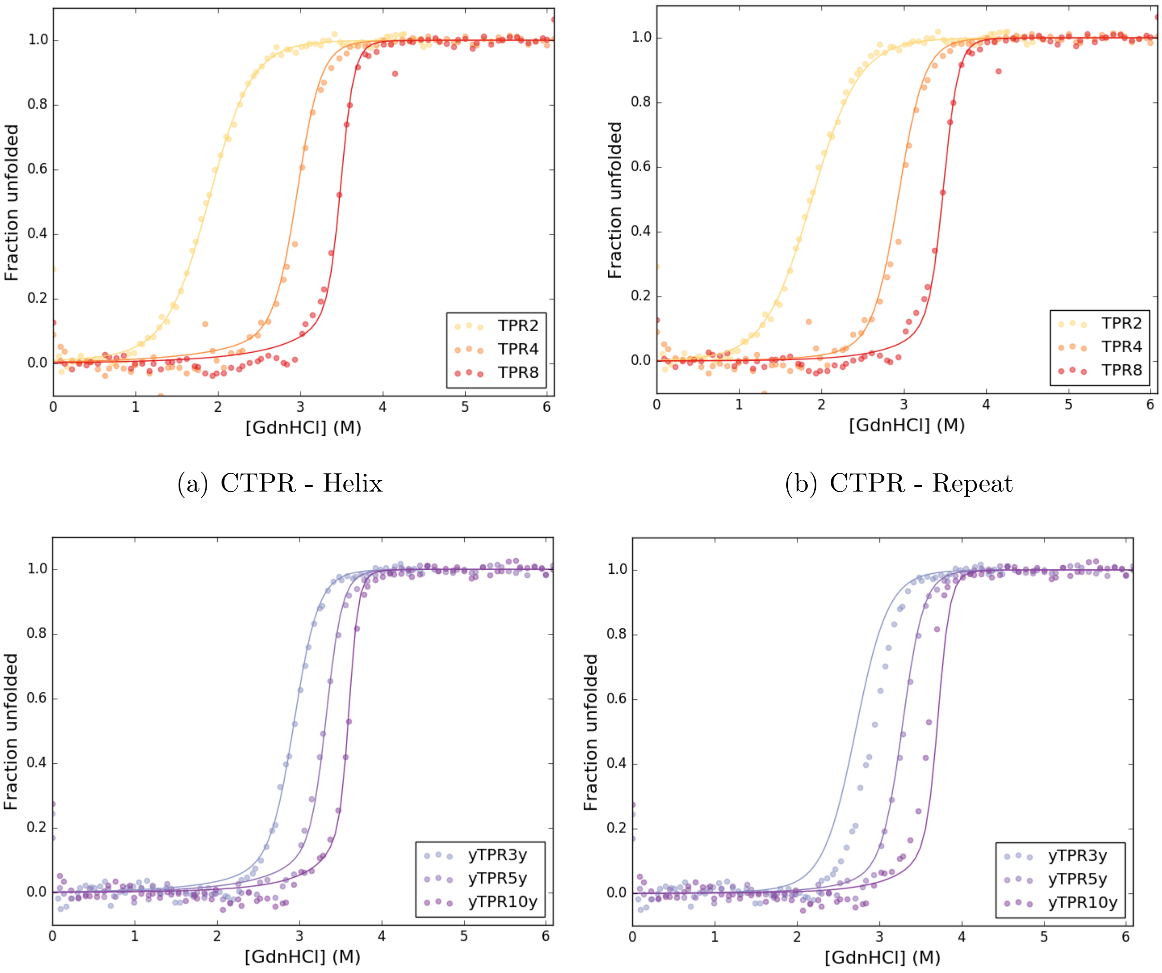
Although a quantitative description using a three-state model is not possible, qualitatively the three variants can be compared. First, the general trend of the unfolding transition is retained, indicating that the protein as a whole is not severely affected [70, 71]. Second, there is no apparent difference in the first unfolding transition between the three proteins, indicating that the C-terminal ybbR-tag does not affect the native stability and that this is independent of whether there is a GS linker or not. Third, PR65 and yPR65y exhibit the same unfolding behaviour in the second transition as well. Since yPR65y and yPR65-GSy have the same N-terminal tag, there must be another reason why the second transition of yPR65-GSy deviates from the other two proteins. The explanation could be the following: First, variations in protein concentration can affect the magnitude of the signal and hence the slope of each transition. Second, PR65 and yPR65y samples included in Figure 6.4b were on one 384-well plate while yPR65-GSy samples were on another plate. Since all curves overlap up until the discontinuity at 5 M, the edge effects may differ between plates.

6.3.3 Ising models of CTPR_RV repeats

To enable a crude understanding of the effects of ybbR-tags, two-state fits of CTPRs were presented in the previous section. However, as is apparent from the m-value and from the many data sets published by other labs, a two-state model is not sufficient to describe CTPR unfolding beyond two repeats. Hence, denaturation data of CTPR_RV proteins were transformed to fraction unfolded, averaged across a triplicate and fitted using Ising models.

Homozipper approximation

First, I attempted to fit the data using the simplest of models, the homozipper approximation. Considering the stabilizing effects of the ybbR-tags, the two sets of proteins were analysed separately. Previous studies on CTPRs had taken the repeating unit to be an individual α-helix because their systems required capping helices for the proteins to be soluble [116, 124, 126–128]. Here however, no such caps were present as it was found recently that TPRs without a capping helix are still soluble and express in high yields [113]. Furthermore, since the A and B helices within a TPR repeat are not identical and do not form the same interactions with their nearest neighbours (which would theoretically require a heteropolymer model), it would be more logical to take a whole



(a) CTPR - Helix

(b) CTPR - Repeat

(c) yCTPRy - Helix

(d) yCTPRy - Repeat

Figure 6.5: Homopolymer Ising models of CTPR_RV proteins.

repeat as the repeating unit. Nevertheless, data of both helix and whole repeat models are presented here, and fits to the data and the corresponding parameters are displayed in Figure 6.5a,b and listed in Table 6.2, respectively. Both models fit the data very well, and show that the destabilizing effects of the mutations in the RV series are primarily due to a change in the intrinsic energy. In the helix model, ΔG_i increased by 0.85 kcal mol^{-1}, whereas in the repeat model it increased by 1.25 kcal mol^{-1}. In contrast, the interaction energies decreased by 0.22 kcal mol^{-1} irrespective of which repeating unit was chosen, and thereby compensate for some of the energetic gain caused by the increase of intrinsic energies. The denaturant dependency of ΔG_i was not affected in either model as values were within error.

In Table 6.3, the free energies of unfolding in water (as measured by the two-state fits) are compared to the conformational energies derived from the two homozipper models. For a 2-repeat protein the two parameters agree within error, as it can be described by either

Table 6.2: Homozipper Ising model fit parameters of CTPR_RV proteins. Fitted using PyFolding. Errors quoted are the errors of the fit.

	Repeat unit: helix		
Protein	**ΔG_i [kcal mol^{-1}]**	**m_i [kcal mol^{-1} M^{-1}]**	**ΔG_{ij} [kcal mol^{-1}]**
CTPR_RV	2.61 ± 0.06	-0.61 ± 0.01	-5.0 ± 0.1
yCTPR_RVy	1.36 ± 0.06	-0.79 ± 0.02	-4.4 ± 0.1
CTPRa [a]	1.76 ± 0.01	-0.63 ± 0.01	-4.78 ± 0.01
	Repeat unit: repeat		
Protein	**ΔG_i [kcal mol^{-1}]**	**m_i [kcal mol^{-1} M^{-1}]**	**ΔG_{ij} [kcal mol^{-1}]**
CTPR_RV	0.18 ± 0.03	-1.05 ± 0.02	-4.3 ± 0.1
yCTPR_RVy	0.00000 ± 0.00003	-1.07 ± 0.05	-4.3 ± 0.2
CTPRa[a]	-1.07 ± 0.03	1.03 ± 0.02	-4.08 ± 0.06

[a] As previously reported by Perez-Riba [139].

Table 6.3: Comparison of free energies derived from two-state and Ising model fits.

	2-state	**Helix**	**Repeat**
Protein	**$\Delta G_{U-N}^{H_2O}$ [kcal mol^{-1}]**[a]	**ΔG_{U-N} [kcal mol^{-1}]**[b]	**ΔG_{U-N} [kcal mol^{-1}]**[b]
CTPR_RV2	4.05 ± 0.07	4.6 ± 0.4	3.9 ± 0.2
CTPR_RV4	10.1 ± 0.6	14.1 ± 0.9	12.2 ± 0.4
CTPR_RV8	16.1 ± 0.3	33 ± 2	28.7 ± 0.8

[a] From Table 6.1.

[b] Calculated using Equations 6.14, 2.5 and 2.7.

two-state or Ising model [124]. For larger arrays the energies tend to vary significantly from those obtained by a two-state fit. It was not possible to compare these data with previously published results of TPR unfolding since these studies used (a) different buffer systems, (b) different consensus sequences, (c) a C-terminal capping helix, and (d) in most cases included that capping helix in the homozipper model although the sequence of the cap was altered slightly from the consensus [113, 116, 124, 126–128].

At first it was attempted to fit the ybbR-tagged proteins with a homozipper as well, based on the (possibly erroneous) assumption that a stabilizing effect of the tag would be "distributed" across the whole repeat array. When considering a helix as a repeating unit, the fit minimzed to parameters that described the denaturation data well enough (Figure 6.5c, Table 6.2). However, when the repeating unit was taken to be a repeat, the optimal fit parameters did not describe the data any longer but instead assumed some average values that fitted the 5-repeat protein, under-estimated the 3-repeat protein and over-estimated the 10-repeat protein. This clearly suggests that a different model is required to examine the effects of the ybbR-tags on the CTPR array.

Heteropolymer model

The ybbR-tag is different from the TPRs and hence should be modelled as a separate topology within the framework of a heteropolymer model. The PyFolding modelling suite has options for various topologies that could be used here:

- **Repeat domain (R),** which is the repeating unit either defined as a helix or whole repeat
- **Helix domain (H)** that has ΔG_i and ΔG_{ij} different from the repeat domain
- **Cap domain (C),** which are the same as the repeating unit just at the C-terminus of an array and as such should only be used when modelling C-terminal deletions [354].

Currently, it is not known whether or not N- and C-terminal ybbR helices stabilize the neighbouring repeats by exactly the same amount. Using I-Tasser and Robetta it was possible to gain insight into possible native states of a ybbR-tagged CTPR_RV (Figure

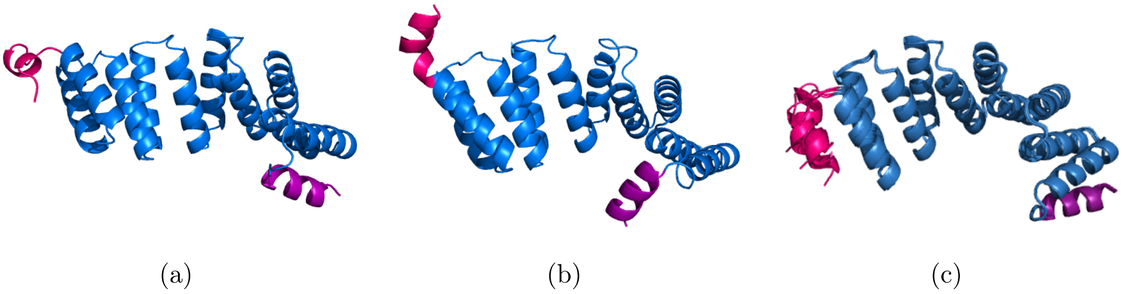


Figure 6.6: I-Tasser model 1 (a) and model 2 (b), and an overlay of the Robetta models (c) of yCTPR_RV5y, where the TPRs are blue, and the N- and C-terminal ybbR-tags are magenta and purple, respectively. (a) The C-terminal helix is packed against the last repeat forming a new interface, while the N-terminal ybbR-tag is partially unfolded to stack against the first repeat. (b) Neither ybbR-tag packed against the repeat, the N-terminal one however forms a continuation of the first repeat helix, in which case it is likely to affect the intrinsic stability of the repeat instead. The top two models were the only models out of the five provided that aligned to a CTPR crystal structure (PDBid 2hyz) with an RMSD of 1.0 Å or less. In all the other models, the repeat arrays were not formed properly, e.g. contained unfolded stretches, which would be observed as a destabilization in an experiment. (c) All Robetta models converged to the same conformation for the TPR repeats (RMSDs of <0.4 Å between all models) and the C-terminal ybbR-tag which docks against the final repeat. Only the N-terminal showed some conformational variability. An alignment of the CTPRs to a crystal (PDBid 2hyz) produced an alignment of 0.748 Å

6.6). Since the tags have been added to the repeat array without any linkers, it is more likely than not that the N- and C-terminal tags actually differ in their effects. That is, at the C-terminus the loop sequence of the preceding repeat could allow the ybbR-helix to stack against the interface of the final repeat (Figure 6.6a,c). This is likely the largest factor contributing to stabilization, as most natural repeat proteins have been found to have C-terminal capping helices [19]. Therefore, it is unlikely that the C-terminal ybbR-tag is decoupled as shown in Figure 6.6. However, no connecting loop is present at the N-terminus, and at least 4 amino acids would be required to connect a helix to the first repeat [19]. Therefore, the N-terminal ybbR-helix may either be decoupled as it can fold on its own and hence it would not affect the repeat array, or it may partially unfold to form a turn, thereby allowing the most N-terminal residues of the helix to stack against the interface of the first repeat (Figure 6.6a,c). In either case, it is assumed that the coupling between repeats is so strong that the repeat itself does not partially unfold. Lastly, it is possible that it simply forms an extension of the A-helix of the first repeat (Figure 6.6b)

Considering these options, the following topologies are possible (see Figure 6.7 for a visualization):

1. **$\mathbf{R_NH}$:** Only the C-terminal ybbR-tag contributes to the increase in stability and it can be modelled using a helix domain.

2. **$\mathbf{HR_NH}$:** Both, the N- and C-terminal ybbR-tags contribute to the increase in stability and are modelled using helix domains of the same intrinsic and interfacial energies.

3. **$\mathbf{HR_NC}$:** Both, the N- and C-terminal ybbR-tag contribute to the increase in stability but are modelled as separate entities. Although modelling the C-terminal ybbR-tag as a cap is theoretically wrong (since it is not a variation of the repeating unit), it is currently the only way to describe the case where N- and C-terminal ybbR-tags contribute by different amounts.

4. **$\mathbf{R_NC}$:** Only the C-terminal ybbR-tag contributes to the increase in stability and it can be modelled using a cap domain.

One problem with exploring all these topologies is that only six different protein constructs and hence only six unique denaturation data sets are available, theoretically allowing only ≤ 6 parameters to be fitted. That is, with the current data, topologies 2 and 3 will most definitely result in over-fitting. Nevertheless, global fits to the data using all four topologies did minimize and the corresponding parameters are listed in Table 6.4. However, independently of whether or not there were more parameters to be fitted than data sets available, the global fitting resulted in a non-unique, poor solution for most cases. Hence, it was not possible to obtain errors of the fit using the encoded method for most models [354]. Constraining the repeat parameters to those obtained from the homozipper did not affect the minimization and error estimation. Only in the helix $\text{HR}_\mathbf{N}\text{C}$ and repeat $\text{HR}_\mathbf{N}\text{H}$ topologies could errors be estimated, albeit they are very large. The fits corresponding to these two topologies, shown in Figure 6.8, are almost identical.

In all Ising models, the energetic parameters of the repeat domain are within error of or close to the homozipper values (Table 6.4). In cases of where the N-terminal ybbR-tag is assumed to leave the array unaffected and the C-terminal tag is modelled as a cap domain or helix domain, the C-terminal ybbR-helix has a small negative ΔG_i, indicating that it could be stably folded on its own. These trends are independent of which unit of repetition was chosen (repeat or helix), suggesting that the exact definition of the final element in the array does not matter. In fact, the values of ΔG_i and m_i obtained from these two types of models are very similar, and it was only the ΔG_{ij} values that differed substantially. Since ΔG_{ij} is defined as the interaction between i and $i+1$, the interfacial energy is irrelevant in this case, as the helix/cap constitutes the last motif.

In models having the $\text{HR}_\text{N}\text{H}$ topology, the fit assigns either a positive ΔG_i (helix) or approximately 0 kcal mol^{-1} (repeat), while the dentaurant dependence, m_i, is similar between models. The interface energy of the ybbR-helix topology, which is affecting the first repeat of the array, is negative for both models and is within error of that of the repeat topology.

When N- and C-terminal tag are modelled as different topologies, the N-terminal ybbR-helix is assigned a large negative ΔG_i and approximately zero ΔG_{ij}, which suggests it is somewhat decoupled from the N-terminal repeat or helix. The C-terminal ybbR-helix has a positive ΔG_i, indicating that the ΔG_{ij} of the preceding repeat or helix is necessary for it to remain folded.

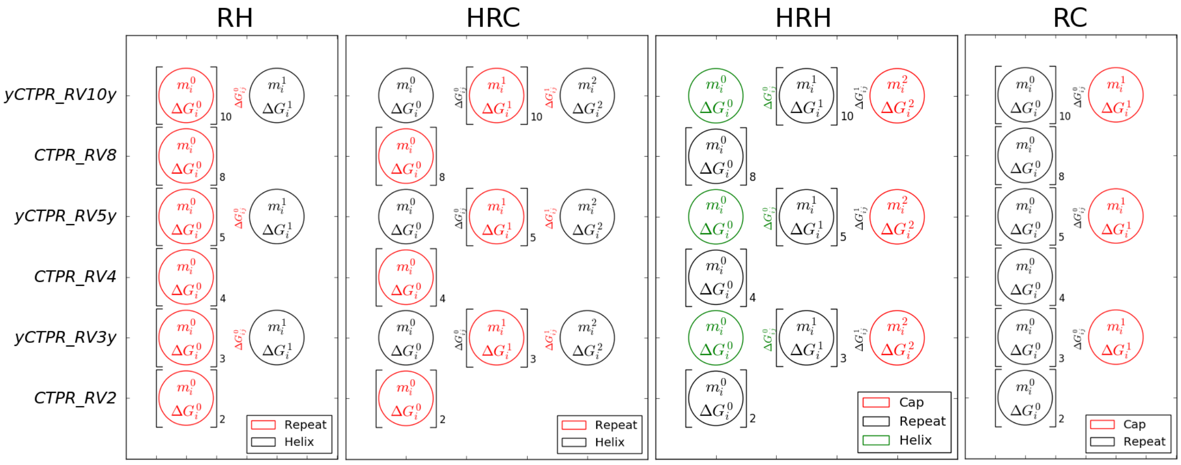


Figure 6.7: Topologies of the different models based on a whole repeat as the repeating unit. Models based on the helix as the repeating unit simply have double the repeats.

Table 6.4: Results of fits of heteropolymer Ising models to CTPR_RV proteins, obtained using PyFolding [354]. Errors of the fit are estimated using a numerical approximation of the covariance matrix using the Jacobian of the fit [354]. If the determinant of this Jacobian is zero, fitting errors will be infinite indicating a non-unique solution. Nevertheless, all models produce fits to the data with $R^2 > 0.99$.

Repeat unit: helix

Topology	$\mathrm{R_N}^a$	$\mathrm{R_NH}$	$\mathrm{HR_NH}$	$\mathrm{HR_NC}$	$\mathrm{R_NC}$
H ΔG_i [kcal mol^{-1}]	-	-0.45 ± inf	1.53 ± inf	-8.1 ± 0.3	-
H m_i [kcal mol^{-1} M^{-1}]	-	-1.11 ± inf	-0.97 ± inf	-2.5 ± 0.5	-
H ΔG_{ij} [kcal mol^{-1}]	-	-2.30 ± inf	-4.94 ± inf	0 ± 4	-
R ΔG_i [kcal mol^{-1}]	2.61 ± 0.06	2.55 ± inf	2.58 ± inf	3 ± 1	2.55 ± inf
R m_i [kcal mol^{-1} M^{-1}]	-0.61 ± 0.01	-0.61 ± inf	-0.61 ± inf	-1 ± 3	-0.61 ± inf
R ΔG_{ij} [kcal mol^{-1}]	-5.0 ± 0.1	-4.90 ± inf	-4.94± inf	- 5 ± 3	-4.90 ± inf
C ΔG_i [kcal mol^{-1}]	-		-	1 ± 3	-0.45 ± inf
C m_i [kcal mol^{-1} M^{-1}]	-		-	-1 ± 8	-1.11 ± inf

Repeat unit: repeat

Topology	$\mathrm{R_N}^a$	$\mathrm{R_NH}$	$\mathrm{HR_NH}$	$\mathrm{HR_NC}$	$\mathrm{R_NC}$
H ΔG_i [kcal mol^{-1}]	-	-0.39 ± inf	0 ± 3	-9.49 ± inf	-
H m_i [kcal mol^{-1} M^{-1}]	-	-1.22 ± inf	-1 ± 9	-1.01 ± inf	-
H ΔG_{ij} [kcal mol^{-1}]	-	-6.59 ± inf	-4 ± 2	-0.01 ± inf	-
R ΔG_i [kcal mol^{-1}]	0.18 ± 0.03	0.15 ± inf	0 ± 3	0.19 ± inf	0.15 ± inf
R m_i [kcal mol^{-1} M^{-1}]	-1.05 ± 0.02	-1.04 ± inf	-1 ± 2	-1.09 ± inf	-1.04 ± inf
R ΔG_{ij} [kcal mol^{-1}]	-4.3 ± 0.1	-4.22 ± inf	-4 ± 1	-4.47 ± inf	-4.22 ± inf
C ΔG_i [kcal mol^{-1}]	-		-	0.58 ± inf	-0.39 ± inf
C m_i [kcal mol^{-1} M^{-1}]	-		-	-1.01 ± inf	-1.22 ± inf

[a] Homozipper model of CTPR_RV2,4 and 8

6.4 Discussion

An isolated ybbR peptide has previously been shown to be α-helical in solution [322]. Hence, if its structure is retained in another protein context, it should add to that protein's molar ellipticity in a proportional amount. However, in the majority of the current CD data of CTPR proteins and PR65 it is not possible to resolve its contribution unequivocally due to experimental uncertainties. Only for yCTPR_RV3y it appeared as if the ybbR-tag added to the α-helical content of the final construct as expected. Due to time limitations, it was not possible to produce CTPR constructs of the same number of repeats without the ybbR-tag. These constructs, as well as multiple experimental repeats, will be necessary to examine in what extend the ybbR-tag adds to the α-helical content of CTPRs and PR65.

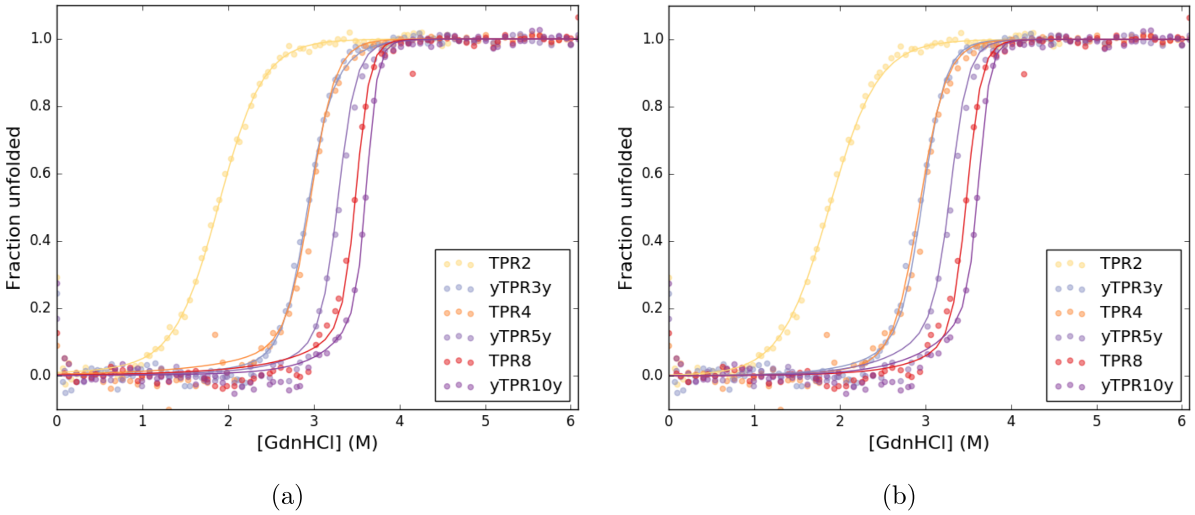


Figure 6.8: Heteropolymer Ising models of CTPR_RV variants corresponding to fits where the jacobian was non-zero. (a) Fits to the data based on a helix model of the topology HR_NC. (b) Fits to the data based on a repeat model of the topology HR_NH. Fit parameters are given in Table 6.4. Both models produce fits to the data with $R^2 > 0.99$.

In previous experiments, it was found that the signal of a PR65 denaturation curve was very noisy when dispensed into the same 96-well plates as those used for the CTPRs. This was most likely due to non-specific binding of PR65 to the plastic surface of the well since signal-to-noise ratios were comparable to fluorimeter data when dentaturations were performed in low-binding 384-well plates, the only non-stick plates available at the time. However, due to the edge effects observed with 384-well plates, PR65 denaturation should be performed in non-stick 96-well plates in the future. Nevertheless, the data obtained here suggests that the ybbR-tags did not affect PR65 unfolding and hence are likely to be folded independent of the array.[1] Since the ybbR helix does not contain tryptophans or tyrosines, an unfolding of such a short and therefore chemically weak helix would not be detected in this assay. Instead, if a ybbR-dependent difference in the ellipticity signal can be confirmed in tagged constructs, both transitions, that of PR65 and ybbR-helix unfolding, could be monitored using CD.

Denaturation data for various tagged and un-tagged CTPR_RV constructs could be obtained. The conservative re-shaping mutations of the RV series were meant to affect the interaction of two A-helices and hence the interface between consecutive repeats. As it was already known from data based on CTPR_RV4, these mutations resulted in a

[1] This experiment was repeated post-submission using non-stick 96-well plates and the corresponding data can be found in Appendix B, Figure B.1. A significant difference beyond the usual experimental variation could not be detected indicating that N- and C-terminal ybbR-tags have only minimal or no effect on the unfolding of PR65.

destabilization compared to the CTPRa series. However, the data presented here are the first that enabled Ising analysis on this series and thereby allowed to determine how these mutations affect ΔG_i and ΔG_{ij}. Ising model-derived energies compared to those calculated from a two-state fit exhibited a considerable difference for larger constructs. In some buffer systems, repeat arrays of up to $N = 6$ have been shown to obey a two-state unfolding transition [128], but clearly that is not the case here. The results furthermore show that the mutations cause a slight decrease of ΔG_{ij} while increasing ΔG_i. A possible reason for the unexpected decrease in ΔG_{ij} is that the interface is rearranged, resulting in the different repeat angles described in Chapter 3. That is, the interface repacking has likely occured to "optimize" ΔG_{ij} to form the most stable fold possible, leaving the intrinsic repeat energies to experience most of the destabilization. This observation suggests that any conservative interface alteration similar to those described here might simply result in novel stacking geometries instead of changing the interfacial energies significantly.

Adding a ybbR-tag to CTPR proteins (both RV and the original series) increased the stability to an extent that was similar to adding another repeat.[2] However, whereas the $m -$ value changed proportionally to N, adding two ybbR-tags to a 3-repeat array resulted in the same m- and $D_{50\%}$-value as those of an untagged 4-repeat array. This result might suggest that at least the population of equilibrium intermediates in the unfolding transition is unaffected. Furthermore, the ybbR-tags caused a stabilization that could compensate for the overall destabilization due to re-shaping, albeit in a different mechanism than the mutations as evidenced by varying m- and $D_{50\%}$-values. One possible explanation of why TPR stability is affected by the ybbR-tags and PR65 is not, lies within the nature of their first and final repeats. In PR65, those are natural capping repeats in which some hydrophobic residues have been substituted for polar amino acids. Our CTPRs, however, have no such capping repeats or helices, leaving the interfaces at either end exposed to the solvent and/or interactions with ybbR-tags.

To determine whether it was possible to differentiate contributions of the respective tags at the two ends of the array, Ising analyses were performed on the whole data set of the RV series. Although it was possible to fit the data using a homozipper model based on a helix (but not a repeat) repetition unit, this approach is clearly wrong given that ybbR helices are very different from a TPR helix and a protein BLAST could not detect any significant similarity. Yet, the result highlights how a model based on incorrect underlying assumptions can still fit these types of data very well, that is with $R^2 > 0.99$. I next attempted analysis of the denaturation data set using a heteropolymer approach. Differ-

[2] Un-tagged CTPR_RVs containing 5 and 10 repeats were built and tested post-submission, the data of which can be found in Appendix B, Figure B.2. Combined, the N- and C-terminal ybbR-tags cause a stabilisation of CTPR_RV arrays by approximately 4 kcal mol^{-1}.

ent hypothetical topologies were explored, all of which resulted in fits with $R^2 > 0.99$. Interestingly, when the N-terminal ybbR-tag was assumed to leave the repeat array unaffected, the results were independent of whether a helix or capping topology was assigned to the C-terminal ybbR-helix. Furthermore, results for ΔG_i and m_i were similar in both helix and repeat based models, indicating that the capping module may be appropriate even if the repetitive unit is a whole repeat. In models where both N- and C-terminal tags were assigned helix topologies, the interfacial energy of the helix module was the same as that of a repeat-repeat or helix-helix interface. However, this is unlikely to be the case in reality, independent of how each tag interacts with their respective end of the repeat array. Indeed structure predictions obtained using I-Tasser and Robetta show variable conformational states for both C- and N-terminal ybbR helices, confirming that they should be treated separately. However, separate assignment of either tag introduces more variable parameters than data sets to be fitted. Therefore, the results from such models should be treated with care, even if they suggest that the N-terminal ybbR-tag is decoupled from the array while the C-terminal tag interacts with it.

In all heteropolymer models, the estimated errors are either infinite or very large. It is likely that the number data sets provided is not large enough to distinguish between different contributions of the tags at either end. The fact that models lacking an N-terminal topology for the ybbR-tag did not minimize well even though it is theoretically not overfitting the data, suggests that the N-terminal ybbR-tag is involved in the stabilization of the array. Due to these issues, it will be necessary to construct a CTPR_RV series in which proteins of different N have no tag, an N-terminal or a C-terminal ybbR-tag. Furthermore, it would be interesting to build a similar series with the CTPRa repeat to examine whether ybbR stabilization depends on the repeat background.

Since, I lacked data for constructs without ybbR-tags I sought to find equivalent published data sets. However, it was not possible to compare my Ising model derived parameters to published data, for the following reasons:

- Differences in buffer systems have an impact on the final results, as salt content can affect the intrinsic repeat stability [128, 139].
- Differences in consensus repeat sequence can affect ΔG_i and ΔG_{ij} to varying degrees [111, 113, 116, 127, 127].
- The use of a C-terminal capping helix is very common and even if it differs in sequence, the data is still analysed using the homozipper model [116, 127, 128].

Since the only true repetitive unit is a whole repeat, not an individual helix, Ising models should be based on a whole repeat topology. This is already being done for consensus ankyrin repeats [133, 135] and has recently been adopted for TPRs with whole capping repeats [21]. However, some simulations showed that the unfolding of TPR is likely to occur helix by helix instead of in whole repeats [125], which would justify a helix based Ising model albeit by using a heteropolymer model that treats A- and B-helices separately. The choice of which repeating unit to take should be based on whether a whole repeat is stable or unstable. If the intrinsic energy of a whole repeat is positive, the rate-limiting step will be the formation of an interface between two repeats and hence a repeat-based Ising model is appropriate [138]. However, if the intrinsic energy of a whole repeat is negative, the rate limiting step is un-likely to be the formation of a repeat-repeat interface, but the formation of the repeat itself and hence a helix-based heteropolymer Ising model may be more appropriate.

Lastly, the coupling of the final repeat to the ybbR-tag has to be re-defined. In the current models the interfacial energy, or coupling interaction, is defined as that between repeat i and repeat $i + 1$. At the N-terminus, a ybbR-helix that interacts with the first repeat would be assigned its own interfacial energy. However, at the C-terminus the interfacial energy between the last repeat and a capping helix is that associated with the repeat, not the helix. Such a description can only be used when the capping helix and the preceding repeat share the same interactions as any intrinsic repeat with a consecutive A-helix. This was shown to be the case for the conventional CTPR capping helix [19], but is unlikely for a C-terminal ybbR-tag. Indeed, the lack of success in modelling a C-terminal tag and providing reasonable error estimates only underlines the fact that ybbR-helices cannot be treated in the same manner as a conventional cap. Instead, the final repeat could be modelled using the mutant repeat topology available in PyFolding to assign a separate energy term to the interface between final repeat and the ybbR-tag. Given data sets for N- or C-terminally tagged constructs, one would then be able to explore more appropriate Ising models in the future.

Chapter 7

Force-induced unfolding of CTPRs

7.1 Introduction

The force-induced unfolding of different repeat proteins has been studied previously by AFM and by MD simulations [191, 194–197, 200]. Of these studies only Li *et al.* [194] reported data on a consensus ankyrin repeat protein, while all others focussed on natural repeat proteins. Although the first natural and consensus TPR structures have been known since 1998 [356] and 2003 [19], respectively, Kim *et al.* [197] conducted an extensive study examining HEAT, armadillo, LRR and ankyrin repeats but did not include TPRs. Both natural and consensus TPRs form superhelical structures [19] and therefore are the only repeats that assemble into the geometry resembling a physical spring.

The initial force response of a 24-repeat ankyrin protein was shown to be linear, indicating a spring-like stretching of the repeat array before unfolding events of individual repeats were observed [196]. In an MD simulation of Importin-β, amino acids in the hydrophobic core of the repeat array re-arranged to allow the protein to extend under small forces without unfolding of individual repeats [200]. Generally however, repeat proteins unfold one or multiple repeats at a time when subjected to an external force. Consensus ankyrin repeats were shown to unfold at $\sim$ 50 pN, one repeat at a time [194]. In contrast, their temperature or chemical-induced unfolding in bulk was found to be highly cooperative and can be described using Ising models [24, 132, 133, 355]. In ensemble unfolding measurements, TPRs exhibit high cooperativity and can be described using Ising models [116, 355]. The intrinsic and interfacial energies are different for TPRs and ankyrins: while ankyrins have positive intrinsic and negative interfacial energies and a very large mismatch between them, the TPRs from our group were shown to have both negative intrinsic and interfacial energies with a much smaller mismatch (Chapter 6).

After we had obtained preliminary force-spectroscopy data of PR65, I decided to extend my investigation to CTPRs to aid the understanding of PR65 unfolding. Since CTPR force responses prove to be less complex, I chose to discuss it before reporting the

PR65 data. To our knowledge, the work presented in this chapter is the first to describe the force-induced unfolding of TPR proteins. Using repeat arrays of varying size, I can examine their unfolding behaviour under force and also whether their supramolecular shape endows them with flexibility similar to that observed for the 24-repeat ankyrinB [196]. If such a supramolecular response is present, it may be amplified in longer arrays and hence I examine proteins with up to 20 repeats. Furthermore, thermodynamic stability and the folding kinetics of repeat proteins change with increasing number of repeats and therefore it would be interesting to compare the force response of CTPRs with varying repeat number [124, 135]. Finally, by using two different consensus sequences, I can investigate how alterations in geometry, and intrinsic and interfacial energies (3 and 6) translate into changes of the mechanical behaviour.

7.2 Methods

7.2.1 Construct generation

DNA constructs of CTPR_RV proteins were build sequentially from from single/double repeat modules using BamHI/BglII cloning [116]. First, ybbR-tags were introduced by RTH mutagenesis directly adjacent to the repeat sequence either N-terminally or C-terminally of a single repeat, giving rise to yCTPR_RV1 and CTPR_RV1y respectively. Second, the required number of repeats were added to yCTPR_RV1, resulting in yCTPR_RV4, yCTPR_RV9 and yCTPR_RV19. Last, the C-terminally tagged repeat was added to produce constructs with $N = 5, 10, 20$ that contained both N- and C-terminal ybbR-tags.

To facilitate ybbR-tagged construct generation, a pRSET vector was later modified to contain an N-terminal ybbR-tag between a TEV cleavage site and a BamHI restriction site, and a C-terminal ybbR-tag between a HindIII restriction site and the stop codon. The restriction sites give rise to additional amino acids between the individual ybbR-tags and the protein: GS at the N-terminus and KL at the C-terminus. Using this vector, CTPRa5 and CTPRa10 were assembled first before cutting and pasting them with BamHI and HindIII into the ybbR-tag vector. Recombination of CTPRa10 by *E. coli* resulted in a 9 instead of a 10-repeat construct.

The correct length of all constructs was verified by Sanger sequencing and restriction digests.

7.2.2 Sample preparation

Protein-DNA chimeras of the CTPR_RV series were produced in 50 µl volumes of 50 mM sodium phosphate pH 6.5, 150 mM NaCl, 50 mM $MgCl_2$ containing 10 µM protein, 20 µM home made CoA-oligo and 20 µM Sfp-synthase. Samples were incubated over-night at room temperature and flash-frozen. Before force-spectroscopy experiments, all samples were thawed and purified using an YMC Pack Diol-300 equilibrated in 50 mM Tris-HCl pH 7.5, 150 mM NaCl. 10 µl of fractions containing protein attached to two DNA-oligos were incubated with 100 to 200 ng functionalised DNA handles at room temperature for at least 30 min. Only 0.3 to 0.5 µl of that mixture were used to prepare samples for measurements.

Protein-DNA chimeras of the CTPR_RV series were produced in 50 µl volumes of 50 mM HEPES pH 7.5, 10 mM $MgCl_2$ containing 10 µM protein, 20 µM commercial CoA-oligo and 10 µM Sfp-synthase. Samples were incubated over-night at room temperature and purified immediately as described above. Only 4-6 µl of fractions containing protein attached to two DNA-oligos were incubated with 200 ng of functionalised DNA handles and 0.5-1 µl of the mixture were used to prepare samples for measurements.

7.2.3 Estimating unfolding forces of transitions

Due to the nature of their unfolding transition, it was not possible to extract the unfolding forces, which traditionally are the force at which a protein or a subdomain unfolds completely, i.e. the force peak. The force waves were extracted from Igor and analysed using Python. The data of each force curve were binned into a histogram, giving rise to clear peaks corresponding to the 0-force baseline and the unfolding plateau (Figure 7.1a). The positions of these peaks was extracted from the histogram using a sum of two Gaussian functions and a linear dependence of the background noise on force (force clamping):

$$P(F) = mF + c + a_1 e^{\frac{1}{2}\left(\frac{F-\mu_1}{\sigma_1}\right)^2} + a_2 e^{\frac{1}{2}\left(\frac{F-\mu_2}{\sigma_2}\right)^2}, \tag{7.1}$$

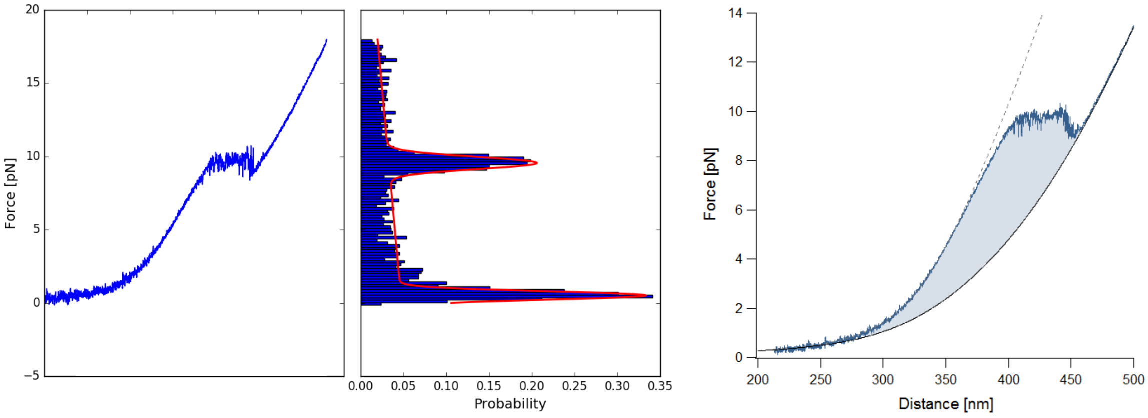
where $P(F)$ is the probability density of force values, m and c are the slope and intercept of the noise level, and a the scaling factor, μ the mean and σ the standard deviation of the gaussian.

7.2.4 Calculating free energies of unfolding

Force-extension curves taken at 10 nm s^{-1} were fitted with WLC models for both the DNA and fully extended protein. The non-equilibrium energies were then extracted from force-distance curves, which is simply the difference between the unfolding trace, $F(x)$ and the contour of the fully extended protein, $C(x)$:

$$\Delta G_{D-N} = \int_{x_1}^{x_2} F(x)\,\mathrm{d}x - \int_{x_1}^{x_2} C(x)\,\mathrm{d}x, \tag{7.2}$$

which corresponds to the area between those two curves (Figure 7.1b).



(a) (b)

Figure 7.1: Calculating the forces and energies of TPR unfolding transitions. (a) The mean unfolding force is extracted by fitting a Gaussian function (red) to a histogram of forces (right) which was derived from the raw data (left, plotted as force against its index array). (b) The non-equilibrium energies of unfolding are simply the area (shaded light blue) between the unfolding curve and the contour of the fully extended construct.

Assuming a simple model where each repeat has the same energy and the minimal cooperative unit, which has to be formed before any other repeats can fold using this seed, the energy of unfolding can be expressed as

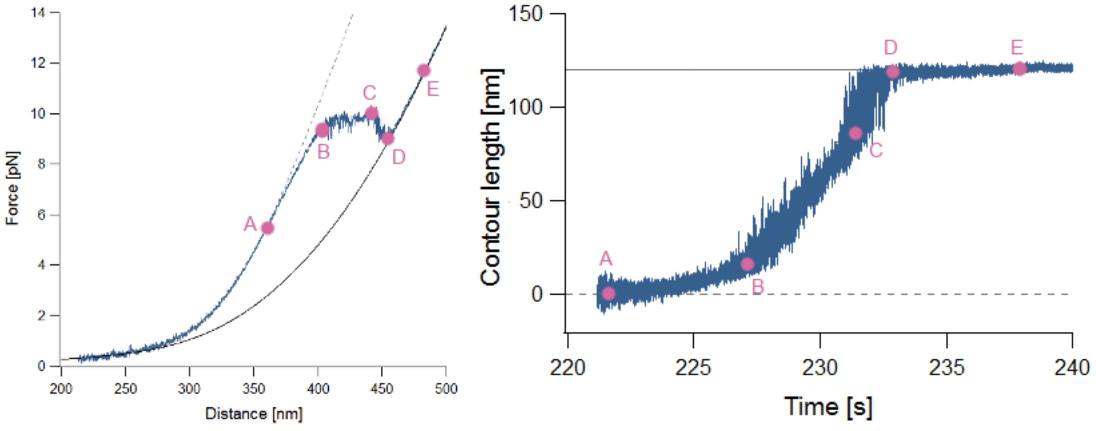
$$\Delta G_{D-N} = N\Delta G_r - \Delta G_s, \tag{7.3}$$

where N is the number of repeats, ΔG_r is the energy per repeat, and ΔG_s is the energetic cost of forming the minimal cooperative unit. For a collection of repeat proteins, Equation 7.3 can either be used to solve a linear system of equations when the energies of at least two proteins with different N are known, or it can be used to optimize ΔG_r and ΔG_s when the energies of three or more proteins with different N are available using

$$\begin{bmatrix} N_1 & -1 \\ N_2 & -1 \\ \vdots & \vdots \\ N_M & -1 \end{bmatrix} \begin{bmatrix} \Delta G_r \\ \Delta G_s \end{bmatrix} = \begin{bmatrix} \Delta G_{D-N,1} \\ \Delta G_{D-N,2} \\ \vdots \\ \Delta G_{D-N,M} \end{bmatrix} \tag{7.4}$$

7.2.5 Contour-length transformation

Contour-length transformations (CLTs) were performed as described by Puchner *et al.* [357] using an Igor procedure written by Markus Jahn (Rief group). In brief, due to variation between experiments (such as altered external conditions or random fluctuation) DNA and protein parameters can differ, making a comparison between two force-extension traces difficult. Therefore, data are first fitted with WLC models to extract the persistence and contour-lengths, which are then used to transform the data from force-extension space into contour-length space (Figure 7.2). That is, Equations 2.1 and 2.2, which describe the data as $F(L_{protein}, L_{DNA}, \xi)$ are solved for the contour-lengths, resulting in an expression of the form of $L_{protein}(F, L_{DNA}, \xi)$. By performing CLTs, dependencies on the individual experimental contexts are removed, allowing comparisons between multiple experiments of the same protein or of different proteins.



(a)

(b)

Figure 7.2: Contour-length transformation from force-distance space (a) to contour-length-time space (b). Positions of points A-E are approximate.

7.3 Results

7.3.1 Qualitative description of the CTPRa and CTPR_RV force behaviour

Force-extension data were collected at varying pulling speeds for all five proteins. Representative force curves for different TPR constructs, representing consecutive stretch and relax cycles of the same molecule, are shown in Figure 7.3. As extension increases, the DNA handles are stretched first before the protein experiences any force. Surprisingly, instead of unfolding one repeat at a time, which would give rise to consecutive force peaks, all TPRs somehow "melt" apart. This constant force plateau is nearly unrecognisable in constructs with five repeats (Figure 7.3), but is very striking in the longer proteins. Furthermore, it is very noisy which suggests that many fast unfolding and refolding events are occurring, possibly in multiple instances all over the repeat array. At the end of the plateau, a rupture is observed in all proteins, in which the folded remainder of the pro-

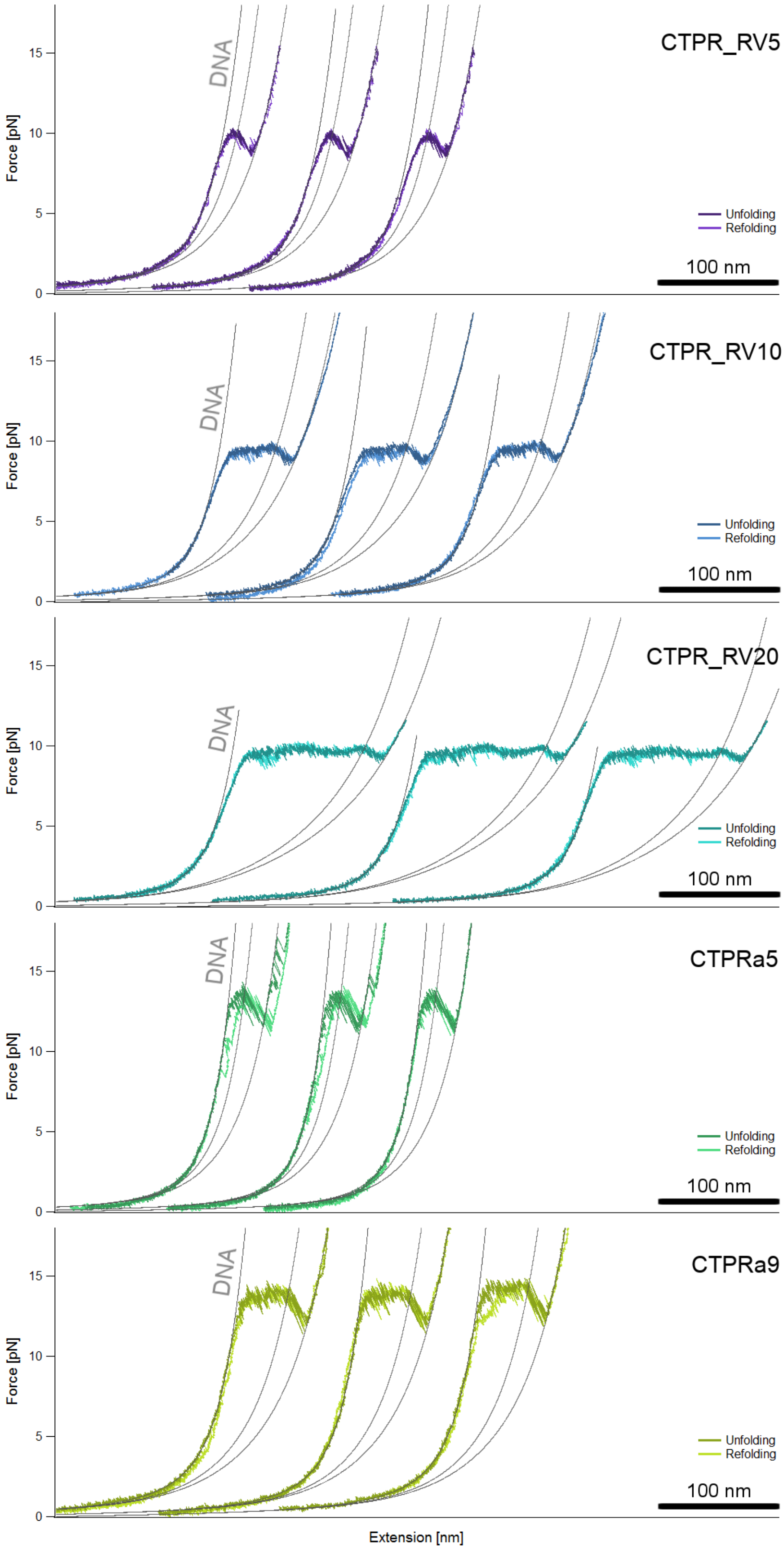


Figure 7.3: Representative force-extension curves for each TPR protein, where unfolding and refolding traces are coloured in darker and lighter shades, respectively. WLC fits to the DNA stretching as well as the rupture point and the full extension of the construct are shown in grey. All axes have been scaled equally to facilitate comparison.

tein unfolds in one large transition. This final rupture is not a single transition but also contains some noise, suggesting that parts of the protein can temporarily refold.

Figure 7.3 also shows the refolding curves which overlay nearly perfectly with unfolding curves of the same stretch-relax cycle, i.e. TPRs refold without hysteresis. In cases where the refolding and unfolding curves do not perfectly overlap, this difference can often be attributed either to drift (trace 2 of CTPR_RV10, Figure 7.3), or to DNA slippage (traces 1 and 2 of CTPRa5, Figure 7.3). The latter is known to occur in some attachments and/or DNA handle batches. These artefacts were seen to some extend in all CTPRa5.

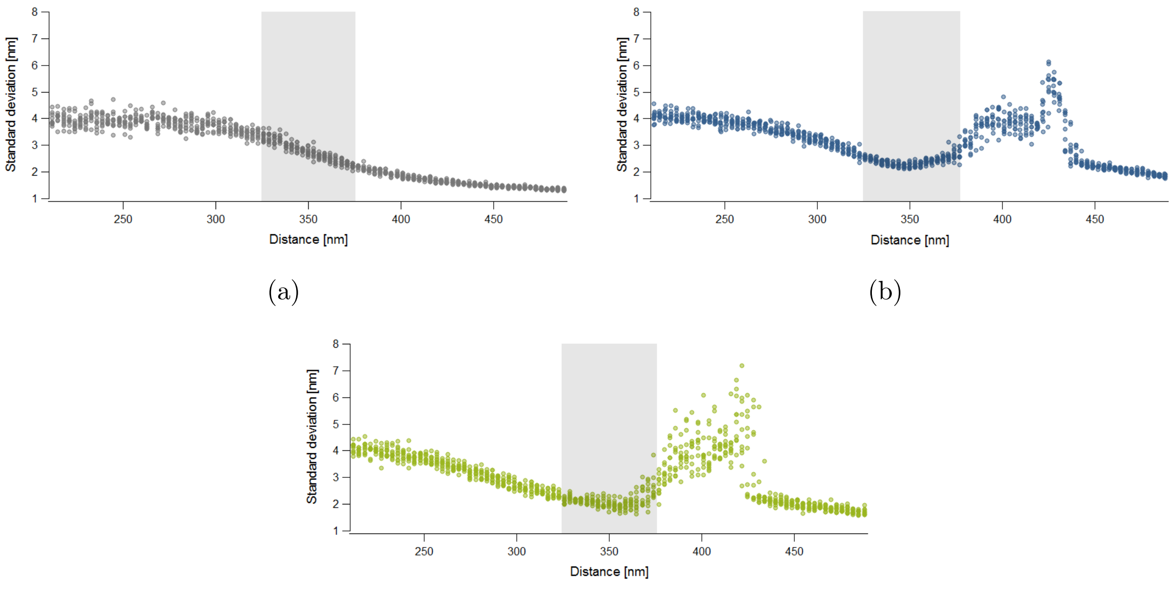
The unfolding and refolding behaviour of CTPRas and CTPR_RVs differs in three ways:

1. CTPRas unfold at higher forces than CTPR_RVs.
2. The rupture of CTPRas is larger in magnitude, i.e. the drop in the force upon unfolding is larger
3. In proteins of the RV series a clear deviation from the DNA WLC is observed prior to the unfolding plateau.

To examine the pre-transition further, the standard deviations of the bead distances (i.e. the raw data of bead fluctuations) were calculated for 10 traces of CTPRa9, CTPR_RV10 as well as a double-handle DNA construct without any protein (Figure 7.4). When a DNA molecule is stretched the standard deviation decreases as the increase in tension reduces Brownian motion of the beads (Figure 7.4a). In CTPR_RV10 curves this decrease is present only until about a distance of 320 to 330 nm, indicating that the protein starts to experience force at that point (Figure 7.4b). Instead of continued force clamping, the standard deviation gradually reaches a minimum and then increases until the much higher standard deviations of the force plateau. This means that the protein is being stretched although this is unlikely to involve the same molecular mechanism as in the plateau. In contrast, CTPRa constructs do not exhibit this first protein response at all (Figure 7.4c).

7.3.2 Contour-lengths and determination of the minimal cooperative unit

Although data were acquired at varying pulling speeds, traces taken at speeds of both 10 nm s^{-1} and 100 nm s^{-1} were only used to fit WLC models for DNA. To fit rupture point and fully extended protein (grey traces in Figure 7.8) only 10 nm s^{-1} traces were used as it was easier to distinguish the first, gradual force response of the protein as well as the rupturing point. The number of molecules measured as well as the total number of traces for each pulling speed are shown in Table 7.1.



(a)

(b)

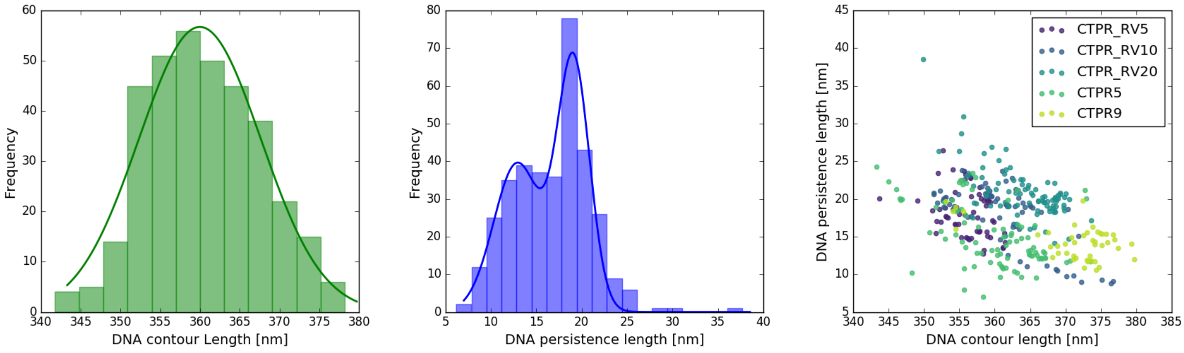
(c)

Figure 7.4: Standard deviation of the bead distance fluctuations for 10 extensions of (a) a DNA dimer, (b) CTPR_RV10 and (c) CTPRa9. The grey-shaded rectangle highlights region where the pre-unfolding force-response of CTPR_RV10 occurs. The differences of the DNA standard deviation between different constructs is likely due to experimental variation, which can influence parameters such as the DNA persistence and contour-lengths.

Table 7.1: Overview of the number of CTPR molecules examined and the data obtained at different pulling velocities. N_{10} and N_{100} refer to the number of traces taken at 10 nm s^{-1} and 100 nm s^{-1}, respectively.

Protein	**Molecules**	$\mathbf{N_{10}}$	$\mathbf{N_{100}}$
yTPR_RV5y	4	8	50
yTPR_RV10y	5	8	58
yTPR_RV20y	4	8	45
yCTPRa5y	11	15	68
yCTPRa9y	3	7	38

Due to the pattern observed in the standard deviation of force traces, DNA WLCs were fitted only up to 4 pN. On average across all CTPRa data, DNA contour-lengths were 360 ± 8 nm, while persistence lengths fluctuated between 13 ± 3 and 19 ± 2 nm depending on the sample set up (Figure 7.5a,b). A scatter plot of both parameters shows that these parameters varied for all proteins to a similar extent. The optimal value for the DNA stiffness, K, was explored manually in the high force regime where it dominates the DNA stretching response [164], and then set as constant for all cycles of the same



(a) (b) (c)

Figure 7.5: Summary of parameters obtained from DNA WLC fits to unfolding curves of CTPRs: (a) contour-length histogram (Gaussian fit: $\mu = 360$ nm, $\sigma = 8$ nm), (b) histogram of the DNA persistence lengths (Gaussian fit: $\mu_1 = 13$ nm, $\sigma_1 = 3$ nm, $\mu_2 = 19$ nm, $\sigma_2 = 2$ nm), and (c) DNA persistence length plotted against contour-length for each TPR protein.

molecule. Usual values were between $K = 400$ and $K = 800$.

Histograms of fully extended protein contour-lengths and the rupturing point are shown in Figure 7.6, and the results from the corresponding Gaussian fits are listed in Table 7.2. The expected length of each protein construct, L_{tot}, was calculated assuming 0.36 nm per amino acid and includes contour-length contributions of ybbR-tags at either end of the protein. Also given in Table 7.2 are the end-to-end distances for each protein, which were calculated using the respective crystal structures [126, 139]. Since TPRs form elongated superhelices, these distances are quite considerable and hence the expected contour-lengths have to be adjusted to reflect this initial state prior to unfolding (L^*_{tot}). If ybbR-tags did not associate with the protein but were instead expected to unfold with the DNA, their length would also have to be deducted (L^{**}_{tot}). Comparing both L^*_{tot} and L^{**}_{tot} to the measured contour-lengths of the fully extended proteins, L_p, reveals that in most cases L_p is smaller than expected. This is particularly clear for CTPRas, of which the expected contour-lengths with and without tag are outside one standard deviation of the measured contour-lengths. For CTPR_RVs a trend is less clear because DNA fits tended to vary more for these constructs, resulting in larger variation of the final protein contour.

In the full extension histogram of CTPR_RV10 some values cluster into a clear subsidiary peak at a contour much less than the expected length, at roughly 107 nm. This would correspond to a protein of 9 repeats ($L^*_{tot} = 108.2$ nm, $|\Delta\vec{r}| = 6.28$ nm). Since all TPRs are built up from genetically identical DNA sequences recombination products of smaller repeat number may be observed. Although the major species expressed is still the full length construct, recombination increases with repeat array size. For example,

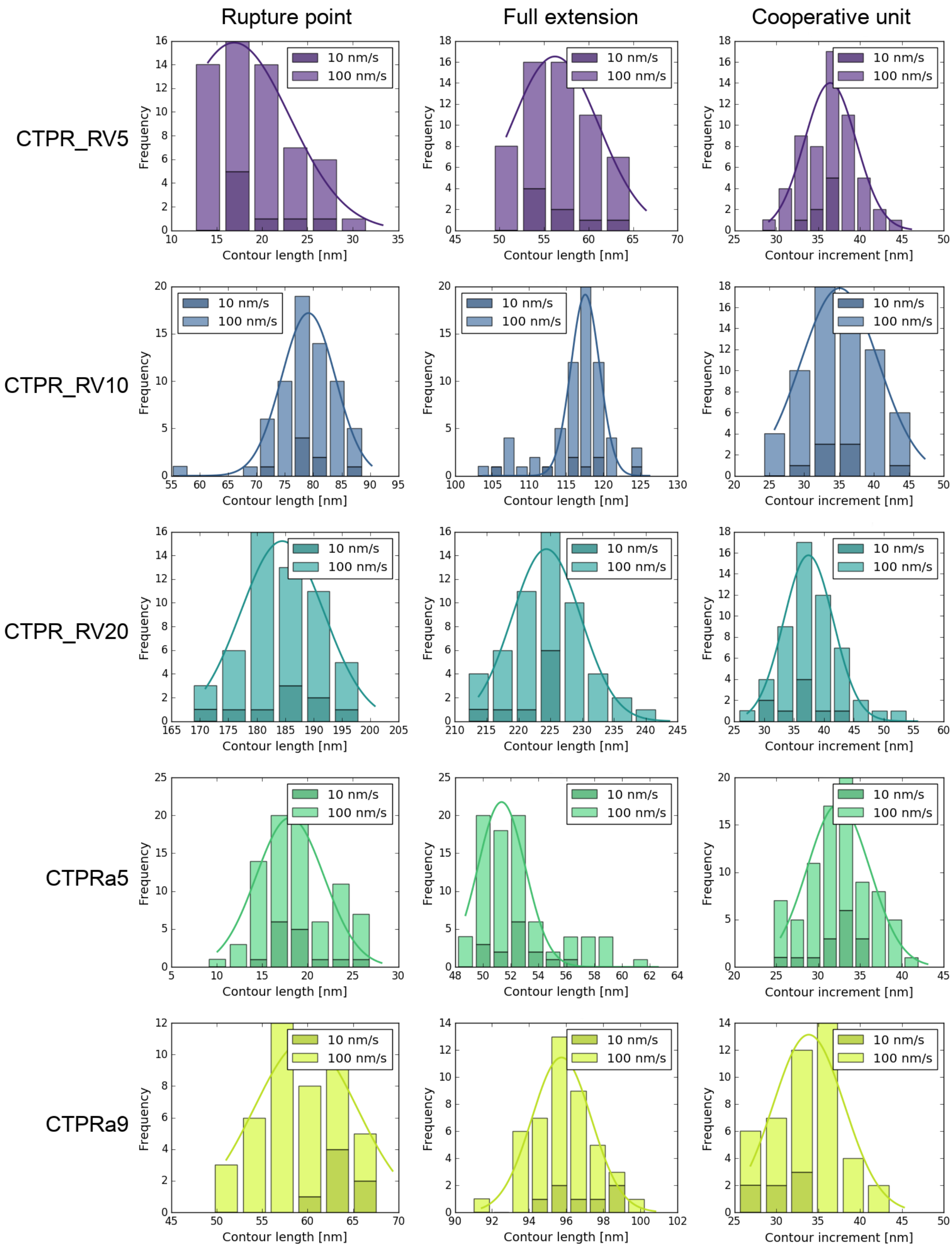


Figure 7.6: Contour-length histograms of WLC fits to unfolding transitions of TPRs. The left column shows the contours of the rupture point, the middle column shows the full extension data, and the last column are histograms of the differences of the contour-length change from the rupture point to full extension. Means and standard deviations of the Gaussian fits are listed in Table 7.2

a 5-repeat construct only combines minimally and smaller species are $< 5\%$, whereas recombination of a 20-repeat construct is so high that only approximately 50% of the protein extracted from *E. coli* is full length. Although mass-spectrometry of CTPR_RV10 showed that the major species had the correct number of repeats, a small amount of recombined protein must be present as this is obvious in a clearly separated peak in the histogram. Some CTPRa5 molecules were observed to unfold to slightly longer contours of ~ 57 nm, which would agree with L^{**}_{tot}, but those are in the minority and do not form a clear subsidiary peak.

After contours were fitted to the fully extended construct and the rupturing point, their difference was calculated for each unfolding event to determine the contour-length of the substructure which unfolds in the final transition (Figure 7.6). Due to the noisy unfolding plateau the placement of the rupture point contour is somewhat arbitrary which is reflected in the higher standard deviations of the corresponding Gaussians fits (Table 7.2). The increase in length from the rupturing point to the fully extended contour was very similar for all proteins, especially for those of the same repeat type. As a reference, 37 nm correspond to approximately 3 repeats while 32 nm are equivalent to 2.6 repeats. That is, in the final rupture secondary structure equivalent to 3 repeats unfold in one, more or less cooperative transition indicating that this is the minimal folding unit of TPRs under force.

7.3.3 Unfolding forces are independent of pulling speed

The force response of proteins like titin I27 is loading rate dependent [151], that is the higher the pulling speeds the higher the unfolding force. However, for some proteins

Table 7.2: Expected and measured contour-lengths of CTPRa proteins. End-to-end distances are measured between the C_α atoms of the first and last amino acids. The exact length of the ybbR-tags differ between CTPR_RV (12 amino acids) and CTPRa (16 amino acids) constructs due to molecular cloning. All values are in nm. Theoretical values were calculated assuming 0.36 nm per amino acid.

Protein	$\mathbf{L_{tot}}$	$\lvert\Delta\vec{r}\rvert$	$\mathbf{L_{tag}}$	$\mathbf{L^{*}_{tot}}$ [a]	$\mathbf{L^{**}_{tot}}$ [b]	$\mathbf{L_R}$	$\mathbf{L_P}$	$\mathbf{L_{CU}}$
yTPR_RV5y	65.52	4.19	4.32	61.33	57.01	17 ± 6	56 ± 5	36 ± 3
yTPR_RV10y	126.72	7.22	4.32	119.5	115.18	79 ± 5	118 ± 2	35 ± 6
yTPR_RV20y	249.12	14.59	4.32	234.53	230.21	184 ± 7	224 ± 5	37 ± 4
yCTPRa5y	66.96	4.69	5.76	62.27	56.51	18 ± 4	51 ± 2	32 ± 4
yCTPRa9y	115.92	7.65	5.76	108.27	102.51	60 ± 6	96 ± 2	34 ± 4

[a] $L^{*}_{tot} = L_{tot} - |\Delta\vec{r}|$

[b] $L^{**}_{tot} = L_{tot} - |\Delta\vec{r}| - L_{tag}$

the unfolding forces are loading rate-independent and the energy of folding equals the area enclosed by the force extension curve [358]. Such proteins are said to unfold at equilibrium. All TPRs were subjected to varying pulling velocities higher than 10 and 100 nm s^{-1}. For many of the five constructs, speeds up to at least 1 µm s^{-1}, if not 5 µm s^{-1}, were tested (Figure 7.7). For a very few molecules I also increased the speeds up to 50 µm s^{-1}, but due to the sampling rates used and the time resolution of the equipment (∼ 10 µs) traces obtained at those velocities were of very low data density and hence have been excluded here. In all cases, independent of the loading rate, forces of unfolding and the general unfolding behaviour remained the same (Figure 7.7). Most importantly, after a velocity screen into the µm s^{-1} range, subsequent traces taken at 10 or 100 nm s^{-1} were virtually identical to traces before the speed was increased. In some instances, hysteresis in the refolding traces could be observed at higher speeds, but the unfolding forces always overlapped indicating that the molecule was completely refolded before the next extension even if hysteresis was present. Furthermore, events which could indicate misfolding were never observed.

The exact unfolding forces (i.e. height of the plateau) were extracted from 10 and 100 nm s^{-1} traces and are identical within one standard deviation (Table 7.3). Unfolding

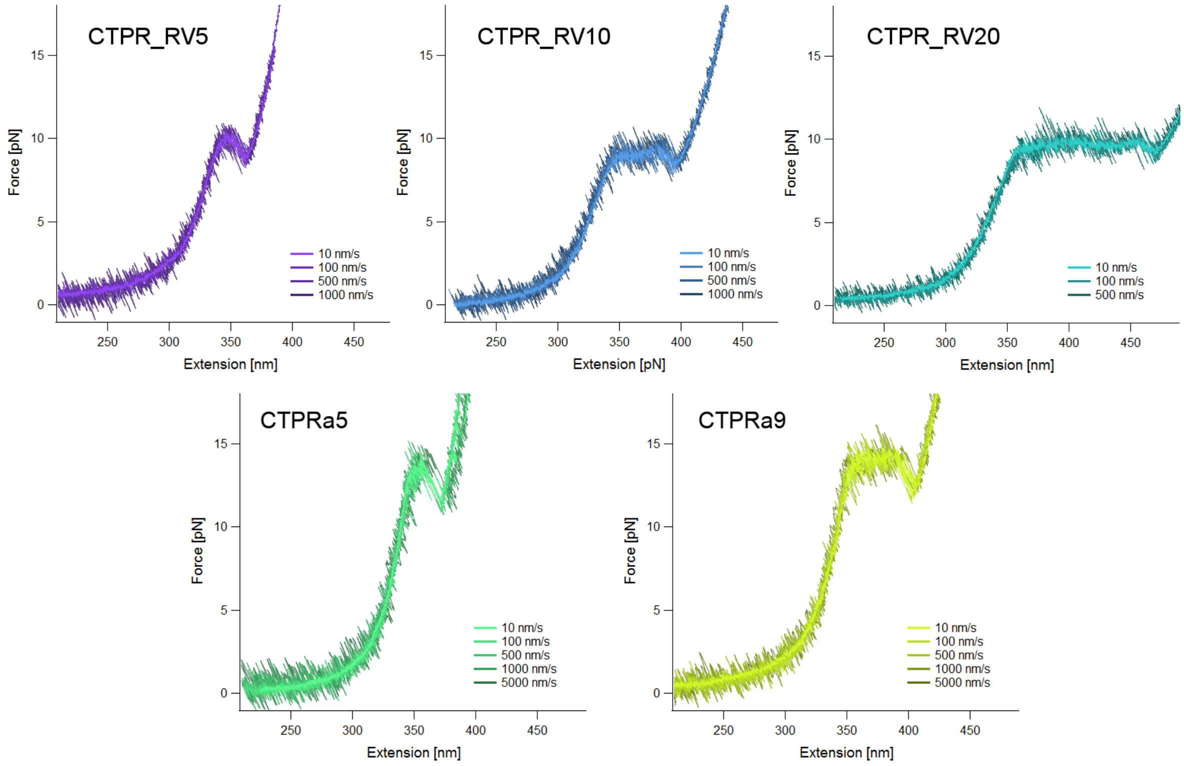


Figure 7.7: Force-extension curves of the same molecule taken at different pulling velocities. Overlaying the traces of each molecule shows that the unfolding behaviour is not affected by the loading rate.

Table 7.3: Unfolding forces and non-equilibrium energies of CTPR proteins. Listed are averages ± standard deviation.

Protein	**F_{10} [pN]**	**F_{100} [pN]**	**ΔG [k_BT]**	**ΔG [kcal mol^{-1}]**
yTPR_RV5y	9.19 ± 0.09	9.2 ± 0.1	45 ± 2	27 ± 1
yTPR_RV10y	9.4 ± 0.1	9.4 ± 0.2	96 ± 5	57 ± 3
yTPR_RV20y	9.5 ±0.1	9.2 ± 0.2	180 ± 4	107 ± 2
yCTPRa5y	12.7 ± 0.2	12.8 ± 0.3	66 ± 2	39 ± 1
yCTPRa9y	13.8 ± 0.2	13.7 ± 0.2	140 ± 3	83 ± 2

forces of different CTPR_RV proteins are largely the same, whereas those of CTPRa5 and CTPRa9 differ by ~ 1 pN. Histograms of the extracted forces from all constructs except CTPRa9 appear to be normally distributed. Considering that unfolding forces of the same construct can fluctuate between experiments due to variations in bead size, sample chamber prepartion and hence calibration, more independent data sets will be necessary to confirm whether the difference observed between CTPRa5 and CTPRa9 is significant.

7.3.4 Free energy of unfolding

The energy of unfolding was calculated for 10 nm s^{-1} extension traces of all proteins, and the average values and their standard deviations are listed in Table 7.3. It was previously observed for a system at equilibrium that the free energy of unfolding due to force is the same as that due to a chemical denaturant [168]. When compared with energies obtained from either two-state or Ising models of ensemble measurements, the values obtained from single-molecule measurements are significantly larger. Since the ybbR-tag was found to be stabilizing, but it is yet unclear to what extend, an Ising model derived free energy will most likely lie between the energy of a 5-repeat array and a 6-repeat array, which are 16.3 and 20.42 kcal mol^{-1}, respectively. Even if the stability of a yCTPR_RV5y is equivalent to a CTPR_RV6, the energies of unfolding obtained from force spectroscopy still differ by 7 kcal mol^{-1}. In comparison, the Ising model derived free energies of unfolding of a CTPRa5 and CTPRa6 are 21.67 and 26.82 kcal mol^{-1}, both of which are also much lower than those obtained from single molecule experiments of yCTPRay.

Using the energies of unfolding from differently sized constructs, one can test whether the free energy has a linear dependence on the number of repeats by taking into account that the final transition within each trace has an energetic barrier associated with it. In such a model, intrinsic repeat stability and interfacial contributions are not distinguished. One can solve a system of linear equations based on Equations 7.3 and 7.4 using the values for each protein given in Table 7.3. A fit of equation 7.4 to all three CTPR_RV data points gives $G_r = 5.3 \pm 0.2$ kcal mol^{-1} and $G_s = -2 \pm 3$ kcal mol^{-1}. Since energies of only two constructs are available for the CTPRa series, the equation system cannot be fitted but can only be solved manually, yielding $G_r = 10.9 \pm 0.6$ kcal mol^{-1} and $G_s = -15 \pm 3$ kcal mol^{-1}.

7.3.5 Comparing the unfolding behaviour of proteins of different repeat number

To qualitatively compare proteins to each other, representative traces were transformed into contour-length space to remove any variations between different experiments. Figure 7.8 contains one trace each in the same coordinate system. If contour-length is displayed as a function of time, the unfolding plateau transforms into a continuous slope which shows that the contour-length increase is linearly proportional to time until the rupturing point. These slopes appear to be independent of repeat type, although a thorough quantitative analysis has yet to be conducted. The kinetics of this transition on a molecular scale is so fast that it is impossible to resolve folding sub-states. In a contour-length histogram, only the DNA contour and the contour of the fully extended construct would exhibit clear peaks, while the plateau itself forms a region of constant noise. The pre-unfolding response of CTPR_RVs, which had some curvature in force-extension curves, is approximately linear in CLTs. Although this feature has also yet to be analysed quantitatively, it appears that the slope of this part of the force-response slightly decreases with increasing repeat size.

It is important to note that the sampling rates at which the data were acquired differ

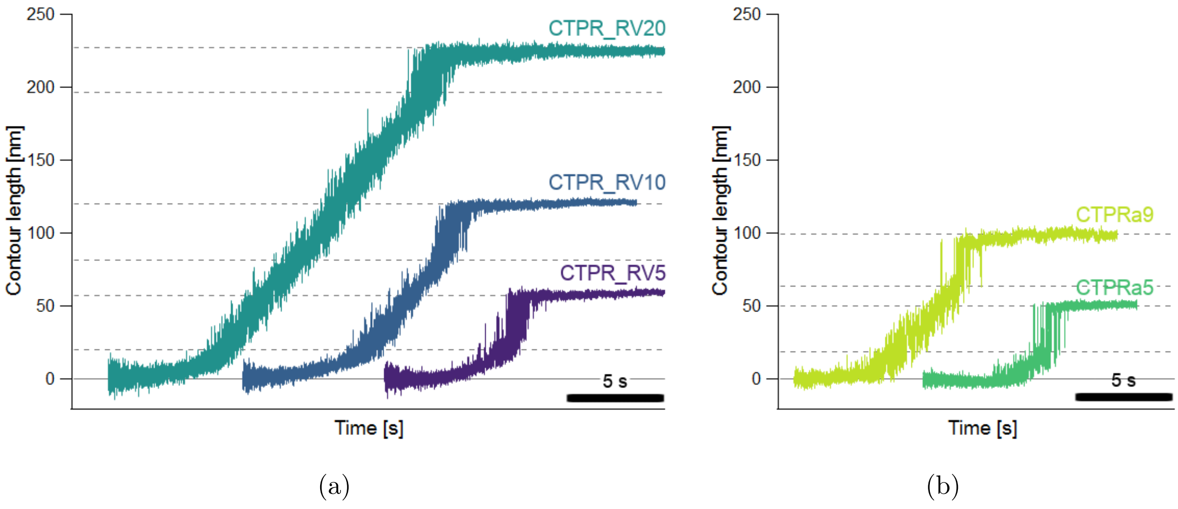


Figure 7.8: Contour-length transformation of CTPR_RVs (a) and CTPRas (b). The DNA contour is represented as a solid line, while dashed lines are the contours of rupture points and fully extended proteins.

for two out of the five molecules shown: the CTPRa9 and CTPR_RV20 data shown were collected at 10 and 30 kHz, respectively. All other data shown was acquired at 20 kHz. To make the noise levels roughly comparable, the smoothing factor was adjusted to compensate for the different sampling rates but the data density was left unaffected. A last clear difference between the repeat types, lies in the final rupture. The frequency of unfolding and refolding events in the CTPRas is much lower than that in the CTPR_RVs (Figure 7.8). Although only CTPR_RV5, CTPR_RV10 and CTPRa5 are directly comparable, the same trend is observed in both CTPR_RV20 and CTPRa9 transitions. Quantitative analyses of these transitions have yet to be conducted to confirm the observed differences and to determine for example rate constants of folding and unfolding in this transition. Lastly, when multiple traces of a single molecule are compared, no rupturing transition is exactly like the other highlighting that underlying molecular mechanism of this part of the transition must be extremely probabilistic.

7.4 Discussion

In contrast to previously reported force-extension data of repeat proteins, TPRs do not exhibit force peaks that correspond to the unfolding of individual or a set of multiple repeats. Instead their unfolding and refolding behaviour is characterised by a single transition at constant force before approximately the final three folded repeats open in a single rupture. Interestingly, this number corresponds to the folding correlation length of TPRs determined computationally [125]. Ferreiro *et al.* [125] furthermore observe multiple folding domains once the repeat number exceeds twice the correlation length. Although, unfolding or folding of more than three repeats is never observed, the high number of fluctuations in the rupture could be due to multiple sets of three repeats attempting to fold at once. Successful folding of a multiple 3-repeat domains is unlikely since the protein would have to overcome the energy barrier of forming more than one folding unit simultaneously while fully stretched.

Extension and refolding traces of all CTPR proteins overlap without hysteresis at all pulling velocities tested, indicating that they are at equilibrium. Minimal hysteresis was observed previously for β-catenin and clathrin (armadillo and HEAT repeats, respectively), although both of these proteins exhibit force-peaks for the (un)folding of each repeat [197]. The range of pulling velocities that TPRs were subjected to are similar to those used in previous AFM studies: 400 nm s^{-1} (consensus ankyrin, [194]), 10 to 300 nm s^{-1} (variety of proteins, [197]), 12 to 200 nm s^{-1} (ankyrin-B, [196]), or 50 nm s^{-1} (gankyrin [195]). This suggests that the difference in unfolding behaviour is not due to the loading rate, but instead is an intrinsic characteristic of CTPRs. However, it cannot be excluded that the difference in behaviour is due to the method itself. It would therefore be interesting to either measure CANKs by optical tweezers or TPRs by AFM to confirm that the force response is independent of the set-up used and only due to repeat type.

The overall force-response of the two TPR variants examined here is very similar and scales linearly with the repeat number in that the unfolding plateau is extended but the final rupturing transitions remains unaffected. However, the two variants differ in a number of ways.

1. The greater chemical stability of CTPRas when compared with CTPR_RVs translates into higher mehcanical stability. CTPRas unfold at 13-14 pN, while CTPR_RVs unfold at ~ 9.3 pN.

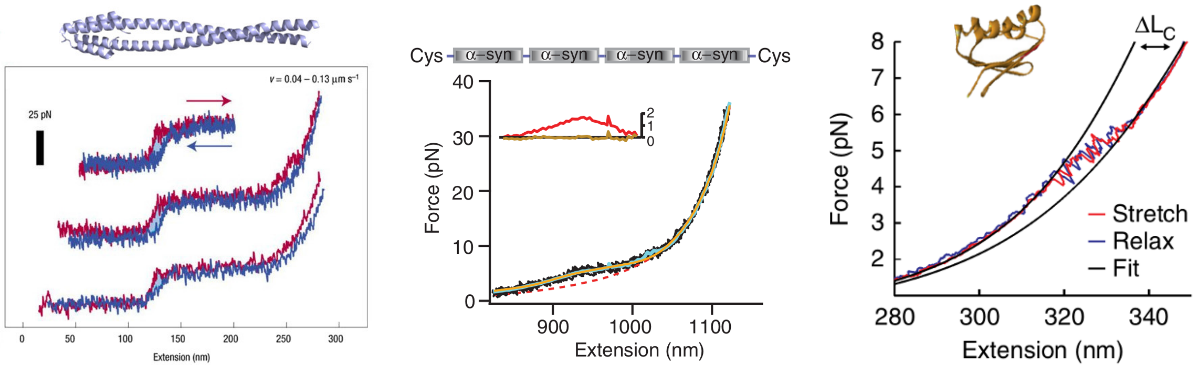
2. CTPR_RVs exhibit some form of pre-unfolding response, which could relate to stretching of the superhelix by rearrangements in the hydrophobic core as suggested by MD simulations of different repeat proteins [199–201]. Since the superhelix of CTPR_RVs is wider and more flexible than that of CTPRas (Chapter 3), it may simply have more freedom to rearrange its interface packing while still remaining overall folded. In contrast, the interfacial energies between individual helices as well as between repeats is slightly larger in CTPR_RVs than that in CTPRas (Chapter 6) possibly giving it more mechanical resilience. At this point, it is not clear whether these two factors (geometry and interfacial coupling) or indeed any others are the source of the pre-unfolding transition. Further investigation, e.g. of CTPRs with even higher interfacial energies, may shed light on this question.

3. The final rupture, although of similar contour-length increase, differs in magnitude (i.e. ΔF between rupture and full extension contours) between both sets of proteins. The free energy of this last transition is larger in CTPRas than in CTPR_RVs. However, it was not possible to get accurate values for this last transition using a simple linear sum of repeat energies and the energetic cost of forming the cooperative unit. In fact, when comparing force-extension traces of CTPR_RVs of different length the magnitude of the rupture is inversely proportional to array length. Such a trend is not obvious in proteins of the CTPRa series.

When WLCs were fitted for both DNA and protein, some measured contour-lengths at full extension varied significantly from those expected. The consistent discrepancies observed for CTPRas however are very large and hence are unlikely due to experimental variation. Furthermore, the standard deviations were quite large and therefore, it is not conclusive whether or not the ybbR-tags are unfolding with the repeat protein or with the DNA. At the moment it is impossible to determine why CTPRs are are shorter than expected and therefore, this issue needs to be subject of further investigation.

A proper comparison of free energies of unfolding between single molecule and ensemble derived data is currently not possible as I lack the appropriate Ising model description of ybbR-tagged proteins. Yet, the energies of tagged variants can be estimated using intrinsic and interfacial energies derived from homozipper models. Even if the ybbR-tags stabilized TPRs to an extent that is equivalent to adding one more repeat, the energies obtained from single molecule data are still significantly larger. This suggests that TPRs retain some form of structure when they are denatured using GdnHCl, whereas under force, the amino acid chain is fully extended such that all the energy is dissipated. Indeed, Cortajarena *et al.* [359] found that CTPRas did not adopt a random coil at high denaturant concentrations but instead were more compact due to extensive polyproline-II helical structure.

Although one cannot estimate the kinetics of any unfolding or refolding transition quantitatively using CLTs, some qualitative conclusions could be drawn. The force response is nearly identical in all TPRs and largely only differs in the length of the unfolding plateau. It may be possible that CTPR_RV proteins of different length differ in their pre-unfolding response but more rigorous analysis is required to prove this hypothesis. Extension in the unfolding plateau is gradual in all TPRs and due to the very fast kinetics, intermediates in the plateau itself cannot be resolved. Contrastingly, in the final rupturing transition the minimal cooperative unit is observed to explore the unfolded and folded states in very quick succession and the frequency of these fluctuation differs between repeat types. Such fast fluctuations were also observed for β-catenin and clathrin and were attributed to the fast unfolding of whole repeats or α-helices [197]. The rupture also resemble the fast helix-coil transitions detected in the unfolding and refolding transitions of the small α-helical protein calmodulin [360]. The difference in kinetics can be due to either different overall stabilities, or due to the mismatch between intrinsic and interfacial energies. CTPRa is more stable than CTPR_RV, and the mismatches are 3.01 and 4.48 kcal mol^{-1} for CTPRas and CTPR_RVs, respectively. In previous work from our group [139, unpublished], we found that CTPRa proteins appear to have slower refolding and slower unfolding rates than proteins of the CTPRn series. CTPRn repeats are more stable than CTPRas and have a mismatch between intrinsic and interfacial energy of 5.14 kcal mol^{-1}. When comparing all three proteins, the difference in the observed kinetics is unlikely to be due to changes in overall stability but more likely to correlate with the mismatch in intrinsic and interfacial energies. Furthermore, for both CTPRa and CTPRn constructs an increase in repeat number correlates with an increase and decrease of refolding and unfolding rates, respectively [124, 139]. This has been proposed to be due to either an increase in stability with increasing size [124, 132], or (b) an increasing number of parallel folding pathways [135]. Although such a dependence of the kinetics on N cannot not be determined qualitatively (or quantitatively) from CLTs, constant distance measurements at increasing extensions around the rupturing point will provide more detail.



(a) (b) (c)

Figure 7.9: Examples of other proteins unfolding at quasi-equilibrium. (a) AFM extension curves of the *Dictyostelium discoideum* myosin coiled-coil tail domain [155] with a structural presentation of the human myosin coiled-coil (residues 1526 to 1571, PDB id 5cj1) [361]. (b) Force-extension curve of the intrinsically disordered protein α-synuclein in form of a tandem-tetramer, the schematic representation of which is above [169]. (c) Forc-extension curve of a ferrodoxin-like peptide, which forms some α-helical substructures [362].

As TPRs unfolding was very distinctive compared with that observed for other repeat proteins, I tried to find examples of non-repeat proteins that unfold in a manner that cannot be described by WLC fits. Three examples are given in Figure 7.9: the myosin tail (coiled-coil), α-synuclein (intrinsically disordered) and a ferrodoxin-like α-helical peptide. Interestingly, the AFM force-extension curves of the myosin tail strongly resemble the TPR unfolding plateau and only lacks the final rupture. These data were well-described by a model that assumed the probabilistic unfolding of individual coiled-coil segments. Both the IDP and the small helical peptide exhibit gradual transition from the DNA to the fully extended contour and in particular the unfolding transition of the ferrodoxin-like peptide is significantly noisier than the trace either side of the transition, indicating very fast unfolding and refolding events that are beyond the resolution of optical tweezers [169, 184]. The unfolding of both systems could be described using a quasi-equilibrium model assuming a sequential, two-state folding behaviour. This equilibrium model can describe TPR force-distance data, but only until the rupturing point. Of course, since TPR unfolding cannot be described by a two-state model, this fitting procedure will have to be adapted to include the Ising model formalism.

Finally, our group previously showed using hydrogen-deuterium exchange, that the end repeats of TPRs are less protected than the internal repeats [139, unpublished]. Therefore, it is quite possible that TPRs start unfolding at both ends of the array when subjected to force, thereby increasing the number of possible unfolding and refolding events. Furthermore, the A-helices of each repeat make contacts to both helices of the preceding repeat [19], which creates the possibility that individual helices unfold, not whole repeats, as the second helix of each repeat would still be stabilized by a neighbouring repeat.

Taken together, all these data can help build a possible folding mechanism for TPRs. At the beginning of the unfolding plateau individual α-helices at either end of the repeat array start fluctuating between native and unfolded states. As the force is increased, terminal helices are less likely to refold and hence expose the adjacent helices. These then start exploring the unfolded state as they now lack the protection of the terminal helices. In this manner the whole array is gradually unravelled until about three repeats remain which unfold in an apparent two-state transition. Once the protein is fully unfolded, refolding is still likely to occur anywhere within the protein - possibly even in multiple places at the same time. With increasing extension, these refolding events become less and less likely, as a stable folded unit can be formed no longer. Such mechanism could explain why consensus TPRs do not unfold repeat by repeat: the individual building blocks, both repeats and single helices, are stable enough that they can adopt their native structure even if only transiently. That is a single TPR repeat does not only have two states, folded and unfolded, but can also adopt intermediate conformations. However, the clear division of the force response into plateau and rupture indicates that the folding mechanisms in these two transitions differ at the molecular level. It has been suggested, and was shown using consensus ankyrin repeats, that for a repeat protein in solution the energy barrier of folding lies within the formation of two repeats without the formation of an interface [63, 135]. Such a barrier is clearly different under force for both ankyrins [194], which fold a single repeat at a time, and TPRs, which have to overcome an energy barrier that involves the folding of three repeats but only at the beginning. Once the minimal cooperative TPR unit has been formed, the folding mechanism is clearly different. Folding events in the plateau could either arise from folding of repeats and/or helices followed by interface formation with the existing folded repeats, or unfolded parts of the protein could use the folded repeats as a template and form the interface upon folding. The current data cannot distinguish between either mechanism, but we hope that future investigations can shed light onto the matter.

Chapter 8

Mechanical characterization of PR65 in the low-force regime

8.1 Introduction

The simplicity of consensus repeat proteins often facilitates the elucidation of their folding and unfolding pathways. In natural repeat proteins, however, sequence variation between repeats can introduce a large amount of complexity. PR65, consisting of 15 HEAT repeats, has been the subject of multiple studies. Our group has studied its unfolding pathway [70, 71] while other groups, such as Grinthal *et al.* [201] and our collaborator Giovanni Settanni (Johannes Gutenberg Universität, Mainz, Germany) examined its force response in MD simulations.

Using chemical equilibrium denaturations and kinetic studies of single point mutations and truncations, a previous member of our group, Maksym Tsytlonok, was able to delineate the unfolding pathway of PR65 in solution (Figure 8.1). With increasing denaturant concentrations PR65 first unfolds to a hyperfluorescent intermediate before unfolding fully at higher denaturant concentrations [71]. In the transition between native and the intermediate states, two possible unfolding pathways exist. Either HEAT14-15 unfold first followed by the unfolding of HEAT3-10, or vice versa. In fact, the unfolding of repeats 3-10 can be further decomposed into two separate event, the unfolding of HEAT8-10 and HEAT3-7, respectively. However, the latter unfolding event is so fast, that it was difficult to resolve experimentally. In the intermediate state, HEAT1-2 and HEAT11-13 are folded. At higher denaturant concentrations these two domains then unfold sequentially, first HEAT1-2 followed by HEAT11-13. Examination of a range of truncations of PR65 furthermore showed these five repeats have a stabilizing function on the central domains (HEAT3-10) and prevent them from aggregating [70]. Repeats 14-15 may not stabilize repeats 11-13 but shield the interface of HEAT13 from oligomerization [70].

In MD simulations, Grinthal *et al.* [201] applied constant pulling and pushing forces

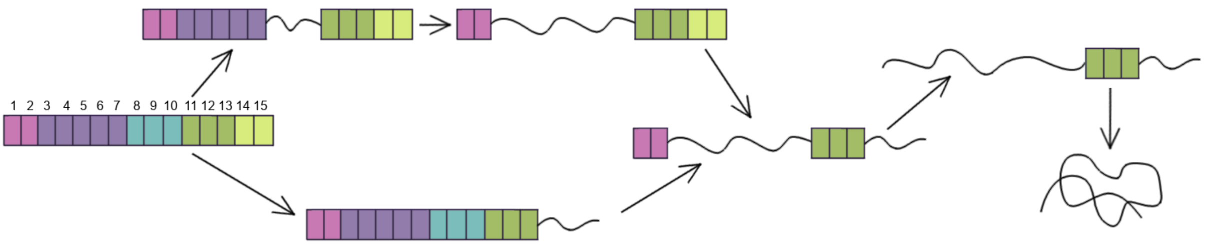


Figure 8.1: The unfolding pathway of PR65 in equilibrium denaturations as described by Tsytlonok *et al.* [71].

to the ends of PR65 without causing any loss of secondary structure. Instead the applied forces were distributed across the array and only distorted the horseshoe-like shape until fracturing at distinct interfaces could be observed in both pushing and pulling experiments. Under pulling forces, this fracture was located between HEAT6 and HEAT7, whereas under compression the interface between HEAT9 and HEAT10 fractured first. If a proline in the inter-repeat loop between HEAT6 and 7 was substituted for a glycine, this interface did not fracture but instead new weak points appeared between HEAT4-5 and HEAT8-9, which fractured at even higher forces than the HEAT6-7 interface in WT protein. These data are consistent with our ensemble studies, as they highlight the low stability of the central repeat interfaces.

Settanni and co-workers performed steered MD simulations to study the force-induced unfolding of PR65 *in silico* (unpublished data). They observed the sequential unfolding of one repeat or one helix at a time. They noticed that the order of repeat unfolding was highly variable. Among 50 simulations, only HEAT15 showed a consistent force response: it almost always unfolded first. The other repeats appear to unfold in any possible order, although small trends could be observed. Unfolding of repeat 15 was often followed by the sequential unfolding of HEAT14 and HEAT13. Sometimes, HEAT10 unfolded early and was followed by the unfolding of HEAT9 and HEAT8. Finally, HEAT11 and HEAT5 often unfold last and second to last, respectively. These results highlight the multitude of pathways accessible to PR65, many of which are likely too sparsely populated to be detected in ensemble experiments.

These two MD simulation studies probe two distinct regimes of the PR65 force response; a pre-unfolding, tertiary conformational change as well as the unfolding of individual repeats. To confirm the computational data by actual experiment, M. Tsytlonok conducted force spectroscopy experiments of PR65 using AFM in the laboratory of Piotr Marszalek (Duke University, USA). He could show that PR65 unfolded and refolded in a step-wise manner, one helix or one repeat at a time (Figure 8.2) [312, unpublished]. Only rarely, the unfolding of more than one repeat at a time was observed. Some hysteresis between extension and retraction curves was present as unfolding forces averaged to around

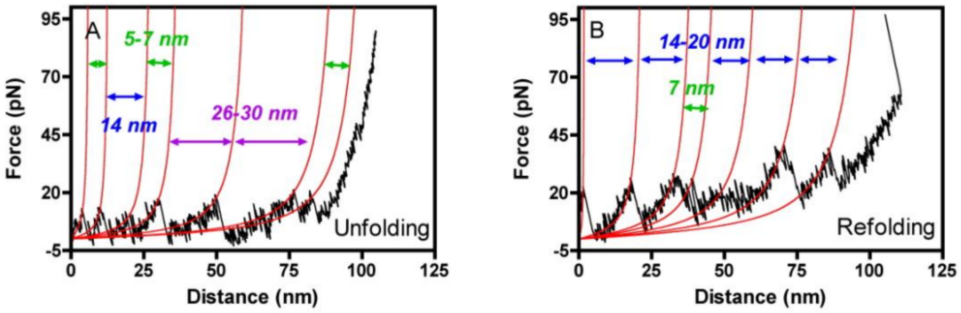


Figure 8.2: Examples of unfolding and refolding force-distance curves of PR65 generated by AFM. WLCs (red) have been fitted to identifiable force peaks which correspond to the (un)folding of helices (green contour length increments), individual repeats (blue contour length increments) and 1.5 to 2 repeats (purple contour length increments). Adapted from [312].

24 pN and refolding peaks occurred at approximately 18 pN.

The original aim of this part of my PhD project was to continue the mechanical characterisation of PR65 and to create mutants and stapled variants with altered mechanical stabilities. However, due to the low signal-to-noise ratios of AFM data, I chose to create constructs that could be examined using optical tweezers instead (Chapter 5). In contrast to AFM, optical tweezers provide much higher resolution, especially in the low force regime. Moreover, the use of a dual-beam tweezer setup abolished the need for surface immobilization and thereby ensures that the force is truly applied where intended.

8.2 Methods

8.2.1 Genetic constructs

The generation of PR65 wild-type attachments involving ybbR-tags or bio-orthogonal chemisties is described in Chapter 5, Section 5.2.1. A variety of N-terminal H_6 tagged point mutants was available from the previous unfolding studies of PR65 [71]. Genes containing the V25A/V65A, V201A, V284A and S445P mutations were amplified by PCR and introduced into pRSET_GST_ybbR using the BamHI and HindIII restriction sites.

8.2.2 Sample preparation

Wild-type and point mutants tagged with ybbR sequences were conjugated to CoA-modified DNA oligomers in 50 µl Sfp-synthase reaction buffer using 10 µM protein, 10 µM Sfp-synthase and 20 µM CoA oligo. Conjugations involving orthogonal chemistries were performed in 50 µl PBS containing 10 µM protein and 20 µM of the corresponding modified DNA oligo. If sufficient functionalised oligo was available, these were scaled up to 20 μM protein and 40 μM oligo. yPR65-277az was conjugated to oligos in the same reaction conditions as ybbR-tagged proteins, using 10 μM of CoA-oligo and 10 μM of DBCO-oligo.

All conjugations were purified by size exclusion using either a YMC Pack Diol-300 or a Superdex S200 10-300 equilibrated in 50 mM Tris-HCl, 150 mM NaCl, 1mM DTT. In the case of ybbR-tagged constructs, 4-6 μl of the YMC fractions containing protein attached to two DNA oligos were incubated with 200 ng of functionalised DNA handles, whereas 5 μl (azPR65az) or 10 μl (cycPR65cyc, yPR65-277az) of S200 fractions corresponding to the correct attachment species were incubated with 100 ng of functionalised DNA handles. Optimizing the protein:handle:beads ratio of all PR65 constructs was more difficult than for CTPR proteins, and the number of successful attachments deteriorated within 2 hrs of measurements, indicating that PR65 causes the beads and/or handles to stick to each other and the chamber surfaces.

8.2.3 Data Analysis

The force-extension trace of PR65-WT and point mutants was sub-divided into 4 phases, three of which were further subdivided into two transitions each, one apparently linear and one non-equilibrium rupture. The force-extension trace of yPR65-277az was subdivided into 3 transitions, one of which was apparently linear. WLCs were fitted to force peaks and boundaries between transitions. Forces at these points were recorded manually and force midpoints were calculated for linear transitions.

8.3 Results

8.3.1 Force response of wild-type PR65

Qualitative description of the PR65 force response

Force-extension data were collected at pulling speeds of 10 and 100 nm s^{-1} for the four different attachments yPR65-GSy, yPR65y, azPR65az and cycPR65cyc. Representative stretch-relax cycles of single molecules are shown in Figure 8.3. In contrast to previous AFM data from our group, I did not observe PR65 unfolding in one repeat or helix at a time. Instead, PR65 unfolded in a series of apparent linear and non-equilibrium transitions which indicate that it unfolds in distinct sub-domains of different mechanical stability. These transitions are defined as follows (Figure 8.3, 8.4):

1. **Transition A:** After the DNA is stretched the protein begins to extend (A1) until some repeats exhibit fast kinetics between the folded and unfolded state (A2). Refolding events from this intermediate to the fully folded state never reach the DNA

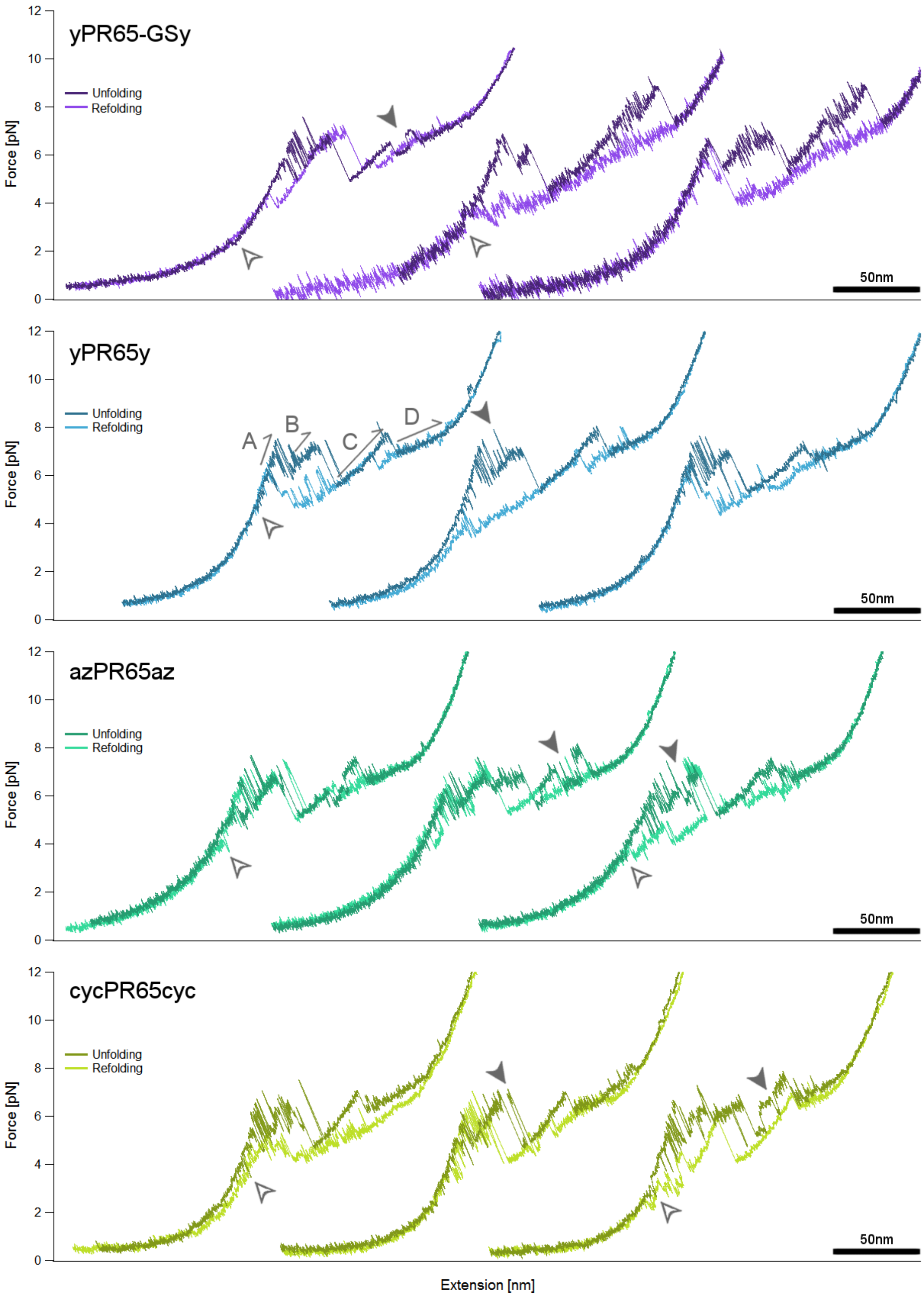


Figure 8.3: Representative full-length force extension curves of the same molecule for each PR65 attachments. The arrows in Trace 1 of yPR65y indicate are to highlight the transitions defined in Figure 8.4. Unfolding traces are coloured in darker shades, while refolding traces are coloured in lighter shades. All but the second and third trace of yPR65-GSy were taken at pulling speeds of 10 nm s^{-1}. Only one full-length 10 nm s^{-1} trace was taken, and therefore two 100 nm s^{-1} traces were included.

contour. Sometimes, partial and transient unfolding events can be observed before this major transition, at forces as low as 2 pN (empty arrowheads in Figure 8.3).

2. **Transition B:** Once the fully folded state becomes inaccessible, refolding attempts are less likely. The rest of the folded molecule continues to be stretched in a noisy, almost linear transition (B1) which is followed by a rupture that signifies the unfolding of several repeats at once (B2). In some cases, transient refolding events to the linear B1 transition can be observed after the rupture (e.g. the first trace of each yPR65y and cycPR65cyc).

3. **Transition C:** The folded remainder of the protein is extended in yet another, but much longer, linear transition (C1) which also ends in a clear rupture (C2).

4. **Transition D:** The final folded repeats of the protein are extended gradually until they are fully unfolded.

It is important to note that none of the linear transitions can be adequately fitted using a WLC model. Furthermore, unfolding and refolding between adjacent states occurs over a range of extensions and forces. That is, although the order of Transitions A-D is always the same, the exact start and end points of each transition differ between traces and hence no unfolding curves is like any other (Figure 8.3). The kinetics between the A- and B-transitions are always very fast but are rarely identical between traces of the same molecule. In contrast, B-C and C-D transitions can exhibit some transient fluctuations but can also persist while the molecule is extended further (filled arrowheads in Figure 8.3). In a small subset of curves, another transition could be observed that occurs after the B-transition (Figure 8.5, traces 2 and 3 of azPR65az and cycPR65cyc in Figure 8.3, respectively). Sometimes, it also appeared that the B-transition ended in this fifth substate (e.g. second trace of yPR65y in Figure 8.3), although it was difficult to define the boundary to the B-transition in those cases.

Figure 8.6a shows stretch-relax cycles of different molecules obtained at two different pulling velocities. The variability between transitions of different molecules is of similar magnitude to that observed between consecutive cycles of the same molecule. Furthermore, the variation increases at pulling velocities of 100 nm s^{-1}, unfolding force peaks increase marginally, and the C-transition can split into two linear parts with fast kinetics at the overlap (e.g. traces 3, 5, 6 in Figure 8.6). Figure 8.6b shows stretch-relax cycles of the same molecule but only to extensions of approximately 50 nm. The A and B1 transitions are captured in this partial unfolding, while the B2 transition occurs less frequently. Even when the protein is only partially extended, different unfolding patterns can be observed.

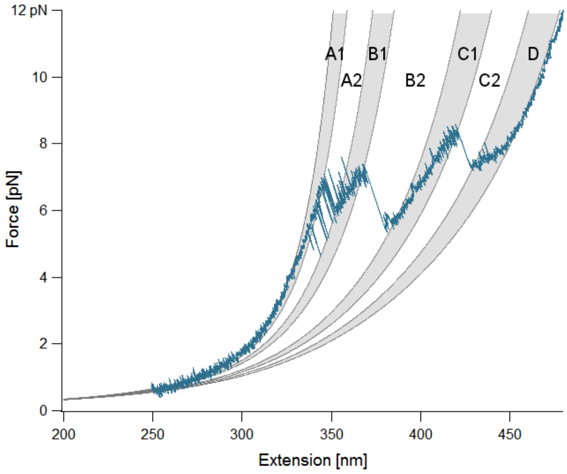


Figure 8.4: Defining the unfolding pattern of PR65. The three force peaks constitute the unfolding events A-C, while the final gradual extension is defined as a fourth transition, D. Since events A-C are clearly separated into an almost linear extension of folded repeats which ends in a distinct rupture, they have been subdivided into two transitions each.

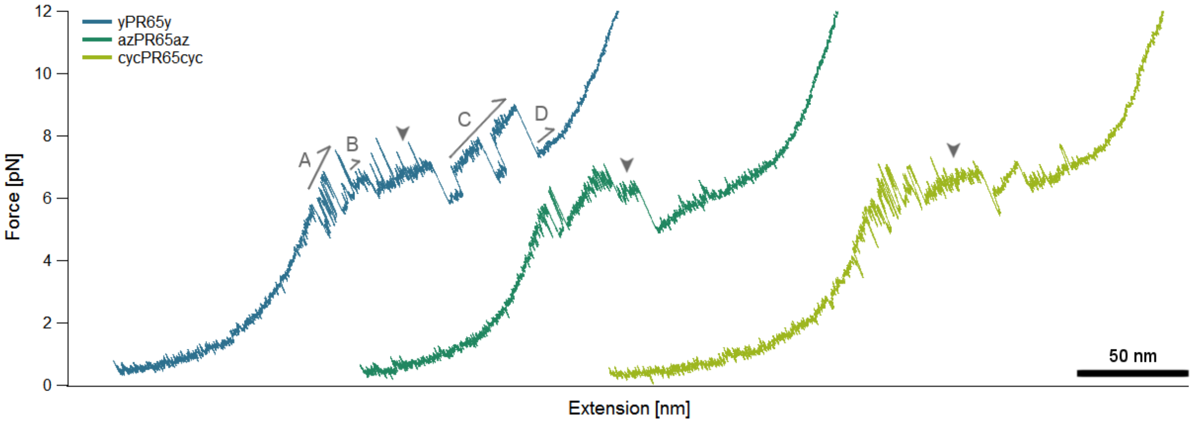
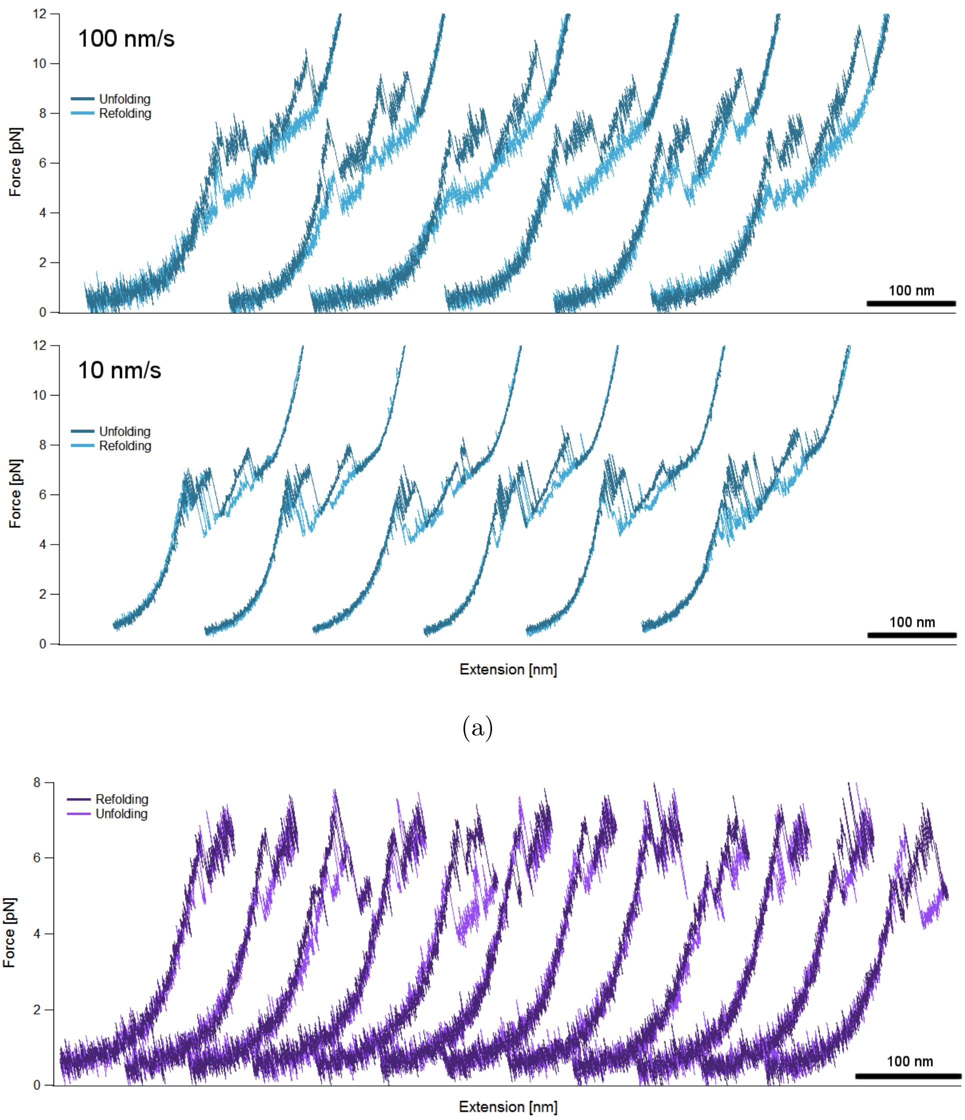


Figure 8.5: Example unfolding traces of three molecules that show the presence of a rare fifth stretching event between the B and C transitions (arrowheads).

The variation in unfolding is reflected in a similar variation of the refolding behaviour of PR65. The refolding trace rarely overlaps completely with the unfolding trace leading to a variable amount of hysteresis. In most cases, PR65 refolds through an extended D-transition even if this transition was diminished or absent in the unfolding trace (e.g. 100 nm s^{-1} traces of yPR65-GSy in Figure 8.3 or yPR65y in Figure 8.6). The point at which the molecule refolds into the C1 transition often occurs at lower forces than the peak of the C-transition in the unfolding curve. The refolding curve then retraces the C1 transition and usually continues along it without ever refolding into the conformation of the B1 transition (e.g. yPR65y traces and third azPR65az trace in Figure 8.3). Sometimes



(a)

(b)

Figure 8.6: Traces highlighting the variation in unfolding and refolding patterns of PR65. (a) Representative traces of 6 different molecules taken at 100 nm s^{-1} (top) and 10 nm s^{-1} (bottom). Variability of the PR65 force response between different molecules is of similar magnitude as observed within a single molecule. Increasing pulling velocities results in even larger variation as well as higher force peaks especially in the C-transition. (b) Representative traces from over 60 stretch-relax cycles of the same molecule which was only extended by approximately 50 nm. The pulling velocity here was 100 nm s^{-1}.

the rest of the molecule refolds in a single step such that unfolding and refolding traces of the A-transitions overlap (e.g. third yPR65y and second cycPR65cyc trace in Figure 8.3). At other times, the final refolding event occurs at much lower forces (e.g. second traces of yPR65-GSy and yPR65y, and third trace of azPR65a in Figure 8.3) and can even be absent (e.g. first trace of cycPR65cyc in Figure 8.3). Higher pulling velocities appear to leave the refolding forces unaffected, resulting in larger hysteresis, but affect the likelihood of the molecule folding through the individual transitions and mostly refold gradually from full extension to a point at which the molecule refolds to the native state in one sudden transition (Figure 8.6).

When PR65 was not pulled to full extension, it was possible to obtain more than 50 stretch-relax cycles of the same molecule. However, it was not possible to pull a given molecule to full extension for more than 10 cycles before the force response started to deviate strongly from that described above. These differences were, in true PR65-fashion, also highly variable. Often, any unfolding peaks observed in these cycles were at much higher forces, i.e. >10 pN. Taken together, these observations suggest that parts of PR65 misfold into conformations that are more stable than the native structure. It was attempted to unfold these misfolded molecules by increasing the pulling velocity and the maximal extension. However, this was not successful with the wild-type protein to date.

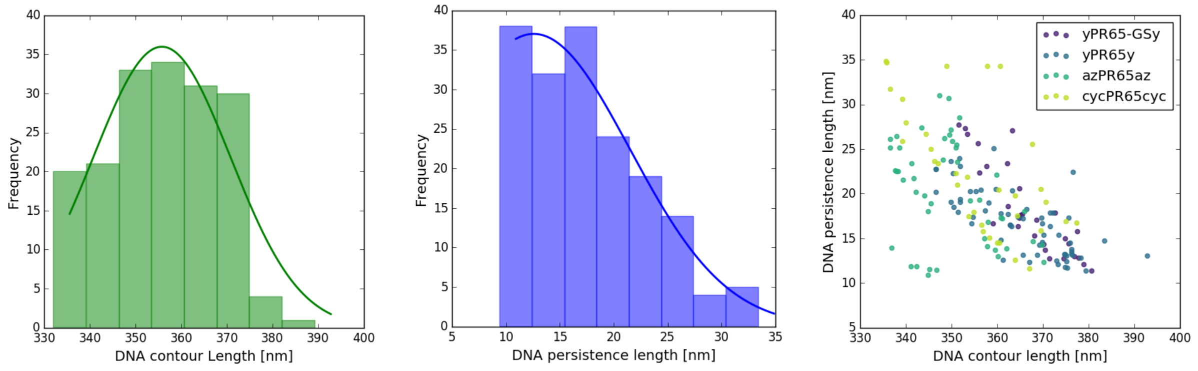
Characterisation of the unfolding intermediates

Extension traces obtained at pulling velocities of both 10 and 100 nm s^{-1} were used to measure the contour length of the fully unfolded protein. However, due to the increased variation observed at 100 nm s^{-1}, fitting of WLC contours to each transition point was limited to 10 nm s^{-1} traces in which all four transitions were clearly identifiable. The few traces containing the fifth transition were only used to fit the fully unfolded contour. The number of molecules as well as the total number of cycles obtained at each speed are listed in Table 8.1.

DNA WLCs were fitted to extension curves only up to 4 pN or up to where the first transient unfolding event could be observed and the results are displayed in Figure 8.7. In PR65 constructs, DNA handles were measured to have a contour length of 356 ± 13 nm and a persistence length of 13 ± 9 nm. Molecules with persistence length less than 10 nm were excluded from the analysis. The optimal value for the DNA stiffness was usually

Table 8.1: Overview of the number of PR65 molecules examined and the data obtained at different pulling velocities. N_{10} and N_{100} refer to the number of traces taken at 10 nm s^{-1} and 100 nm s^{-1}, respectively, while N^*_{10} refers to the number of traces which were used for the analysis of the individual transitions.

Protein	**Molecules**	$\mathbf{N_{10}}$	$\mathbf{N^*_{10}}$	$\mathbf{N_{100}}$
yPR65-GSy	4	2	1	28
yPR65y	26	39	34	24
azPR65az	10	15	8	32
cycPR65cyc	8	12	6	22



(a) (b) (c)

Figure 8.7: Summary of parameters obtained from DNA WLC fits to unfolding curves of all four PR65 wild-type attachments: (a) contour length histogram (Gaussian fit: $\mu = 356$ nm, $\sigma = 15$ nm), (b) histogram of the DNA persistence lengths (Gaussian fit: $\mu = 13$ nm, $\sigma = 9$ nm), and (c) DNA persistence length plotted against contour length for each construct.

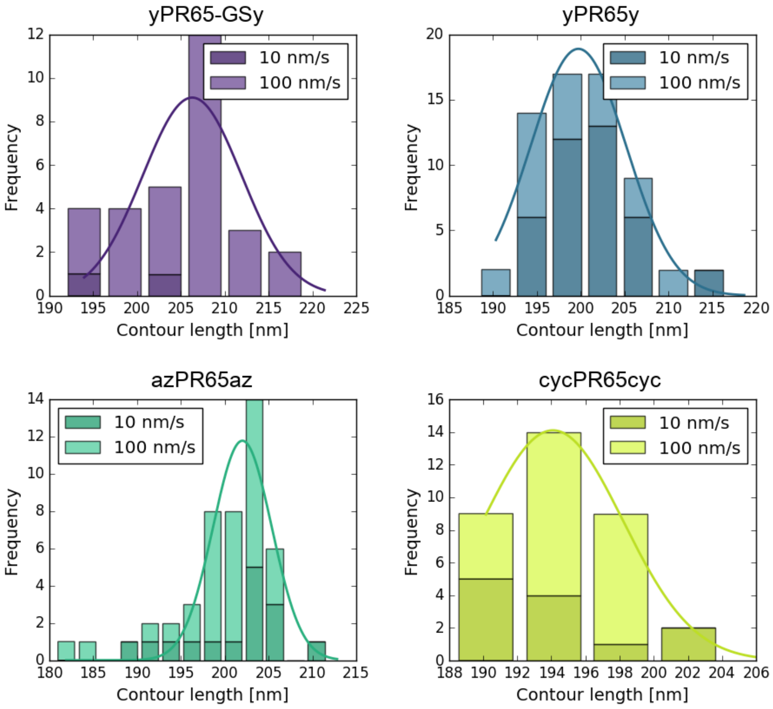


Figure 8.8: Histograms of full extension contour lengths of the four different PR65 attachments. The results of the corresponding Gaussian fits are listed in Table 8.2

.

between $K = 400$ and $K = 1000$.

Histograms of fully unfolded contours are shown in Figure 8.8 and the results of the corresponding Gaussian fits are listed in Table 8.2. The end-to-end distance of PR65 was taken to be that of the *apo* structure and was subtracted from the total expected

Table 8.2: Contour lengths at full extension of wild-type PR65 constructs. End-to-end distances are measured between the C_α atoms of the first and last amino acids of the *apo* structure (PDBid: 1b3u). All values are in nm. Theoretical values were calculated assuming 0.36 nm per amino acid.

Protein	$\mathbf{L_{tot}}$	$\|\Delta\vec{r}\|$	$\mathbf{L_{tag}}$	$\mathbf{L^*_{tot}}$ [a]	$\mathbf{L^{**}_{tot}}$ [b]	$\mathbf{L_c}$
yPR65-GSy	218.88	9.83	6.84	209.05	202.21	206±6
yPR65y	216.36		4.32	206.53		200±5
azPR65az	210.24		-	200.41	-	202±3
cycPR65cyc			-		-	194±4

[a] $L^*_{tot} = L_{tot} - |\Delta\vec{r}|$

[b] $L^{**}_{tot} = L_{tot} - |\Delta\vec{r}| - L_{tag}$

Table 8.3: Contour length increments, otherwise called gains, of the individual sub-transitions in PR65. All values are in nm represent averages and standard deviations of each data set.

Protein	ΔL_A	ΔL_{A1}	ΔL_{A1}	ΔL_B	ΔL_{B1}	ΔL_{B2}	ΔL_C	ΔL_{C1}	ΔL_{C2}	ΔL_D
yPR65-GSy	34	12	22	77	16	61	54	30	24	38
yPR65y	36±3	11±4	26±3	70±6	15±6	56±6	64±13	31±12	33±3	33±10
azPR65az	36±5	9±3	26±3	85±13	23±11	62±8	45±15	16±10	30±7	37±12
cycPR65cyc	34±3	9±3	25±3	74±3	16±3	58±2	46±10	18±5	27±6	43±11

Table 8.4: Forces at ruptures and transition boundaries in unfolding traces of PR65. All values are given in pN and represent averages and standard deviations of the individual data sets.

Protein	$F_{DNA,A}$	F_A	$F_{A,B}$	F_B	$F_{B,C}$	F_C	$F_{C,D}$	$F_{D,ext}$
yPR65-GSy	5.7	7.4	5.6	6.7	4.9	7.0	6.5	8.1
yPR65y	4±1	6.9±0.4	5.8±0.3	6.8±0.4	5.2±0.2	7.9±0.4	6.9±0.3	8.3±0.3
azPR65az	4±1	6.2±0.7	5.2±0.6	6.7±0.3	5.0±0.3	7.1±0.7	6.3±0.5	7.5±0.4
cycPR65cyc	4.2±0.7	6.7±0.2	5.5±0.4	6.6±0.2	4.8±0.2	7.1±0.5	6.2±0.3	7.7±0.3

contour length. For both ybbR-tagged constructs, both L^*_{tot} and $L^*_{tot}*$ are within one standard deviation of the measured values. The azPR65az construct is measured to have the expected contour length within one standard deviation, whereas cycPR65cyc is measured to be shorter. Contour lengths of individual molecules that varied by ± 10 nm from the mean are likely due to the uncertainty in the DNA WLC. However, it could also be possible that a helix or a whole repeat is already unfolded prior to the A-transition, resulting in shorter protein contour lengths.

For all of the constructs other than yPR65y, the amount of data for each construct

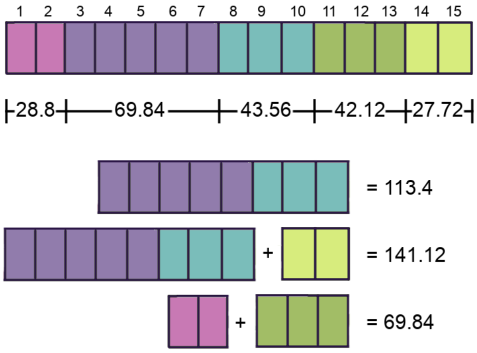


Figure 8.9: Theoretical contour lengths of the repeat domains (in nm) identified in equilibrium denaturations [71]. When folded, a single repeat has an end-to-end distance of approximately 1 nm.

was not sufficient to produce meaningful histograms for forces and contour lengths of each sub-transition. Therefore, only averages and standard deviations were calculated. Average contour-length increases for each transition and the respective sub-transitions are listed in Table 8.3. The forces at which these contours intersect the unfolding force-extension trace are listed in Table 8.4. The variation observed qualitatively in the previous section is reflected in the large standard deviations of the contour length increments, which are between 10% and 60% of the average. In contrast, the forces of ruptures and transition boundaries appear to be less variable as percentage errors tend to be less than 10%. Figure 8.9 shows the domain boundaries that were identified using chemical denaturation [70, 71], the expected contour lengths of each domain as well as the contour lengths of domains that were shown to unfold in the same equilibrium denaturation transition. None of these domain contour lengths correlate well with any of the main four transitions observed. Considering the variation in start and end point of each unfolding transition, it is likely that these domain boundaries are "flexible" within a single molecule. However, given the results from ensemble measurements, transitions A and B probably correspond to the unfolding of some of the chemically weaker repeats (e.g. HEAT3-10), while transitions C and D correspond to the unfolding of more stable repeats.

Since simple averages cannot capture the variations of the data set, the force midpoints of the linear transitions and the force peaks were plotted against their respective contour-length increase (Figure 8.10)[1]. The data in both A-transitions is too spread out

[1] The WLC model cannot describe the apparently linear transitions which are likely to arise from a continuum of states that interchange rapidly within the dead time of the measurement (approximately 10 µs). The majority of the force response of PR65 is not at quasi-equilibrium (i.e. there is significant hysteresis) and hence models such as those used to describe the dynamics of the villin head piece [168] cannot be employed meaningfully either. Although fitting WLC models to the data as defined in Figure 8.4 is technically incorrect, such procedures can provide rough quantitative estimates on the contour lengths which could otherwise also be determined manually after CLT of the data.

to draw meaningful conclusions. In the B-transitions clustering appears to be similar in all constructs. In the C-transition some differences between ybbR and orthogonal chemistry attachments may be present, as the majority of ybbR-tagged molecules clusters at higher peak and midpoint forces than the majority of constructs without the ybbR-tag. This separation is even clearer in the D-transition. The data were unsuitable to perform statistical tests since data sets were not normally distributed and not of equal variance. To conclusively determine whether this difference is real, more data on the orthogonal chemistry attachments are required.

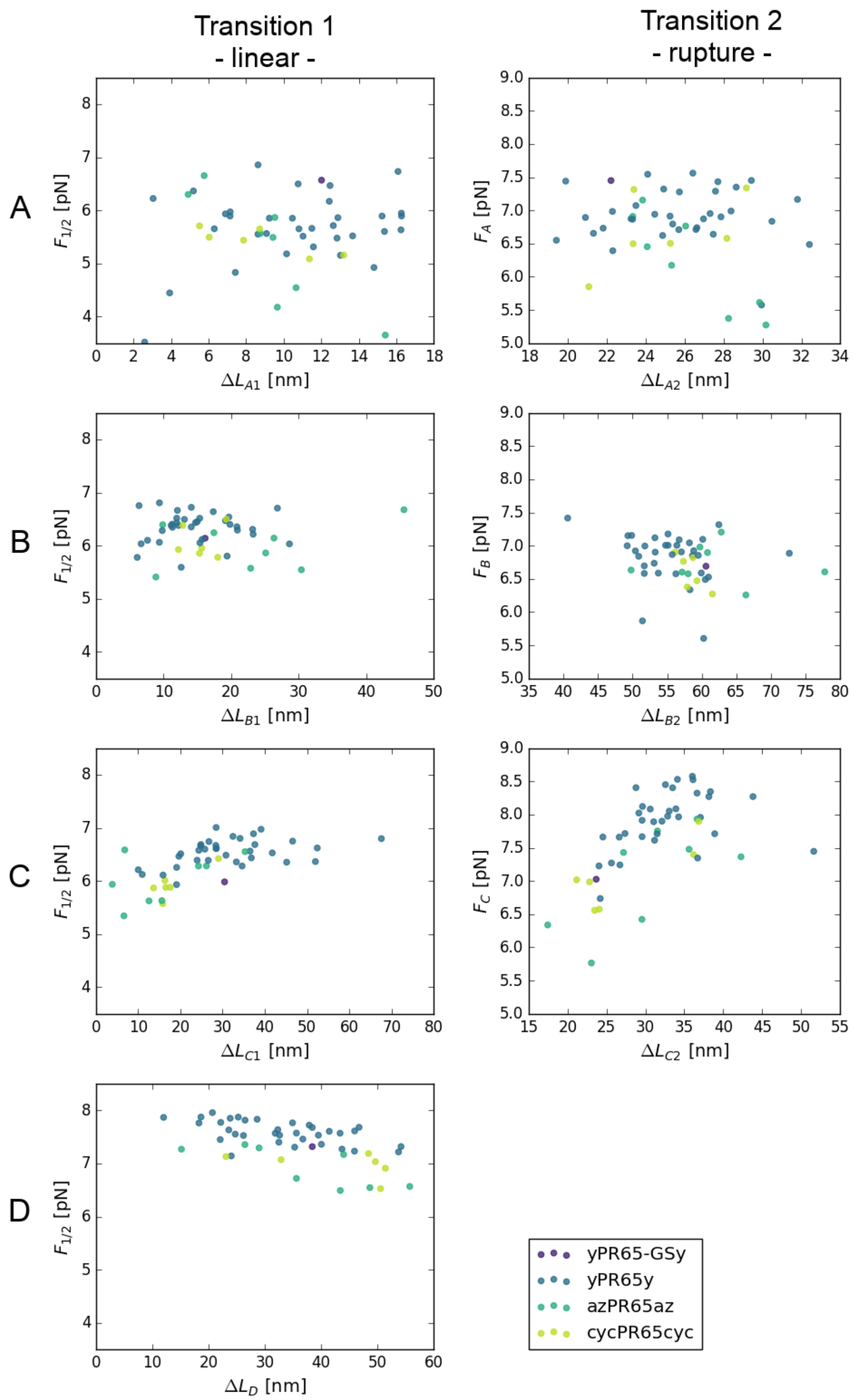


Figure 8.10: Scatter plots of the force midpoints of linear transitions (left) and the peak forces of ruptures (right) as a function of their respective contour-length increase.

8.3.2 Single point mutations give insight into the order of unfolding

If single point mutations affect the stability of the repeat in which they are localised, it will cause a change in the corresponding unfolding force. I chose to investigate the following five mutations across the different domains, all of which have previously been examined by chemical denaturation (Figure 8.11a, [71]):

- **V25A and V65A** in HEAT1 and 2, each of which alone do not affect the stability of PR65, but together might cause a detectable change. V25A changes in the unfolding kinetics.
- **V201A** in HEAT6, which destabilizes the native state.
- **V284A** in HEAT8, which moderately destabilizes the native state.
- **S445P** in HEAT12, which does not affect the native state but destabilises the intermediate relative to the unfolded state.
- **V536A** in HEAT 14, which significantly destabilizes the native state.

Representative force-extension curves of ybbR-tagged molecules containing the V201A, V284A, S445P mutations are presented in Figure 8.12. Data of the V25A/V65A double mutation has been obtained but was not yet analysed, and the V536A mutation has not yet been measured. Table 8.5 summarizes the number of molecules and traces analysed. Only few measurements were obtained for the V283A mutant and data of only two out of four molecules were of sufficient quality for further analysis.

At a first inspection, the unfolding and refolding behaviour of all mutants is very similar to the wild-type data which confirms that these single point mutants do not significantly alter the overall unfolding behaviour of PR65. Some subtle differences, however, could be observed for the V201A and V284A mutants. While in wild-type protein, peaks

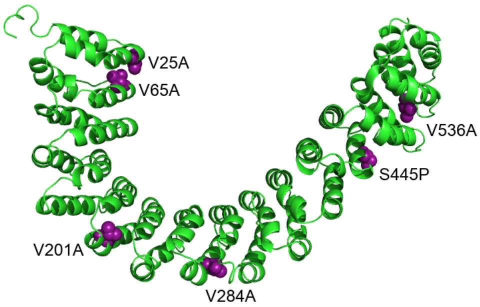


Figure 8.11: Location of the residues selected for mutagenesis within the PR65 *apo* crystal structure (PDBid: 1b3u, [16])

,

Table 8.5: Overview of the number of molecules examined for each single point mutant and the total data obtained at different pulling velocities. Abbreviations are the same as in Table 8.1.

Protein	**Molecules**	$\mathbf{N_{10}}$	$\mathbf{N^*_{10}}$	$\mathbf{N_{100}}$
V201A	13	24	23	48
V284A	2	3	3	5
S445P	12	15	15	51

of the A- and B-transitions are often quite similar in force, some V201A and all V284A molecules showed a reduction in force of the A-transition (arrowheads in Figure 8.12). In other V201A molecules this difference was absent (second molecule in Figure 8.12). Visually, the S445P mutant appeared to exhibit an unaltered force-response compared to the wild-type. However, in multiple cases it was possible to successfully refold S445P after a misfolding event had occurred. Consecutive traces of one such molecule are shown in Figure 8.13. The first cycle and the second unfolding trace exhibit a normal force-response. The refolding trace in the second cycle already indicates that an abnormal folding event has occurred. In the four subsequent cycles the unfolding behaviour of the molecule is very different until in the sixth trace the refolding pattern is yet again normal. The last two cycles then show the usual folding behaviour of PR65.

The measured DNA contour and persistence lengths for these data were 363 ± 9 nm and 16 ± 5 nm, respectively. For each mutant, histograms of the protein contour lengths at full extension are shown in Figure 8.14 and the results of the corresponding Gaussian fits are listed in Table 8.6. There was not sufficient data for V284A to fit a Gaussian, and hence the average and standard deviation were calculated instead. Two of the three proteins are shorter than expected if one assumes the ybbR-tag unfolds with the protein, but they are within one standard deviation of the expected length if the ybbR-tag unfolds with the DNA prior to the protein.

Table 8.6: Contour-lengths at full extension of PR65 single point mutants. End-to-end distances are the same as for the wild-type, but due to a different cloning strategy the ybbR-tag is now separated from the protein by 2 amino acids at either of the termini. All values are in nm.

Protein	$\mathbf{L_{tot}}$	$\lvert\boldsymbol{\Delta}\vec{r}\rvert$	$\mathbf{L_{tag}}$	$\mathbf{L^*_{tot}}^{a}$	$\mathbf{L^{**}_{tot}}^{a}$	$\mathbf{L_c}$
V201A	217.8	9.83	5.76	208.0	202.21	194±9
V284A						204±5
S445P						201±6

[a] As defined in Table 8.2.

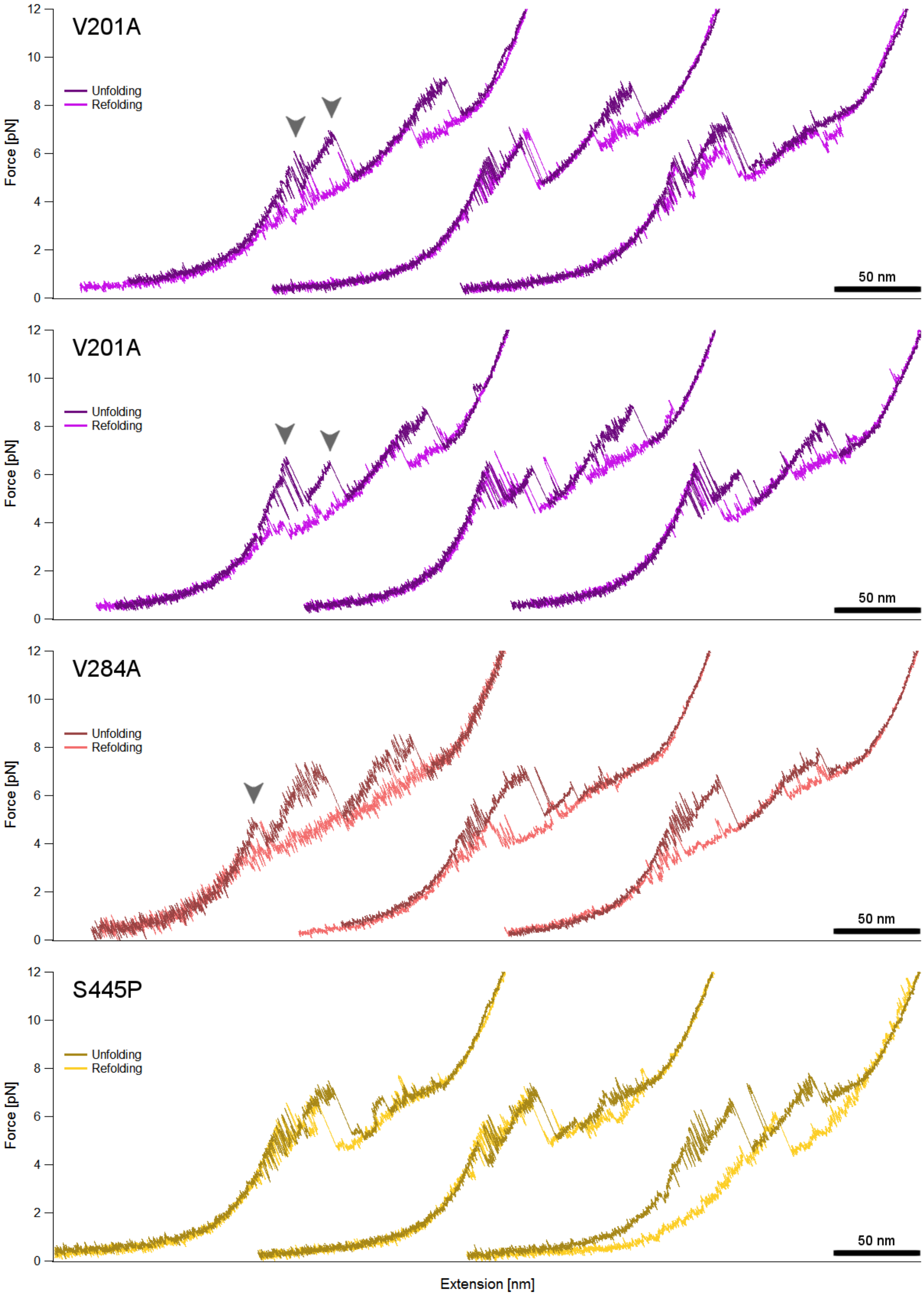


Figure 8.12: Example full length force-extension curves of the same molecule for each of the three single point mutants. Unfolding traces are coloured in darker shades, while refolding traces are coloured in lighter shades. Arrowhead indicate the transitions where mutants differ from wild-type PR65. All data shown were obtained at pulling velocities of 10 nm s^{-1}.

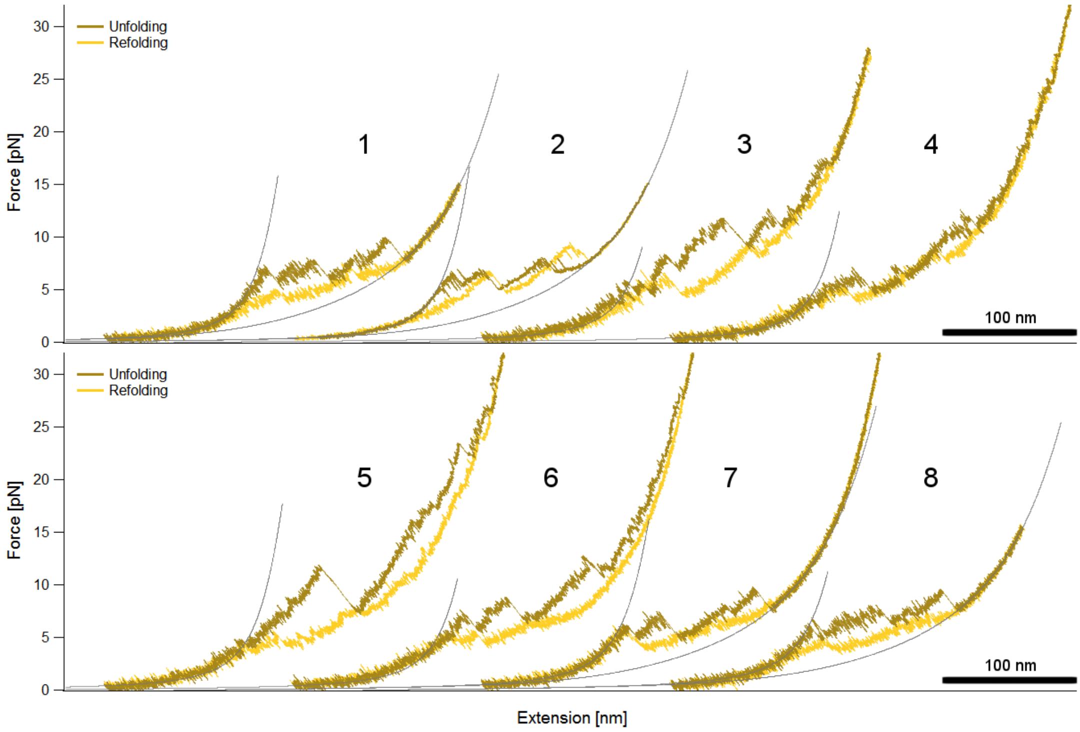


Figure 8.13: Misfolding of the S445P mutant can be recovered in some cases by applying large forces. Shown are consecutive stretch-relax cycles of the same molecule. Misfolding occurs in the refolding trace of cycle 2. DNA contours of cycle 2 are shown in cycles 3-5 to highlight that those parts of the protein which unfold in the A and B transition never refold properly. The refolding trace in cycle 5 is then again normal and is followed by the usual force response in cycles 6 and 7.

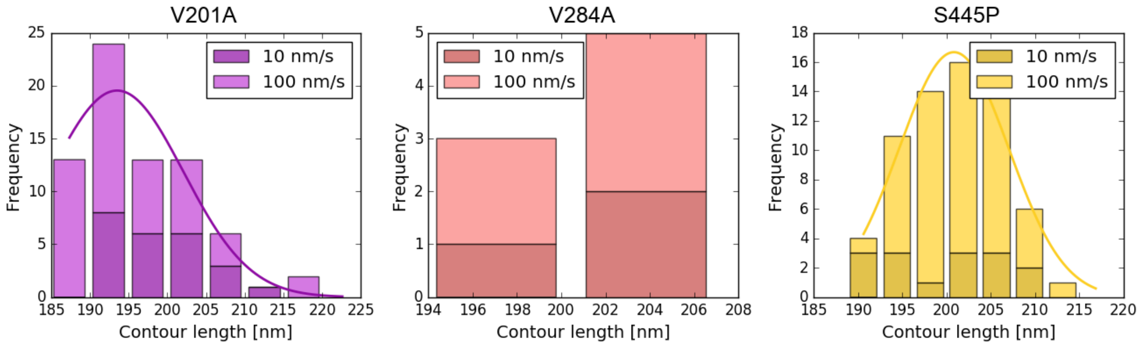


Figure 8.14: Histograms of full extension contour lengths of the three PR65 mutants. The results of the corresponding Gaussian fits for the V201A and S445P mutations are listed in Table. The amount of data obtained for the V284A mutant was not sufficient to fit a Gaussian.

The contour-length increments of each transition and the corresponding peak and boundary forces are listed in Table 8.6 and Table 8.8, respectively. On average, the individual transitions of the mutants appear to have a similar contour-length increase as

Table 8.7: Contour length increments of sub-transitions in PR65 single point mutants. All values are in nm, and represent averages and standard deviations.

Protein	ΔL_A	ΔL_{A1}	ΔL_{A1}	ΔL_B	ΔL_{B1}	ΔL_{B2}	ΔL_C	ΔL_{C1}	ΔL_{C2}	ΔL_D
V201A	42±8	10±3	32±8	71±8	17±7	54±7	63±14	26±9	36±7	27±14
V284A	34±3	9±2	25±2	82±4	23±4	59±4	50±14	22±13	28±7	41±13
S445P	38±4	11±4	26±3	73±6	18±7	55±4	59±18	26±13	33±8	30±11
yPR65y	36±3	11±4	26±3	70±6	15±6	56±6	64±13	31±12	33±3	33±10

Table 8.8: Forces at ruptures and transition boundaries of PR65 single point mutants. All values are given in pN, and represent averages and standard deviations.

Protein	$F_{DNA,A}$	F_A	$F_{A,B}$	F_B	$F_{B,C}$	F_C	$F_{C,D}$	$F_{D,ext}$
V201A	3.9±0.8	6.2±0.5	5.0±0.3	6.6±0.5	5.0±0.3	8.1±0.3	6.9±0.5	8.2±0.3
V284A	3.6±0.6	5.6±0.3	4.8±0.3	6.8±0.2	5.0±0.3	7.3±0.6	6.4±0.5	8.1±0.1
S445P	4.3±0.7	6.6±0.7	5.8±0.5	6.9±0.7	5.2±0.2	7.8±0.7	6.9±0.5	7.9±0.3
yPR65y	4±1	6.9±0.4	5.8±0.3	6.8±0.4	5.4±0.6	7.9±0.4	6.9±0.3	8.3±0.3

was observed for the wild-type. Only the A-transition of V201A tend to larger contour-length increases. The average force peaks and transition boundaries show that the peak force of the V284A A-transition is clearly lower than that of the wild-type. The boundary between A- and B-transitions occurs at lower forces in both V201A and V284A.

These differences can be better visualized when midpoint and peak forces are plotted against their respective contour-length increase (Figure 8.15). In transition A1 of all V284A molecules and about 50% of V201A molecules the transition midpoint occurs at lower forces than that of the wild-type. For the S445P mutant a trend can also be seen, but it is less well defined. In the A2-transition the V201A data cluster into two different regions. Approximately one half of the molecules have similar contour-length increases and rupturing forces as the V284A mutant (first molecule in Figure 8.12). The other half has rupturing forces only slightly smaller than the wild-type but of much longer contour-length increases (second molecule in Figure 8.12). Since the rupture of transition A of V201A and V284A occurs at smaller forces, the force of the B1 midpoint is lowered as well. In transition B2, V201A ruptures spread over a wider force range, whereas V284A, S445P and the wild-type cluster tightly in one region on the plot. No clear differences between the mutants and the WT are observed for the other transitions, suggesting that these are unaffected by any of the mutations examined here.

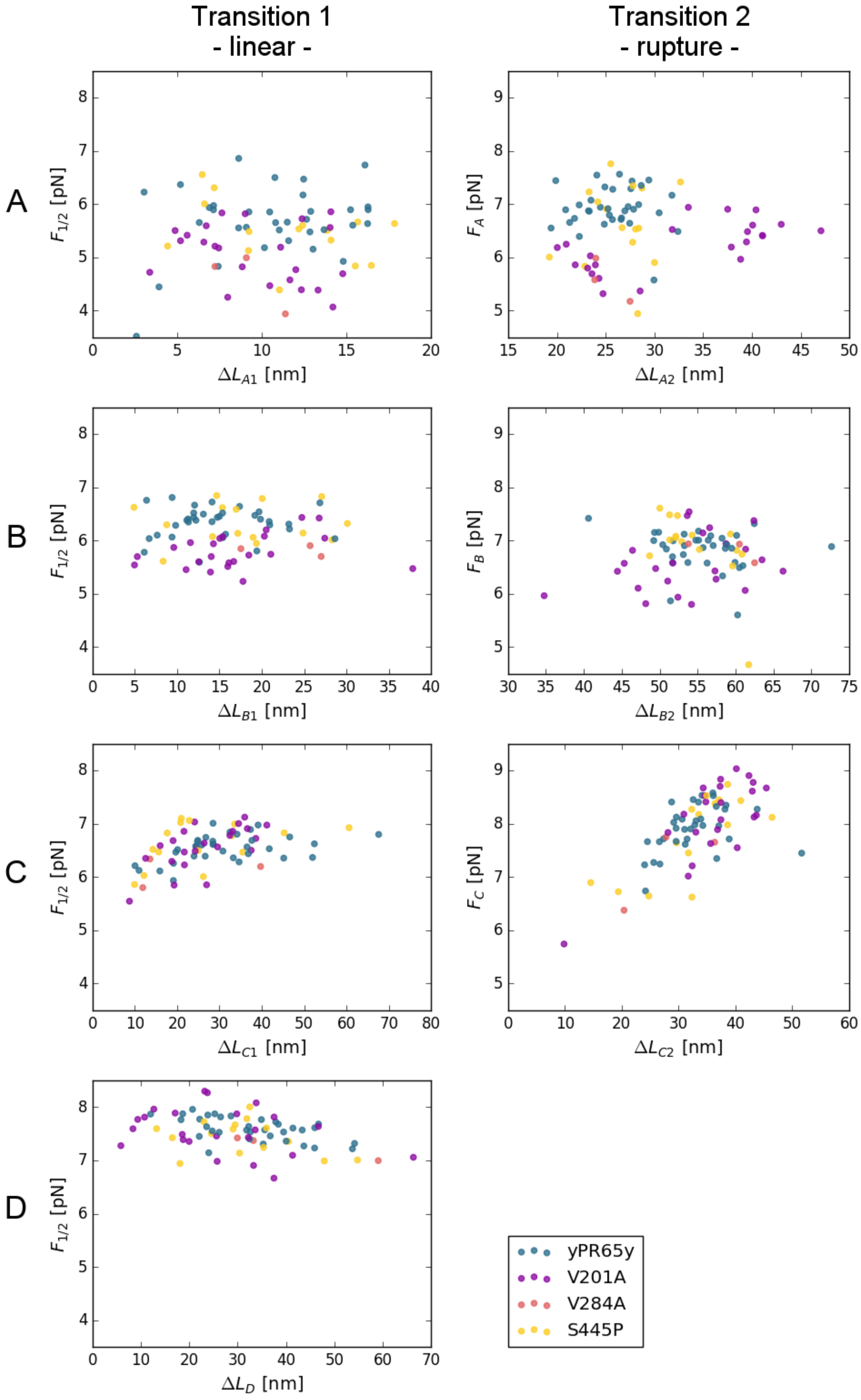


Figure 8.15: Scatter plots of the mutant force midpoints of linear transitions (left) and the peak forces of ruptures (right) as a function of their respective contour-length increase. The wild-type data of yPR65y are included for ease of comparison.

8.3.3 Changing the pulling geometry: yPR65-277az

To reduce the complexity of the unfolding pattern, one can apply forces to only a subset of repeats. This can be achieved either by using truncation mutants or by attaching one of the handles at an internal repeat. Since many PR65 truncations are unstable, I chose to introduce an unnatural amino acid at the position of E277 for an internal attachment located just after the inter-repeat loop between HEAT7 and HEAT8, at the beginning of the HEAT8 A-helix. (Figure 8.16).

Two stretch-relax cycles of four molecules are shown in Figure 8.17. Surprisingly, the unfolding of the N-terminal half on its own is very different from any transition observed for the whole molecule. Three transitions, not all of which are present in every trace, are defined in Figure 8.18. The first rupture of the molecule can occur at a range of different forces, from 6 pN to almost 10 pN (Figure 8.17). Sometimes fast unfolding and refolding dynamics to a second state can be observed (Transition 1), before the majority of the N-terminal half unfolds in one non-equilibrium transition (Transition 2). The final unfolding event (Transition 3) resembles the D-transition but is often much shorter or

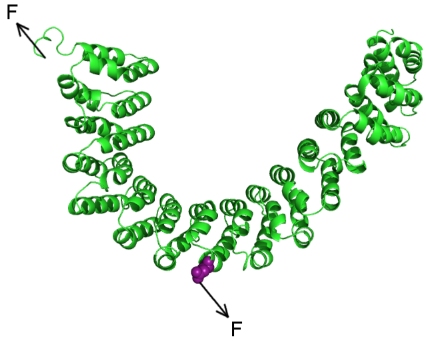


Figure 8.16: Pulling geometry in the yPR65-227az variant. Forces are applied to the N-terminus and to the residue at position 277.

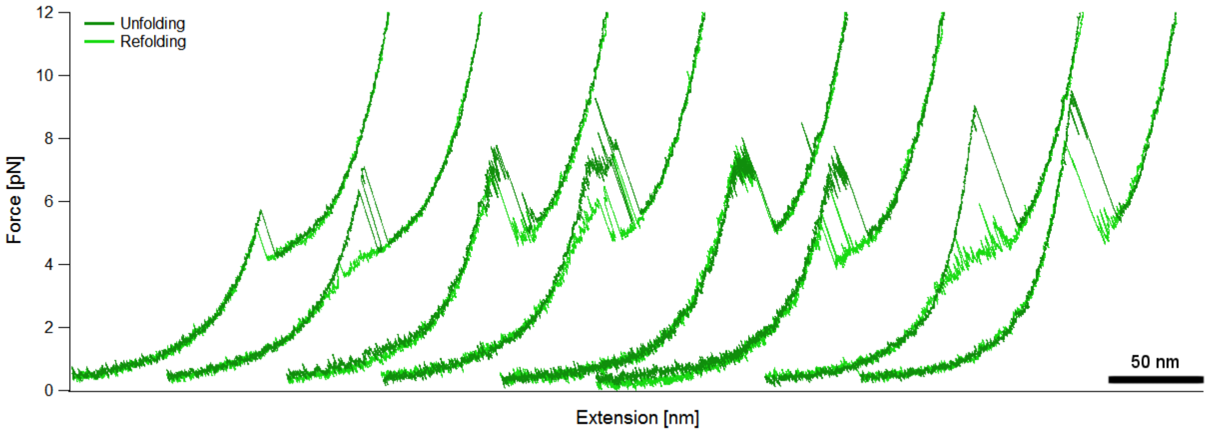


Figure 8.17: Representative stretch-relax cycles of four yPR65-277az molecules. All data shown were obtained at pulling velocities of 10 nm s^{-1}

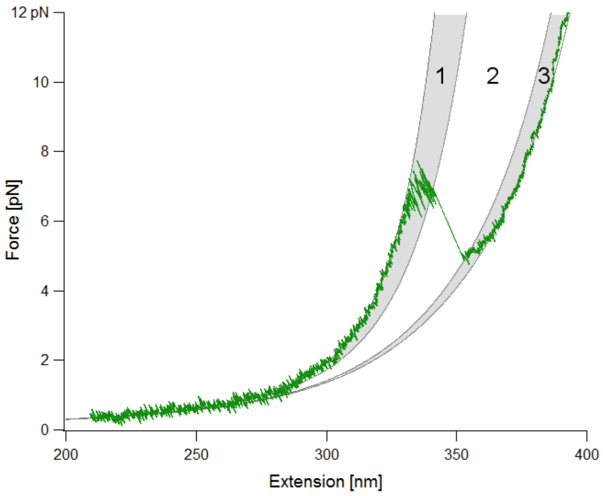


Figure 8.18: Sub-division of the yPR65-277az force response into three transitions. These were not always clearly defined in all force-extension curves; some molecules or traces of the same molecule only exhibited a single transition or two out of the three.

nearly absent (Figure 8.17). The refolding curves show variations comparable to the wild-type. Similar to curves of the end-to-end pulling geometry, the refolding traces follow the final transition for longer until they refold fully, either in a gradual or sudden transition. Although variations between individual stretch-relax cycles of the same molecule are still present, misfolding was never observed and sometimes data of more than 100 cycles could be obtained for the same molecule. Furthermore, it was possible to monitor the unfolding at a variety of pulling velocities (Figure 8.19). Whereas the unfolding forces appear to increase slightly with higher velocities, the refolding transitions become more likely to be gradual and less likely to exhibit distinct force peaks with higher velocities. After velocities of up to 1 µm s^{-1} were tested, the unfolding behaviour at the usual 10 nm s^{-1}or 100 nm s^{-1} was the same as before the velocity was increased.

DNA WLC fits resulted in contour and persistence lengths of 356 ± 3 nm and 21 ± 5 nm, respectively, and the spread is similar to all other constructs tested (Figure 8.20a). A histogram of the full extension contour is shown in Figure 8.20b and the fitting result is listed in Table 8.9. The measured contour of the fully extended N-terminal half is approximately 10 nm shorter than expected even if the N-terminal ybbR-tag is considered independent of the repeat array. Some molecules were subjected to very high forces (>30 pN), but further unfolding events could never be observed.

Average contour-length increases and forces of each peak and transition boundary are listed in Table 8.10, while Figure 8.21 shows the peak forces of transitions 1 and 2 as well as the force midpoint of transition 3 plotted against the respective contour-length increases. The data are compared with the A-, B- and D-transitions of whole molecules. Since a state resembling the C-transition was never observed, a comparison to these data

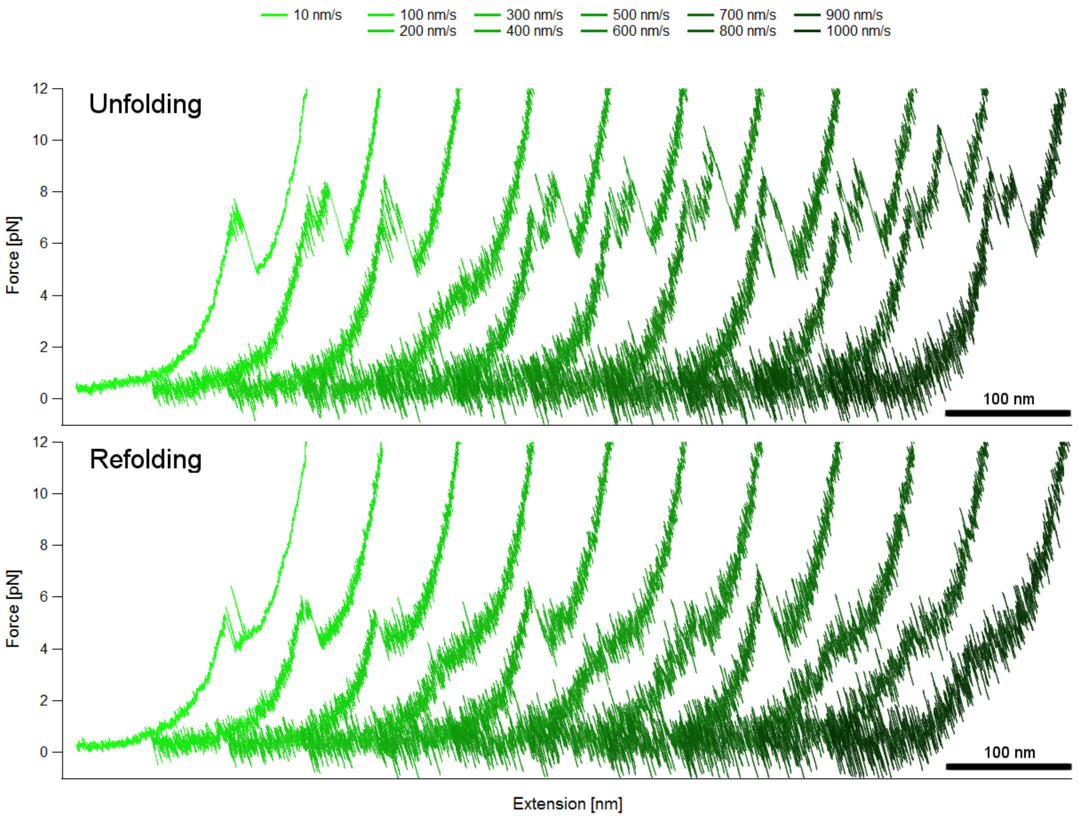
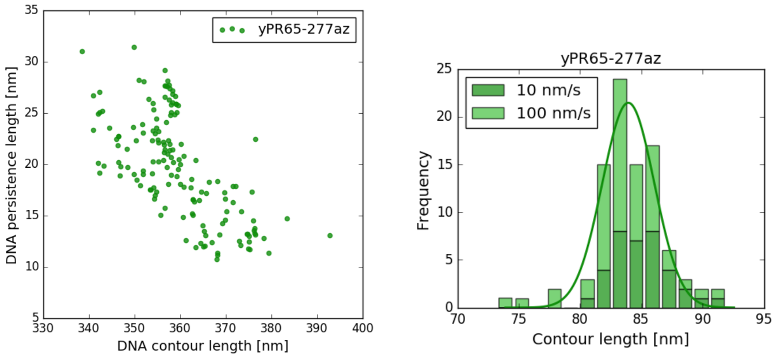


Figure 8.19: Force-extension curves of the same yPR65-277az molecule taken at different pulling velocities, starting with 10 nm s^{-1} (light green) and increasing up to 1 µm s^{-1} (dark green).



(a) (b)

Figure 8.20: WLC parameters of the DNA (a) and the fully extended yPR65-277az protein (b). Data are from 10 molecules, of which a total of 45 and 72 traces were obtained at 10 nm s^{-1} and 100 nm s^{-1} respectively.

was not performed. The data presented in Figure 8.21 confirms that unfolding of only one half of the molecule is yet again different that unfolding the whole molecule.

Table 8.9: Contour lengths at full extension of yPR65-277az. End-to-end distances are between the first residue and E277 in the *apo* crystal. All values are in nm.

Protein	$\mathbf{L_{tot}}$	$\|\Delta\vec{r}\|$	$\mathbf{L_{tag}}$	$\mathbf{L^{*}_{tot}}$ [a]	$\mathbf{L^{**}_{tot}}$ [a]	$\mathbf{L_c}$
yPR65-277az	103.32	7.33	3.6	96.83	93.32	84±2

[a] As defined in Table 8.2.

Table 8.10: Contour-length increments and forces at ruptures and transition boundaries of yPR65-277az. All lengths are in nm while forces are in pN. The data represent averages of 42 10 nm s^{-1} traces.

Protein	ΔL_1	ΔL_2	ΔL_3	$F_{peak,1}$	$F_{peak,2}$	$F_{2,3}$	$F_{3,ext}$
yPR65-277az	22±5	52±5	17±8	7.8±0.9	7.2±0.4	5.0±0.3	6.0±0.4

8.4 Discussion

8.4.1 AFM and optical tweezers each resolve a very different force response of PR65

Considering that we had previously obtained AFM data of PR65, it was very surprising that I did not observe unfolding of PR65 in a step-wise manner, repeat by repeat. Instead I could detect only four, sometimes five, unfolding events, which suggests that the protein is unfolding in larger domains. At first, we assumed that this was due to the change in experimental set-up. Due to surface immobilization in AFM experiments, PR65 was likely never observed to unfold fully [312]. Furthermore, it is unsure at which pulling velocities traces as presented in Figure 8.2 were obtained. I examined the unfolding of PR65 at velocities higher than 100 nm s^{-1} (e.g. 200-500 nm s^{-1}) but found that the protein response was even more variable and when combined with its misfolding propensity, these data were more difficult to analyse. Lastly, in conjugation reactions and denaturations performed in buffers at lower pH (i.e. 6.5 or 6.0) without salt, protein was lost over time, most likely due to sticking to plastic surfaces or aggregation. In an optical tweezers set up, the functionalised beads and the chamber can provide surfaces for non-specific interactions even after blocking. Therefore, I performed all experiments in Tris-HCl pH 7.5 with 150 mM NaCl to ensure solubility of PR65 over longer periods of time at room temperature. However, the AFM data was acquired in MES pH 6.5 without salt. PR65 contains a conserved R-D ladder along one side of the HEAT repeat array, which stabilizes interactions between neighbouring repeats [16]. Screening of these electrostatic interactions by ions in the buffer might therefore alter the overall stability of the molecule and hence could result in a different force response. However, previously

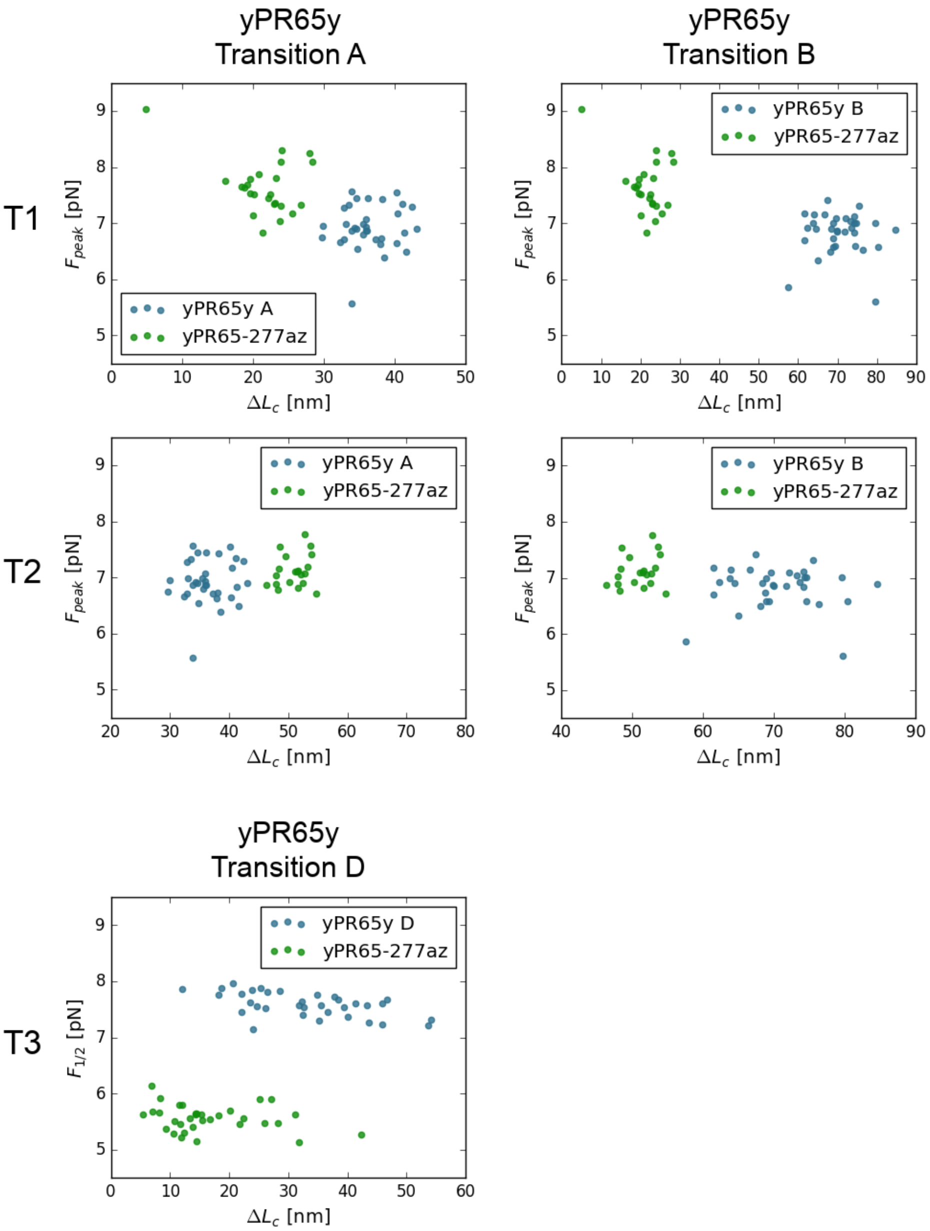


Figure 8.21: Scatter plots of the peak forces of transitions 1 and 2, and the force midpoint of transition 3 transitions against their respective contour-length increase. Each transition is compared to the full A- and B-transitions of yPR65y.

published AFM data of repeat proteins were acquired in a range of buffers, most of which contained salt [194–197]. For example, Serquera *et al.* [195] found that the characteristic unfolding pattern of gankyrin was independent of the sample buffer used. Moreover, clathrin, another HEAT repeat protein, was found to unfold in a step-wise manner in PBS. Therefore, I have yet to confirm whether the difference in force response is due to the buffer composition, or due to the pulling velocities employed.[2] We are also in process

[2] Post-submission, force-extension data were obtained for azPR65az measured in both 50 mM Tris-HCl pH 7.5, 150 mM NaCl and 50 mM MES pH 6.5. These show no dependence of the force response on buffer (see Appendix B, Figure B.3).

of obtaining the original AFM data from our collaborators for in depth comparison.

8.4.2 Destabilizing point mutations provide insight into the unfolding order

Although the four unfolding events defined here were present in most traces, their boundaries were prone to shift. Using simple averages of the contour-length increase in each (sub)transition, it was not possible to assign the unfolding of particular repeat domains to transitions in the force-extension curve. More data is required to produce reliable contour length histograms for each transition, and to examine whether the transition boundaries shift in discrete intervals, i.e. by the contour length of one helix or repeat. Furthermore, if discrete intervals are present, it will be possible to identify clearly separated clusters in force versus contour length plots.

Data for three single point mutants were presented, two of which (V201A, V284A) showed subtle, but significant changes in the unfolding response, and one of which (V201A) produced a subpopulation of molecules with different contour-length increase. The V284A mutation significantly affects the first transition and causes that domain of the protein to unfold at even lower forces. However, the fast dynamics of unfolding and refolding in that transition remain unaffected. V284A is part of the hydrophobic core at the interface between HEAT8 and HEAT9. Since in the data presented here, it only lowers the unfolding force, but leaves the contour-length increase unaffected, breaking of this interface must be included in the first transition. On its own however, the V284A mutation cannot identify which repeats exactly are unravelling. In molecules containing the V201A mutation, two distinct events could be observed. About half of the molecules exhibited a similarly altered unfolding as the V284A mutant. In the other half, however, the first transition exhibited a higher contour-length increase corresponding to approximately one additional repeat unfolding, and only slightly decreased in unfolding force. The additional contour-length increase is of the order of one more repeat or two more helices (one from each of two repeats) unfolding. In these molecules, the second transition subsequently occurs at lower forces although a change in contour-length increase could not be detected. V201A resides in the first helix of HEAT6 where it is involved in forming hydrophobic interactions between HEAT5 and HEAT6. On average the first unfolding event corresponds to a contour-length increase equivalent to 2-3 repeats. Given the results from these two mutants, it is likely that the unfolding of helices from HEAT6-8 occurs in the A-transition. While more mutations across different repeats are required to test this hypothesis, it nevertheless is supported by the data obtained from unfolding only the N-terminal half.

As it was intended, pulling PR65 from the N-terminus and an internal attachment point reduced the complexity of the force response as only the unfolding and refolding of seven repeats could be probed. However, the unfolding pattern of the yPR65-277az attachment was again different from and could not be compared directly to the full extension of PR65. Although in some cases, similar dynamics to that of the A-transition could be observed, it occurred at higher forces and corresponded to a contour-length increase of only 1-2 repeats. Since the attachment in this variant is at residue 277, just after the inter-repeat loop between HEAT7 and HEAT8, the majority of HEAT8 does not experience any force. As the unfolding shifts to higher forces, this suggests that the unfolding of HEAT8 is required for HEAT7 and HEAT6 to unfold. If the V201A mutation is present, it becomes more likely that HEAT5 (or parts of it) unfolds together with these repeats. Given the observation about the A-transition, the B-transition is likely to involve the unfolding of HEAT3-5 as well as HEAT9-10. Due to the presence of more stable repeats adjacent to HEAT3 and HEAT10, the exact domain boundary could vary probabilistically, causing them to unfold in the B-transition in some traces and in the C-transition in others, which could give rise to the large variations observed for these transitions. Currently, it is not possible to assign the unfolding of a particular set of repeats to either the C- or D-transition. However, in comparison to data of ybbR-tagged constructs, other attachments tended to exhibit a decrease in unfolding forces of these last two transitions. Although this observation has yet to be confirmed, the data hint towards a slight stabilising effect of the ybbR-tags on the terminal repeats.

It is important to note that the results presented here contrast with the findings from MD simulations performed by Settanni and co-workers, which show that HEAT15 is most likely to unfold first. Of course, such an event could be possible if HEAT15 unfolds concurrently with 1-2 central repeats. Then, a mutation in HEAT6 or 8 would simply shift the probability of which of the central repeats unfold first. Future investigations, e.g. of the V536A mutant, will hopefully provide clarity. The scenario in which parts of HEAT6-8 unfold first does however include the weak point between repeats 6 and 7 that was identified by Grinthal *et al.* [201]. Although data form pulling on the N-terminal half implies that the unfolding of HEAT 8 precedes the unfolding of HEAT6-7, which is similar to what has been observed in equilibrium denaturations, the very fast unfolding and refolding dynamics of the first transition are still present in both mutants. This suggests that the source of these dynamics lies between HEAT6 and HEAT8. Destabilization of the HEAT5/6 and HEAT8/9 interfaces shifts these dynamics to lower forces but might leave the kinetics unaffected. Further analysis on the kinetics of this transition as well as investigations of mutations in HEAT7 and the HEAT6/7 interface would be required to confirm this qualitative observation. In any case, the data presented here hint towards a coupling of the unfolding of HEAT8 to that of HEAT6-7, and future studies may be able to elucidate the details.

8.4.3 Misfolding is likely to arise in the most stable HEAT repeats

In contrast to mutations in the central, weaker repeats, the S445P mutation did not show any alteration in the unfolding behaviour of PR65. Considering, the variation that is intrinsic to the force response, more data is required to confirm this observation. However, this mutant could be refolded successfully several times, which was not yet observed with the wild-type protein. HEAT12 is one of only three repeats that lack the conserved proline in the A-helix [16], suggesting that a substitution to the consensus proline could have a considerable effect as it was evolutionary selected against. Furthermore, the observation that partial unfolding (i.e. unfolding of only the central repeats) and unfolding of the N-terminal half never resulted in misfolding events, suggests that the source of misfolding lies in the C-terminal proportion of the protein. Misfolding occurred at high forces, when the more stable domains of PR65, which unfold in C- and D-transition, can refold. Misfolding in these regions would also leave the other part of the protein unfolded. Together these observations indicate that misfolding occurs somewhere in HEAT11-13. When these repeats are not correctly folded, their "gate-keeper" function for the folding of the central repeats is lost and unfolded remainder is then not able to refold correctly. The S445P mutation may not prevent misfolding, but it might destabilize the misfolded substructure such that once it is broken up the whole protein can refold into the native structure. However, it is doubtful whether such misfolding events would occur in a biological context and even if they did, whether they would be of any relevance. Indeed, it is thought that PR65 mostly exist in complex with the C-subunit *in vivo*, which binds to HEAT11-15 [206, 219]. Such a binding event is expected to significantly stabilise these repeats and decrease the likelihood of them being unfolded (by force or otherwise). This scenario would also support the fly-casting mechanism proposed for PR65 by Tsytlonok *et al.* [71], in which the unfolding of the central repeats would enable PR65 to sample a larger space in search for B-subunits. A stably folded C-terminal domain would therefore be crucial such that the central repeats can refold for the enzyme to perform its function.

8.4.4 A hypothetical mechanism that could give rise to a linear force response

Each of the four unfolding transitions of PR65 contained a sub-transition that was approximately linear in appearance. In the A-transition this linearity is mostly observed in refolding events as it is difficult to distinguish from the DNA response. However, the B- and C-transitions contain clear linear sub-transitions that are followed by a rupture, whereas the D-transition is linear only. In contrast to the D-transition, I would hesitate to claim that the B- and C-transitions are equilibrium transitions similar to that observed for CTPRs or other α-helical proteins [155, 168, 362]. Significant hysteresis between the unfolding and refolding traces can be present, which may be particularly large in the B-transition. Nevertheless, the underlying mechanism is probably similar. Grinthal *et al.* [201] and Kappel *et al.* [200] showed in simulations that mechanical stress is distributed along the HEAT repeat arrays of both PR65 and Importin-β, causing a deformation of the tertiary structure while leaving the secondary structure intact. In the apparent linear transitions such a global response could be combined with fast kinetics of unfolding and refolding of parts of, or whole α-helices and repeats at either end of the folded remainder of the protein. Specifically the start of the C-transition is often similar in noise as the DNA prior to the A-transition, indicating that if fluctuations in the secondary structure are present, they must be of a small amplitude. Furthermore, the slopes of the linear transitions decrease with each unfolding event: in the A-transition linearity is nearly invisible (and its presence can be debated), while the slope of the D-transition is closer to horizontal compared to any of the others. If one considered this slope as the spring constant similar to that of a Hookean spring, it at least appears to be proportional to the number of folded repeats. Unfortunately, at this point, the true underlying mechanism within these transitions remains only speculative.

8.4.5 Conclusion

In summary, PR65 unfolding and refolding is highly variable in the conditions tested here. Instead of a step-wise unfolding repeat by repeat, it was shown to unfold in four transitions, three of which contain linear sub-transitions prior to a non-equilibrium unfolding event. The variability observed in all traces suggests that the unfolding of a given repeat is probabilistic to some extend and as such it is not possible to assign clear domain boundaries (Figure 8.22). Using single point mutations I could assign the unfolding of (parts of) HEAT6-8 to the first transition and suggest that the repeats immediately

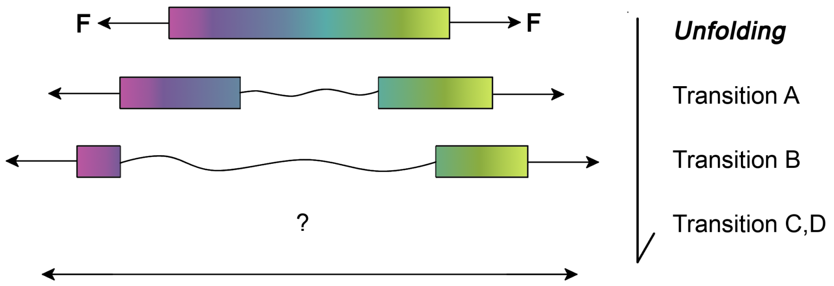


Figure 8.22: Schematic representation of a possible unfolding mechanism of PR65 under force. Although domains within the HEAT repeat array has a different stability, PR65 exhibits a variable unfolding response which indicates that the boundaries between domains are fluent.

adjacent to this domain unfold in the second transition (Figure 8.22). Although I lack direct evidence, indirect support from misfolding events as well as comparisons between different attachment methods indicate that HEAT1-2 and HEAT11-15 unfold in the later transitions at higher forces. Two of the point mutations tested shifted the unfolding events to lower forces without major alterations in the overall force-response, confirming that they could be used to modulate the mechanical stability of PR65, and particularly the integrity of its central repeats, in future phosphatase activity assays. At this time, the refolding and misfolding data has yet to be analysed, which I expect to provide further insights. However, due to the variations observed in each data set, it is clear that more force data of mutants are required to confirm the trends and hypothesis presented here.

Chapter 9

Final conclusions and future work

9.1 DNA-protein chimeras and their applications

In previous single-molecule experiments our group and others showed that α-solenoid repeat proteins unfold at relatively low forces which are difficult to detect by AFM. I therefore chose optical tweezers instead. However, this required the development of a method with which to site-specifically attach DNA handles to a protein containing 14 cysteines: I used amber codon expression to produce alkyne-, azide-, and cyclopropene-bearing PR65 variants, and I attached DNA oligos with complementary functional groups.

Copper-catalysed click between protein and DNA proved to be very inefficient, most likely due to sequestration of copper by the DNA. However, copper-independent approaches worked as well as, if not better than, other attachment methods and do not produce side products such as maleimide and thiol dimers. An alternative method - the Sfp-mediated attachment of CoA to the ybbR-tag - was also applied in two of the repeat-protein systems, PR65 and CTPRs. Surprisingly, the tag greatly stabilised CTPR arrays, possibly due to interaction of the tag with at the repeat interfaces on either end of the array. PR65 stability was not affected to the same degree. Lastly, the intrinsic propensity of the ybbR-tag to form helical structure will limit its application at internal sites.

The two methods enabled reliable attachments of several different proteins, and together they can provide researchers with a set of tools to be used for site-specific labelling and DNA attachment of proteins containing cysteines in force spectroscopy applications and beyond. Indeed, a current member of the Itzhaki group will soon start applying these methods in experiments that use protein-DNA chimeras based on TPRs as an anti-cancer approach to target the Androgen receptor variants.

9.2 Resolving issues with DNA fitting

When WLCs were fitted for both DNA and protein, some of the contour lengths at full extension were found to be significantly smaller than the calculated values based on the length of the polypeptide chain. The magnitude of the discrepancy varied and appeared to be independent of the attachment method used. It was largest for constructs that exhibited the largest unfolding responses, such as CTPRs and yPR65-277az. Shorter contour lengths may arise from weak or unstructured regions that unfold alongside the DNA at very low forces or that are already unfolded prior to the application of force [363]. Considering the varied force response and misfolding events observed for PR65, this is likely to be the source of the discrepancy in contour lengths. However, it is unlikely to be the explanation in the case of the CTPRs, as they are very stable in solution. Lastly, both CTPR_RV and PR65 exhibited some form of force response or unfolding before the main transition started, and hence DNA WLCs were only fitted up to 4-5 pN. At these forces the fitting procedure can be unreliable, and hence variability (or error) can be introduced at this stage. We are currently in the process of addressing these issues and hope that an improvement of the fitting procedure can resolve them.

9.3 Connecting structure and shape to mechanics and dynamics

9.3.1 Functional relevance of dynamics and mechanics

In several instances, the mechanical properties of repeat proteins as well as their ability to be flexible has been implied to be crucial to biological function [6, 71, 197, 199, 201, 203, 279]. The elasticity of repeat proteins can be defined on two levels: in their ability to change shape, and to unfold and refold under force.

Here, I investigated two repeat-protein systems in more detail. In the Rap protein family binding of a peptide results in a compaction at one end of the repeat array, which leads to a conformational change at the other end. Hence, the flexibility of the repeat array is required for the transmission of the allosteric signal. In PP2A, the repeat-protein scaffold PR65 is observed to be highly flexible and its ability to change shape is proposed to be necessary for subunit exchange and catalytic activity. Furthermore, PR65 has been suggested to function as a transmitter of allosteric information between the regulatory and catalytic PP2A subunits.

ENMs can be a powerful tool to assess the functional relevance of dynamics in allosteric mechanisms [245]. However, they are only a harmonic approximation of the much more complex motion that we expect to find in reality and under physiological conditions the normal mode amplitudes are thought to be severely damped [287, 364]. Nevertheless, since the time of the conception of ENMs a large amount of indirect evidence has been collected that somehow normal modes must be relevant for biological function [246–249, 259, 263, 265]. To be able to use ENM for predictive analyses it is first necessary to prove that they can describe the motions observed in reality. Here, I showed conformational changes of Rap and PR65 (as identified in crystal structures) are indeed accessible through vibrational dynamics. Furthermore, by comparing to data from MD simulations, I showed that ENMs can capture the major motions sufficiently well, even though they cannot account for the anharmonicity of the system.

Although one ENM on its own cannot model a whole conformational change, a combination of ENMs or ENMs combined with other approaches can model such processes [258, 364]. Furthermore, ENMs can actually predict mechanically weak spots in a structure and thereby sites at which forces arising from the fluctuations can cause conformational changes or induce unfolding [364, 365]. Using single-molecule force spectroscopy, I was able to show that the chemically weak central repeats of PR65 unfold before any other repeats and can do so with high reversibility. The motions described by ENMs that are lowest in frequency and affect the protein globally all pivot about these central repeats. These results therefore raise the very exciting possibility that they could put strain onto the central domain and cause its unfolding. Such an unfolding event was proposed to be functionally relevant by Tsytlonok *et al.* [71], as it could widen the search radius of PR65 to find regulatory subunits while still bound to the catalytic subunit. Furthermore, an uncoupling of the N-terminal half from the C-terminal repeats may facilitate binding of the large variety of B-subunits as it would be more accessible. Since the refolding of the central repeats can occur under forces of 6 pN, PR65 could then pull B-subunits back into proximity of the catalytic subunit, even if the viscosity of the surrounding medium (such as the cytosol) is very high. In fact, HEAT7 and HEAT8 are subject to multiple phosphory- lations in cardiac PR65, and hence their stability might be particularly important for PP2A function *in vivo* [366]. Phosphorylation of several residues was shown to reduce PP2A activity, possibly due to destabilising effects resulting from the introduction of additional charges [366].

Lastly, the localisation of a substrate-binding subunit and a catalytic subunit to opposite ends of a repeat-protein scaffold is not limited to PP2A. For instance, in the SCF$^{\mathrm{Skp2}}$ ubiquitin ligase, the substrate-binding complex is recruited to one end of a cullin-repeat domain and the E2 ubiquitin-conjugating enzyme binds to the other end (Figure 9.1). The Cullins are highly flexible repeat proteins, and their ability to change shape is thought to be crucial for orchestrating consecutive cycles of substrate ubiquitination [368, 369]. Considering the similar architectures of the SCF and PP2A enzyme complexes, it seems plausible that there is a common underlying mechanism exploiting flexible repeat-protein

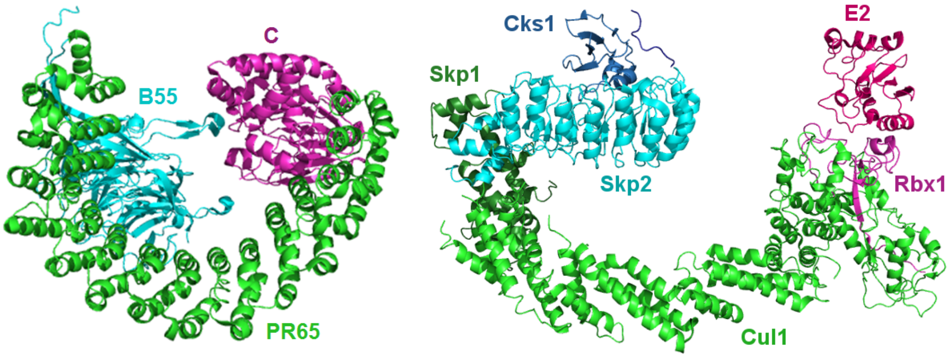


Figure 9.1: Repeat proteins linking function in multi-protein complexes. (a) PP2A (PDB 3dw8) consisting of PR65, the catalytic subunit C, and the B55 regulatory subunit. (b) Modelled structure of the SCF$^{\mathrm{Skp2}}$ ubiquitin ligase consisting of three core subunits Skp1, cullin-repeat containing protein Cul1 and Rbx1, and substrate-recognition subunit Skp2 with accessory protein Cks1, which together recruit the substrate p27. Also shown is the E2 ubiquitin-conjugating enzyme, which is recruited to the SCF by Rbx1, together forming the catalytic entity (PDBs 2AST, 1LDK and 4Q5E [367]).

connectors for such catalytic processes.

9.3.2 Relating Ising models to mechanical and dynamic properties of CTPRs

The use of consensus repeat proteins has been an immense benefit to the repeat protein folding field. As they comprise repeat arrays of varying length containing exactly the same repeat sequence, it is possible to decompose their conformational free energy into intrinsic repeat and inter-repeat interfacial stabilities. These analyses have provided many insights into the length-scale and magnitude of cooperativity in repeat protein, and how it depends on the stability of each repeat as well as the interactions between them [125, 133].

Here, I explored the dynamic potential and mechanical characteristics of two CTPR variants which differ significantly in shape and thermodynamic stability. ENM models showed that the thermodynamically weaker CTPR_RV variant is more favourable entropically, and this raises the question as to how a protein's shape is related to its overall stability. Is the difference in stability between CTPR_RV and CTPRa only due to changes in repeat stability and interfacial interactions, or does the increase in entropy due to changes in shape contribute as well?

In equilibrium chemical denaturation studies, the stability of CTPRs increases with increasing number of repeats. Under force, the presence of more repeats does not increase the unfolding force but instead only extend the range by which the protein unfolds. However, a change in mechanical stability can be observed, and hence, in CTPRs it must correlate with the intrinsic and interfacial energies and not with overall stability. Further investigations into how these two energetic parameters relate to mechanical stability should therefore enable the development of a model that can predict the spring constants of CTPRs and, thereby, the rational design of repeat proteins with customised mechanical properties.

9.3.3 The scope of future investigations

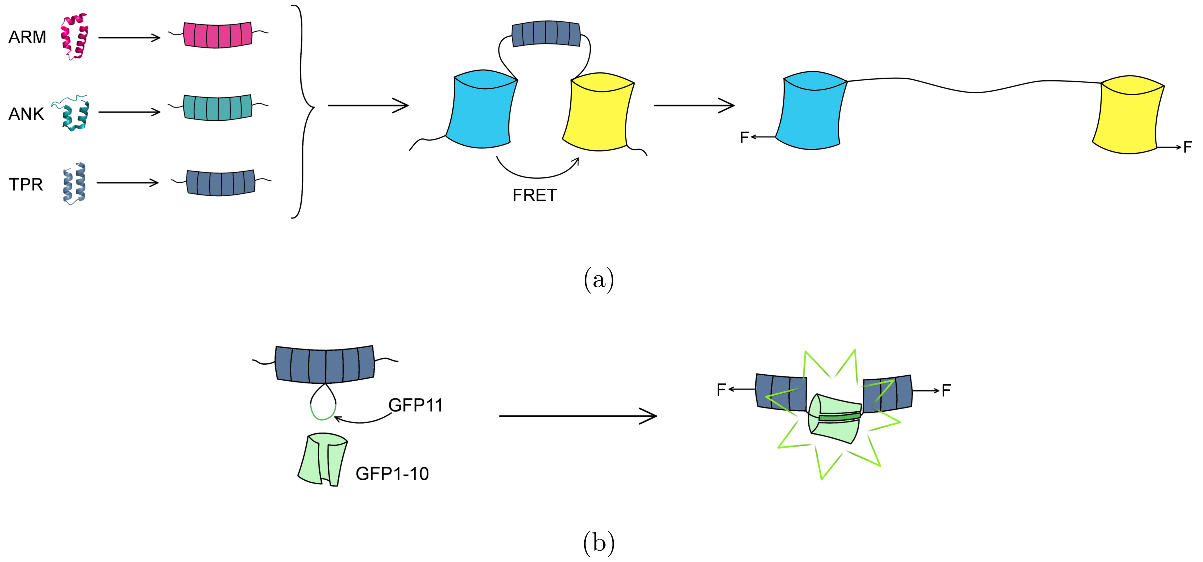
To examine the effects of mechanically altered PR65 variants, I started to set up phosphatase assays involving three different *holo*-enzymes. However, the production of a functional catalytic subunit requires baculovirus-mediated expression in insect cells. Unfortunately, I have not yet been able to produce sufficient protein for enzymatic assays due to genetic instability of the subunit gene in the virus. Furthermore, I have used a different amber suppression system to produce substrates that incorporate phospho-serines site-specifically in response to a stop-codon [223]. Phosphorylation of PP2A substrates is achieved with a specific kinase and usually results in variable phosporylation when multiple serines and threonines are present [209, 217]. The ability to engineer in and vary the positions of the phosphoserines will therefore be important in understanding the catalytic process and will support the future examination of key aspects of catalysis such as processivity of an enzyme against multiply phosphorylated substrates.

My investigations of the physical properties of PP2A, Rap proteins and CTPRs have merely scratched the surface and have raised many more questions than they provided answers. It became clear that the scope of the project went well beyond what I had initially envisaged, and hence we recently established a collaboration with Ivet Bahar (University of Pittsburgh) and Reuven Gordon (University of Victoria, Canada). Bahar has extensive experience in modelling protein dynamics and indeed, has herself established the ENM approach. Gordon recently developed a technique, extraordinary acoustic Raman (EAR) spectroscopy, that can directly measure the vibrational resonances of single molecules in solution [370, 371]. This technique is currently the only one that can probe vibrational motions in solutions. By using a sub-set of proteins as well as computational analyses of the whole solenoid repeat protein class, we intend to explore how repeat type, shape, interface packing and stability translate into mechanics and dynamics, and subsequently function.

9.4 Future application of consensus-designed repeat proteins in mechanotransduction research

Genetically encoded molecular tension sensors have allowed the first quantification of pN forces within cells and even in worms and flies (Figure 9.2a). However, these sensors cover only part of the force range that is thought to be relevant for the regulation of many intracellular processes [362, 372]. From my experiments on CTPRs I know that chemical stability translates approximately into mechanical stability. By varying repeat type and consensus sequence one could therefore produce FRET sensors that probe a broader range of forces than currently possible (Figure 9.2a). Such proteins would fulfil all four criteria of a force sensor (unfolding at low pN force, 100% reversible, unfolding is independent of loading rate, and refolding occurs without hysteresis). Furthermore, their force response is independent of the number of repeats which would allow for customization of the sensor by selecting the length that gives the best signal for each specific application.

Additionally, I can envision a novel strategy using split-GFP [373]. Previous studies from our group have shown that the interface between adjacent repeats can be destabilised in a tune-able way by insertion of a long loop [114]. Thus, an interface that is weakened by the insertion of the 11th β-strand of the GFP β-barrel is expected to break under force before the other repeats unfold, enabling the GFP-11 peptide to complement GFP1-10



(a)

(b)

Figure 9.2: Repeat protein-based force sensors. (a) FRET-based repeat protein force sensors. Since repeat proteins are mechanically weaker than GFP, they will unfold first thereby decreased the FRET between the fluorophores. (b) Split-GFP repeat-protein scaffolds fluorescent force sensors. Upon an application of force the repeat array splits at the loop position (or unfolds fully) which enables the complementation of fluorescent GFP protein.

and form the intact fluorescent protein (Figure 9.2b). When the force is released, the loop will re-form and the fluorescence will be lost. If necessary, the interface at the site of the loop introduction can be further weakened using directed mutagenesis to further weaken the interface. In contrast to FRET-sensors, this system will provide an on-off switch-like response and is furthermore much smaller in size.

The unique mechanical behaviour makes repeat proteins ideal for force-sensing applications and should provide a more precise readout of force than is possible with current sensors. Furthermore, their design is straightforward, making it possible to customise such sensors to create bespoke tools for researchers wishing to explore distinct force regimes both at as well as beyond current limitation.

Appendix A

Protein sequences

A.1 Amino acid sequences

H_6-PR65

MHHHHHHMAAADGDDSLYPIAVLIDELRNEDVQLRLNSIKKLSTIALALGVERTR
SELLPFLTDTIYDEDEVLLALAEQLGTFTTLVGGPEYVHCLLPPLESLATVEETVV
RDKAVESLRAISHEHSPSDLEAHFVPLVKRLAGGDWFTSRTSACGLFSVCYPRVS
SAVKAELRQYFRNLCSDDTPMVRRAAASKLGEFAKVLELDNVKSEIIPMFSNLAS
DEQDSVRLLAVEACVNIAQLLPQEDLEALVMPTLRQAAEDKSWRVRYMVADKF
TELQKAVGPEITKTDLVPAFQNLMKDCEAEVRAAASHKVKEFCENLSADCRENV
IMSQILPCIKELVSDANQHVKSALASVIMGLSPILGKDNTIEHLLPLFLAQLKDECP
EVRLNIISNLDCVNEVIGIRQLSQSLLPAIVELAEDAKWRVRLAIIEYMPLLAGQLG
VEFFDEKLNSLCMAWLVDHVYAIREAATSNLKKLVEKFGKEWAHATIIPKVLAM
SGDPNYLHRMTTLFCINVLSEVCGQDITTKHMLPTVLRMAGDPVANVRFNVAKS
LQKIGPILDNSTLQSEVKPILEKLTQDQDVDVKYFAQEALTVLSLA

GST-PR65-H_6

MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPN
LPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIA
YSKDFETLKVDFLSKLPEMLKMFEDRLCHKTYLNGDHVTHPDFMLYDALDVVL
YMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPP
KSDLVPRGSMAAADGDDSLYPIAVLIDELRNEDVQLRLNSIKKLSTIALALGVERT
RSELLPFLTDTIYDEDEVLLALAEQLGTFTTLVGGPEYVHCLLPPLESLATVEETV
VRDKAVESLRAISHEHSPSDLEAHFVPLVKRLAGGDWFTSRTSACGLFSVCYPRV
SSAVKAELRQYFRNLCSDDTPMVRRAAASKLGEFAKVLELDNVKSEIIPMFSNLA
SDEQDSVRLLAVEACVNIAQLLPQEDLEALVMPTLRQAAEDKSWRVRYMVADK
FTELQKAVGPEITKTDLVPAFQNLMKDCEAEVRAAASHKVKEFCENLSADCREN
VIMSQILPCIKELVSDANQHVKSALASVIMGLSPILGKDNTIEHLLPLFLAQLKDEC

PEVRLNIISNLDCVNEVIGIRQLSQSLLPAIVELAEDAKWRVRLAIIEYMPLLAGQL
GVEFFDEKLNSLCMAWLVDHVYAIREAATSNLKKLVEKFGKEWAHATIIPKVLA
MSGDPNYLHRMTTLFCINVLSEVCGQDITTKHMLPTVLRMAGDPVANVRFNVA
KSLQKIGPILDNSTLQSEVKPILEKLTQDQDVDVKYFAQEALTVLSLAHHHHHH

yPR65y

MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPN
LPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIA
YSKDFETLKVDFLSKLPEMLKMFEDRLCHKTYLNGDHVTHPDFMLYDALDVVL
YMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPP
KSDLVPRGSDSLEFIASKLAMAAADGDDSLYPIAVLIDELRNEDVQLRLNSIKKLST
IALALGVERTRSELLPFLTDTIYDEDEVLLALAEQLGTFTTLVGGPEYVHCLLPPL
ESLATVEETVVRDKAVESLRAISHEHSPSDLEAHFVPLVKRLAGGDWFTSRTSAC
GLFSVCYPRVSSAVKAELRQYFRNLCSDDTPMVRRAAASKLGEFAKVLELDNVK
SEIIPMFSNLASDEQDSVRLLAVEACVNIAQLLPQEDLEALVMPTLRQAAEDKSW
RVRYMVADKFTELQKAVGPEITKTDLVPAFQNLMKDCEAEVRAAASHKVKEFC
ENLSADCRENVIMSQILPCIKELVSDANQHVKSALASVIMGLSPILGKDNTIEHLLP
LFLAQLKDECPEVRLNIISNLDCVNEVIGIRQLSQSLLPAIVELAEDAKWRVRLAIIE
YMPLLAGQLGVEFFDEKLNSLCMAWLVDHVYAIREAATSNLKKLVEKFGKEWA
HATIIPKVLAMSGDPNYLHRMTTLFCINVLSEVCGQDITTKHMLPTVLRMAGDPV
ANVRFNVAKSLQKIGPILDNSTLQSEVKPILEKLTQDQDVDVKYFAQEALTVLSL
ADSLEFIASKLAHHHHHH

yPR65-GSy

MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPN
LPYYIDGDVKLTQSMAIIRYIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIA
YSKDFETLKVDFLSKLPEMLKMFEDRLCHKTYLNGDHVTHPDFMLYDALDVVL
YMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHPP
KSDLVPRGSDSLEFIASKLAMAAADGDDSLYPIAVLIDELRNEDVQLRLNSIKKLST
IALALGVERTRSELLPFLTDTIYDEDEVLLALAEQLGTFTTLVGGPEYVHCLLPPL
ESLATVEETVVRDKAVESLRAISHEHSPSDLEAHFVPLVKRLAGGDWFTSRTSAC
GLFSVCYPRVSSAVKAELRQYFRNLCSDDTPMVRRAAASKLGEFAKVLELDNVK
SEIIPMFSNLASDEQDSVRLLAVEACVNIAQLLPQEDLEALVMPTLRQAAEDKSW
RVRYMVADKFTELQKAVGPEITKTDLVPAFQNLMKDCEAEVRAAASHKVKEFC
ENLSADCRENVIMSQILPCIKELVSDANQHVKSALASVIMGLSPILGKDNTIEHLLP
LFLAQLKDECPEVRLNIISNLDCVNEVIGIRQLSQSLLPAIVELAEDAKWRVRLAIIE
YMPLLAGQLGVEFFDEKLNSLCMAWLVDHVYAIREAATSNLKKLVEKFGKEWA
HATIIPKVLAMSGDPNYLHRMTTLFCINVLSEVCGQDITTKHMLPTVLRMAGDPV

ANVRFNVAKSLQKIGPILDNSTLQSEVKPILEKLTQDQDVDVKYFAQEALTVLSL
ASGSGSGSDSLEFIASKLAHHHHHH

H$_6$-c1WT

MHHHHHHMAAADGDDSLYPIAVLIDELRNEDVQLRLNSIKKLSTIALALGVERTR
SELLPFLTDTIYDEDEVLLALAEQLGTFTTLVGGPEYVHCLLPPLESLATVEETVV
RDKAVESLRAISHEHSPSDLEAHFVPLVKRLAGGDWFTSRTSACGLFSVCYPRVS
SAVKAELRQYFRNLCSDDTPMVRRAAASKLGEFAKVLELDNVKSEIIPMFSNLAS
DEQDSVRLLAVEACVNIAQLLPQEDLEALVMPTLRQAAEDKSWRVRYMVADKF
TELQKAVGPEITKTDLVPAFQNLMKDCEAEVRAAASHKVKEFCENLSADCRENV
IMSQILPCIKELVSDANQHVKSALASVIMGLSPILGKDNTIEHLLPLFLAQLKDECP
EVRLNIISNLDCVNEVIGIRQLSQSLLPAIVELAEDAKWRVRLAIIEYMPLLAGQLG
VEFFDEKLNSLCMAWLVDHVYAIREAATSNLKKLVEKFGKEWAHATIIPKVLAM
SGDPNYLHRMTTLFCINVLSEVCGQDITTKHMLPTVLRMAGDPVANVRFNVAKS
LQKIGPILDNSTLQSEVKPILEKLTQDQDVDVKYFAQEALTVLSLAKDVLKLVEA
RPMIHELLTEGRRSKTNKAKTLATWATKELRKLKNQA

H$_6$-c1int

MHHHHHHMAAADGDDSLYPIAVLIDELRNEDVQLRLNSIKKLSTIALALGVERTR
SELLPFLTDTIYDEDEVLLALAEQLGTFTTLVGGPEYVHCLLPPLESLATVEETVV
RDKAVESLRAISHEHSPSDLEAHFVPLVKRLAGGDWFTSRTSACGLFSVCYPRVS
SAVKAELRQYFRNLCSDDTPMVRRAAASKLGEFAKVLELDNVKSEIIPMFSNLAS
DEQDSVRLLAVEACVNIAQLLPQEDLEALVMPTLRQAAEDKSWRVRYMVADKF
TELQKAVGPEITKTDLVPAFQNLMKDCEAEVRAAASHKVKEFCENLSADCRENV
IMSQILPCIKELVSDANQHVKSALASVIMGLSPILGKDNTIEHLLPLFLAQLKDECP
EVRLNIISNLDCVNEVIGIRQLSQSLLPAIVELAEDAKWRVRLAIIEYMPLLAGQLG
VEFFDEKLNSLCMAWLVDHVYAIREAATSNLKKLVEKFGKEWAHATIIPKVLAM
SGDPNYLHRMTTLFCINVLSEVCGQDITTKHMLPTVLRMAGDPVANVRFNVAKS
LQKIGPILDNSTLQSEVLPILLKLLQDQDVDVKYFAAEALTVLSLAKDVLKLVEAR
PMIHELLTEGRRSKTNKAKTLATWATKELRKLKNQA

H$_6$-c2WT

MHHHHHHMAAADGDDSLYPIAVLIDELRNEDVQLRLNSIKKLSTIALALGVERTR
SELLPFLTDTIYDEDEVLLALAEQLGTFTTLVGGPEYVHCLLPPLESLATVEETVV
RDKAVESLRAISHEHSPSDLEAHFVPLVKRLAGGDWFTSRTSACGLFSVCYPRVS
SAVKAELRQYFRNLCSDDTPMVRRAAASKLGEFAKVLELDNVKSEIIPMFSNLAS
DEQDSVRLLAVEACVNIAQLLPQEDLEALVMPTLRQAAEDKSWRVRYMVADKF
TELQKAVGPEITKTDLVPAFQNLMKDCEAEVRAAASHKVKEFCENLSADCRENV

IMSQILPCIKELVSDANQHVKSALASVIMGLSPILGKDNTIEHLLPLFLAQLKDECP
EVRLNIISNLDCVNEVIGIRQLSQSLLPAIVELAEDAKWRVRLAIIEYMPLLAGQLG
VEFFDEKLNSLCMAWLVDHVYAIREAATSNLKKLVEKFGKEWAHATIIPKVLAM
SGDPNYLHRMTTLFCINVLSEVCGQDITTKHMLPTVLRMAGDPVANVRFNVAKS
LQKIGPILDNSTLQSEVKPILEKLTQDQDVDVKYFAQEALTVLSLADVMLVQPRV
EFILSFIDHIAGDEDHTDGVVACAAGLIGDLCTAFGKDVLKLVEARPMIHELLTEG
RRSKTNKAKTLATWATKELRKLKNQA

$\mathbf{H_6}$-c2int

MHHHHHHMAAADGDDSLYPIAVLIDELRNEDVQLRLNSIKKLSTIALALGVERTR
SELLPFLTDTIYDEDEVLLALAEQLGTFTTLVGGPEYVHCLLPPLESLATVEETVV
RDKAVESLRAISHEHSPSDLEAHFVPLVKRLAGGDWFTSRTSACGLFSVCYPRVS
SAVKAELRQYFRNLCSDDTPMVRRAAASKLGEFAKVLELDNVKSEIIPMFSNLAS
DEQDSVRLLAVEACVNIAQLLPQEDLEALVMPTLRQAAEDKSWRVRYMVADKF
TELQKAVGPEITKTDLVPAFQNLMKDCEAEVRAAASHKVKEFCENLSADCRENV
IMSQILPCIKELVSDANQHVKSALASVIMGLSPILGKDNTIEHLLPLFLAQLKDECP
EVRLNIISNLDCVNEVIGIRQLSQSLLPAIVELAEDAKWRVRLAIIEYMPLLAGQLG
VEFFDEKLNSLCMAWLVDHVYAIREAATSNLKKLVEKFGKEWAHATIIPKVLAM
SGDPNYLHRMTTLFCINVLSEVCGQDITTKHMLPTVLRMAGDPVANVRFNVAKS
LQKIGPILDNSTLQSEVLPILLKLLQDQDVDVKYFAAEALTVLSLADVMLVQPRVE
FILSFIDHIAGDEDHTDGVVACAAGLIGDLCTAFGKDVLKLVEARPMIHELLTEGR
RSKTNKAKTLATWATKELRKLKNQA

y(CTPR_QKRV)$_{\mathbf{N}}$y

MRGSHHHHHHGLVPRGSDSLEFIASKLA(AEALNNLGNVYREQGDYQKAIEYYQK
ALELDPRS)$_N$DSLEFIASKLA

y(CTPRa_QK)$_{\mathbf{N}}$y

MRGSHHHHHHNNNNNNNNNNENLYFQGDSLEFIASKLAGS(AEAWYNLGNAYYK
QGDYQKAIEYYQKALELDPRS)$_N$SKLDSLEFIASKLA

A.2 Sequences used for geometric calcuations

3dw8

```
#repeat start, repeat end
8 42
```

```
45 76
83 117
121 156
159 195
198 234
238 272
276 312
319 355
358 393
397 433
436 473
475 511
514 550
553 586
```

1qgk

```
#repeat start, repeat end
1 32
33 68
85 120
129 161
170 205
212 248
253 290
314 360
364 397
402 439
449 481
500 539
544 590
600 636
644 681
686 724
732 776
787 829
833 873
```

CTPR20

```
#repeat_start, repeat_end
```

1 30

35 64

69 98

103 132

137 166

171 200

205 234

239 268

273 302

307 336

341 370

375 404

409 438

443 472

477 506

511 540

545 574

579 608

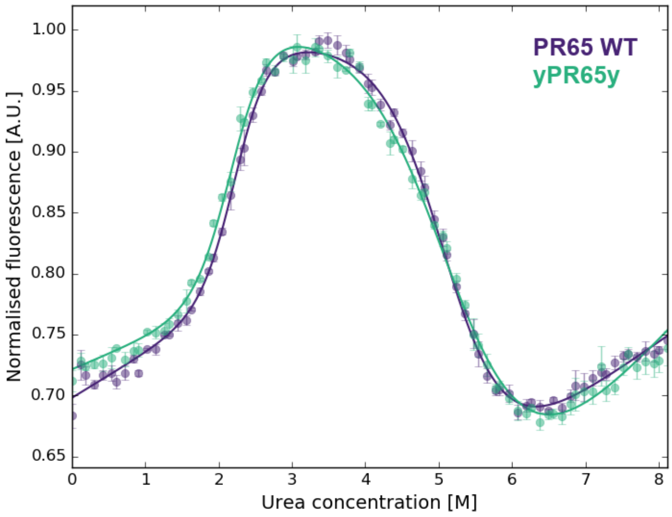
613 642

647 676

Appendix B

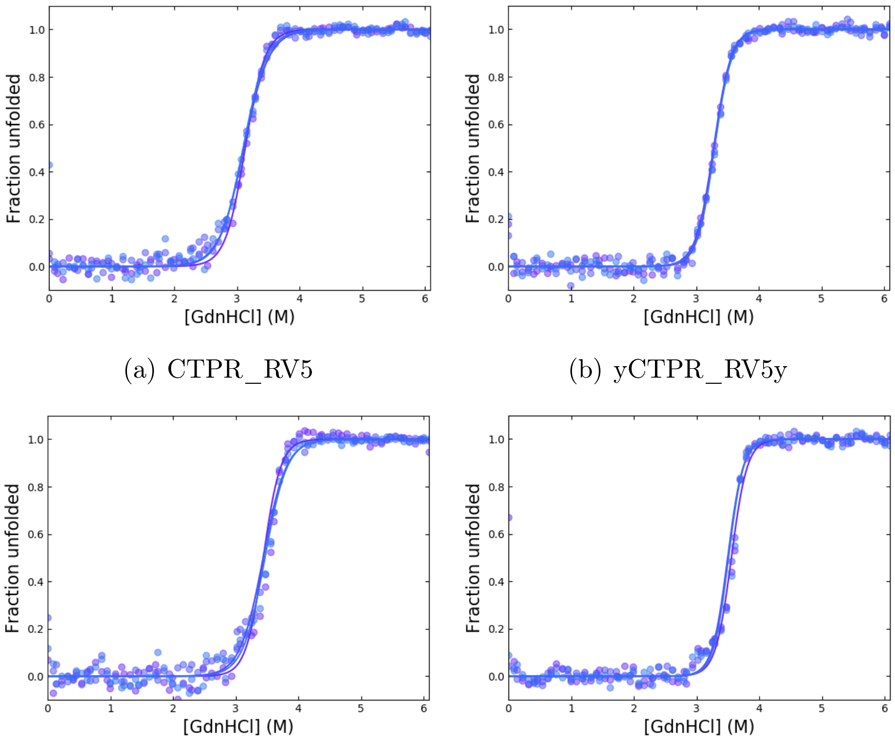
Post-submission data

B.1 Effect of ybbR-tags on protein stability



Protein	**D$_{50\%-1}$ [M]**	***m*$_1$ [kcal mol^{-1} M^{-1}]**	**D$_{50\%-2}$ [M]**	***m*$_1$ [kcal mol^{-1} M^{-1}]**
PR65	2.24 ± 0.06	2.51 ± 0.09	5.13 ± 0.03	1.40 ± 0.02
yPR65y	2.14 ± 0.06	2.48 ± 0.06	5.20 ± 0.05	1.16 ± 0.06

Figure B.1: Equilibrium denaturations of PR65 WT (purple) and terminally ybbR-tagged PR65 (green). The data represent the average and standard error of the mean of three technical replicates for each experiment that were fitted with Equation 6.11. Tryptophans were excitated at 295 nm and fluorescence was monitored at 340 nm.



(a) CTPR_RV5

(b) yCTPR_RV5y

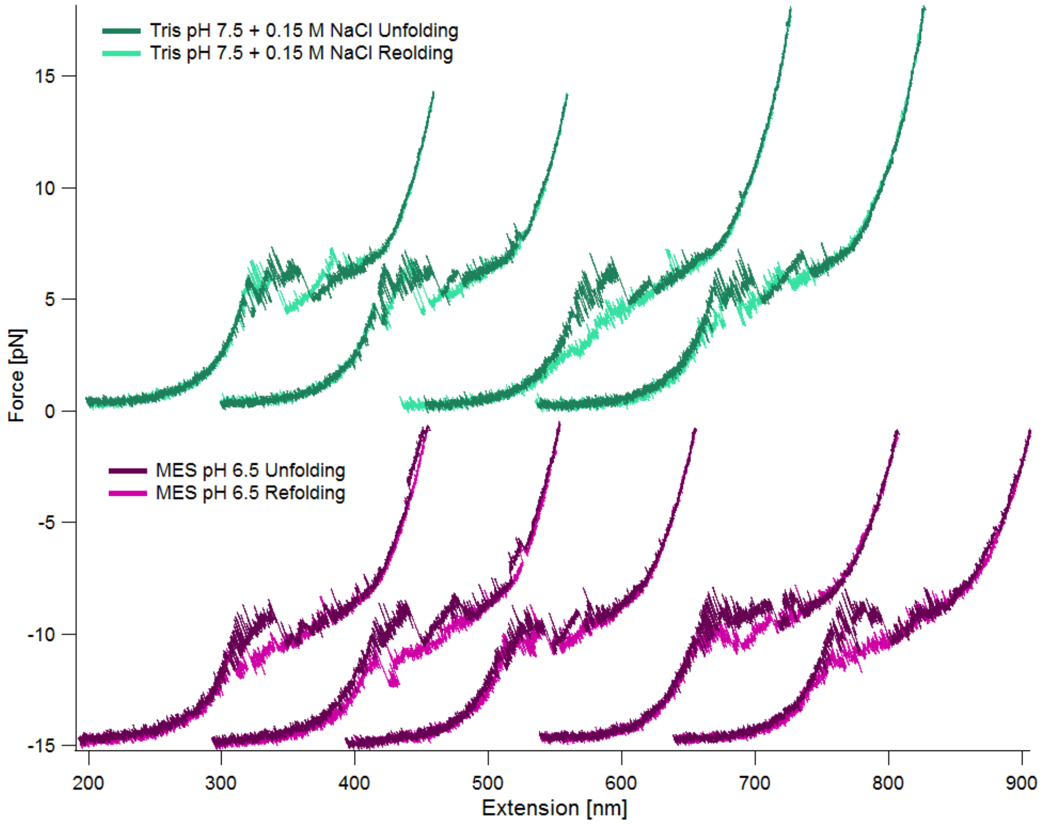
(c) CTPR_RV10

(d) yCTPR_RV10y

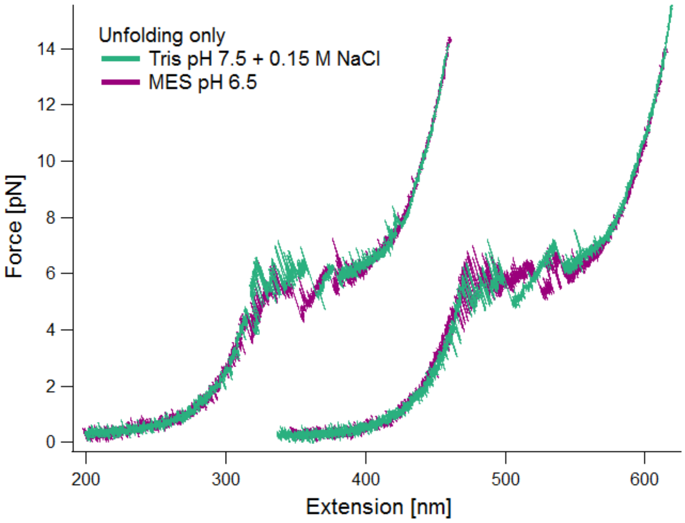
Protein	**D$_{50\%}$ [M]**	***m*-value [kcal mol^{-1} M^{-1}]**	**$\Delta G_{U-N}^{H_2O}$ [kcal mol^{-1}]**
CTPR_RV5	3.114 ± 0.008	3.3 ± 0.2	10.3 ± 0.6
yCTPR_RV5y	3.287 ± 0.003	4.26 ± 0.07	14.0 ± 0.2
CTPR_RV10	3.449 ± 0.007	3.6 ± 0.3	12.5 ± 0.9
yCTPR_RV10y	3.52 ± 0.01	4.82 ± 0.06	17.0 ± 0.2

Figure B.2: Equilibrium denaturation of untagged and ybbR-tagged CTPR_RV proteins with 5 and 10 repeats. Since the denaturation profile of larger CTPR arrays deviates from a simple two-state unfolding transition, such a fit can estimate the m-value wrongly by compensating for the deviation by adjusting baseline parameters. Therefore, the baselines were fitted first using a straight-line equation to obtain values for α_N, α_U, β_N and β_U. These were then fixed to obtain the apparent $D_{50\%}$ and m-value using Equation 6.3. The values listed represent the mean and standard error of the three technical replicates shown in each figure.

B.2 Force-induced unfolding of PR65 - MES vs Tris



(a)



(b)

Figure B.3: Representative force extension cycles of azPR65az in either 50 mM Tris pH 7.5, 150 mM NaCl or 50 mM MES pH 6.5. Both buffer conditions contained 1mM DTT. (a) Sequential force extension cycles of 2 molecules each, taken at 10 nm s^{-1}. (b) Two traces taken in each buffer condition overlaid.

Appendix C

Publications

PHILOSOPHICAL TRANSACTIONS B

rstb.royalsocietypublishing.org

Opinion piece

Check for updates

Cite this article: Perez-Riba A, Synakewicz M, Itzhaki LS. 2018 Folding cooperativity and allosteric function in the tandem-repeat protein class. *Phil. Trans. R. Soc. B* **373**: 20170188.
http://dx.doi.org/10.1098/rstb.2017.0188

Accepted: 17 January 2018

One contribution of 17 to a discussion meeting issue 'Allostery and molecular machines'.

Subject Areas:
biophysics, molecular biology, structural biology, computational biology

Keywords:
repeat proteins, protein cooperativity, allostery, molecular switches, Rap proteins, elastic network models

Author for correspondence:
Laura S. Itzhaki
e-mail: laura.itzhaki@gmail.com

Folding cooperativity and allosteric function in the tandem-repeat protein class

Albert Perez-Riba[†], Marie Synakewicz[†] and Laura S. Itzhaki

Department of Pharmacology, University of Cambridge, Tennis Court Road, Cambridge CB2 1PD, UK

AP-R, 0000-0002-4659-0320; MS, 0000-0003-0256-2712; LSI, 0000-0001-6504-2576

The term allostery was originally developed to describe structural changes in one binding site induced by the interaction of a partner molecule with a distant binding site, and it has been studied in depth in the field of enzymology. Here, we discuss the concept of action at a distance in relation to the folding and function of the solenoid class of tandem-repeat proteins such as tetratricopeptide repeats (TPRs) and ankyrin repeats. Distantly located repeats fold cooperatively, even though only nearest-neighbour interactions exist in these proteins. A number of repeat-protein scaffolds have been reported to display allosteric effects, transferred through the repeat array, that enable them to direct the activity of the multi-subunit enzymes within which they reside. We also highlight a recently identified group of tandem-repeat proteins, the RRPNN subclass of TPRs, recent crystal structures of which indicate that they function as allosteric switches to modulate multiple bacterial quorum-sensing mechanisms. We believe that the folding cooperativity of tandem-repeat proteins and the biophysical mechanisms that transform them into allosteric switches are intimately intertwined. This opinion piece aims to combine our understanding of the two areas and develop ideas on their common underlying principles.

This article is part of a discussion meeting issue 'Allostery and molecular machines'.

1. Tandem-repeat protein: folding for function?

Tandem-repeat domains are one of the most common protein architectures. The frequency of these arrays is probably a result of replication slippage and recombination events on the DNA [1,2]. These mechanisms are considered sources of hypermutability and have given rise to a high polymorphism rate compared with the background rate of point mutations [2–4]. Tandem-repeat proteins have been grouped into different classes according to the size (number of amino acids) of the individual repeats [1,2,5]. In this work, we will focus on the solenoid class comprising repeats of approximately 12–40 amino acids. Individual repeats are not independently stable and a minimum of two or three repeats is required for a stable unit.

The simplest solenoid proteins contain repeats of two secondary structure elements: α/α, α/β or β/β. More complex repeats have three or four secondary elements [5,6]. In all cases, the secondary structure elements and their relative arrangement give rise to a variety of tertiary structures whose geometries can readily be described using the three angles between the repeat planes: curvature, twist and lateral bending [7,8]. The 'solenoid' term originally referred to a coil wound into a tightly packed helix. The repeats pack to form superhelices that differ greatly in their geometries, dependent on the structural class: some fold into planar, horseshoe-like structures, others form spring-like helices

[†] These authors contributed equally to the study.

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.4041455.

THE ROYAL SOCIETY
PUBLISHING

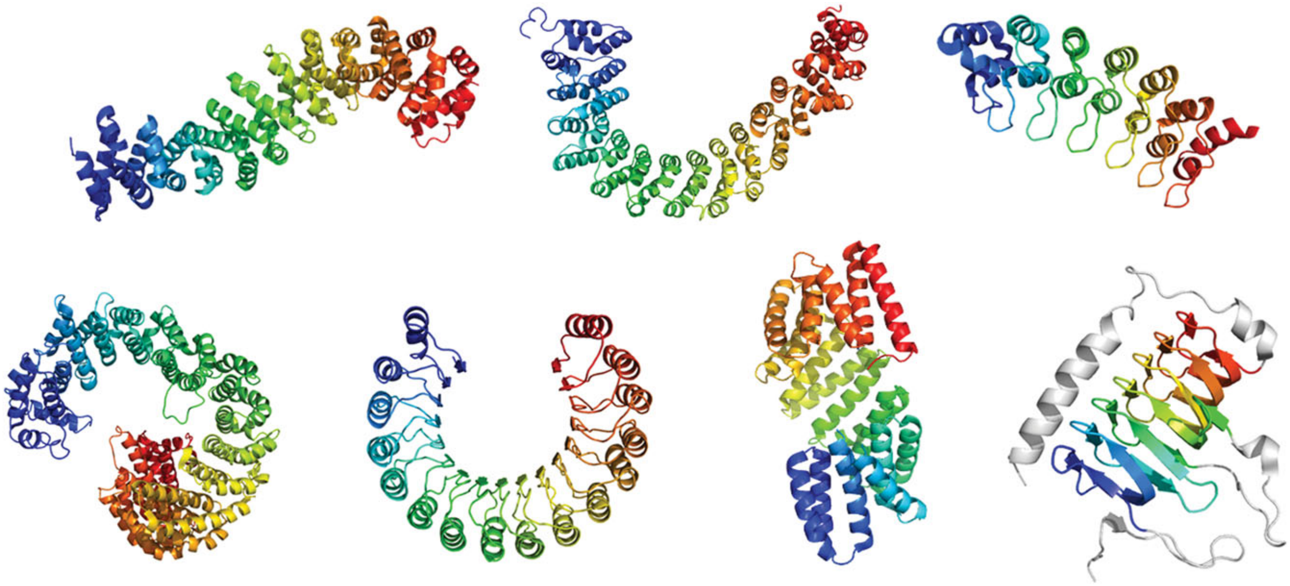


Figure 1. Secondary structure representations of solenoid tandem-repeat proteins. From top left to bottom right (PDB identifiers in parenthesis): ARM-repeat protein β-catenin (2Z6H), HEAT-repeat protein PR65 (1B3U), Ankyrin-repeat protein gankyrin (1UOH), HEAT-repeat protein Importin-β (3ND2), leucine-rich repeat (LRR) protein Ribonuclease Inhibitor (1BNH), TPR protein RapI (4I1A), β-helical repeat protein carbonic anhydrase (1QRE).

and others are very linear (figure 1). All share a common feature in that their architectures create elongated interfaces for molecular recognition, mostly to other proteins but also for some subclasses to DNA and RNA [9–13]. The α-solenoids, the focus of this opinion piece, comprise repeats of a hairpin of antiparallel α-helices [2]. Armadillo repeats [14], HEAT repeats [15,16] and the tetratricopeptide repeats (TPR) [17] are the most common members of this class.

2. Cooperativity in the folding of tandem-repeat proteins: relationship to function?

Tandem-repeat protein structures are stabilized exclusively by local interactions either within a repeat or between adjacent repeats. By contrast, the stability of globular proteins originates from the high cooperativity between sequence-distant interactions and the burial of a large hydrophobic surface area. Nevertheless, small repeat proteins (approx. 100 amino acids) such as p16, myotrophin and the Notch ankyrin domain show two-state unfolding at equilibrium, like globular proteins of similar sizes [18–20]. Protein engineering analysis of p16, myotrophin and gankyrin mapped out their kinetic folding and unfolding pathways and revealed polarized transition states in which structure was localized to a subset of repeats at one end of the protein [21–25], whereas in Notch the central repeats were structured in the transition state [26]. It was shown that the order in which the repeats fold is governed by their relative stabilities, with the most stable repeats folding first, and consequently, the folding pathways can be redirected relatively straightforwardly by manipulating the stability distribution across the repeat array [23,25–27]. It follows also that under any given set of conditions there may be flux through multiple alternative pathways [23], as originally predicted by Wolynes and co-workers [28]. Moreover, the cooperativity of the folding process (both at equilibrium and under kinetic conditions) can also be readily tuned using appropriate mutations [29,30].

High cooperativity is not always desirable, as non-cooperative folding may be important for biological function of some repeat proteins. A striking example is the interaction between the transcription factor NFκB and the 6-ankyrin-repeat protein IκBα, which regulates NFκB by sequestering it in the cytoplasm. The two C-terminal ankyrin repeats of IκBα are intrinsically unfolded and only fold upon binding to NFκB. Not only was the folding process shown to be critical for high-affinity binding, but the large difference in intracellular stability of un-complexed IκBα compared with the NFκB-bound form was also shown to play an essential role in transcriptional regulation. Un-complexed IκBα with its unfolded repeats 5–6 is degraded in a ubiquitin-independent manner with a very short half-life, whereas NFκB-bound IκBα is stable in the cytoplasm and requires triggered ubiquitin-mediated proteolysis for its degradation and the subsequent release of NFκB [31].

Another example of the relationship between folding cooperativity and function is the 15-HEAT repeat protein PR65. PR65 is the scaffold subunit of the heterotrimeric enzyme protein phosphatase 2A (PP2A). The catalytic subunit and the substrate-bound regulatory subunit bind at opposite ends of the PR65 repeat array, and it has been proposed that rather than being a rigid scaffold for these two subunits, PR65 functions as an elastic connector that coordinates cycles of catalysis of multiply phosphorylated substrates [32]. Our analysis suggests that the non-cooperative unfolding of the HEAT repeats, arising from the very heterogeneous distribution of stabilities across the repeat array, might also facilitate PR65's connector function [33].

3. The nearest-neighbour description of repeat protein folding

The simple topology of the repeat-protein architecture has enabled the use of a one-dimensional Ising model description to define the energetic values of each repeat under the assumption of all repeats being coupled. The Ising model was originally developed to describe interactions of atomic dipole spins in ferromagnetic materials. In such a material, the atomic dipoles can adopt one of two states (spin +1 or −1). However, their states are coupled to their nearest neighbours through an exchange interaction, a potential that favours parallel alignment between states [34]. Owing to this coupling, flipping of one spin can result in cascades, or so-called 'spin-waves' [35]. In early work on the ankyrin-repeat domain of Notch, Barrick and co-workers [18] recognized that the protein's stability follows a simple additive rule. Regan and co-workers [36] applied the Ising model to so-called 'consensus-designed repeat proteins', comprising identical repeats containing the most conserved residues in a protein family. A single value of intrinsic stability of the repeats (ΔG_i) and of interfacial stability between repeats (ΔG_{ij}) was shown to be sufficient to describe the folding of a series of consensus-designed repeat proteins of increasing size. These energetic values are additive, and the Gibbs free energy of unfolding of a repeat protein comprising identical repeats thus follows the equation:

$$\Delta G_{D-N} = n\Delta G_i + (n-1)\Delta G_{ij} = -RT\ln\kappa^n\tau^{(n-1)}$$

where n is the number of repeats, κ the intrinsic stability and τ the interfacial stability [36–39]. Several families of repeat proteins have been found to follow the Ising model both at equilibrium and under kinetic conditions [36,40–42]. With this description, one can see that the origin of cooperativity of repeat-protein folding lies in the mismatch between the intrinsic and interfacial repeat stabilities.

Folding cooperativity of repeat proteins breaks down above approximately 100–150 amino acids, similar to the cooperativity limit of globular proteins [43,44]. Moreover, for repeat proteins both large and small, an array with (i) fewer intrinsically stable repeats, (ii) high interfacial stability relative to intrinsic stability and (iii) a more homogeneous distribution of stabilities across the array length will tend to unfold more cooperatively. Indeed, when such conditions are met, the folding of even giant repeat proteins of 300 or more amino acids has, strikingly, been shown to approximate a two-state behaviour [29,45,46].

Hydrogen-exchange experiments have shown that the internal repeats of consensus-designed repeat proteins are more protected from exchange than the terminal repeats [47–49]. That is, even when the repeats are identical in sequence they are not all equally stable. The probabilistic nature of the Ising model and the higher stabilities for a greater number of repeats can be explained in simple terms. For a repeat to unfold, it requires its neighbouring repeats to unfold also. Thus, the terminal repeats of the array are the most likely to unfold, as they have only one neighbour. In natural repeat proteins, however, the simple additivity of internal and interfacial energies becomes more difficult to dissect because repeats have different sequences and therefore different stabilities. For example, analysis of the unfolding pathway of PR65 showed that this giant repeat protein has weak central repeats, which unfold before the N- and C-terminal subdomains [33,50].

We recently showed that extending the length of a single inter-repeat loop in a consensus-designed TPR protein (CTPR) can have a large effect on stability depending on the number of repeats in the array (A.P.-R., L.S.I. *et al*., under revision). This is in contrast with the small effects observed upon insertion of a long loop into consensus-designed ankyrin-repeat proteins and β-helical repeat proteins [51,52]. Although further investigation is required, there does appear to be a trend of a greater energetic cost of loop extensions when the repeat type has a smaller mismatch between intrinsic and interfacial stability as is the case for TPRs. In other words, short inter-repeat loops are required for a repeat protein that has weak inter-repeat interfaces, possibly because of low enthalpic and high entropic contribution to the overall stability.

In summary, the rules governing the cooperativity of repeat protein structures are now well understood. More and more, we are starting to see that repeat proteins are not static rods and that the natural functions of many repeat proteins require highly dynamic conformational properties. In this opinion piece, we question the relationship between the folding cooperativity and the function of repeat proteins and whether cooperativity plays a role in controlling the transmission of information across the repeat array.

4. The RRNPP family of molecular switches

In recent years, a new family of bacterial regulators has been gaining recognition. Known as the RRNPP family, the name of these cytosolic peptide-sensing regulators refers to the founding members of the family, Rap-Rgg-NprR-PlcR-Prgx [53]. They all have the same domain organization: an N-terminal three-helix bundle, a flexible helical linker and a C-terminal TPR capable of binding short peptides of five to eight residues. Notably, the N-terminal domain and the helical linker form a four-helix bundle that resembles a pair of TPRs. These proteins share the following mechanism: peptide binding to the C-terminal domain triggers a conformational change that propagates to the N-terminal domain. Here, we examine the Rap proteins of *Bacillus subtilis*, cytosolic aspartate phosphatases that affect downstream gene expression upon binding of the quorum-sensing peptide to their C-terminal TPR domain.

Rap phosphatases and their peptide activators were originally described in *B. subtilis* by Perego and co-workers [54–59]. There are 16 Rap homologues in *B. subtilis*. As an example, RapH acts as phosphatase of Spo0F in its Apo form (peptide unbound) and prevents downstream sporulation, whereas RapF binds and inhibits gene regulators such as ComA [53]. RapH and RapF were co-crystallized with Spo0F and ComA, respectively, and both Rap homologues showed the same overall conformation when bound to their partner molecule (figure 2) [61]. In another study, the crystal structure of Apo-RapI was compared with that of RapJ in complex with the PhrC peptide. The solenoid structure of the RapJ–PhrC complex showed a higher degree of compactness relative to Apo-RapI [60] (figure 2).

Lacking a complete set of crystal structures of the same Rap homologue in three conformational states, Parashar and co-workers [60] used homologous structures to propose a mechanism of action for signal transduction and concluded that quorum-sensing peptides inhibit Rap function via an allosteric mechanism (figure 2). The compact solenoid was described as the inactive configuration and the extended solenoid as the active one. In its active configuration, the N-terminal helix

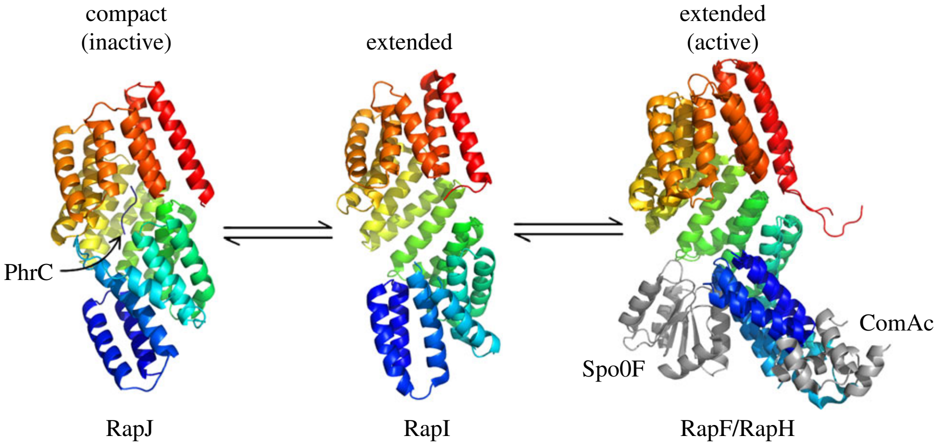


Figure 2. Structures of different Rap proteins (C-terminus in red) depicting a possible mode of action. When the TPR domain binds to a signalling peptide, it causes the Rap protein to adopt a compact, or 'closed' conformation. Upon binding an interaction partner, however, conformational changes in the TPR domain are minimal, whereas the N-terminal three-helix bundle flips by approximately 180° [60].

bundle is capable of exposing the Spo0F- or ComA-binding regions of RapF and RapH, respectively. Peptide-bound Rap proteins undergo a conformational change locking the N-terminal domain in a compact configuration in which its binding sites are inaccessible.

Rap proteins are very different from their artificial counterparts, the CTPRs. Both types of TPRs are capable of forming large cooperatively folded repeat arrays, but somehow the amino acid sequence of Rap proteins encodes the additional ability to generate a repeat array with extreme flexibility and consequent dynamic switch-like behaviour capable of transmitting the information of an input to an output across the array. The nearest-neighbour cooperativity between repeats appears to have increased its complexity. Ultimately, allostery necessarily requires a dynamic system, providing further evidence that repeat proteins are not simply rigid rods.

5. Do intrinsic dynamics of Rap proteins form the underlying basis for allostery?

Given the functional relevance of the conformational changes seen for the Rap proteins crystallographically, we have conducted an extensive analysis to investigate whether they arise from the intrinsic dynamics of each protein. To model the vibrational dynamics, we have chosen to use elastic network models (ENMs) for their strong dependence on the shape of the overall structure instead of atomistic detail (see electronic supplementary material, Methods, extensively reviewed in [62] and references therein; [63]). In an ENM, protein dynamics are decomposed into different motions with specific directions, the normal modes. The lowest three normal modes of RapI are shown in figure 3. The predominant motion is that of bending, followed by a screw-like twist of the TPR helix and more localized motions of the N- and C-terminal three-helix bundle and TPR repeats, respectively. We compared normal modes of different Rap configurations using ENMs of a structure-based sequence alignment (electronic supplementary material, figure S1). The dynamics we observe are very similar in all four proteins, and the lowest normal modes tend to involve the collective motion of approximately 40–80% of the protein (electronic supplementary material, figure S2). Owing to the high structural similarity between RapH and RapF, both proteins explore a nearly identical motional space (electronic supplementary material, figure S3). The major difference between the extended conformation, RapI, and the active conformations is the orientation of the N-terminal three-helix bundle, and hence the normal modes of all three proteins are very similar, especially when the TPR domains are modelled independently (electronic supplementary material, figure S3). By contrast, when the Rap protein adopts the compact and inactive conformation, only the motion of a very few of the lowest normal modes remains conserved to some degree, most of which are dominated by the motion of the TPR domain (electronic supplementary material, figure S3). Comparing both peptide-bound and -unbound ENMs of RapJ reveals that most changes in dynamics of the lower modes are not due to the presence of the bound peptide but simply due to the compacted conformation (electronic supplementary material, figure S4).

We further examined whether the motions observed in the ENMs can account for some of the conformational changes observed in crystallography by measuring how well the lowest five normal modes of a protein overlap with the conformational transition to another protein (figure 4). The transition between the extended and compact conformations is very well described by the normal modes of either RapI (to 0.87) or RapJ (to 0.75). Transitions between either compact or extended conformation and the active forms (RapH and RapF) are less well described by the lowest five normal modes due to the rotation of the N-terminal three-helix bundle and linker domain (figure 4). This observation is not entirely unexpected, as more localized motions are captured by normal modes of higher order. When mapping the normal modes to the conformational changes seen within the independently modelled TPR domains (small numbers in figure 4), agreements are greater than 0.7 for most.

Considering these comparisons, a mechanism emerges by which at least the TPR domain of Rap proteins on its own could potentially explore the different conformations observed. The peptide, substrate or transcription factor could then simply trap the protein energetically in a given conformation.

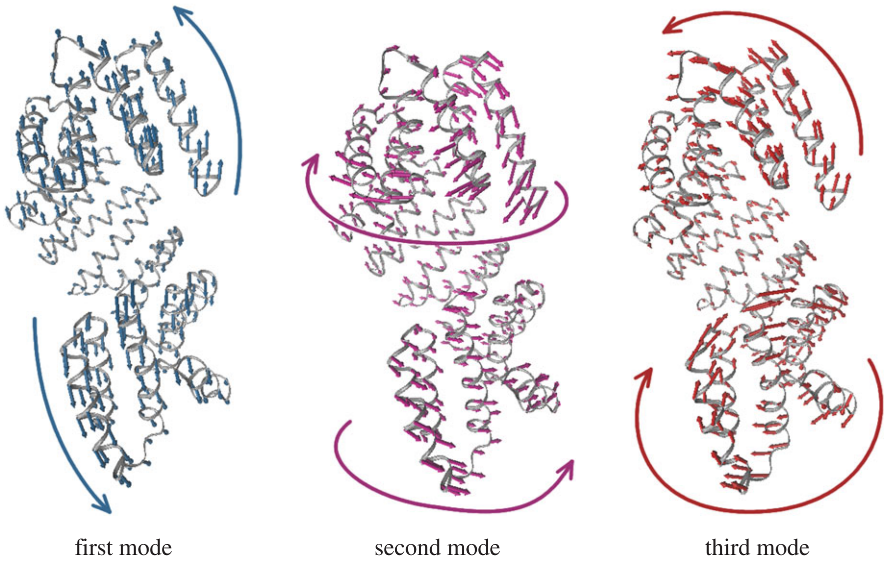


Figure 3. The three lowest normal modes of RapI. The first mode (teal) describes a bending motion that alters the distance between the N- and C-termini; the second mode (magenta) tightens the superhelix in a screw-like motion; the third mode (red) twists the N-terminal three-helix bundle and the C-terminal TPR relative to the repeat array superhelix. Models were generated using the NMWIZ plugin for VMD [64]. Movies of these modes can be found with the electronic supplementary material.

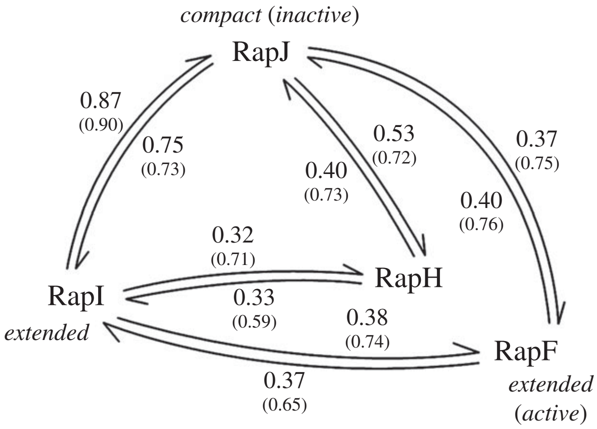


Figure 4. Quantitative comparison between the lowest five normal modes and conformational changes between different Rap proteins. The arrows represent the conformational change vector, and the values equal the corresponding cumulative overlap between the vector and the ENM of the starting structure (see electronic supplementary material, Methods). For example, the first five normal modes of RapI can account for 0.87 of the conformational change between RapI and RapJ, while the first five normal modes of RapJ only describe 0.75. Numbers in brackets correspond to the cumulative overlaps between the dynamics of truncated and independently modelled TPR domains and their respective conformational changes (see electronic supplementary material, Methods).

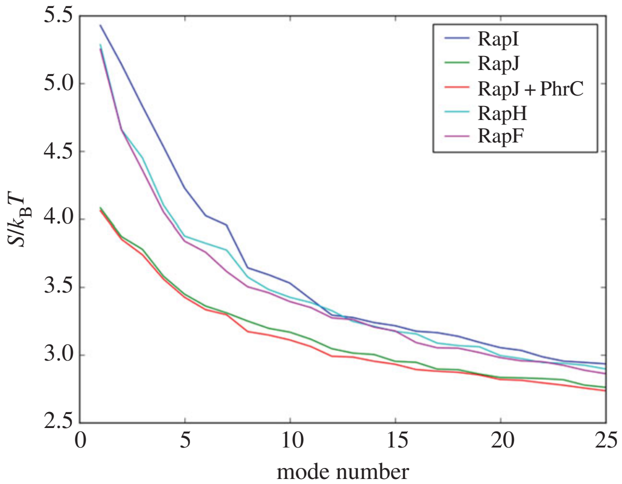


Figure 5. Entropy contributions of each normal mode to the total motion. The closed conformation was modelled both with and without the PhrC peptide. As the differences between both models are only small (electronic supplementary material, figure S5), the effect of peptide binding on the entropy of the system is negligible compared with the entropic cost of the conformational change.

In fact, when considering the entropic contributions of each normal mode (figure 5), the extended conformation is entropically the most favourable. The compact conformation of RapJ comes along with a considerable entropic cost, making it energetically unfavourable (electronic supplementary material, figure S5), for which the enthalpic contributions of multiple contacts between peptide and TPRs need to compensate [60].

Lastly, by analysing the correlation of motion between different residues (figure 6), we can obtain an insight into why the three-helix bundle in the compact conformation has very little potential for rotating to bind partner proteins. The TPRs exhibit correlated motion only with their nearest neighbours, giving rise to the distinctive pattern of squares along the diagonal [45]. The binding of the peptide marginally increases nearest-neighbour correlation at the centre (purple box in figure 6) which understandably arises from the cross-correlations of higher modes (electronic supplementary material, figures S4 and S6). Movements of the rotated N-terminal three-helix bundle, linker domain and first TPRs repeat (blue box in figure 6) are strongly correlated, suggesting that they form a subdomain relative to the rest of the TPR repeats. Some nearest-neighbour correlations are reduced in the extended conformation of RapI, whereas they are either further reduced or even reversed in the

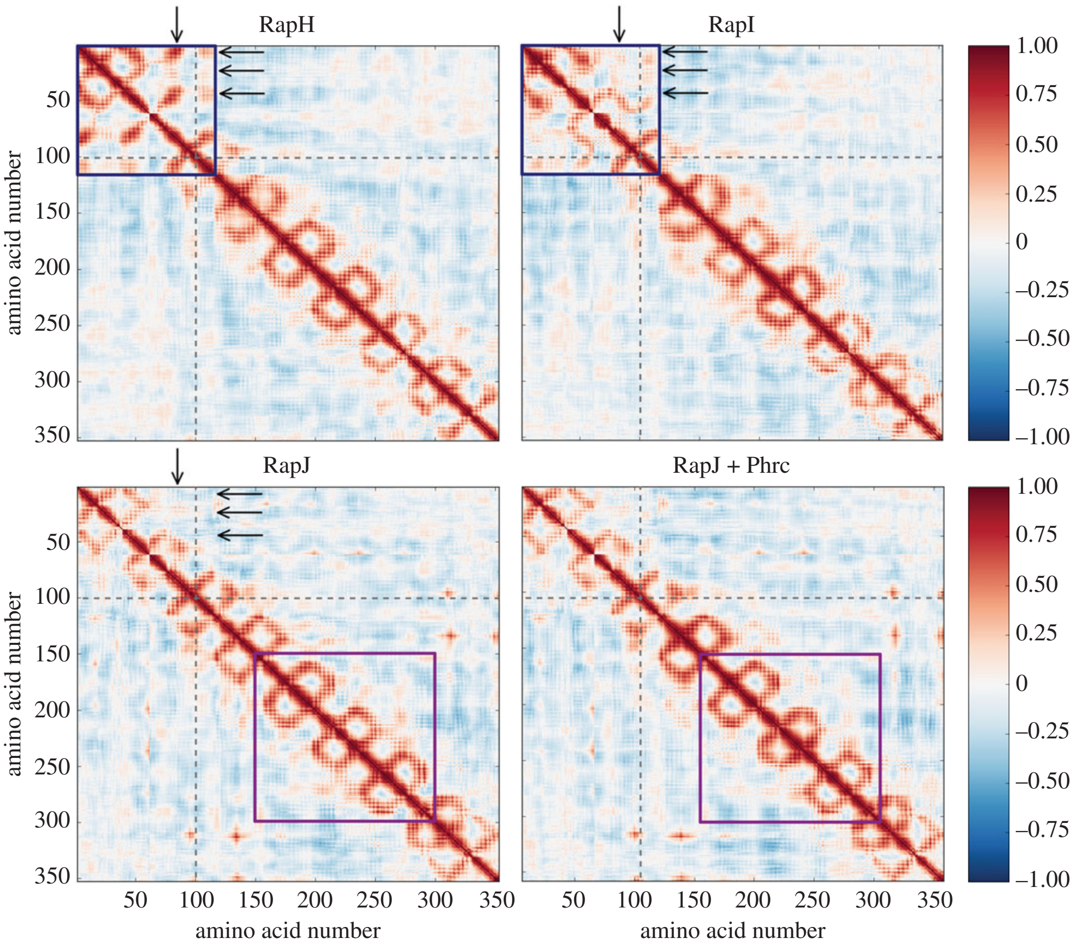


Figure 6. Representative cross-correlation maps for the partner-bound, open and peptide-bound conformations. Cross-correlation between residues is a measure of how much these residues move in the same direction, where values of 1 and −1 represent perfectly correlated and anti-correlated motions, respectively [62]. The TPR repeats exhibit correlated motions only with their nearest neighbours, giving rise to the distinctive pattern of squares along the diagonal. Movements of the rotated N-terminal three-helix bundle, linker domain and first TPR motif (blue box) are non-TPR-like, exhibiting non-nearest-neighbour correlations, suggesting that they form a subdomain relative to the rest of the TPR repeats. Some of these correlations are reduced in the open conformation, or even reversed, once a continuous TPR array is formed (arrows) and the distinction of this domain is lost. The global movement of peptide binding TPRs (purple box) and neighbouring repeats is only minimally affected in the presence of the peptide, which only causes a slight increase in the nearest-neighbour correlations. The N-terminal helix bundle and TPR repeats are divided by grey dashed lines and correlations are mirrored across the diagonal for clarity. The cross-correlation was summed over the lowest 25 modes, for correlation maps of lower modes (see electronic supplementary material, figure S6).

compact conformation (arrows in figure 6). It is notable that the presence of the peptide does not directly influence the correlated motion of the N-terminal domain, indicating that locking of the three-helix bundle to the TPRs is entirely due to the conformational change. However, the peptide could induce indirect or allosteric effects by stabilizing the TPR domain in the compact conformation. These effects could then be ‘transmitted’ through the array via the interaction potential between repeats, that is via the altered cooperativity of individual repeat interfaces due to the TPR rearrangement in the compact conformation.

In summary, the compact and extended conformations of Rap proteins have different supramolecular geometries, arising from differences in the inter-repeat packing. Consequently, they must have different values of the interfacial repeat stability. The ENMs showed that both conformations are easily accessible through the motions of the TPR domain, albeit that an extended conformation of the array may be preferred owing to the entropic cost of the compact state. Ultimately, the intrinsic flexibility of the TPR array may allow for the existence of two functionally different conformations that can be locked by their respective binding partners.

6. Relating conformational flexibility to the allosteric mechanisms of ‘banana-shaped’ repeat proteins in multi-protein enzyme complexes

When we look across the repeat protein class, the Rap proteins are not the only example where the repeat scaffold may contribute to allosteric mechanisms due to its dynamic flexibility. In quite a few systems, the repeat protein must change its conformation to bind to a variety of partners that all differ in shape and size. We are currently investigating proteins of different repeat types to examine whether their experimentally observed dynamics can also be described by ENM normal modes. As global motions are largely determined by the over-all shape of a molecule [62], one of our leading questions is whether the dynamics of two different proteins with the same tertiary shape will exhibit the same motions, independent of repeat type.

One such protein is PR65, and, as mentioned earlier, we are interested in understanding how it regulates the activity of the heterotrimeric PP2A enzyme (figure 7*a*). From crystal

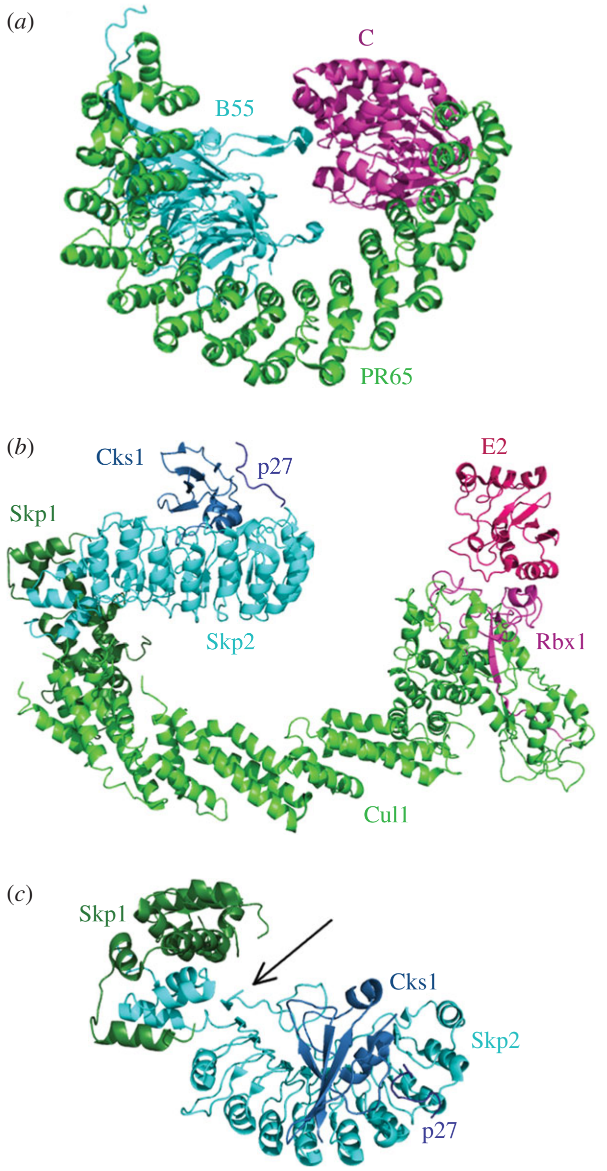


Figure 7. Repeat proteins linking function in multi-protein complexes. (*a*) PP2A (PDB 3dw8) consisting of the HEAT-repeat scaffold subunit PR65, the catalytic subunit C bound to the C-terminus of PR65, and a regulatory subunit (B55) bound to the N-terminus of PR65. (*b*) Modelled structure of the SCF$^{\mathrm{Skp2}}$ ubiquitin ligase consisting of three core subunits Skp1, cullin-repeat containing protein Cul1 and Rbx1, and substrate-recognition subunit Skp2 with accessory protein Cks1, which together recruit the substrate p27. Also shown is the E2 ubiquitin-conjugating enzyme, which is recruited to the SCF by Rbx1, together forming the catalytic entity (PDBs 2AST, 1LDK and 4Q5E [65]). Thus, in both complexes, the substrate-recognition subunit is bound to one end of the repeat protein and the catalytic subunit to the other end. (*c*) Top view of the Skp2 bound to Skp1, Cks1 and p27, highlighting the insertion of the Skp2 C-terminal tail at its N-terminus.

structures of PR65 in complex with the catalytic C-subunit and different regulatory B-subunits, it is clear that PR65 needs to be highly flexible structurally to be able to form the multitude of PP2A complexes that are present in the cell [66–68]. Biophysical analysis has shown that binding of the catalytic subunit to the C-terminal repeats of PR65 increases by an order of magnitude the affinity of the N-terminal repeats of PR65 for an inhibitor, the SV40 small t antigen [69]. However, there are no obvious direct contacts between the small t antigen and the catalytic subunit [69], suggesting a process by which PR65 functions as an allosteric transmitter of catalytic-subunit binding, though the underlying mechanism involved remains to be resolved.

A number of years ago, we demonstrated such an allosteric effect in the LRR protein Skp2, which is one of many variable substrate-binding subunits of the multi-subunit SCF (Skp1-Cullin-F-box) ubiquitin ligases (figure 7*b*) [70]. Skp2 has an F-box motif, with which it binds to the Skp1 subunit, thereby connecting it to the Cullin subunit and the rest of the SCF ligase. The C-terminal 'tail' of Skp2 is unstructured and folds back onto the concave face of the LRR domain (figure 7*c*) [65]. Binding of the accessory subunit Cks1 to the C-terminus of the LRR domain results in hydrogen–deuterium exchange protection of the N-terminal LRR repeats without any direct contacts between Cks1 and N-terminal Skp2 repeats [70]. We therefore proposed that binding of Cks1 decreases fluctuations in the C-terminal tail of Skp2, thereby stabilizing residues in the tail that form a β-sheet between the first LRR and the F-box. In the absence of this β-sheet, the linker between the LRR domain and the F-box may constitute a hinge, which could account for deprotection of the N-terminal LRRs when Cks1 is not bound. Thus, the hinge may function as a sensor of substrate binding, tightening of which could reduce the motions of Skp2 and thereby allow for efficient ubiquitination and/or this binding event could be translated allosterically through the Cullin subunit to the E2 ubiquitin-conjugating enzyme. The Cullins themselves are highly flexible repeat proteins, and their ability to change shape is thought to be crucial for orchestrating consecutive cycles of substrate ubiquitination [71]. Considering the similar architectures of the SCF and PP2A enzyme complexes, we hypothesize that there is a common underlying mechanism exploiting flexible repeat-protein scaffolds for such catalytic processes. At this point, it is not clear how exactly this scaffold flexibility arises and how it depends on the repeat types. Moreover, it remains to be seen whether repeat types with different packing interactions, interfacial energies and cooperativities will exhibit correspondingly different dynamics and macromolecular flexibility.

7. Bridging tandem-repeat cooperativity and allosteric transmission

In summary, we have explored the relationship between the stability and cooperativity of repeat arrays and the functional transmission of information along them. For example, Rap proteins are capable of transforming a high concentration of quorum-sensing peptides into a signalling response to downstream effectors [56,61,72,73]. This implies that Rap proteins display higher affinity for their binding partners than for the peptides because the bound conformation is only favoured at the high concentrations associated with quorum-sensing. The Rap proteins are an example of a system where nearest-neighbour interactions in a repeat array can cause allosteric inactivation. Our ENM results showed that the first five normal modes of the extended conformation could account for most of the conformational changes between the extended and compact form, suggesting an equilibrium between the two that favours the extended form in the absence of the peptide. Owing to different packing interactions in the extended and compact conformations, the N-terminal domain displays a varying degree of correlated movement relative to the TPR domain. This observation supports the idea of an N-terminal helix bundle reaching an energy minimum when cooperatively interacting with the compact TPR domain and thus becoming incapable of exploring partner-binding configurations.

In addition to these insights from the Rap ENMs, the examples of banana-shaped proteins discussed here suggest that repeat arrays involved in diverse cellular processes have the potential to function as allosteric modulators in multi-protein complexes and are not simply a molecular-recognition platform for multiple binding partners. In most cases, these repeat proteins are not rigid, rod-like entities, but rather they need to be flexible to function in a biological context. ENMs are a simplistic but efficient way for us to gain insights into the conformational space explored by repeat proteins. Using them, we can identify structural points of allosteric significance, such as hinges or weak points, and design experiments accordingly. Naturally, we think that repeat stability and cooperativity can be linked to distinct mechanical characteristics and therefore function as a transmission pathway for information to travel through the repeat array. Any local event, such as binding of a partner molecule or alterations of repeat packing in TPR arrays, should therefore modulate repeat stability and shape, and this change could be transmitted to nearest neighbours by way of the interaction potential, similar to the mechanism that gives rise to spin-waves in ferromagnets. Hence we suggest that context-dependent changes in cooperativity between repeats must, at least partly, be the basis for allosteric effects in tandem-repeat proteins, and, as such, any repeat protein in itself could function as a switch.

Ultimately, the question of how distantly located repeats can fold cooperatively, how Rap proteins change their super-helical structure upon binding and how information is transmitted through multi-protein complexes via a repeat protein may be different manifestations of the same physical mechanism, namely that underlying the Ising model. The two parameters of intrinsic repeat stability and the interaction potential (i.e. interfacial stability) are straightforward to quantify in consensus repeat arrays but are not easily determined in natural repeat proteins owing to the different sequences of the repeats. Nevertheless, we believe that this parametrization will still hold true but will just result in a model that is mathematically non-trivial. It is crucial to carefully dissect the relationship between repeat protein cooperativity and their ability to function as switches such that we can tune them artificially, thereby translating the peptide-sensing capability to the biotech industry. Last but not least, repeat proteins make up nearly one-third of the human proteome [74], and, given their widespread involvement in key signalling cascades, an understanding of allostery in repeat proteins is also necessary to shed light on the transmission of information in central cellular processes.

Data accessibility. All data presented in this article are based on calculations of data from public repositories, using publicly available software and the methods described in detail in the electronic supplementary material.

Authors' contributions. A.P.-R., M.S. and L.S.I. conceived the opinion piece. M.S. carried out the ENMs. A.P.-R., M.S. and L.S.I. wrote the manuscript.

Competing interests. We declare we have no competing interests.

Funding. L.S.I. acknowledges the support of a Senior Fellowship from the UK Medical Research Foundation. A.P.-R. was supported by a BBSRC Doctoral Training Programme scholarship and an Oliver Gatty Studentship. M.S. was supported by a BBSRC Doctoral Training Programme scholarship.

References

1. Kajava AV. 2001 Review: proteins with repeated sequence structural prediction and modeling. *J. Struct. Biol.* **134**, 132–144. (doi:10.1006/jsbi.2000.4328)
2. Kajava AV. 2012 Tandem repeats in proteins: from sequence to structure. *J. Struct. Biol.* **179**, 279–288. (doi:10.1016/j.jsb.2011.08.009)
3. Buard J, Vergnaud G. 1994 Complex recombination events at the hypermutable minisatellite CEB1 (D2S90). *EMBO J.* **13**, 3203–3210.
4. Ellegren H. 2000 Microsatellite mutations in the germline: implications for evolutionary inference. *Trends Genet.* **16**, 551–558. (doi:10.1016/S0168-9525(00)02139-9)
5. Kobe B, Kajava AV. 2000 When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. *Trends Biochem. Sci.* **25**, 509–515. (doi:10.1016/S0968-0004(00)01667-4)
6. Kajava AV. 2002 What curves α-solenoids? Evidence for an α-helical toroid structure of Rpn1 and Rpn2 proteins of the 26 S proteasome. *J. Biol. Chem.* **277**, 49 791–49 798. (doi:10.1074/jbc.M204982200)
7. Bublitz M, Holland C, Sabet C, Reichelt J, Cossart P, Heinz DW, Bierne H, Schubert W-D. 2008 Crystal structure and standardized geometric analysis of InlJ, a listerial virulence factor and leucine-rich repeat protein with a novel cysteine ladder. *J. Mol. Biol.* **378**, 87–96. (doi:10.1016/j.jmb.2008.01.100)
8. Forwood JK, Lange A, Zachariae U, Marfori M, Preast C, Grubmüller H, Stewart M, Corbett AH, Kobe B. 2010 Quantitative structural analysis of importin-β flexibility: paradigm for solenoid protein structures. *Structure* **18**, 1171–1183. (doi:10.1016/j.str.2010.06.015)
9. Filipovska A, Rackham O. 2012 Modular recognition of nucleic acids by PUF, TALE and PPR proteins. *Mol. Biosyst.* **8**, 699–708. (doi:10.1039/c2mb05392f)
10. Liu S, Melonek J, Boykin LM, Small I, Howell KA. 2013 PPR-SMRs: ancient proteins with enigmatic functions. *RNA Biol.* **10**, 1501–1510. (doi:10.4161/rna.26172)
11. Yin P *et al.* 2013 Structural basis for the modular recognition of single-stranded RNA by PPR proteins. *Nature* **504**, 168–171. (doi:10.1038/nature12651)
12. Barkan A, Small I. 2014 Pentatricopeptide repeat proteins in plants. *Annu. Rev. Plant Biol.* **65**, 415–442. (doi:10.1146/annurev-arplant-050213-040159)
13. Ke J *et al.* 2013 Structural basis for RNA recognition by a dimeric PPR-protein complex. *Nat. Struct. Mol. Biol.* **20**, 1377–1382. (doi:10.1038/nsmb.2710)
14. Tewari R, Bailes E, Bunting KA, Coates JC. 2010 Armadillo-repeat protein functions: questions for little creatures. *Trends Cell Biol.* **20**, 470–481. (doi:10.1016/j.tcb.2010.05.003)
15. Andrade MA, Petosa C, O'Donoghue SI, Müller CW, Bork P. 2001 Comparison of ARM and HEAT protein repeats. *J. Mol. Biol.* **309**, 1–18. (doi:10.1006/jmbi.2001.4624)
16. Andrade MA, Perez-Iratxeta C, Ponting CP. 2001 Protein repeats: structures, functions, and evolution. *J. Struct. Biol.* **134**, 117–131. (doi:10.1006/jsbi.2001.4392)
17. Allan RK, Ratajczak T. 2011 Versatile TPR domains accommodate different modes of target protein recognition and function. *Cell Stress Chaperones* **16**, 353–367. (doi:10.1007/s12192-010-0248-0)
18. Zweifel ME, Leahy DJ, Hughson FM, Barrick D. 2003 Structure and stability of the ankyrin domain of the *Drosophila* Notch receptor. *Protein Sci.* **12**, 2622–2632. (doi:10.1110/ps.03279003)
19. Zhang B, Peng Z. 1996 Defective folding of mutant p16(INK4) proteins encoded by tumor-derived alleles. *J. Biol. Chem.* **271**, 28734. (doi:10.1074/jbc.271.46.28734)
20. Lowe AR, Itzhaki LS. 2007 Biophysical characterisation of the small ankyrin repeat protein myotrophin. *J. Mol. Biol.* **365**, 1245–1255. (doi:10.1016/j.jmb.2006.10.060)

21. DeVries I, Ferreiro DU, Sánchez IE, Komives EA. 2011 Folding kinetics of the cooperatively folded subdomain of the IκBα ankyrin repeat domain. *J. Mol. Biol.* **408**, 163 – 176. (doi:10.1016/j.jmb.2011.02.021)

22. Hutton RD, Wilkinson J, Faccin M, Sivertsson EM, Pelizzola A, Lowe AR, Bruscolini P, Itzhaki LS. 2015 Mapping the topography of a protein energy landscape. *J. Am. Chem. Soc.* **137**, 14 610 – 14 625. (doi:10.1021/jacs.5b07370)

23. Lowe AR, Itzhaki LS. 2007 Rational redesign of the folding pathway of a modular protein. *Proc. Natl Acad. Sci. USA* **104**, 2679 – 2684. (doi:10.1073/pnas.0604653104)

24. Tang KS, Fersht AR, Itzhaki LS. 2003 Sequential unfolding of ankyrin repeats in tumor suppressor p16. *Structure* **11**, 67 – 73. (doi:10.1016/S0969-2126(02)00929-2)

25. Werbeck ND, Rowling PJE, Chellamuthu VR, Itzhaki LS. 2008 Shifting transition states in the unfolding of a large ankyrin repeat protein. *Proc. Natl Acad. Sci. USA* **105**, 9982 – 9987. (doi:10.1073/pnas.0705300105)

26. Tripp KW, Barrick D. 2008 Rerouting the folding pathway of the Notch Ankyrin domain by reshaping the energy landscape. *J. Am. Chem. Soc.* **130**, 5681 – 5688. (doi:10.1021/ja0763201)

27. Sivertsson EM, Itzhaki LS. 2014 Protein folding: when ribosomes pick the structure. *Nat. Chem.* **6**, 378 – 379. (doi:10.1038/nchem.1926)

28. Ferreiro DU, Cho SS, Komives EA, Wolynes PG. 2005 The energy landscape of modular repeat proteins: topology determines folding mechanism in the ankyrin family. *J. Mol. Biol.* **354**, 679 – 692. (doi:10.1016/j.jmb.2005.09.078)

29. Werbeck ND, Itzhaki LS. 2007 Probing a moving target with a plastic unfolding intermediate of an ankyrin-repeat protein. *Proc. Natl Acad. Sci. USA* **104**, 7863 – 7868. (doi:10.1073/pnas.0610315104)

30. Street TO, Bradley CM, Barrick D. 2007 Predicting coupling limits from an experimentally determined energy landscape. *Proc. Natl Acad. Sci. USA* **104**, 4907 – 4912. (doi:10.1073/pnas.0608756104)

31. Truhlar SME, Mathes E, Cervantes CF, Ghosh G, Komives EA. 2008 Pre-folding IκBα Alters control of NF-κB signaling. *J. Mol. Biol.* **380**, 67 – 82. (doi:10.1016/j.jmb.2008.02.053)

32. Grinthal A, Adamovic I, Weiner B, Karplus M, Kleckner N. 2010 PR65, the HEAT-repeat scaffold of phosphatase PP2A, is an elastic connector that links force and catalysis. *Proc. Natl Acad. Sci. USA* **107**, 2467 – 2472. (doi:10.1073/pnas.0914073107)

33. Tsytlonok M, Craig PO, Sivertsson E, Serquera D, Perrett S, Best RB, Wolynes PG, Itzhaki LS. 2013 Complex energy landscape of a giant repeat protein. *Structure* **21**, 1954 – 1965. (doi:10.1016/j.str.2013.08.028)

34. Bowley R, Sánchez M. 1999 *Introductory statistical mechanics*. Oxford, UK: Clarendon Press.

35. Kittel C. 2005 *Introduction to solid state physics*, 8th edn. New York, NY: John Wiley & Sons, Inc.

36. Kajander T, Cortajarena AL, Main ERG, Mochrie SGJ, Regan L. 2005 A new folding paradigm for repeat proteins. *J. Am. Chem. Soc.* **127**, 10 188 – 10 190. (doi:10.1021/ja0524494)

37. Aksel T, Barrick D. 2009 *Biothermodynamics, part A*. Amsterdam, The Netherlands: Elsevier.

38. Aksel T, Majumdar A, Barrick D. 2011 The contribution of entropy, enthalpy, and hydrophobic desolvation to cooperativity in repeat-protein folding. *Structure* **19**, 349 – 360. (doi:10.1016/j.str.2010.12.018)

39. Millership C, Phillips JJ, Main ERG. 2016 Ising model reprogramming of a repeat protein's equilibrium unfolding pathway. *J. Mol. Biol.* **428**, 1804 – 1817. (doi:10.1016/j.jmb.2016.02.022)

40. Cortajarena AL, Mochrie SGJ, Regan L. 2011 Modulating repeat protein stability: the effect of individual helix stability on the collective behavior of the ensemble. *Protein Sci.* **20**, 1042 – 1047. (doi:10.1002/pro.638)

41. Aksel T, Barrick D. 2014 Direct observation of parallel folding pathways revealed using a symmetric repeat protein system. *Biophys. J.* **107**, 220 – 232. (doi:10.1016/j.bpj.2014.04.058)

42. Kloss E, Barrick D. 2009 C-terminal deletion of leucine-rich repeats from YopM reveals a heterogeneous distribution of stability in a cooperatively folded protein. *Protein Sci.* **18**, 1948 – 1960. (doi:10.1002/pro.205)

43. Löw C, Weininger U, Zeeb M, Zhang W, Laue ED, Schmid FX, Balbach J. 2007 Folding mechanism of an ankyrin repeat protein: scaffold and active site formation of human CDK inhibitor p19INK4d. *J. Mol. Biol.* **373**, 219 – 231. (doi:10.1016/j.jmb.2007.07.063)

44. Low C, Weininger U, Neumann P, Klepsch M, Lilie H, Stubbs MT, Balbach J. 2008 Structural insights into an equilibrium folding intermediate of an archaeal ankyrin repeat protein. *Proc. Natl Acad. Sci. USA* **105**, 3779 – 3784. (doi:10.1073/pnas.0710657105)

45. Kloss E, Courtemanche N, Barrick D. 2008 Repeat-protein folding: new insights into origins of cooperativity, stability, and topology. *Arch. Biochem. Biophys.* **469**, 83 – 99. (doi:10.1016/j.abb.2007.08.034)

46. Courtemanche N, Barrick D. 2008 The leucine-rich repeat domain of Internalin B folds along a polarized N-terminal pathway. *Structure* **16**, 705 – 714. (doi:10.1016/j.str.2008.02.015)

47. Wetzel SK, Ewald C, Settanni G, Jurt S, Plückthun A, Zerbe O. 2010 Residue-resolved stability of full-consensus ankyrin repeat proteins probed by NMR. *J. Mol. Biol.* **402**, 241 – 258. (doi:10.1016/j.jmb.2010.07.031)

48. Main ERG, Stott K, Jackson SE, Regan L. 2005 Local and long-range stability in tandemly arrayed tetratricopeptide repeats. *Proc. Natl Acad. Sci. USA* **102**, 5721 – 5726. (doi:10.1073/pnas.0404530102)

49. Cortajarena AL, Mochrie SGJ, Regan L. 2008 Mapping the energy landscape of repeat proteins using NMR-detected hydrogen exchange. *J. Mol. Biol.* **379**, 617 – 626. (doi:10.1016/j.jmb.2008.02.046)

50. Tsytlonok M, Itzhaki LS. 2012 The how's and why's of protein folding intermediates. *Arch. Biochem. Biophys.* **531**, 14 – 23. (doi:10.1016/j.abb.2012.10.006)

51. Schilling J, Schöppe J, Plückthun A. 2014 From DARPins to LoopDARPins: novel LoopDARPin design allows the selection of low picomolar binders in a single round of ribosome display. *J. Mol. Biol.* **426**, 691 – 721. (doi:10.1016/j.jmb.2013.10.026)

52. MacDonald JT, Kabasakal BV, Godding D, Kraatz S, Henderson L, Barber J, Freemont PS, Murray JW. 2016 Synthetic beta-solenoid proteins with the fragment-free computational design of a beta-hairpin extension. *Proc. Natl Acad. Sci. USA* **113**, 10 346 – 10 351. (doi:10.1073/pnas.1525308113)

53. Do H, Kumaraswami M. 2016 Structural mechanisms of peptide recognition and allosteric modulation of gene regulation by the RRNPP family of quorum-sensing regulators. *J. Mol. Biol.* **428**, 2793 – 2804. (doi:10.1016/j.jmb.2016.05.026)

54. Perego M, Hoch JA. 1996 Cell-cell communication regulates the effects of protein aspartate phosphatases on the phosphorelay controlling development in *Bacillus subtilis*. *Proc. Natl Acad. Sci. USA* **93**, 1549 – 1553. (doi:10.1073/pnas.93.4.1549)

55. Tzeng Y-L, Feher VA, Cavanagh J, Perego M, Hoch JA. 1998 Characterization of interactions between a two-component response regulator, Spo0F, and its phosphatase, RapB. *Biochemistry* **37**, 16 538 – 16 545. (doi:10.1021/bi981340o)

56. Perego M. 2001 A new family of aspartyl phosphate phosphatases targeting the sporulation transcription factor Spo0A of *Bacillus subtilis*. *Mol. Microbiol.* **42**, 133 – 143. (doi:10.1046/j.1365-2958.2001.02611.x)

57. Core LJ, Ishikawa S, Perego M. 2001 A free terminal carboxylate group is required for PhrA pentapeptide inhibition of RapA phosphatase. *Peptides* **22**, 1549 – 1553. (doi:10.1016/S0196-9781(01)00491-0)

58. Bongiorni C, Stoessel R, Perego M. 2007 Negative regulation of *Bacillus anthracis* sporulation by the Spo0E family of phosphatases. *J. Bacteriol.* **189**, 2637 – 2645. (doi:10.1128/JB.01798-06)

59. Diaz AR, Core LJ, Jiang M, Morelli M, Chiang CH, Szurmant H, Perego M. 2012 *Bacillus subtilis* RapA phosphatase domain interaction with its substrate, phosphorylated Spo0F, and its inhibitor, the PhrA peptide. *J. Bacteriol.* **194**, 1378 – 1388. (doi:10.1128/JB.06747-11)

60. Parashar V, Jeffrey PD, Neiditch MB. 2013 Conformational change-induced repeat domain expansion regulates rap phosphatase quorum-sensing signal receptors. *PLoS Biol.* **11**, e1001512. (doi:10.1371/journal.pbio.1001512)

61. Baker MD, Neiditch MB. 2011 Structural basis of response regulator inhibition by a bacterial anti-activator protein. *PLoS Biol.* **9**, e1001226. (doi:10.1371/journal.pbio.1001226)

62. Bahar I, Lezon TR, Yang L-W, Eyal E. 2010 Global dynamics of proteins: bridging between structure and function. *Annu. Rev. Biophys.* **39**, 23 – 42. (doi:10.1146/annurev.biophys.093008.131258)

63. Rodgers TL, Burnell D, Townsend PD, Pohl E, Cann MJ, Wilson MR, McLeish TCB. 2013 ΔΔPT: a

comprehensive toolbox for the analysis of protein motion. *BMC Bioinformatics* **14**, 1–9. (doi:10.1186/1471-2105-14-183)

64. Bakan A, Meireles LM, Bahar I. 2011 ProDy: protein dynamics inferred from theory and experiments. *Bioinformatics* **27**, 1575–1577. (doi:10.1093/bioinformatics/btr168)

65. Hao B, Zheng N, Schulman BA, Wu G, Miller JJ, Pagano M, Pavletich NP. 2005 Structural basis of the Cks1-dependent recognition of p27Kip1 by the SCFSkp2 ubiquitin ligase. *Mol. Cell* **20**, 9–19. (doi:10.1016/j.molcel.2005.09.003)

66. Cho US, Xu W. 2007 Crystal structure of a protein phosphatase 2A heheterotrimer holoenzyme. *Nature* **445**, 53–57. (doi:10.1038/nature05351)

67. Xu Y, Chen Y, Zhang P, Jeffrey PD, Shi Y. 2008 Structure of a protein phosphatase 2A holoenzyme: insights into B55-mediated tau dephosphorylation. *Mol. Cell* **31**, 873–885. (doi:10.1016/j.molcel.2008.08.006)

68. Wlodarchak N, Guo F, Satyshur KA, Jiang L, Jeffrey PD, Sun T, Stanevich V, Mumby MC, Xing Y. 2013 Structure of the Ca^{2+}-dependent PP2A heterotrimer and insights into Cdc6 dephosphorylation. *Cell Res.* **23**, 931–946. (doi:10.1038/cr.2013.77)

69. Chen Y, Xu Y, Bao Q, Xing Y, Li Z, Lin Z, Stock JB, Jeffrey PD, Shi Y. 2007 Structural and biochemical insights into the regulation of protein phosphatase 2A by small t antigen of SV40. *Nat. Struct. Mol. Biol.* **14**, 527–534. (doi:10.1038/nsmb1254)

70. Yao Z-P, Zhou M, Kelly SE, Seeliger MA, Robinson CV, Itzhaki LS. 2006 Activation of ubiquitin ligase SCF(Skp2) by Cks1: insights from hydrogen exchange mass spectrometry. *J. Mol. Biol.* **363**, 673–686. (doi:10.1016/j.jmb.2006.08.032)

71. Liu J, Nussinov R. 2011 Flexible cullins in cullin-RING E3 ligases allosterically regulate ubiquitination. *J. Biol. Chem.* **286**, 40 934–40 942. (doi:10.1074/jbc.M111.277236)

72. Parashar V, Mirouze N, Dubnau DA, Neiditch MB. 2011 Structural basis of response regulator dephosphorylation by Rap phosphatases. *PLoS Biol.* **9**, e1000589. (doi:10.1371/journal.pbio.1000589)

73. Mirouze N, Parashar V, Baker MD, Dubnau DA, Neiditch MB. 2011 An atypical Phr peptide regulates the developmental switch protein RapH. *J. Bacteriol.* **193**, 6197–6206. (doi:10.1128/JB.05860-11)

74. Pellegrini M, Marcotte EM, Yeates TO. 1999 A fast algorithm for genome-wide analysis of proteins with repeated sequences. *Proteins Struct. Funct. Bioinforma.* **35**, 440–446. (doi:10.1002/(SICI)1097-0134(19990601)35:4<440::AID-PROT7>3.0.CO;2-Y)

Bibliography

[1] A. Perez-Riba, M. Synakewicz, and L. S. Itzhaki. Folding cooperativity and allosteric function in the tandem-repeat protein class. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 373(1749), 2018.

[2] A. V. Kajava. Tandem repeats in proteins: From sequence to structure. *Journal of Structural Biology*, 179(3):279 – 288, 2012.

[3] M. Pellegrini, E. M. Marcotte, and T. O. Yeates. A fast algorithm for genome-wide analysis of proteins with repeated sequences. *Proteins: Structure, Function, and Bioinformatics*, 35(4):440–446, 1999.

[4] Y. Javadi and L. S. Itzhaki. Tandem-repeat proteins: regularity plus modularity equals design-ability. *Current Opinion in Structural Biology*, 23(4):622–631, 2013.

[5] B. Kobe and A. V. Kajava. When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. *Trends in Biochemical Sciences*, 25 (10):509–515, 2000.

[6] J. K. Forwood, A. Lange, U. Zachariae, M. Marfori, C. Preast, H. Grubmüller, M. Stewart, A. H. Corbett, and B. Kobe. Quantitative structural analysis of importin-β flexibility: Paradigm for solenoid protein structures. *Structure*, 18(9): 1171–1183, 2010.

[7] T. J. Brunette, F. Parmeggiani, P. S. Huang, G. Bhabha, D. C. Ekiert, S. E. Tsutakawa, G. L. Hura, J. A. Tainer, and D. Baker. Exploring the repeat protein universe through computational protein design. *Nature*, 528(7583):580–584, 2015.

[8] U. S. Cho and W. Xu. Crystal structure of a protein phosphatase 2A heheterotrimer holoenzyme. *Nature*, 445:53–57, 2007.

[9] H. Gao, X. Wu, J. Chai, and Z. Han. Crystal structure of a TALE protein reveals an extended N-terminal DNA binding region. *Cell Research*, 22:1716, 2012.

[10] S. Krzywda, A. M. Brzozowski, H. Higashitsuji, J. Fujita, R. Welchman, S. Dawson, R. J. Mayer, and A. J. Wilkinson. The Crystal Structure of Gankyrin, an Oncoprotein Found in Complexes with Cyclin-dependent Kinase 4, a 19 S Proteasomal ATPase Regulator, and the Tumor Suppressors Rb and p53. *Journal of Biological Chemistry*, 279(2):1541–1545, 2004.

[11] V. Parashar, P. D. Jeffrey, and M. B. Neiditch. Conformational change-induced repeat domain expansion regulates rap phosphatase quorum-sensing signal receptors. *PLOS Biology*, 11(3):1–15, 2013.

[12] B. Kobe and J. Deisenhofer. Mechanism of ribonuclease inhibition by ribonuclease inhibitor protein based on the crystal structure of its complex with ribonuclease A. *Journal of Molecular Biology*, 264(5):1028 – 1043, 1996.

[13] Y. Xing, K.-I. Takemaru, J. Liu, J. D. Berndt, J. J. Zheng, R. T. Moon, and W. Xu. Crystal structure of a full-length β-catenin. *Structure*, 16(3):478–487, 2008.

[14] E. K. Leinala, P. L. Davies, and Z. Jia. Crystal structure of β-helical antifreeze protein points to a general ice binding model. *Structure*, 10(5):619–627, 2002.

[15] B. Kobe, T. Gleichmann, J. Horne, I. G. Jennings, P. D. Scotney, and T. Teh. Turn up the HEAT. *Structure*, 7(5):R91–R97, 1999.

[16] M. R. Groves, N. Hanlon, P. Turowski, B. A. Hemmings, and D. Barford. The structure of the protein phosphatase 2A PR65/A subunit reveals the conformation of its 15 tandemly repeated HEAT motifs. *Cell*, 96(1):99–110, 1999.

[17] M. A. Andrade, C. Petosa, S. I. O'Donoghue, C. W. Müller, and P. Bork. Comparison of ARM and HEAT protein repeats. *Journal of Molecular Biology*, 309(1): 1–18, 2001.

[18] L. D. D'Andrea and L. Regan. TPR proteins: the versatile helix. *Trends in Biochemical Sciences*, 28(12):655 – 662, 2003.

[19] E. R. Main, Y. Xiong, M. J. Cocco, L. D'Andrea, and L. Regan. Design of stable α-helical arrays from an idealized TPR motif. *Structure*, 11(5):497–508, 2003.

[20] H. Do and M. Kumaraswami. Structural mechanisms of peptide recognition and allosteric modulation of gene regulation by the rrnpp family of quorum-sensing regulators. *Journal of Molecular Biology*, 428(14):2793 – 2804, 2016.

[21] J. Marold, J. Kavran, G. Bowman, and D. Barrick. A naturally occurring repeat protein with high internal sequence identity defines a new class of tpr-like proteins. *Structure*, 23(11):2055 – 2065, 2015.

[22] K. Geiger-Schuller and D. Barrick. Broken TALEs: Transcription Activator-like Effectors Populate Partly Folded States. *Biophysical Journal*, 111(11):2395–2403, 2016.

[23] S. Kay, S. Hahn, E. Marois, G. Hause, and U. Bonas. A bacterial effector acts as a plant transcription factor and induces a cell size regulator. *Science*, 318(5850): 648–651, 2007.

[24] H. Binz, M. T. Stumpp, P. Forrer, P. Amstutz, and A. Plückthun. Designing repeat proteins: Well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. *Journal of Molecular Biology*, 332(2):489 – 503,

2003.

[25] P. Bork. Hundreds of ankyrin-like repeats in functionally diverse proteins: mobile modules that cross phyla horizontally? *Proteins*, 17(4):363–374, 1993.

[26] P. Michaely, D. R. Tomchick, M. Machius, and R. G. Anderson. Crystal structure of a 12 ANK repeat stack from human ankyrinR. *EMBO J.*, 21(23):6387–6396, 2002.

[27] A. Kajava. Structural diversity of leucine-rich repeat proteins11edited by f. cohen. *Journal of Molecular Biology*, 277(3):519–527, 1998.

[28] A. V. Kajava and A. C. Steven. β-rolls, β-helices, and other β-solenoid proteins. In *Fibrous Proteins: Amyloids, Prions and Beta Proteins*, volume 73 of *Advances in Protein Chemistry*, pages 55 – 96. Academic Press, 2006.

[29] M. Karplus. Behind the folding funnel diagram. *Nature Chemical Biology*, 7:401, 2011.

[30] J. Kubelka, J. Hofrichter, and W. A. Eaton. The protein folding 'speed limit'. *Current Opinion in Structural Biology*, 14:76–88, 2004.

[31] T. S. Kang and R. M. Kini. Structural determinants of protein folding. *Cellular and Molecular Life Sciences*, 66:2341–2361, 2009.

[32] E. R. G. Main, K. Stott, S. E. Jackson, and L. Regan. Local and long-range stability in tandemly arrayed tetratricopeptide repeats. *Proceedings of the National Academy of Sciences*, 102(16):5721–5726, 2005.

[33] J. Kubelka, T. K. Chiu, D. R. Davies, W. A. Eaton, and J. Hofrichter. Submicrosecond protein folding. *Journal of Molecular Biology*, 359(3):546–553, 2006.

[34] P. Wolynes, J. Onuchic, and D. Thirumalai. Navigating the folding routes. *Science*, 267(5204):1619–1620, 1995.

[35] K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl. The protein folding problem. *Annual Review of Biophysics*, 37(1):289–316, 2008.

[36] A. R. Fersht. Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. *Proceedings of the National Academy of Sciences*, 92(24):10869–10873, 1995.

[37] P. Li, F. Y. Oliva, A. N. Naganathan, and V. Muñoz. Dynamics of one-state downhill protein folding. *Proceedings of the National Academy of Sciences*, 106(1):103–108, 2009.

[38] M. Tsytlonok and L. S. Itzhaki. The how's and why's of protein folding intermediates. *Archives of Biochemistry and Biophysics*, 531(1-2):14–23, 2013.

[39] I. E. Sanchez and T. Kiefhaber. Non-linear rate-equilibrium free energy relationships and Hammond behavior in protein folding. *Biophysical Chemistry*, 100(1-3):397–407, 2003.

[40] C. F. Wright, K. Lindorff-Larsen, L. G. Randles, and J. Clarke. Parallel protein-unfolding pathways revealed and mapped. *Nature Structural Biology*, 10(8):658–662, 2003.

[41] K. A. Dill and J. L. MacCallum. The protein-folding problem, 50 years on. *Science*, 338:1042–1046, 2012.

[42] H. Li, C. Tang, and N. S. Wingreen. Nature of driving force for protein folding: A result from analyzing the statistical potential. *Physical Review Letters*, 79:765–768, 1997.

[43] J. N. Onuchic and P. G. Wolynes. Theory of protein folding. *Current Opinion in Structural Biology*, 14:70–75, 2004.

[44] K. A. Dill. Dominant forces in protein folding. *Biochemistry*, 29(31):7133–7155, 1990.

[45] J.-H. Cho, S. Sato, J.-C. Horng, B. Anil, and D. P. Raleigh. Electrostatic interactions in the denatured state ensemble: Their effect upon protein folding and protein stability. *Archives of Biochemistry and Biophysics*, 469:20–28, 2008.

[46] A. E. Mirsky and L. Pauling. On the Structure of Native, Denatured, and Coagulated Proteins. *Proceedings of the National Academy of Sciences*, 22(7):439–447, 1936.

[47] J. S. Yang, W. W. Chen, J. Skolnick, and E. I. Shakhnovich. All-atom ab initio folding of a diverse set of proteins. *Structure*, 15:53–63, 2006.

[48] J. Chen and W. E. Stites. Packing is a key selection factor in the evolution of protein hydrophobic cores. *Biochemistry*, 50:15280 – 15289, 2001.

[49] A. R. Fersht, A. Matouschek, and L. Serrano. The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *Journal of Molecular Biology*, 224(3):771–782, 1992.

[50] J.-H. Cho and D. P. Raleigh. Electrostatic interactions in the denatured state and in the transition state for protein folding: Effects of denatured state interactions on the analysis of transition state structure. *Journal of Molecular Biology*, 359: 1437–1446, 2006.

[51] G. Haran. How, when and why proteins collapse: the relation to folding. *Current Opinion in Structural Biology*, 22:14–20, 2012.

[52] J. Clarke and A. R. Fersht. Engineered disulfide bonds as probes of the folding pathway of barnase: increasing the stability of proteins against the rate of denaturation. *Biochemistry*, 32(16):4322–4329, 1993.

[53] H. Nishi, A. Shaytan, and A. R. Panchenko. Physicochemical mechanisms of protein regulation by phosphorylation. *Frontiers in Genetics*, 5:270, 2014.

[54] A. Bah, R. M. Vernon, Z. Siddiqui, M. Krzeminski, R. Muhandiram, C. Zhao, N. Sonenberg, L. E. Kay, and J. D. Forman-Kay. Folding of an intrinsically disordered protein by phosphorylation as a regulatory switch. *Nature*, 519:106, 2014.

[55] O. O. Sogbein, D. A. Simmons, and L. Konermann. Effects of ph on the kinetic reaction mechanism of myoglobin unfolding studied by time-resolved electrospray ionization mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 11(4):312 – 319, 2000.

[56] C. Tanford. Protein denaturation: Part C. theoretical models for the mechanism of denaturation. volume 24 of *Advances in Protein Chemistry*, pages 1 – 95. Academic Press, 1970.

[57] S. E. Jackson and A. R. Fersht. Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. *Biochemistry*, 30(43):10428–10435, 1991.

[58] M. Oliveberg and P. G. Wolynes. The experimental survey of protein-folding energy landscapes. *Quarterly Review of Biophysics*, 38(3):245–288, 2005.

[59] J. L. Neira. Nmr as a tool to identify and characterize protein folding intermediates. *Archives of Biochemistry and Biophysics*, 531(1):90 – 99, 2013.

[60] B. A. Schulman, P. S. Kim, C. M. Dobson, and C. Redfield. A residue-specific NMR view of the non-cooperative unfolding of a molten globule. *Nature Structural Biology*, 4:630, 1997.

[61] A. Miranker, C. Robinson, S. Radford, R. Aplin, and C. Dobson. Detection of transient protein folding populations by mass spectrometry. *Science*, 262(5135): 896–900, 1993.

[62] B. Schuler and W. A. Eaton. Protein folding studied by single-molecule FRET. *Current Opinion in Structural Biology*, 18(1):16–26, 2008.

[63] E. Kloss, N. Courtemanche, and D. Barrick. Repeat-protein folding: new insights into origins of cooperativity, stability, and topology. *Archives of Biochemistry and Biophysics*, 469(1):83–99, 2008.

[64] Y. Ueda, H. Taketomi, and N. Gō. Studies on protein folding, unfolding, and fluctuations by computer simulation. II. A. three-dimensional lattice model of lysozyme. *Biopolymers*, 17(6):1531–1548, 1978.

[65] A. Šali, E. Shakhnovich, and M. Karplus. How does a protein fold? *Nature*, 369: 248, 1994.

[66] H. Lammert, A. Schug, and J. N. Onuchic. Robustness and generalization of structure-based models for protein folding and function. *Proteins: Structure, Function, and Bioinformatics*, 77(4):881–891, 2009.

[67] R. B. Best. Atomistic molecular simulations of protein folding. *Current Opinion in Structural Biology*, 22(1):52 – 61, 2012.

[68] S. Piana, K. Lindorff-Larsen, and D. E. Shaw. Protein folding kinetics and thermodynamics from atomistic simulation. *Proceedings of the National Academy of Sciences*, 109(44):17845–17850, 2012.

[69] A. L. Cortajarena and L. Regan. Ligand binding by tpr domains. *Protein Science*, 15(5):1193–1198, 2006.

[70] M. Tsytlonok, P. Sormanni, P. J. E. Rowling, M. Vendruscolo, and L. S. Itzhaki. Subdomain architecture and stability of a giant repeat protein. *The Journal of Physical Chemistry B*, 117(42):13029–13037, 2013.

[71] M. Tsytlonok, P. O. Craig, E. Sivertsson, D. Serquera, S. Perrett, R. B. Best, P. G. Wolynes, and L. S. Itzhaki. Complex energy landscape of a giant repeat protein. *Structure*, 21(11):1954–1965, 2013.

[72] A. Tevelev, I. J. Byeon, T. Selby, K. Ericson, H. J. Kim, V. Kraynov, and M. D. Tsai. Tumor suppressor p16INK4A: structural characterization of wild-type and mutant proteins by NMR and circular dichroism. *Biochemistry*, 35(29):9475–9487, 1996.

[73] B. Zhang and Z. yu Peng. A minimum folding unit in the ankyrin repeat protein p16$^{\mathrm{ink4a}}$. *Journal of Molecular Biology*, 299(4):1121 – 1132, 2000.

[74] K. S. Tang, B. J. Guralnick, W. K. Wang, A. R. Fersht, and L. S. Itzhaki. Stability and folding of the tumour suppressor protein p16. *Journal of Molecular Biology*, 285(4):1869 – 1886, 1999.

[75] D. U. Ferreiro, S. S. Cho, E. A. Komives, and P. G. Wolynes. The energy landscape of modular repeat proteins: topology determines folding mechanism in the ankyrin family. *Journal of Molecular Biology*, 354(3):679–692, 2005.

[76] A. R. Lowe and L. S. Itzhaki. Rational redesign of the folding pathway of a modular protein. *Proceedings of the National Academy of Sciences*, 104(8):2679–2684, 2007.

[77] M. E. Zweifel, D. J. Leahy, F. M. Hughson, and D. Barrick. Structure and stability of the ankyrin domain of the Drosophila Notch receptor. *Protein Sci.*, 12(11):2622–2632, 2003.

[78] C. C. Mello, C. M. Bradley, K. W. Tripp, and D. Barrick. Experimental characterization of the folding kinetics of the notch ankyrin domain. *Journal of Molecular Biology*, 352(2):266 – 281, 2005.

[79] R. D. Hutton, J. Wilkinson, M. Faccin, E. M. Sivertsson, A. Pelizzola, A. R. Lowe, P. Bruscolini, and L. S. Itzhaki. Mapping the Topography of a Protein Energy Landscape. *Journal of the American Chemical Society*, 137(46):14610–14625, 2015.

[80] C. M. Bradley and D. Barrick. The notch ankyrin domain folds via a discrete, centralized pathway. *Structure*, 14(8):1303–1312, 2006.

[81] K. W. Tripp and D. Barrick. Rerouting the folding pathway of the notch ankyrin

domain by reshaping the energy landscape. *Journal of the American Chemical Society*, 130(17):5681–5688, 2008.

[82] M. Zeeb, H. Rosner, W. Zeslawski, D. Canet, T. A. Holak, and J. Balbach. Protein folding and stability of human CDK inhibitor p19(INK4d). *Journal of Molecular Biology*, 315(3):447–457, 2002.

[83] C. Löw, U. Weininger, M. Zeeb, W. Zhang, E. D. Laue, F. X. Schmid, and J. Balbach. Folding mechanism of an ankyrin repeat protein: scaffold and active site formation of human CDK inhibitor p19(INK4d). *Journal of Molecular Biology*, 373 (1):219–231, 2007.

[84] C. Löw, U. Weininger, P. Neumann, M. Klepsch, H. Lilie, M. T. Stubbs, and J. Balbach. Structural insights into an equilibrium folding intermediate of an archaeal ankyrin repeat protein. *Proceedings of the National academy of Sciences of the United States of America*, 105(10):3779–3784, Mar 2008.

[85] C. Löw, N. Homeyer, U. Weininger, H. Sticht, and J. Balbach. Conformational switch upon phosphorylation: Human cdk inhibitor p19ink4d between the native and partially folded state. *ACS Chemical Biology*, 4(1):53–63, Jan 2009.

[86] D. U. Ferreiro, C. F. Cervantes, S. M. Truhlar, S. S. Cho, P. G. Wolynes, and E. A. Komives. Stabilizing IkappaBalpha by "consensus" design. *Journal of Molecular Biology*, 365(4):1201–1216, 2007.

[87] I. DeVries, D. U. Ferreiro, I. E. Sanchez, and E. A. Komives. Folding kinetics of the cooperatively folded subdomain of the IκBα ankyrin repeat domain. *Journal of Molecular Biology*, 408(1):163–176, 2011.

[88] J. A. Lamboy, H. Kim, K. S. Lee, T. Ha, and E. A. Komives. Visualization of the nanospring dynamics of the IkappaBalpha ankyrin repeat domain in real time. *Proceedings of the National Academy of Sciences*, 108(25):10178–10183, 2011.

[89] J. A. Lamboy, H. Kim, H. Dembinski, T. Ha, and E. A. Komives. Single-molecule FRET reveals the native-state dynamics of the iκbα ankyrin repeat domain. *Journal of Molecular Biology*, 425(14):2578 – 2590, 2013.

[90] P. Michaely and V. Bennett. The membrane-binding domain of ankyrin contains four independently folded subdomains, each comprised of six ankyrin repeats. *Journal of Biological Chemistry*, 268(30):22703–9, 1993.

[91] N. D. Werbeck and L. S. Itzhaki. Probing a moving target with a plastic unfolding intermediate of an ankyrin-repeat protein. *Proceedings of the National Academy of Sciences*, 104(19):7863–7868, 2007.

[92] N. D. Werbeck, P. J. Rowling, V. R. Chellamuthu, and L. S. Itzhaki. Shifting transition states in the unfolding of a large ankyrin repeat protein. *Proceedings of the National Academy of Sciences*, 105(29):9982–9987, 2008.

[93] M. Tsytlonok, S. M. Ibrahim, P. J. E. Rowling, W. Xu, M. J. Ruedas-Rama, A. Orte, D. Klenerman, and L. S. Itzhaki. Single-molecule FRET reveals hidden complexity in a protein energy landscape. *Structure*, 23(1):190–198, 2015.

[94] A. Freiberg, M. P. Machner, W. Pfeil, W. D. Schubert, D. W. Heinz, and R. Seckler. Folding and stability of the leucine-rich repeat domain of internalin B from Listeri monocytogenes. *Journal of Molecular Biology*, 337(2):453–461, 2004.

[95] E. Kloss and D. Barrick. Thermodynamics, kinetics, and salt dependence of folding of yopm, a large leucine-rich repeat protein. *Journal of Molecular Biology*, 383(5): 1195 – 1209, 2008.

[96] T. P. Dao, A. Majumdar, and D. Barrick. Capping motifs stabilize the leucine-rich repeat protein PP32 and rigidify adjacent repeats. *Protein Science*, 23(6):801–811, 2014.

[97] N. Courtemanche and D. Barrick. Folding thermodynamics and kinetics of the leucine-rich repeat domain of the virulence factor internalin b. *Protein Science*, 17 (1):43–53, 2008.

[98] N. Courtemanche and D. Barrick. The leucine-rich repeat domain of Internalin B folds along a polarized N-terminal pathway. *Structure*, 16(5):705–714, 2008.

[99] E. Kloss and D. Barrick. C-terminal deletion of leucine-rich repeats from YopM reveals a heterogeneous distribution of stability in a cooperatively folded protein. *Protein Sci.*, 18(9):1948–1960, 2009.

[100] E. F. Vieux and D. Barrick. Deletion of internal structured repeats increases the stability of a leucine-rich repeat protein, yopm. *Biophysical Chemistry*, 159(1):152 – 161, 2011.

[101] T. P. Dao, A. Majumdar, and D. Barrick. Highly polarized c-terminal transition state of the leucine-rich repeat domain of PP32 is governed by local stability. *Proceedings of the National Academy of Sciences*, 112(18):E2298–E2306, 2015.

[102] M. Junker, C. C. Schuster, A. V. McDonnell, K. A. Sorg, M. C. Finn, B. Berger, and P. L. Clark. Pertactin beta-helix folding mechanism suggests common themes for the secretion and folding of autotransporter proteins. *Proceedings of the National Academy of Sciences*, 103(13):4918–4923, 2006.

[103] D. E. Kamen, Y. Griko, and R. W. Woody. The stability, structural organization, and denaturation of pectate lyase C, a parallel beta-helix protein. *Biochemistry*, 39 (51):15932–15943, 2000.

[104] C. F. Wright, S. A. Teichmann, J. Clarke, and C. M. Dobson. The importance of sequence diversity in the aggregation and evolution of proteins. *Nature*, 438(7069): 878–881, 2005.

[105] F. Rousseau, H. Wilkinson, J. Villanueva, L. Serrano, J. W. Schymkowitz, and L. S.

Itzhaki. Domain swapping in p13suc1 results in formation of native-like, cytotoxic aggregates. *Journal of Molecular Biology*, 363(2):496–505, 2006.

[106] L. K. Mosavi, D. L. Minor, and Z. Y. Peng. Consensus-derived structural determinants of the ankyrin repeat motif. *Proceedings of the National Academy of Sciences*, 99(25):16029–16034, 2002.

[107] M. T. Stumpp, P. Forrer, H. K. Binz, and A. Plückthun. Designing repeat proteins: modular leucine-rich repeat protein libraries based on the mammalian ribonuclease inhibitor family. *Journal of Molecular Biology*, 332(2):471–487, 2003.

[108] F. Parmeggiani, R. Pellarin, A. P. Larsen, G. Varadamsetty, M. T. Stumpp, O. Zerbe, A. Caflisch, and A. Plückthun. Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational optimization of the hydrophobic core. *Journal of Molecular Biology*, 376(5):1282–1304, 2008.

[109] A. Urvoas, A. Guellouz, M. Valerio-Lepiniec, M. Graille, D. Durand, D. C. Desravines, H. van Tilbeurgh, M. Desmadril, and P. Minard. Design, production and molecular structure of a new family of artificial alpha-helicoidal repeat proteins (αrep) based on thermostable HEAT-like repeats. *Journal of Molecular Biology*, 404(2):307–327, 2010.

[110] T. J. Magliery and L. Regan. Beyond consensus: statistical free energies reveal hidden interactions in the design of a TPR motif. *Journal of Molecular Biology*, 343 (3):731–745, 2004.

[111] A. L. Cortajarena, T. Kajander, W. Pan, M. J. Cocco, and L. Regan. Protein design to understand peptide ligand recognition by tetratricopeptide repeat proteins. *Protein Engineering, Design and Selection*, 17(4):399–409, 2004.

[112] R. Tamaskovic, M. Simon, N. Stefan, M. Schwill, and A. Plückthun. Designed ankyrin repeat proteins (DARPins) from research to therapy. *Meth. Enzymol.*, 503: 101–134, 2012.

[113] C. Millership, J. J. Phillips, and E. R. Main. Ising Model Reprogramming of a Repeat Protein's Equilibrium Unfolding Pathway. *Journal of Molecular Biology*, 428(9 Pt A):1804–1817, 2016.

[114] A. Perez-Riba, A. R. Lowe, E. R. G. Main, and L. S. Itzhaki. Context-Dependent Energetics of Loop Extensions in a Family of Tandem-Repeat Proteins. *Biophysical Journal*, 114(11):2552–2562, 2018.

[115] V. S. Devi, H. K. Binz, M. T. Stumpp, A. Plückthun, H. R. Bosshard, and I. Jelesarov. Folding of a designed simple ankyrin repeat protein. *Protein Sci.*, 13(11): 2864–2870, 2004.

[116] T. Kajander, A. L. Cortajarena, E. R. G. Main, S. G. J. Mochrie, and L. Regan. A new folding paradigm for repeat proteins. *Journal of the American Chemical*

Society, 127(29):10188–10190, 2005.

[117] W. Lenz. Beitrag zum Verständnis der magnetischen Erscheinungen in festen Körpern. *Physikalische Zeitschrift*, 21:613–615, 1920.

[118] E. Ising. Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Physik*, 31: 253–258, 1925.

[119] J. A. Schellman. The factors affecting the stability of hydrogen-bonded polypeptide structures in solution. *The Journal of Physical Chemistry*, 62(12):1485–1494, 1958.

[120] B. H. Zimm and J. K. Bragg. Theory of the phase transition between helix and random coil in polypeptide chains. *The Journal of Chemical Physics*, 31(2):526–535, 1959.

[121] S. Lifson and A. Roig. On the theory of helix-coil transition in polypeptides. *The Journal of Chemical Physics*, 34(6):1963–1974, 1961.

[122] A. L. Cortajarena, S. G. Mochrie, and L. Regan. Mapping the energy landscape of repeat proteins using nmr-detected hydrogen exchange. *Journal of Molecular Biology*, 379(3):617 – 626, 2008.

[123] A. L. Cortajarena and L. Regan. Calorimetric study of a series of designed repeat proteins: modular structure and modular folding. *Protein Sci.*, 20(2):336–340, 2011.

[124] Y. Javadi and E. R. G. Main. Exploring the folding energy landscape of a series of designed consensus tetratricopeptide repeat proteins. *Proceedings of the National Academy of Sciences*, 106(41):17383–17388, 2009.

[125] D. U. Ferreiro, A. M. Walczak, E. A. Komives, and P. G. Wolynes. The energy landscapes of repeat-containing proteins: topology, cooperativity, and the folding funnels of one-dimensional architectures. *PloS Computational Biology*, 4(5):e1000070, 2008.

[126] T. Kajander, A. L. Cortajarena, S. Mochrie, and L. Regan. Structure and stability of designed tpr protein superhelices: unusual crystal packing and implications for natural tpr proteins. *Acta Crystallographica Section D*, 63(7):800–811, 2007.

[127] A. L. Cortajarena, S. G. J. Mochrie, and L. Regan. Modulating repeat protein stability: The effect of individual helix stability on the collective behavior of the ensemble. *Protein Science*, 20(6):1042–1047, 2011.

[128] J. J. Phillips, Y. Javadi, C. Millership, and E. R. Main. Modulation of the multistate folding of designed TPR proteins through intrinsic and extrinsic factors. *Protein Sci.*, 21(3):327–338, 2012.

[129] G. Interlandi, S. K. Wetzel, G. Settanni, A. Plückthun, and A. Caflisch. Characterization and further stabilization of designed ankyrin repeat proteins by combining molecular dynamics simulations and experiments. *Journal of Molecular Biology*, 375(3):837–854, 2008.

[130] T. Merz, S. K. Wetzel, S. Firbank, A. Plückthun, M. G. Grutter, and P. R. Mittl. Stabilizing ionic interactions in a full-consensus ankyrin repeat protein. *Journal of Molecular Biology*, 376(1):232–240, 2008.

[131] M. A. Kramer, S. K. Wetzel, A. Plückthun, P. R. Mittl, and M. G. Grutter. Structural determinants for improved stability of designed ankyrin repeat proteins with a redesigned C-capping module. *Journal of Molecular Biology*, 404(3):381–391, 2010.

[132] S. K. Wetzel, G. Settanni, M. Kenig, H. K. Binz, and A. Plückthun. Folding and unfolding mechanism of highly stable full-consensus ankyrin repeat proteins. *Journal of Molecular Biology*, 376(1):241 – 257, 2008.

[133] T. Aksel, A. Majumdar, and D. Barrick. The contribution of entropy, enthalpy, and hydrophobic desolvation to cooperativity in repeat-protein folding. *Structure*, 19(3):349–360, 2011.

[134] S. K. Wetzel, C. Ewald, G. Settanni, S. Jurt, A. Plückthun, and O. Zerbe. Residue-resolved stability of full-consensus ankyrin repeat proteins probed by NMR. *Journal of Molecular Biology*, 402(1):241–258, 2010.

[135] T. Aksel and D. Barrick. Direct observation of parallel folding pathways revealed using a symmetric repeat protein system. *Biophysical Journal*, 107(1):220–232, 2014.

[136] P. Alfarano, G. Varadamsetty, C. Ewald, F. Parmeggiani, R. Pellarin, O. Zerbe, A. Plückthun, and A. Caflisch. Optimization of designed armadillo repeat proteins by molecular dynamics simulations and NMR spectroscopy. *Protein Science*, 21(9): 1298–1314, 2012.

[137] F. Parmeggiani, P.-S. Huang, S. Vorobiev, R. Xiao, K. Park, S. Caprari, M. Su, J. Seetharaman, L. Mao, H. Janjua, *et al.* A general computational approach for repeat protein design. *Journal of Molecular Biology*, 427(2):563–575, 2015.

[138] K. Geiger-Schuller, K. Sforza, M. Yuhas, F. Parmeggiani, D. Baker, and D. Barrick. Extreme stability in de novo-designed repeat arrays is determined by unusually stable short-range interactions. *Proceedings of the National Academy of Sciences*, 115(29):7539–7544, 2018.

[139] A. Perez-Riba. *Rational redesign or repeat-protein structure and function for new tools and biomaterials.* PhD thesis, University of Cambridge, 2016.

[140] Y. Chen, S. E. Radford, and D. J. Brockwell. Force-induced remodelling of proteins and their complexes. *Current Opinion in Structural Biology*, 30(0):89–99, 2015.

[141] G. Žoldák and M. Rief. Force as a single molecule probe of multidimensional protein energy landscapes. *Current Opinion in Structural Biology*, 23(1):48–57, 2013.

[142] A. Ashkin. Acceleration and trapping of particles by radiation pressure. *Physical Review Letters*, 24:156–159, 1970.

[143] A. Ashkin, J. M. Dziedzic, J. E. Bjorkholm, and S. Chu. Observation of a single-beam gradient force optical trap for dielectric particles. *Optics Letters*, 11(5):288–290, 1986.

[144] C. Cecconi, E. Shank, F. Dahlquist, S. Marqusee, and C. Bustamante. Protein-DNA chimeras for single molecule mechanical folding studies with the optical tweezers. *European Biophysics Journal*, 37(6):729–738, 2008.

[145] C. Cecconi, E. Shank, S. Marqusee, and C. Bustamante. DNA molecular handles for single-molecule protein-folding studies by optical tweezers. In G. Zuccheri and B. Samorì, editors, *DNA Nanotechnology*, volume 749 of *Methods in Molecular Biology*, pages 255–271. Humana Press, 2011.

[146] J. C. Cordova, D. K. Das, H. W. Manning, and M. J. Lang. Combining single-molecule manipulation and single-molecule detection. *Current Opinion in Structural Biology*, 28(0):142–148, 2014.

[147] L. Tskhovrebova, J. Trinick, J. A. Sleep, and R. M. Simmons. Elasticity and unfolding of single molecules of the giant muscle protein titin. *Nature*, 387:308, 1997.

[148] M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez, and H. E. Gaub. Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science*, 276(5315): 1109–1112, 1997.

[149] J. F. Marko and E. D. Siggia. Statistical mechanics of supercoiled DNA. *Physical Review E, Statistical Physics, Plasmas, Fluids and Related Interdisciplinary Topics*, 52(3):2912–2938, 1995.

[150] M. S. Z. Kellermayer, S. B. Smith, H. L. Granzier, and C. Bustamante. Folding-unfolding transitions in single titin molecules characterized with laser tweezers. *Science*, 276(5315):1112–1116, 1997.

[151] H. Li, A. F. Oberhauser, S. B. Fowler, J. Clarke, and J. M. Fernandez. Atomic force microscopy reveals the mechanical design of a modular protein. *Proceedings of the National Academy of Sciences*, 97(12):6527–6531, 2000.

[152] R. B. Best, S. B. Fowler, J. L. T. Herrera, A. Steward, E. Paci, and J. Clarke. Mechanical unfolding of a titin ig domain: Structure of transition state revealed by combining atomic force microscopy, protein engineering and molecular dynamics simulations. *Journal of Molecular Biology*, 330(4):867 – 877, 2003.

[153] P. M. Williams, S. B. Fowler, R. B. Best, J. Luis Toca-Herrera, K. A. Scott, A. Steward, and J. Clarke. Hidden complexity in the mechanical properties of titin. *Nature*, 422:446, 2003.

[154] M. Carrión-Vázquez, A. F. Oberhauser, S. B. Fowler, P. E. Marszalek, S. E. Broedel, J. Clarke, and J. M. Fernandez. Mechanical and chemical unfolding of a single protein: A comparison. *Proceedings of the National Academy of Sciences*, 96(7):

3694–3699, 1999.

[155] I. Schwaiger, C. Sattler, D. R. Hostetter, and M. Rief. The myosin coiled-coil is a truly elastic protein structure. *Nat Mater*, 1(4):232–235, 2002.

[156] M. Rief, J. Pascual, M. Saraste, and H. E. Gaub. Single molecule force spectroscopy of spectrin repeats: low unfolding forces in helix bundles. *Journal of Molecular Biology*, 286(2):553 – 561, 1999.

[157] J. R. Forman, S. Qamar, E. Paci, R. N. Sandford, and J. Clarke. The remarkable mechanical strength of polycystin-1 supports a direct role in mechanotransduction. *Journal of Molecular Biology*, 349(4):861 – 871, 2005.

[158] L. G. Randles, R. W. Rounsevell, and J. Clarke. Spectrin domains lose cooperativity in forced unfolding. *Biophysical Journal*, 92(2):571–577, 2007.

[159] K. Svoboda, C. F. Schmidt, B. J. Schnapp, and S. M. Block. Direct observation of kinesin stepping by optical trapping interferometry. *Nature*, 365:721, 1993.

[160] B. Milic, A. Chakraborty, K. Han, M. C. Bassik, and S. M. Block. Kif15 nanomechanics and kinesin inhibitors, with implications for cancer chemotherapeutics. *Proceedings of the National Academy of Sciences*, 115(20):E4613–E4622, 2018.

[161] M. I. Molodtsov, C. Mieck, J. Dobbelaere, A. Dammermann, S. Westermann, and A. Vaziri. A Force-Induced Directional Switch of a Molecular Motor Enables Parallel Microtubule Bundle Formation. *Cell*, 167(2):539–552, 2016.

[162] S. Tafoya and C. Bustamante. Molecular switch-like regulation in motor proteins. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 373(1749), 2018.

[163] K. Neupane, D. A. Foster, D. R. Dee, H. Yu, F. Wang, and M. T. Woodside. Direct observation of transition paths during the folding of proteins and nucleic acids. *Science*, 352(6282):239–242, 2016.

[164] M. D. Wang, H. Yin, R. Landick, J. Gelles, and S. M. Block. Stretching DNA with optical tweezers. *Biophysical Journal*, 72(3):1335–1346, 1997.

[165] J. M. Eeftens, S. Bisht, J. Kerssemakers, M. Kschonsak, C. H. Haering, and C. Dekker. Real-time detection of condensin-driven DNA compaction reveals a multistep binding mechanism. *EMBO Journal*, 36(23):3448–3457, 2017.

[166] R. B. Best, B. Li, A. Steward, V. Daggett, and J. Clarke. Can non-mechanical proteins withstand force?: Stretching barnase by atomic force microscopy and molecular dynamics simulation. *Biophysical Journal*, 81(4):2344–2356, 2001.

[167] C. Cecconi, E. A. Shank, C. Bustamante, and S. Marqusee. Direct observation of the three-state folding of a single protein molecule. *Science*, 309(5743):2057–2060, 2005.

[168] G. Zoldák, J. Stigler, B. Pelz, H. Li, and M. Rief. Ultrafast folding kinetics and

cooperativity of villin headpiece in single-molecule force spectroscopy. *Proceedings of the National Academy of Sciences*, 110(45):18156, 2013.

[169] A. Solanki, K. Neupane, and M. T. Woodside. Single-molecule force spectroscopy of rapidly fluctuating, marginally stable structures in the intrinsically disordered protein α-synuclein. *Physical Review Letters*, 112:158103, 2014.

[170] M. Baclayon, P. v. Ulsen, H. Mouhib, M. H. Shabestari, T. Verzijden, S. Abeln, W. H. Roos, and G. J. L. Wuite. Mechanical unfolding of an autotransporter passenger protein reveals the secretion starting point and processive transport intermediates. *ACS Nano*, 10(6):5710–5719, 2016.

[171] D. Sharma, O. Perisic, Q. Peng, Y. Cao, C. Lam, H. Lu, and H. Li. Single-molecule force spectroscopy reveals a mechanically stable protein fold and the rational tuning of its mechanical stability. *Proceedings of the National Academy of Sciences*, 104 (22):9278–9283, 2007.

[172] Q. Peng and H. Li. Domain insertion effectively regulates the mechanical unfolding hierarchy of elastomeric proteins: Toward engineering multifunctional elastomeric proteins. *Journal of the American Chemical Society*, 131(39):14050–14056, 2009.

[173] Q. Li, Z. N. Scholl, and P. E. Marszalek. Capturing the mechanical unfolding pathway of a large protein with coiled-coil probes. *Angewandte Chemie International Edition in English*, 53(49):13429–13433, 2014.

[174] Q. Peng and H. Li. Direct observation of tug-of-war during the folding of a mutually exclusive protein. *Journal of the American Chemical Society*, 131(37):13347–13354, 2009.

[175] E. J. Guinn, B. Jagannathan, and S. Marqusee. Single-molecule chemo-mechanical unfolding reveals multiple transition state barriers in a small single-domain protein. *Nat Commun*, 6:6861, 2015.

[176] E. A. Shank, C. Cecconi, J. W. Dill, S. Marqusee, and C. Bustamante. The folding cooperativity of a protein is controlled by its chain topology. *Nature*, 465(7298): 637–640, 2010.

[177] B. Jagannathan, P. J. Elms, C. Bustamante, and S. Marqusee. Direct observation of a force-induced switch in the anisotropic mechanical unfolding pathway of a protein. *Proceedings of the National Academy of Sciences*, 109(44):17820–17825, 2012.

[178] K. Neupane, A. P. Manuel, and M. T. Woodside. Protein folding trajectories can be described quantitatively by one-dimensional diffusion over measured energy landscapes. *Nature Physics*, 12:700 EP, 2016.

[179] R. A. Maillard, G. Chistol, M. Sen, M. Righini, J. Tan, C. M. Kaiser, C. Hodges, A. Martin, and C. Bustamante. ClpX(P) generates mechanical force to unfold and translocate its protein substrates. *Cell*, 145(3):459–469, 2011.

[180] L. Rognoni, T. Möst, G. Žoldák, and M. Rief. Force-dependent isomerization kinetics of a highly conserved proline switch modulates the mechanosensing region of filamin. *Proceedings of the National Academy of Sciences*, 111(15):5568–5573, 2014.

[181] A. P. Wiita, R. Perez-Jimenez, K. A. Walther, F. Gräter, B. J. Berne, A. Holmgren, J. M. Sanchez-Ruiz, and J. M. Fernandez. Probing the chemistry of thioredoxin catalysis with force. *Nature*, 450:124, 2007.

[182] Y. Cao, M. M. Balamurali, D. Sharma, and H. Li. A functional single-molecule binding assay via force spectroscopy. *Proceedings of the National Academy of Sciences*, 104(40):15677–15681, 2007.

[183] Y. Cao, T. Yoo, S. Zhuang, and H. Li. Protein-protein interaction regulates proteins' mechanical stability. *Journal of Molecular Biology*, 378(5):1132–1141, 2008.

[184] J. Stigler and M. Rief. Calcium-dependent folding of single calmodulin molecules. *Proceedings of the National Academy of Sciences*, 109(44):17814–17819, 2012.

[185] O. D. Broekmans, G. A. King, G. J. Stephens, and G. J. L. Wuite. DNA twist stability changes with magnesium(2+) concentration. *Physical Review Letters*, 116: 258102, 2016.

[186] D. Bauer, S. Meinhold, R. P. Jakob, J. Stigler, U. Merkel, T. Maier, M. Rief, and G. Žoldák. A folding nucleus and minimal ATP binding domain of Hsp70 identified by single-molecule force spectroscopy. *Proceedings of the National Academy of Sciences*, 2018. ISSN 0027-8424.

[187] F. Baumann, M. S. Bauer, M. Rees, A. Alexandrovich, M. Gautel, D. A. Pippig, and H. E. Gaub. Increasing evidence of mechanical force as a functional regulator in smooth muscle myosin light chain kinase. *eLife*, 6:e26473, 2017.

[188] M. Righini, A. Lee, C. Canari-Chumpitaz, T. Lionberger, R. Gabizon, Y. Coello, I. Tinoco, and C. Bustamante. Full molecular trajectories of RNA polymerase at single base-pair resolution. *Proceedings of the National Academy of Sciences*, 115 (6):1286–1291, 2018.

[189] C. A. Meng, F. M. Fazal, and S. M. Block. Real-time observation of polymerase-promoter contact remodeling during transcription initiation. *Nature Communications*, 8(1):1178, 2017.

[190] S. R. K. Ainavarapu, L. Li, C. L. Badilla, and J. M. Fernandez. Ligand binding modulates the mechanical stability of dihydrofolate reductase. *Biophysical Journal*, 89(5):3337–3344, 2005.

[191] G. Settanni, D. Serquera, P. E. Marszalek, E. Paci, and L. S. Itzhaki. Effects of ligand binding on the mechanical properties of ankyrin repeat protein gankyrin. *PLoS Comput Biol*, 9(1):e1002864, 2013.

[192] E. M. Puchner, A. Alexandrovich, A. L. Kho, U. Hensen, L. V. Schäfer, B. Brand-

meier, F. Gräter, H. Grubmüller, H. E. Gaub, and M. Gautel. Mechanoenzymatics of titin kinase. *Proceedings of the National Academy of Sciences*, 105(36):13385–13390, 2008.

[193] C. D. Buckley, J. Tan, K. L. Anderson, D. Hanein, N. Volkmann, W. I. Weis, W. J. Nelson, and A. R. Dunn. The minimal cadherin-catenin complex binds to actin filaments under force. *Science*, 346(6209):600, 2014.

[194] L. Li, S. Wetzel, A. Plückthun, and J. M. Fernandez. Stepwise unfolding of ankyrin repeats in a single protein revealed by atomic force microscopy. *Biophysical Journal*, 90(4):L30–L32, 2006.

[195] D. Serquera, W. Lee, G. Settanni, P. E. Marszalek, E. Paci, and L. S. Itzhaki. Mechanical unfolding of an ankyrin repeat protein. *Biophysical Journal*, 98(7): 1294–1301, 2010.

[196] G. Lee, K. Abdi, Y. Jiang, P. Michaely, V. Bennett, and P. E. Marszalek. Nanospring behaviour of ankyrin repeats. *Nature*, 440:246–249, 2006.

[197] M. Kim, K. Abdi, G. Lee, M. Rabbi, W. Lee, M. Yang, C. J. Schofield, V. Bennett, and P. E. Marszalek. Fast and forceful refolding of stretched α-helical solenoid proteins. *Biophysical Journal*, 98:3086–3092, 2010.

[198] A. Valbuena, A. M. Vera, J. Oroz, M. Menendez, and M. Carrión-Vázquez. Mechanical properties of β-catenin revealed by single-molecule experiments. *Biophysical Journal*, 103(8):1744–1752, 2012.

[199] M. Sotomayor, D. P. Corey, and K. Schulten. In search of the hair-cell gating spring elastic properties of ankyrin and cadherin repeats. *Structure*, 13(4):669–682, 2005.

[200] C. Kappel, U. Zachariae, N. Dölker, and H. Grubmüller. An unusual hydrophobic core confers extreme flexibility to HEAT repeat proteins. *Biophysical Journal*, 99 (5):1596–1603, 2010.

[201] A. Grinthal, I. Adamovic, B. Weiner, M. Karplus, and N. Kleckner. PR65, the HEAT-repeat scaffold of phosphatase PP2A, is an elastic connector that links force and catalysis. *Proceedings of the National Academy of Sciences*, 107(6):2467–2472, 2010.

[202] N. Fukuhara, E. Fernandez, J. Ebert, E. Conti, and D. Svergun. Conformational variability of nucleo-cytoplasmic transport factors. *Journal of Biological Chemistry*, 279(3):2176–2181, 2004.

[203] U. Zachariae and H. Grubmüller. Importin-β: Structural and dynamic determinants of a molecular spring. *Structure*, 16(6):906–915, 2008.

[204] D. Barrick. Biological regulation via ankyrin repeat folding. *ACS Chemical Biology*, 4(1):19–22, Jan 2009.

[205] S. S. Cohen, I. Riven, A. L. Cortajarena, L. De Rosa, L. D. D'Andrea, L. Regan,

and G. Haran. Probing the molecular origin of native-state flexibility in repeat proteins. *Journal of the American Chemical Society*, 137(32):10367–10373, 2015.

[206] C. Lambrecht, D. Haesen, W. Sents, E. Ivanova, and V. Janssens. *Phosphatase Modulators*, chapter 17: Structure, Regulation, and Pharmacological Modulation of PP2A Phosphatases, pages 283–305. Humana Press, 2013.

[207] J.-M. Sontag and E. Sontag. Protein phosphatase 2A dysfunction in Alzheimer's disease. *Frontiers in Molecular Neuroscience*, 7, 2014.

[208] P. J. Eichhorn, M. P. Creyghton, and R. Bernards. Protein phosphatase 2A regulatory subunits and cancer . *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1795(1):1–15, 2009.

[209] Y. Xu, Y. Chen, P. Zhang, P. D. Jeffrey, and Y. Shi. Structure of a protein phosphatase 2A holoenzyme: insights into B55-mediated tau dephosphorylation. *Molecular Cell*, 31(6):873–885, 2008.

[210] C. Liu and J. Götz. How it all started: Tau and protein phosphatase 2A. *Journal of Alzheimer's Disease*, 37(3):482–494, 2013.

[211] J.-M. Sontag and E. Sontag. Regulation of cell adhesion by PP2A and SV40 small tumor antigen: An important link to cell transformation. *Cellular and Molecular Life Sciences CMLS*, 63(24):2979–2991, 2006.

[212] T. S. Kitajima, T. Sakuno, K. ichiro Ishiguro, S. ichiro Iemura, T. Natsume, S. A. Kawashima, and Y. Watanabe. Shugoshin collaborates with protein phosphatase 2A to protect cohesin. *Nature*, 441:46–52, 2006.

[213] A. Espert, P. Uluocak, R. N. Bastos, D. Mangat, P. Graab, and U. Gruneberg. PP2A-B56 opposes Mps1 phosphorylation of Knl1 and thereby promotes spindle assembly checkpoint silencing. *The Journal of Cell Biology*, 206(7):833–842, 2014.

[214] K. Wassmann. Sister chromatid segregation in meiosis II: Deprotection through phosphorylation. *Cell Cycle*, 12(9):1352–1359, 2013.

[215] Z. Xu, B. Cetin, M. Anger, U. S. Cho, W. Helmhart, K. Nasmyth, and W. Xu. Structure and function of the PP2A-shugoshin interaction. *Molecular Cell*, 35(4): 426–441, 2009.

[216] J. Lee, T. S. Kitajima, Y. Tanno, K. Yoshida, T. Morita, T. Miyano, M. Miyake, and Y. Watanabe. Unified mode of centromeric protection by shugoshin in mammalian oocytes and somatic cells. *Nature Cell Biology*, 10:42–52, 2008.

[217] N. Wlodarchak, F. Guo, K. A. Satyshur, L. Jiang, P. D. Jeffrey, T. Sun, V. Stanevich, M. C. Mumby, and Y. Xing. Structure of the Ca^{2+} -dependent PP2A heterotrimer and insights into Cdc6 dephosphorylation. *Cell Research*, 23:931–946, 2013.

[218] A. J. Davis, Z. Yan, B. Martinez, and M. C. Mumby. Protein phosphatase 2A is

targeted to cell division control protein 6 by a calcium-binding regulatory subunit. *Journal of Biological Chemistry*, 283(23):16104–16114, 2008.

[219] Y. Chen, Y. Xu, Q. Bao, Y. Xing, Z. Li, Z. Lin, J. B. Stock, P. D. Jeffrey, and Y. Shi. Structural and biochemical insights into the regulation of protein phosphatase 2A by small t antigen of SV40. *Nature Structural & Molecular Biology*, 14(6):527–534, 2007.

[220] Y. Xing, Y. Xu, Y. Chen, P. D. Jeffrey, Y. Chao, Z. Lin, Z. Li, S. Strack, J. B. Stock, and Y. Shi. Structure of protein phosphatase 2A core enzyme bound to tumor-inducing toxins. *Cell*, 127(2):341–353, 2006.

[221] G. Pósfai, G. Plunkett, T. Fehér, D. Frisch, G. M. Keil, K. Umenhoffer, V. Kolisnychenko, B. Stahl, S. S. Sharma, M. De Arruda, *et al.* Emergent properties of reduced-genome escherichia coli. *Science*, 312(5776):1044–1046, 2006.

[222] S. Wagner, M. M. Klepsch, S. Schlegel, A. Appel, R. Draheim, M. Tarry, M. Högbom, K. J. van Wijk, D. J. Slotboom, J. O. Persson, and J.-W. de Gier. Tuning escherichia coli for membrane protein overexpression. *Proceedings of the National Academy of Sciences*, 105(38):14371–14376, 2008.

[223] D. T. Rogerson, A. Sachdeva, K. Wang, T. Haq, A. Kazlauskaite, S. M. Hancock, N. Huguenin-Dezot, M. M. Muqit, A. M. Fry, R. Bayliss, and J. W. Chin. Efficient genetic encoding of phosphoserine and its nonhydrolyzable analog. *Nature Chemical Biology*, 11(7):496–503, 2015.

[224] K. Wang, A. Sachdeva, D. J. Cox, N. M. Wilf, K. Lang, S. Wallace, R. A. Mehl, and J. W. Chin. Optimized orthogonal translation of unnatural amino acids enables spontaneous protein double-labelling and FRET. *Nature Chemistry*, 6:393–403, 2014.

[225] C. Li, A. Wen, B. Shen, J. Lu, Y. Huang, and Y. Chang. FastCloning: a highly simplified, purification-free, sequence- and ligation-independent PCR cloning method. *BMC Biotechnology*, 11(92), 2011.

[226] A. Hemsley, N. Arnheim, M. D. Toney, G. Cortopassi, and D. J. Galas. A simple method for site-directed mutagenesis using the polymerase chain reaction. *Nucleic Acids Research*, 17(16):6545–6551, 1989.

[227] S. Moore. 'round the horn site-directed mutagenesis. URL `https://openwetware.org/wiki/%27Round-the-horn_site-directed_mutagenesis`.

[228] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel, and A. Bairoch. *Protein Identification and Analysis Tools on the ExPASy Server*, pages 571–607. Humana Press, 2005.

[229] Y. von Hansen, A. Mehlich, B. Pelz, M. Rief, and R. R. Netz. Auto- and cross-power spectral analysis of dual trap optical tweezer experiments using bayesian inference.

Review of Scientific Instruments, 83(9):095116, 2012.

[230] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. 2015.

[231] W. Humphrey, A. Dalke, and K. Schulten. VMD: visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 1996.

[232] A. Bakan, L. M. Meireles, and I. Bahar. ProDy: Protein dynamics inferred from theory and experiments. *Bioinformatics*, 27(11):1575–1577, 2011.

[233] K. J. Millman and M. Aivazis. Python for scientists and engineers. *Computing in Science & Engineering*, 13(2):9–12, 2011.

[234] T. E. Oliphant. Python for scientific computing. *Computing in Science & Engineering*, 9(3):10–20, 2007.

[235] S. v. d. Walt, S. C. Colbert, and G. Varoquaux. The numpy array: A structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30, 2011.

[236] E. Jones, T. Oliphant, P. Peterson, *et al.* SciPy: Open source scientific tools for Python, 2001. http://www.scipy.org/.

[237] J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

[238] T. L. Rodgers, D. Burnell, P. D. Townsend, E. Pohl, M. J. Cann, M. R. Wilson, and T. C. McLeish. ΔΔPT: a comprehensive toolbox for the analysis of protein motion. *BMC Bioinformatics*, 14(183):1–9, 2013.

[239] I. G. Hughes and T. P. Hase. *Measurements and their Uncertainties.* Oxford University Press, 2010.

[240] Z. Bu and D. J. Callaway. Chapter 5 - Proteins MOVE! protein dynamics and long-range allostery in cell signaling. In R. Donev, editor, *Protein Structure and Diseases*, volume 83 of *Advances in Protein Chemistry and Structural Biology*, pages 163–221. Academic Press, 2011.

[241] I. Bahar and A. Rader. Coarse-grained normal mode analysis in structural biology. *Current Opinion in Structural Biology*, 15:586–592, 2005.

[242] M. Karplus and J. Kuriyan. Molecular dynamics and protein function. *Proceedings of the National Academy of Sciences*, 102(19):6679–6685, 2005.

[243] S. Maguid, S. Fernández-Alberti, G. Parisi, and J. Echave. Evolutionary conservation of protein backbone flexibility. *Journal of Molecular Evolution*, 63(4):448–457, 2006.

[244] A. Leo-Macias, P. Lopez-Romero, D. Lupyan, D. Zerbino, and A. R. Ortiz. An analysis of core deformations in protein superfamilies. *Biophysical Journal*, 88: 1291–1299, 2005.

[245] I. Bahar, T. R. Lezon, L.-W. Yang, and E. Eyal. Global dynamics of proteins: Bridging between structure and function. *Annual Review of Biophysics*, 39(1):23–42, 2010.

[246] A. Dutta, J. Krieger, J. Y. Lee, J. Garcia-Nafria, I. H. Greger, and I. Bahar. Cooperative Dynamics of Intact AMPA and NMDA Glutamate Receptors: Similarities and Subfamily-Specific Differences. *Structure*, 23(9):1692–1704, 2015.

[247] T. Perica, Y. Kondo, S. P. Tiwari, S. H. McLaughlin, K. R. Kemplen, X. Zhang, A. Steward, N. Reuter, J. Clarke, and S. A. Teichmann. Evolution of oligomeric state through allosteric pathways that mimic ligand binding. *Science*, 346(6216), 2014.

[248] K. A. Niessen, M. Xu, A. Paciaroni, A. Orecchini, E. H. Snell, and A. G. Markelz. Moving in the right direction: Protein vibrations steering function. *Biophysical Journal*, 112(5):933 – 942, 2017.

[249] R. Brüschweiler. Collective protein dynamics and nuclear spin relaxation. *The Journal of Chemical Physics*, 102(8):3396–3403, 1995.

[250] R. J. Hawkins and T. C. B. McLeish. Coupling of global and local vibrational modes in dynamic allostery of proteins. *Biophysical Journal*, 91:2055–2062, 2006.

[251] M. Kovermann, P. Rogne, and M. Wolf-Watz. Protein dynamics and function from solution state NMR spectroscopy. *Quarterly Reviews of Biophysics*, 49:e6, 2016.

[252] J. R. Engen. Analysis of protein conformation and dynamics by Hydrogen/Deuterium Exchange MS. *Analytical Chemistry*, 81(19):7870–7875, 2009.

[253] A. G. Kikhney and D. I. Svergun. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Letters*, 589 (19PartA):2570–2577, 2015.

[254] N. Go, T. Noguti, and T. Nishikawa. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proceedings of the National Academy of Sciences*, 80(12):3696–3700, 1983.

[255] F. Tama and Y.-H. Sanejouand. Conformational change of proteins arising from normal mode calculations. *Protein Engineering*, 14(1):1–6, 2001.

[256] J. Y. Lee, Z. Feng, X.-Q. Xie, and I. Bahar. Allosteric modulation of intact γ-secretase structural dynamics. *Biophysical Journal*, 113(12):2634–2649, 2017.

[257] Z. Yang, P. Majek, and I. Bahar. Allosteric transitions of supramolecular systems explored by network models: application to chaperonin GroEL. *PloS Computational Biology*, 5(4):e1000360, 2009.

[258] M. Gur, J. D. Madura, and I. Bahar. Global transitions of proteins explored by a multiscale hybrid methodology: Application to adenylate kinase. *Biophysical Journal*, 105:1643–1652, 2013.

[259] L. Yang, G. Song, A. Carriquiry, and R. L. Jernigan. Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes. *Structure*, 16(2):321–330, 2008.

[260] S. Hayward and B. de Groot. Normal modes and essential dynamics. In A. Kukol, editor, *Molecular Modeling of Proteins*, volume 443 of *Methods Molecular Biology*^{TM}, pages 89–106. Humana Press, 2008.

[261] L. Yang, G. Song, and R. L. Jernigan. Protein elastic network models and the ranges of cooperativity. *Proceedings of the National Academy of Sciences*, 106(30): 12347–12352, 2009.

[262] I. Bahar, A. R. Atilgan, and B. Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 2: 173–181, 1997.

[263] I. Bahar, M. Cheng, J. Lee, C. Kaya, and S. Zhang. Structure-encoded global motions and their role in mediating protein-substrate interactions. *Biophysical Journal*, 109(6):1101–1109, 2015.

[264] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal*, 80(1):505–515, 2001.

[265] I. Bahar, A. Wallqvist, D. G. Covell, and R. L. Jernigan. Correlation between native-state hydrogen exchange and cooperative residue fluctuations from a simple model. *Biochemistry*, 37(4):1067–1075, 1998.

[266] M. H. Kim, S. Seo, J. I. Jeong, B. J. Kim, W. K. Liu, B. S. Lim, J. B. Choi, and M. K. Kim. A mass weighted chemical elastic network model elucidates closed form domain motions in proteins. *Protein Science*, 22(5):605–613, 2013.

[267] W. Zheng. Accurate flexible fitting of high-resolution protein structures into cryo-electron microscopy maps using coarse-grained pseudo-energy minimization. *Biophysical Journal*, 100(2):478–488, 2011.

[268] N. Leioatts, T. D. Romo, and A. Grossfield. Elastic Network Models are Robust to Variations in Formalism. *J Chem Theory Comput*, 8(7):2424–2434, 2012.

[269] Y. Xu, Y. Xing, Y. Chen, Y. Chao, Z. Lin, E. Fan, J. W. Yu, S. Strack, P. D. Jeffrey, and Y. Shi. Structure of the protein phosphatase 2A holoenzyme. *Cell*, 127 (6):1239–1251, 2006.

[270] J. Huhn, P. D. Jeffrey, K. Larsen, T. Rundberget, F. Rise, N. R. Cox, V. Arcus, Y. Shi, and C. O. Miles. A structural basis for the reduced toxicity of dinophysistoxin-2. *Chemical Research in Toxicology*, 22(11):1782–1786, 2009.

[271] C. G. Wu, A. Zheng, L. Jiang, M. Rowse, V. Stanevich, H. Chen, Y. Li, K. A. Satyshur, B. Johnson, T. J. Gu, Z. Liu, and Y. Xing. Methylation-regulated de-

commissioning of multimeric PP2A complexes. *Nat Commun*, 8(1):2272, 2017.

[272] M. Perego and J. A. Hoch. Cell-cell communication regulates the effects of protein aspartate phosphatases on the phosphorelay controlling development in Bacillus subtilis. *Proceedings of the National Academy of Sciences*, 93(4):1549–1553, 1996. ISSN 0027-8424.

[273] Y.-L. Tzeng, V. A. Feher, J. Cavanagh, M. Perego, and J. A. Hoch. Characterization of interactions between a two-component response regulator, Spo0F, and its phosphatase, RapB. *Biochemistry*, 37(47):16538–16545, 1998.

[274] M. Perego. A new family of aspartyl phosphate phosphatases targeting the sporulation transcription factor Spo0A of Bacillus subtilis. *Molecular Microbiology*, 42(1): 133–143, 2008.

[275] L. J. Core, S. Ishikawa, and M. Perego. A free terminal carboxylate group is required for PhrA pentapeptide inhibition of RapA phosphatase. *Peptides*, 22(10):1549 – 1553, 2001.

[276] C. Bongiorni, R. Stoessel, and M. Perego. Negative regulation of Bacillus anthracis sporulation by the Spo0E family of phosphatases. *Journal of Bacteriology*, 189(7): 2637–2645, 2007.

[277] A. R. Diaz, L. J. Core, M. Jiang, M. Morelli, C. H. Chiang, H. Szurmant, and M. Perego. Bacillus subtilis RapA phosphatase domain interaction with its substrate, phosphorylated Spo0F, and its inhibitor, the PhrA peptide. *Journal of Bacteriology*, 194(6):1378–1388, 2012.

[278] M. D. Baker and M. B. Neiditch. Structural basis of response regulator inhibition by a bacterial anti-activator protein. *PloS Biology*, 9(12):1–16, 2011.

[279] V. Parashar, N. Mirouze, D. A. Dubnau, and M. B. Neiditch. Structural basis of response regulator dephosphorylation by Rap phosphatases. *PLoS Biol.*, 9(2): e1000589, 2011.

[280] T. Z. Grove, L. Regan, and A. L. Cortajarena. Nanostructured functional films from engineered repeat proteins. *Journal of the Royal Society Interface*, 10(83):20130051, 2013.

[281] M. L. Teodoro, G. N. Phillips, Jr., and L. E. Kavraki. A dimensionality reduction approach to modeling protein flexibility. In *Proceedings of the Sixth Annual International Conference on Computational Biology*, RECOMB '02, pages 299–308. ACM, 2002.

[282] M. Delarue and Y.-H. Sanejouand. Simplified normal mode analysis of conformational transitions in DNA-dependent polymerases: the elastic network model. *Journal of Molecular Biology*, 320:1011–1024, 2002.

[283] T. Ichiye and M. Karplus. Collective motions in proteins: a covariance analysis of

atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins: Structure, Function, and Bioinformatics*, 11(3):205–217, 1991.

[284] M. Tsytlonok and L. S. Itzhaki. Using FlAsH to probe conformational changes in a large HEAT repeat protein. *ChemBioChem*, 13(8):1199–1205, 2012.

[285] J. Pei, B.-H. Kim, and N. V. Grishin. PROMALS3D: a tool for multiple sequence and structure alignment. *Nucleic Acids Research*, 36(7):2295–2300, 2008.

[286] A. Šali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3):779–815, 1993.

[287] J. Ma. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure*, 13:373–380, 2005.

[288] A. Cooper and D. T. F. Dryden. Allostery without conformational change. *European Biophysics Journal*, 11:103–109, 1984.

[289] T. L. Rodgers, P. D. Townsend, D. Burnell, M. L. Jones, S. A. Richards, T. C. B. McLeish, E. Pohl, M. R. Wilson, and M. J. Cann. Modulation of global low-frequency motions underlies allosteric regulation: Demonstration in CRP/FNR family transcription factors. *PLoS Biology*, 11(9):e1001651, 2013.

[290] C. C. Mello and D. Barrick. An experimentally determined protein folding energy landscape. *Proceedings of the National Academy of Sciences*, 101(39):14102–14107, 2004.

[291] K. W. Tripp and D. Barrick. The tolerance of a modular protein to duplication and deletion of internal repeats. *Journal of Molecular Biology*, 344(1):169 – 178, 2004.

[292] S. J. de Vries, M. van Dijk, and A. M. Bonvin. The HADDOCK web server for data-driven biomolecular docking. *Nat Protoc*, 5(5):883–897, 2010.

[293] M. F. Lensink, S. Velankar, A. Kryshtafovych, S. Y. Huang, D. Schneidman-Duhovny, A. Sali, J. Segura, N. Fernandez-Fuentes, S. Viswanath, R. Elber, S. Grudinin, P. Popov, E. Neveu, H. Lee, M. Baek, S. Park, L. Heo, G. Rie Lee, C. Seok, S. Qin, H. X. Zhou, D. W. Ritchie, B. Maigret, M. D. Devignes, A. Ghoorah, M. Torchala, R. A. Chaleil, P. A. Bates, E. Ben-Zeev, M. Eisenstein, S. S. Negi, Z. Weng, T. Vreven, B. G. Pierce, T. M. Borrman, J. Yu, F. Ochsenbein, R. Guerois, A. Vangone, J. P. Rodrigues, G. van Zundert, M. Nellen, L. Xue, E. Karaca, A. S. Melquiond, K. Visscher, P. L. Kastritis, A. M. Bonvin, X. Xu, L. Qiu, C. Yan, J. Li, Z. Ma, J. Cheng, X. Zou, Y. Shen, L. X. Peterson, H. R. Kim, A. Roy, X. Han, J. Esquivel-Rodriguez, D. Kihara, X. Yu, N. J. Bruce, J. C. Fuller, R. C. Wade, I. Anishchenko, P. J. Kundrotas, I. A. Vakser, K. Imai, K. Yamada, T. Oda, T. Nakamura, K. Tomii, C. Pallara, M. Romero-Durana, B. Jimenez-Garcia, I. H. Moal, J. Fernandez-Recio, J. Y. Joung, J. Y. Kim, K. Joo, J. Lee, D. Kozakov, S. Vajda, S. Mottarella, D. R. Hall, D. Beglov, A. Mamonov, B. Xia, T. Bohnuud,

C. A. Del Carpio, E. Ichiishi, N. Marze, D. Kuroda, S. S. Roy Burman, J. J. Gray, E. Chermak, L. Cavallo, R. Oliva, A. Tovchigrechko, and S. J. Wodak. Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment. *Proteins*, 84 Suppl 1:323–348, 2016.

[294] S. J. de Vries, A. S. J. Melquiond, P. L. Kastritis, E. Karaca, A. Bordogna, M. van Dijk, J. a. P. G. L. M. Rodrigues, and A. M. J. J. Bonvin. Strengths and weaknesses of data-driven docking in critical assessment of prediction of interactions. *Proteins: Structure, Function, and Bioinformatics*, 78(15):3242–3249, 2010.

[295] J. P. G. L. M. Rodrigues, M. Trellet, C. Schmitz, P. Kastritis, E. Karaca, A. S. J. Melquiond, and A. M. J. J. Bonvin. Clustering biomolecular complexes by residue contacts similarity. *Proteins: Structure, Function, and Bioinformatics*, 80(7):1810–1817, 2012.

[296] M. F. Lensink, S. Velankar, M. Baek, L. Heo, C. Seok, and S. J. Wodak. The challenge of modeling protein assemblies: the CASP12-CAPRI experiment. *Proteins: Structure, Function, and Bioinformatics*, 86(S1):257–273, 2017.

[297] A. Roy, A. Kucukural, and Y. Zhang. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc*, 5(4):725–738, 2010.

[298] Y. Song, F. DiMaio, R. Y. Wang, D. Kim, C. Miles, T. Brunette, J. Thompson, and D. Baker. High-resolution comparative modeling with RosettaCM. *Structure*, 21(10):1735–1742, 2013.

[299] D. E. Kim, D. Chivian, and D. Baker. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research*, 32(suppl 2):W526–W531, 2004.

[300] R. Srivatsan, V. Robert, T. James, T. Michael, S. Ruslan, P. Jimin, K. David, K. Elizabeth, D. Frank, L. Oliver, K. Lisa, S. Will, K. Bong-Hyun, D. Rhiju, G. N. V., and B. David. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins: Structure, Function, and Bioinformatics*, 77(S9):89–99, 2009.

[301] S. Ovchinnikov, D. E. Kim, R. Y. Wang, Y. Liu, F. DiMaio, and D. Baker. Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta. *Proteins*, 84 Suppl 1:67–75, 2016.

[302] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, and Y. Zhang. The I-TASSER Suite: protein structure and function prediction. *Nature Methods*, 12:7–8, 2015.

[303] Y. Zhang. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, 9:40, 2008.

[304] S. Ovchinnikov, H. Park, D. E. Kim, F. DiMaio, and D. Baker. Protein structure prediction using rosetta in casp12. *Proteins: Structure, Function, and Bioinformatics*, 86(S1):113–121, 2017.

[305] G. Cingolani, C. Petosa, K. Weis, and C. W. Müller. Structure of importin-β bound

to the IBB domain of importin-α. *Nature*, 399(6733):221–229, 1999.

[306] P. Emsley, B. Lohkamp, W. G. Scott, and K. Cowtan. Features and development of coot. *Acta Crystallographica Section D - Biological Crystallography*, 66:486–501, 2010.

[307] F. H. Niesen, H. Berglund, and M. Vedadi. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nature Protocols*, 2:2212 EP, 2007.

[308] G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner. WebLogo: a sequence logo generator. *Genome Res.*, 14(6):1188–1190, 2004.

[309] T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, 18(20):6097–6100, 1990.

[310] E. Krissinel and K. Henrick. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D*, 60(12 Part 1):2256–2268, 2004.

[311] I. N. Shindyalov and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*, 11(9):739–747, 1998.

[312] M. Tsytlonok. *Biophysical Characterization of Linear Repeat Proteins*. PhD thesis, University of Cambridge, 2012.

[313] A. Perez-Riba and L. S. Itzhaki. A method for rapid high-throughput biophysical analysis of proteins. *Scientific Reports*, 7(1):9071, 2017.

[314] S. J. Lee, Y. Matsuura, S. M. Liu, and M. Stewart. Structural basis for nuclear import complex dissociation by RanGTP. *Nature*, 435(7042):693–696, 2005.

[315] S. J. Lee, T. Sekimoto, E. Yamashita, E. Nagoshi, A. Nakagawa, N. Imamoto, M. Yoshimura, H. Sakai, K. T. Chong, T. Tsukihara, and Y. Yoneda. The structure of importin-beta bound to SREBP-2: nuclear import of a transcription factor. *Science*, 302(5650):1571–1575, 2003.

[316] S. M. Liu and M. Stewart. Structural basis for the high-affinity binding of nucleoporin Nup1p to the Saccharomyces cerevisiae importin-beta homologue, Kap95p. *Journal of Molecular Biology*, 349(3):515–525, 2005.

[317] N. Stephanopoulos and M. B. Francis. Choosing an effective protein bioconjugation strategy. *Nature Chemical Biology*, 7:876, 2011. Review Article.

[318] W. Ott, M. A. Jobst, C. Schoeler, H. E. Gaub, and M. A. Nash. Single-molecule force spectroscopy on polyproteins and receptor-ligand complexes: The current toolbox. *Journal of Structural Biology*, 197(1):3 – 12, 2017.

[319] J. Cordova, A. Olivares, Y. Shin, B. Stinson, S. Calmat, K. Schmitz, M.-E. Aubin-

Tam, T. Baker, M. Lang, and R. Sauer. Stochastic but highly coordinated protein unfolding and translocation by the ClpXP proteolytic machine. *Cell*, 158(3):647 – 658, 2014.

[320] D. A. Pippig, F. Baumann, M. Strackharn, D. Aschenbrenner, and H. E. Gaub. Protein-DNA chimeras for nano assembly. *ACS Nano*, 8(7):6551–6555, 2014.

[321] E. Durner, W. Ott, M. A. Nash, and H. E. Gaub. Post-Translational Sortase-Mediated Attachment of High-Strength Force Spectroscopy Handles. *ACS Omega*, 2(6):3064–3069, 2017.

[322] J. Yin, P. D. Straight, S. M. McLoughlin, Z. Zhou, A. J. Lin, D. E. Golan, N. L. Kelleher, R. Kolter, and C. T. Walsh. Genetically encoded short peptide tag for versatile protein labeling by Sfp phosphopantetheinyl transferase. *Proceedings of the National Academy of Sciences*, 102(44):15815–15820, 2005.

[323] J. Yin, A. J. Lin, D. E. Golan, and C. T. Walsh. Site-specific protein labeling by Sfp phosphopantetheinyl transferase. *Nature Protocols*, 1:280, 2006.

[324] T. Nojima, H. Konno, N. Kodera, K. Seio, H. Taguchi, and M. Yoshida. Nano-scale alignment of proteins on a flexible DNA backbone. *PLoS ONE*, 7(12):1–7, 2012.

[325] M. W. Popp, J. M. Antos, and H. L. Ploegh. Site-specific protein labeling via sortase-mediated transpeptidation. *Curr Protoc Protein Sci*, Chapter 15:Unit 15.3, 2009.

[326] M. Jahn, J. Buchner, T. Hugel, and M. Rief. Folding and assembly of the large molecular machine Hsp90 studied in single-molecule experiments. *Proceedings of the National Academy of Sciences*, 113(5):1232–1237, 2016. ISSN 0027-8424.

[327] A. Mukhortava and M. Schlierf. Efficient formation of site-specific protein-dna hybrids using copper-free click chemistry. *Bioconjugate Chemistry*, 27(7):1559–1563, 2016.

[328] K. Lang and J. W. Chin. Cellular incorporation of unnatural amino acids and bioorthogonal labeling of proteins. *Chemical Reviews*, 114(9):4764–4806, 2014.

[329] K. Lang and J. W. Chin. Bioorthogonal reactions for labeling proteins. *ACS Chemical Biology*, 9(1):16–20, 2014.

[330] D. M. Patterson, L. A. Nazarova, and J. A. Prescher. Finding the right (bioorthogonal) chemistry. *ACS Chemical Biology*, 9(3):592–605, 2014.

[331] R. Huisgen. 1,3-dipolar cycloadditions. past and future. *Angewandte Chemie International Edition in English*, 2(10):565–598, 1963.

[332] S. I. Presolski, V. P. Hong, and M. Finn. Copper-catalyzed azide-alkyne click chemistry for bioconjugation. *Current Protocols in Chemical Biology*, 3:153–162, 2011.

[333] C. W. Tornøe, C. Christensen, and M. Meldal. Peptidotriazoles on solid phase: [1,2,3]-triazoles by regiospecific copper(I)-catalyzed 1,3-dipolar cycloadditions of terminal alkynes to azides. *Journal of Organic Chemistry*, 67(9):3057 – 3064, 2002.

[334] V. V. Rostovtsev, L. G. Green, V. V. Fokin, and K. B. Sharpless. A stepwise Huisgen cycloaddition process: Copper(I)-catalyzed regioselective "ligation" of azides and terminal alkynes. *Angewandte Chemie*, 114(14):2708–2711, 2002.

[335] M. Simon, U. Zangemeister-Wittke, and A. Plückthun. Facile double-functionalization of designed ankyrin repeat proteins using click and thiol chemistries. *Bioconjugate Chemistry*, 23(2):279–286, 2012.

[336] A. Sachdeva, K. Wang, T. Elliott, and J. W. Chin. Concerted, rapid, quantitative, and site-specific dual labeling of proteins. *Journal of the American Chemical Society*, 136(22):7785–7788, 2014.

[337] Y. H. Lau, P. de Andrade, S.-T. Quah, M. Rossmann, L. Laraia, N. Skold, T. J. Sum, P. J. E. Rowling, T. L. Joseph, C. Verma, M. Hyvonen, L. S. Itzhaki, A. R. Venkitaraman, C. J. Brown, D. P. Lane, and D. R. Spring. Functionalised staple linkages for modulating the cellular activity of stapled peptides. *Chemical Science*, 5:1804–1809, 2014.

[338] D. M. Abdeljabbar, F. J. Piscotta, S. Zhang, and A. James Link. Protein stapling via azide-alkyne ligation. *Chemical Communications*, 50:14900–14903, 2014.

[339] T. S. Elliott, F. M. Townsley, A. Bianco, R. J. Ernst, A. Sachdeva, S. J. Elsässer, L. Davis, K. Lang, R. Pisa, S. Greiss, , K. S. Lilley, and J. W. Chin. Proteome labelling and protein identification in specific tissues and at specific developmental stages in an animal. *Nature biotechnology*, 32(5):465–472, 2014.

[340] J. Yang, J. Šečkutė, C. M. Cole, and N. K. Devaraj. Live-cell imaging of cyclopropene tags with fluorogenic tetrazine cycloadditions. *Angewandte Chemie International Edition*, 51(30):7476–7479, 2012.

[341] D. M. Patterson, L. A. Nazarova, B. Xie, D. N. Kamber, and J. A. Prescher. Functionalized cyclopropenes as bioorthogonal chemical reporters. *Journal of the American Chemical Society*, 134(45):18638–18643, 2012.

[342] W. Liu, A. Brock, S. Chen, S. Chen, and P. G. Schultz. Genetic incorporation of unnatural amino acids into proteins in mammalian cells. *Nature methods*, 4(3): 239–244, 2007.

[343] J. W. Chin. Expanding and reprogramming the genetic code of cells and animals. *Annual Review of Biochemistry*, 83(1):379–408, 2014.

[344] S. M. Hancock, R. Uprety, A. Deiters, and J. W. Chin. Expanding the genetic code of yeast for incorporation of diverse unnatural amino acids via a pyrrolysyl-tRNA synthetase/tRNA pair. *Journal of the American Chemical Society*, 132(42):

14819–14824, 2010.

[345] J. W. Chin, S. W. Santoro, A. B. Martin, D. S. King, L. Wang, and P. G. Schultz. Addition of p-azido-l-phenylalanine to the genetic code of escherichia coli. *Journal of the American Chemical Society*, 124(31):9026–9027, 2002.

[346] H. Neumann, K. Wang, L. Davis, M. Garcia-Alai, and J. W. Chin. Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome. *Nature*, 464:441–444, 2010.

[347] O. Rackham and J. W. Chin. A network of orthogonal ribosome· mRNA pairs. *Nature Chemical Biology*, 1(3):159–166, 2005.

[348] K. Wang, H. Neumann, S. Y. Peak-Chew, and J. W. Chin. Evolved orthogonal ribosomes enhance the efficiency of synthetic genetic code expansion. *Nature biotechnology*, 25(7):770–777, 2007.

[349] A. Deiters and P. G. Schultz. In vivo incorporation of an alkyne into proteins in escherichia coli. *Bioorganic & Medicinal Chemistry Letters*, 15(5):1521–1524, 2005.

[350] D. P. Nguyen, H. Lusic, H. Neumann, P. B. Kapadnis, A. Deiters, and J. W. Chin. Genetic encoding and labeling of aliphatic azides and alkynes in recombinant proteins via a pyrrolysyl-tRNA synthetase/tRNA$_{CUA}$ pair and click chemistry. *Journal of the American Chemical Society*, 131(25):8720–8721, 2009.

[351] K. C. Neuman and A. Nagy. Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy. *Nature Methods*, 5(6):491–505, 2008.

[352] V. Hong, S. Presolski, C. Ma, and M. Finn. Analysis and optimization of copper-catalyzed azide-alkyne cycloaddition for bioconjugation. *Angewandte Chemie International Edition*, 48(52):9879–9883, 2009.

[353] P. P. Constantinides and J. M. Steim. Solubility of palmitoyl-coenzyme a in acyl-transferase assay buffers containing magnesium ions. *Archives of Biochemistry and Biophysics*, 250(1):267 – 270, 1986.

[354] A. R. Lowe, A. Perez-Riba, L. S. Itzhaki, and E. R. Main. Pyfolding: Open-source graphing, simulation, and analysis of the biophysical properties of proteins. *Biophysical Journal*, 114(3):511–521, 2018.

[355] T. Aksel and D. Barrick. Analysis of repeat-protein folding using nearest-neighbor statistical mechanical models. In M. L. Johnson, J. M. Holt, and G. K. Ackers, editors, *Biothermodynamics, Part A*, volume 455 of *Methods in Enzymology*, chapter 4, pages 95–125. Academic Press, 2009.

[356] A. K. Das, P. W. Cohen, and D. Barford. The structure of the tetratricopeptide repeats of protein phosphatase 5: implications for TPR-mediated protein-protein interactions. *EMBO J.*, 17(5):1192–1199, 1998.

[357] E. M. Puchner, G. Franzen, M. Gautel, and H. E. Gaub. Comparing proteins by their unfolding pattern. *Biophysical Journal*, 95(1):426–434, 2008.

[358] M. Rief and H. Grubmüller. Force spectroscopy of single biomolecules. *Chemphyschem*, 3(3):255–261, 2002.

[359] A. L. Cortajarena, G. Lois, E. Sherman, C. S. O'Hern, L. Regan, and G. Haran. Non-random-coil behavior as a consequence of extensive PPII structure in the denatured state. *Journal of Molecular Biology*, 382(1):203–212, 2008.

[360] J. P. Junker, F. Ziegler, and M. Rief. Ligand-dependent equilibrium fluctuations of single calmodulin molecules. *Science*, 323(5914):633–637, 2009.

[361] E. N. Korkmaz, K. C. Taylor, M. P. Andreas, G. Ajay, N. T. Heinze, Q. Cui, and I. Rayment. A composite approach towards a complete model of the myosin rod. *Proteins*, 84(1):172–189, 2016.

[362] P. Ringer, A. Weissl, A. L. Cost, A. Freikamp, B. Sabass, A. Mehlich, M. Tramier, M. Rief, and C. Grashoff. Multiplexing molecular tension sensors reveals piconewton force gradient across talin-1. *Nature Methods*, 14(11):1090–1096, 2017.

[363] M. Jahn, K. Tych, H. Girstmair, M. Steinmaßl, T. Hugel, J. Buchner, and M. Rief. Folding and domain interactions of three orthologs of Hsp90 studied by single-molecule force spectroscopy. *Structure*, 26(1):96 – 105.e4, 2018.

[364] O. Miyashita, J. N. Onuchic, and P. G. Wolynes. Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proceedings of the National Academy of Sciences*, 100(22):12570–12575, 2003.

[365] M. Habibi, S. S. Plotkin, and J. Rottler. Soft vibrational modes predict breaking events during force-induced protein unfolding. *Biophysical Journal*, 114(3):562–569, 2018.

[366] K. Kotlo, Y. Xing, S. Lather, J. M. Grillon, K. Johnson, R. A. Skidgel, R. J. Solaro, and R. S. Danziger. PR65A phosphorylation regulates PP2A complex signaling. *PLoS ONE*, 9(1):e85000, 2014.

[367] B. Hao, N. Zheng, B. A. Schulman, G. Wu, J. J. Miller, M. Pagano, and N. P. Pavletich. Structural basis of the Cks1-dependent recognition of p27Kip1 by the SCFSkp2 ubiquitin ligase. *Molecular Cell*, 20(1):9 – 19, 2005.

[368] J. Liu and R. Nussinov. Molecular dynamics reveal the essential role of linker motions in the function of cullin-RING E3 ligases. *Journal of Molecular Biology*, 396(5):1508 – 1523, 2010.

[369] J. Liu and R. Nussinov. Flexible cullins in cullin-RING E3 ligases allosterically regulate ubiquitination. *Journal of Biological Chemistry*, 286(47):40934–40942, 2011.

[370] S. Wheaton, R. M. Gelfand, and R. Gordon. Probing the raman-active acoustic vibrations of nanoparticles with extraordinary spectral resolution. *Nature Photonics*,

9:68, 2014.

[371] T. DeWolf and R. Gordon. Theory of acoustic raman modes in proteins. *Physical Review Letters*, 117:138101, 2016.

[372] A. Freikamp, A.-L. Cost, and C. Grashoff. The piconewton force awakens: Quantifying mechanics in cells. *Trends in Cell Biology*, 26(11):838 – 847, 2016. Special Issue: Future of Cell Biology.

[373] D. Kamiyama, S. Sekine, B. Barsi-Rhyne, J. Hu, B. Chen, L. A. Gilbert, H. Ishikawa, M. D. Leonetti, W. F. Marshall, J. S. Weissman, and B. Huang. Versatile protein tagging in cells with split fluorescent protein. *Nat Commun*, 7:11046, 2016.