Comparative analysis of the folding dynamics and kinetics of an engineered knotted protein and its variants derived from HP0242 of *Helicobacter pylori*

Liang-Wei Wang,[1,2,3] Yu-Nan Liu,[3] Ping-Chiang Lyu,[3] Sophie E. Jackson[4] and Shang-Te Danny Hsu[1,2,3,*]

1. Institute of Biological Chemistry, Academia Sinica, Taipei 11529, Taiwan
2. Institute of Biochemical Science, National Taiwan University, Taipei 116, Taiwan
3. Institute of Bioinformatics and Structural Biology, National Tsing Hua University, Hsinchu 30013, Taiwan
4. Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom

* Corresponding author: sthsu@gate.sinica.edu.tw

Abstract

Understanding the mechanism by which a polypeptide chain thread itself spontaneously to attain a knotted conformation has been a major challenge in the field of protein folding. HP0242 is a homodimeric protein from *Helicobacter pylori* with intertwined helices to form a unique pseudo-knotted folding topology. A tandem HP0242 repeat has been constructed to become the first engineered trefoil-knotted protein. Its small size renders it a model system for computational analyses to examine its folding and knotting pathways. Here we report a multi-parametric study on the folding stability and kinetics of a library of HP0242 variants, including the trefoil-knotted tandem HP0242 repeat, using far-UV circular dichroism and fluorescence spectroscopy. Equilibrium chemical denaturation of HP0242 variants shows the presence of highly populated dimeric and structurally heterogeneous folding intermediates. Such equilibrium folding intermediates retain significant amount of helical structures except those at the N- and C-terminal regions in the native structure. Stopped-flow fluorescence measurements of HP0242 variants show that spontaneous refolding into knotted structures can be achieved within seconds, which is several orders of magnitude faster than previously observed for other knotted proteins. Nevertheless, the complex chevron plots indicate that HP0242 variants are prone to misfold into kinetic traps, leading to severely rolled-over refolding arms. The experimental observations are in general agreement with the previously reported molecular dynamics simulations. Based on our results, kinetic folding pathways are proposed to qualitatively describe the complex folding processes of HP0242 variants.

1. Introduction

About two decades ago, knotted proteins were inconceivable for many biochemists, when Mansfield first identified the presence of knotted backbone structure in human carbonic anhydrase B out of 400 protein structures in the protein database (PDB) [1]. Since then, a number of algorithms have been developed [2-5] and systematic surveys of PDB entries have identified hundreds of knotted proteins [6-9]. While the biological functions of these protein knots have not been firmly established, the question of "how do proteins form topological knots?" has attracted a lot of attentions from biophysicists [10-12]. It has been demonstrated by using a reconstituted *in vitro* protein translation system that knotted proteins can attain their natively knotted conformations states *de novo* without the help of molecular chaperones [13]. These include YibK from *H. influenzae* [14] and YbeA from *E. coli* [15], both of which contain a deep trefoil (3_1) knot. Furthermore, a number of *in vitro* refolding experiments showed that, in addition to YibK and YbeA, purified human ubiquitin C-terminal hydrolyases, UCH-L1 [16] and UCH-L3 [17], both of which contain a Gordian (5_2) knot, and more recently MJ0366 from MJ0366 from *M. jannaschii,* which is the smallest trefoil-knotted protein known to date [18], can refold efficiently and spontaneously from chemically denatured states without auxiliary factors. Importantly, a recent study demonstrated that YibK and YbeA in fact remain knotted in their chemically denatured states [19] despite the lack of secondary and tertiary structure according to fluorescence, far-UV circular dichroism (CD), and solution state nuclear magnetic resonance (NMR) spectroscopy [20, 21]. This raised the question as to how highly disordered polypeptides manage to maintain their knotted backbone topologies under denaturing conditions.

The folding pathways of some of the knotted proteins have been delineated experimentally. Although YibK and YbeA share the same type of deep trefoil knot and are of similar size, the former exhibits complex bifurcated folding pathways with multiple kinetic folding intermediates while the latter exhibits a simple linear folding pathway with a single kinetic folding intermediate [15]. Meanwhile, UCH-L3 exhibits two parallel folding pathways with distinct hyper-fluorescent intermediates [17]. UCH-L1 exhibits multiple unfolding kinetic phases similar to those of UCH-L3, but its folding pathway has not been fully established [16]. Although the number of experimentally characterized knotted proteins is limited, all these knotted proteins exhibit well-defined kinetic folding intermediates. Additionally, their folding kinetics is much slower than the typical values of proteins of similar sizes, suggesting that the formation of knotted structure may generate kinetic traps that hinder the native state formation.

Computational analyses of the folding pathways of several knotted proteins have been reported. While initial attempts required biased force fields to facilitate successful folding events [22], improvements in force fields and computation power have enabled detailed characterizations of the folding pathways of a number of knotted proteins [23-27]. According to these simulations, many knotted proteins exhibit parallel folding pathways. Additionally, slip-knotting is often observed in folding intermediates, which requires multiple back-tracking steps to overcome such a misfolding kinetic trap. Consequently, correct knot formation often occurs in the transition state at which significant native contacts (Q) are formed.

Of many knotted proteins that have been subjected to detailed simulation studies, HP0242 from *Helicobacter pylori* is one of the most studied model systems [25, 27]. The open reading frame of HP0242 is located downstream of napA

(HP0243), which is a ferritin-like protein involving in iron storage and heme metabolism. It was therefore postulated to be associated with iron metabolism [28]. The crystal structure of HP0242 has been determined independently by two structural genomics initiatives (PDB ID: 2BO3 and 4U12) [29]. HP0242 homo-dimerises through leucine-zipper mediated hydrophobic interactions between the two long helices H2 while helices H1 and H4 intertwine to form a pseudoknot around the only tryptophan residue, W18. While most of its surface is highly negatively charged, the knotting region is positively charged at neutral pH values (Figure 1). Therefore, folding stability of the local structure may be pH-sensitive. Note that HP0242 further assembles into a higher order tertiary structure in the crystalline state by forming a dimer of dimers through electrostatic interactions between the homodimers. In solution, however, HP0242 exists predominantly as homodimers with only a small fraction existing as tetramers, as demonstrated by analytical ultracentrifugation [28]. The unique folding topology was recognized by the SCOP (structural classification of proteins) database [30] as an HP0242-like fold of its own right. Yeates and co-workers have recently constructed an engineered tandem HP0242 variant by introducing a covalent linkage between the C-terminus of one monomer to the N-terminus of the other monomer, which represents the first example of an engineered protein knot (Figure 1) [31]. Using structure-based coarse-grain simulations, two recent studies independently showed that tandem HP0242 exhibits multiple folding intermediates en route to the natively knotted conformation and that these folding intermediates correspond to structures that contain either slip-knots or wrong chirality [25, 27].

Equilibrium unfolding of HP0242 by chemical denaturation exhibits different unfolding transition patterns when monitored by intrinsic fluorescence and far-UV CD

spectroscopy: the former identified a single transition, corresponding to the native-to-intermediate (N-I) transition while the latter identified an additional transition that correspond to the intermediate-to-denatured (I-D) transition, indicating the highly populated intermediate retains significantly amount of residual helical structure [31]. A similar three-state transition was observed for tandem HP0242, but its intermediate state is much more stable than that of wild type (wt), leading to a significant shift in the transition point that corresponds to the N-I and I-D transitions. Furthermore, all HP0242 variants exhibit strongly rolled-over refolding arms, indicative of the presence of folding intermediates. Lastly, the observed unfolding rates of tandem HP0242 are significantly faster than those of wt. While the transition point of the I-D transition point shifted to higher denaturant concentration with higher wt protein concentration indicated that the intermediate remains dimeric, the exact folding pathways of HP0242 variants have not been established.

Here we report the use of a multi-parametric approach to examine in detail the folding equilibria and kinetics of HP0242 and its variants. Using site-directed mutagenesis, we demonstrate that HP0242 unfolds in two stages during which the first helix (H1) and part of the last helix (H4) unfold first followed by dissociation of the two monomers. Our results further indicate that the folding pathway of HP0242 involves the formation of multiple non-productive, misfolded intermediates that exist in equilibrium with the productive folding intermediate, which is separated from the denatured state by a much higher kinetic barrier such that conformational rearrangements between different intermediate states are required before successful refolding into native conformations. These detailed experimental observations are consistent with previously reported models [25, 27].

2. Material and Methods

2.1. Preparation of recombinant HP0242 variants

The DNA sequence encoding the open reading frame of HP0242 has been cloned into a pET30a vector as described previously [29]. Site-directed mutagenesis of F57W, F57W/W18F and F60W/W18F were generated by PCR. The expression and purification of recombinant HP0242 variants, namely wt, F57W, F57W/W18F and F60W/W18F, were carried out following the protocol as described previously [28]. Briefly, recombinant HP0242 variants were expressed in BL21 (DE3) *E. coli* strain and purified by Ni-NTA resin (GE health care, U.S.A) followed by Factor-X (Novagen) cleavage of the N-terminal fusion tag and finally by size-exclusion chromatography (Superdex 75 26/60, GE health care, U.S.A). The purity was assessed by SDS-PAGE and was estimated to be higher than 95% by visual inspection. For tandem HP0242, the DNA sequence encoding for the open reading frame of the 2ouf-fold [30] was synthesized (GenScript, USA) and subcloned into a modified pET-z2 vector between NcoI and KpnI sites, resulting a fusion with an His-tag at the N-terminus followed by a z2-domain fusion tag, a TEV protease cleavage site and the MJ0366 reading frame. The plasmid was transformed into the BL21(DE3) E. *coli* strain. Purification of tandem HP0242 follows the same procedure as that for wt except that the use of Factor-X for tag removal was replaced by TEV protease.

2.2. Analytical size-exclusion chromatography coupled with multiangle light scattering (SEC-MALS)

The oligomeric state of HP0242 was determined by SEC/MASL using a Wyatt Dawn Heleos II multiangle light scattering detector (Wyatt Technology) coupled to an AKTA Purifier UPC10 FPLC protein purification system using a Superdex 75 5/150 GL size-exclusion column (GE Healthcare) [18, 32]. Ten microlitres of HP0242 (2 mg/ml) in the absence or presence of 4M GdnHCl a buffer containing 10 mM sodium phosphate (pH 6.5) were used for SEC-MALS analyses. BSA (2mg/mL) was used as a standard to calibrate the system. The absolute molecular weight of the observed elution peak was determined by combining the multiangle static light scattering profiles with the corresponding refractive index determined by a refractometer (Wyatt Optilab rEX, connected downstream of the LS detector). A standard value of refractive index, dn/dc = 0.185 mL/g, was used for proteins and the buffer viscosity η = 1.0257 cP at 25 ºC was calculated using SEDNTERP [32]. The value of reference refractive index, 1.3458 RIU, was taken directly from the measurement of the Wyatt Optilab rEX when buffer only passing through the reference cell. The viscosity and reference refractive index of buffer containing 4 M GdnHCl were derived and applied by the same ways.

2.3. Intrinsic fluorescence spectroscopy

The intrinsic fluorescence of HP0242 was monitored by exciting the samples at 260 nm and recording the emission spectra between 300 and 450 nm using a temperature-controlled fluorimeter (FP8500, Jasco, Japan). For equilibrium unfolding of HP0242 variants by GdnHCl, 41 aliquots of protein solution were prepared containing a gradient of GdnHCl (0-8M) concentrations with a linear increment step of 2.5 %. A two-channel liquid dispenser (Hamilton, USA) was used to generate the

41-point denaturant gradient to minimize pipetting errors. The fluorescence anisotropy of HP0242 variants as a function GdnHCl concentration was determined by using the same fluorimeter equipped with two linear polarizing filters to generate polarized excitation light that is vertical or horizontal to the polarizer, and to record polarized emission light that is vertical or horizontal to the polarizer.

2.4. Far-UV CD spectroscopy

Protein stock solution was diluted to 10-15 μM and 300 μL in volume for CD measurements using a quartz cuvette with a path-length of 0.1 cm (Hellma, Germany). The far-UV CD spectra were recorded between 200 and 260 nm using a CD spectrometer (J-815, JASCO, Japan). The samples were temperature-controlled at 25 ºC and the data were collected with a bandwidth of 1 nm, a data interval of 0.5 nm, and an averaging time of one second. The molar ellipticity ($deg \cdot cm^2 \cdot dmol^{-1}$) was calculated by using the following equation: $[\theta] = \theta / (10 \times C \times l \times n)$, where θ is the CD signal in the unit of millidegree, C the molar concentration in M, l the cell path-length in cm, and n the number of amino acid.

2.5. ANS fluorescence spectroscopy

To probe the presence of molten globule that contains clusters of hydrophobic patches of partially unfolded protein structures, 8-Anilinonaphthalene-1-sulfonic acid (ANS) was added to the protein solution at a final concentration of 93 μM in the presence of 11.25 μM HP0242 variants, which yielded a dye to protein ratio of 9:1.

2.6. Analysis of equilibrium unfolding data

All equilibrium unfolding data were subjected to singular value decomposition (SVD) analysis using MATLAB (MATLAB and Statistics Toolbox Release 2012b, The MathWorks, U.S.A.) to determine the number of states associated with the

equilibrium unfolding. The titration series were used to generate an m x n matrix, M, as inputs for SVD analysis, of which m corresponds to the number of recorded wavelengths and n corresponds to the number of titration points, *i.e.*, 41 denaturant concentrations. After SVD, a m x m unitary rotation matrix U_r, a n x n unitary rotation matrix V_r and a m x n diagonal singular value matrix S could be obtained to reconstruct the input matrix M as follows

$$M = U_r S V_r^*$$

where U_r corresponds to the basis function of the observed spectral patterns, and V_r corresponds to the basis coefficient, *i.e.*, changes of individual SVD components, as a function of denaturant concentration. To assess the number of significant components that is necessary for reconstructing the input matrix, the normalized correlation coefficients of individual singular values were calculated and a minimum threshold of 0.8 was used for filtering. Importantly, the SVD algorithm will identify the number of distinct spectra. If two or more states have very similar spectra it may not necessarily identify them as separate states, which is the case for dome of the datasets (Figure 3).

All equilibrium unfolding data were subjected to a two-state or three-state equilibrium-unfolding model as described in ref [16] depending on the number of significant SVD components. During the fitting, $m_{N_2 \leftrightarrow I_2}$, $[D]_{50\%,N_2 \leftrightarrow I_2}$ and $m_{I_2 \leftrightarrow 2D}$ were shared while $[D]_{50\%,I_2 \leftrightarrow 2D}$, which are protein concentration-dependent, were fit individually to different protein concentrations. The intrinsic fluorescence and far-UV CD data were fit separately, and the results are tabulated in Table 1. Additionally, the

changes of maximum fluorescence emission wavelength as a function of denaturant concentration were fitted to the same model for comparison.

2.7. Stopped-flow fluorescence spectroscopy

The folding kinetics of HP0242 variants was monitored by using a stopped-flow spectrometer in fluorescence detection mode (SX18 stopped-flow spectrometer, Applied Photophysics, UK) as described previously [34]. Changes in total fluorescence were monitored using an excitation wavelength of 280 nm with a 320 nm cut-off filter. All experiments were carried out at 298K. HP0242 variants (the protein concentration was 20 μM unless specified otherwise) were buffered in 10 mM sodium phosphate (pH 4.8 or pH 6.5), or 50 mM Tris-HCl (pH 8.0) with or without 8M GdnHCl for unfolding or refolding measurements, respectively. For single-jump kinetic measurements, the folding reactions were triggered by rapidly mixing the folding or unfolding buffer to protein solution with an asymmetric mixing ratio of 10:1. After the 11-fold dilution, the final protein concentration was 1.8 μM. For double-jump kinetic measurements, three different modes were employed, namely, interrupted refolding, interrupted unfolding or sequential refolding kinetic measurement.

For the interrupted refolding (D-N-D) measurements, 8M GdnHCl-denatured protein solution was first mixed with refolding buffer to initiate refolding over different aging periods before the second mixing step with denaturing buffer immediately before kinetic fluorescence measurements under denaturing conditions. For the interrupted unfolding (N-I-N) experiments, native protein solution was mixed with unfolding buffer incubate the protein solution under intermediate conditions before the second mixing step with refolding buffer to allow kinetic measurements under native conditions. For the sequential refolding (D-I-N) experiments, 7M GdnHCl-denatured protein solution was mixed with refolding buffer to allow intermediates to

build up during the aging periods, followed by a second step of refolding buffer dilution for kinetic fluorescence measurements under native conditions. In all three different modes of double-jump measurements, the first mixing step was achieved by mixing protein solutions with unfolding or refolding buffer with a 1:5 mixing ratio. After the aging period, a second mixing step was carried out by mixing the aged protein solution with equal amount of unfolding or refolding buffer (1:1 mixing ratio) immediately before kinetic fluorescence measurements. The resulting kinetic traces were subjected to global fitting in which the folding rates were shared amongst all the kinetic traces of different aging times, and the amplitudes corresponding to different kinetic phases were shown as a function of aging time and the rates at which the corresponding amplitudes built up were extracted by single exponential fitting to the observed amplitude as a function of aging time.

For all kinetic analyses, the observed reaction rates were extracted by fitting the kinetic traces to a single, double or triple exponential function with an offset using the software package GraphPad Prism (GraphPad Software, USA). The choice of model, *i.e.*, number of kinetic phases, was decided using the F-test statistics by Prism.

2.8. Chevron plot analysis

The observed reaction rates in the stopped-flow fluorescence measurements were fit to a simple two-state folding model or an off-pathway three-state folding model to extract the associated kinetic and thermodynamic parameters. For the two-state model, the observed reaction rates, k_{obs}, with a linear refolding arm and a linear unfolding arm, was fit to the following equation [34]:

$$k_{obs} = k_f^{H_2O} \cdot e^{\left(-m_f \cdot [D]_{50\%} / RT\right)} + k_u^{H_2O} \cdot e^{\left(-m_u \cdot [D]_{50\%} / RT\right)}$$

where $k_f^{H_2O}$ and $k_u^{H_2O}$ are the folding and unfolding rate in the water, *i.e.*, in the absence of denaturant, m_f and m_u are the m-values associated with folding and unfolding, R is the gas constant and T is the sample temperature, which is 298 K. Theses parameters are subsequently used to calculate the free energy of unfolding associated with this reaction is follows:

$$\Delta G_{N-D} = -RT\ln\left(\frac{k_u^{H_2O}}{k_f^{H_2O}}\right) = -RT(\ln k_u^{H_2O} - \ln k_f^{H_2O})$$

For a chevron plot with a strongly rolled-over refolding arm, the observed reaction rates, k_{obs}, were fit to the following three-state folding model with an off-pathway folding intermediate [35, 36, 37]:

$$k_{obs} = \frac{k_f^{H_2O} \cdot e^{\left(-m_f \cdot [D]_{50\%} / RT\right)}}{1+K} + k_u^{H_2O} \cdot e^{\left(-m_u \cdot [D]_{50\%} / RT\right)}$$

where K is the equilibrium constant between the denatured and intermediate states.

3. Results

3.1. Equilibrium unfolding of HP0242 variants

Equilibrium unfolding of wt and tandem HP0242 has been characterised by intrinsic fluorescence and far-UV CD spectroscopy by Yeates and co-workers [31]. Both variants were found to exhibit two distinct transitions during GdnHCl-induced chemical denaturation when monitored by far-UV CD spectroscopy; however, only the first transition, which is attributed to the N-I transition, could be monitored by intrinsic fluorescence spectroscopy. Since HP0242 contains only one tryptophan residue, W18, at the N-terminal part of the primary sequence, which is positioned in close proximity to the intertwined knotted region (Figure 1), it is likely that the intermediate maintain substantial amount of residual structure that is highly helical

while the knotted element becomes unfolded. It was further demonstrated that the second transition point for wt is protein concentration-dependent, implying that the intermediate is dimeric and only dissociates into monomers when the protein is fully denatured [31]. Using analytical size-exclusion chromatography-coupled with multiangle light scattering (SEC-MALS), confirm that HP0242 is indeed dimeric in the absence and presence of 4M GdnHCl, *i.e.*, the intermediate is dimeric (Figure 2). In order to extract further information regarding the thermodynamics parameters and conformational changes associated with the equilibrium unfolding of HP0242 variants, additional mutants were generated for multi-parametric analyses.

We first created a double-tryptophan variant of HP0242 by introducing a phenylalanine-to-tryptophan mutation of at position 57 (F57W), which is located on the opposite side of the three-dimension (3D) structure of HP0242 to monitor tertiary structural changes in this part of the structure by intrinsic fluorescence (Figure 1). We subsequently replaced the endogenous tryptophan at position 18 with a phenylalanine (W18F) in addition to the F57W mutation to generate a single-tryptophan double mutant, F57W/W18F; a similar double mutant, F60W/W18F, was generated by introducing a different phenylalanine-to-tryptophan mutation at position 60 (F60W) in addition to the W18F mutation. Together with wt and tandem HP0242, we have in total five variants for spectroscopic characterisations.

Consistent with the previously reported data, intrinsic fluorescence showed that wt and tandem HP0242 exhibit a single transition at *ca.* 2 M GdnHCl during equilibrium unfolding (Figure 3 and Table 1). Their intrinsic fluorescence was quenched upon the addition of GdnHCl, which was accompanied by significant red shift of the maximum emission wavelength (λ_{max}), indicative of solvent exposure of the chromophore due to unfolding. In the case of F57W, however, its intrinsic

fluorescence exhibited two distinct transitions during chemical denaturation: the second transition was protein concentration-dependent, consistent with the results obtained by far-UV CD spectroscopy (Figure 4 and Table 1).

The result indicates that further unfolding around the additional tryptophan side-chain at position 57 is associated with the I-D transition. Indeed, F57W/W18F exhibited significant fluorescence changes between 3 and 6 M GdnHCl, which were also protein concentration-dependent; the fluorescence-derived transition points followed a similar trend as those derived from far-UV CD although their values were consistently smaller (Figure 4 and Table 1). Unlike the endogenous W18, however, their fluorescence was not quenched upon chemical denaturation; instead, both double mutants exhibited increased fluorescence in their denatured states. Additionally, while the intrinsic fluorescence of F57W/W18F exhibited an apparently two-state transition according to singular value decomposition (SVD) analysis, that of F60W/W18F exhibited a well-defined three-state transition process. This can be rationalised structurally. While the side-chain of F57 is positioned at the dimer interface, encompassing the two long helices H2, that of F60 is in contact with helices H2, H3 and H4. Therefore, the engineered tryptophan side-chain at position 60, F60W, may be sensitive to the N-I transition, which affects the structure of helix 4, as well as the I-D transition, which requires the dissociation of helix H2. Note that the changes in λ_{max} are also very sensitive to the unfolding events in HP0242 variants. A simple analysis of the observed λ_{max} as a function of denaturant concentration yielded very similar two- and three-state transition profiles as those derived from a more sophisticated SVD analyses (Figure 5 and Table 3).

As the folding intermediates are dimeric, except for tandem HP0242, the three-state model equilibrium-unfolding model [16], which assumed a unimolecular

three-state equilibrium, is not the exact functional form for describing the dimeric intermediate unfolding process. We therefore fitted the same dataset to a three-state dimeric intermediate model as defined in Scheme 2 in reference [14], which consists of a native dimer (N_2) into a dimeric intermediate (I_2) and unfolded monomer (2D):

$$N_2 \xleftrightarrow{K_1} I_2 \xleftrightarrow{K_2} 2D$$

Global fitting of the protein concentration-dependent equilibrium unfolding dataset, however, requires well-defined baselines, which can be difficult for most of the current datasets. The fitting results hence showed significantly larger deviations, particularly for F57W/W18F and F60W/W18F (Table 3 and Supporting Information Figure 1). In contrast, the fitting results of wt derived from far-UV CD are of excellent quality. Despite the uncertainties in those tryptophan variants of HP0242, the fitting results indicate that the dimeric intermediate is highly stable for all HP0242 variants. In the case of wt, the free energy of unfolding an intermediate dimer, $\Delta G^{H_2O}_{I_2 \leftrightarrow 2D}$, is *ca.* 8 kcal mol^{-1}, which is almost twice as high as the free energy of unfolding a native dimer into an intermediate dimer, $\Delta G^{H_2O}_{N_2 \leftrightarrow I_2}$ (Table 3). Additionally, while the W18F mutation significantly reduces the stability of the native dimer, subsequent unfolding of the intermediate dimer requires similar amount of free energy for all tryptophan mutants as that required for wt.

In order to probe the local dynamics around the tryptophan side-chains, we measured the fluorescence anisotropy of all five HP0242 variants as a function of denaturant concentration (third column in Figure 3). Interestingly, the intrinsic fluorescence anisotropy of wt revealed a second transition at around 5 M GdnHCl in addition to the first transition at around 2.5 M GdnHCl. Meanwhile, the intrinsic fluorescence anisotropy of tandem HP0242 appeared to be sensitive only to the I-D transition but not the N-I transition. In the case of F57W, two distinct transitions were

observed in its intrinsic fluorescence anisotropy while that of F57W/W18F and F60W/W18F was only sensitive to the I-D transition. Together, these results indicate that while the degree of solvent exposure for the wt tryptophan side-chain levelled off after the N-I transition, its dynamics in the intermediate state is different from that in the denatured state, and that the dynamics for the second tryptophan side-chain in F57W changes in the same trend as that of the endogenous tryptophan, W18, implying a highly cooperative unfolding process for the I-D transition. It is unclear why the I-D transitions of F57W/W18F and F60W/W18F are significantly different from each other in terms of their transition points, $[D]_{50\%}$, and cooperativity, m-value.

In the case of F60W/W18F, its intrinsic fluorescence anisotropy was significantly higher than those of the other variants under native and intermediate conditions, suggesting that the mobility of the tryptophan side-chain is much lower than that of the other variants. In other words, the triangular structural motif formed by helices H2, H3 and H4 may be highly stable even under intermediate conditions. An alternative explanation to the increase of fluorescence anisotropy is that the shortening of fluorescence lifetime due to quenching by a charged group, which is probable as the side-chain F60W is in close proximity to the positively charged surface in the native state (Figure 1). Taken together, the use of fluorescence anisotropy enables direct access to the internal dynamics of HP0242 variants during their equilibrium unfolding processes and the results can provide additional structural information that is inaccessible to intrinsic fluorescence, which is similar but not exactly the same as the changes in secondary structures as reflected in far-UV CD spectroscopy (fourth and fifth columns in Figure 3). According to intrinsic fluorescence, unfolding of the intermediate tertiary structure precedes that of the residual helical structure as monitored by far-UV CD spectroscopy, leading to a

significant shift in the apparent I-D transition points between the two structural probes by 0.3-1 M GdnHCl (Figure 4).

To further characterise the intermediate and denatured states of HP042 variants, ANS fluorescence was used to probe residual structure that may be molten globule-like (sixth and seventh columns in Figure 3). While the GdnHCl titration series of ANS fluorescence of wt, tandem repeat and F57W were very similar with a simple increase in ANS fluorescence without significant blue or red shifts, those of F57W/W18F and F60W/W18F were distinctly different. F57W/W18F exhibited significant loss in ANS fluorescence when the concentration of GdnHCl was much lower than what was required for the I-D transition. This was accompanied by a large red shift of *ca.* 100 nm while F60W/W18F exhibited marginal increase in ANS fluorescence also before the I-D transition, which was accompanied by a smaller red shift of *ca.* 50 nm. These results suggest that the replacement of the endogenous tryptophan side-chain with a phenylalanine results in substantial structural perturbation within the knotted region as evidenced by the marked changes in ANS fluorescence in F57W/W18F and F60W/W18F. According to far-UV CD spectroscopy, the loss of the W18F point mutation leads to significant partial unfolding of F57W/W18F and F60W/W18F even in the absence of chemical denaturant (last column in Figure 3). Furthermore, while the N-I transition of tandem HP0242 is similar to that of wt, it requires much more GdnHCl to reach the I-D transition, indicating that the covalent linkage between the two monomers, which results in a closed loop through which the N-terminal helix H1 threads through to form an authentic trefoil protein knot, contributes substantially to its folding stability.

3.2. Folding kinetics of HP0242 variants

3.2.1 Single-jump stopped-flow fluorescence analysis

To delineate the kinetic folding pathways of HP0242 variants, we first carried out single-jump stopped-flow fluorescence measurements of all HP0242 variants. In line with previous findings [30], we observed multiple refolding kinetic phases for both wt and tandem HP0242 with a strong rollover effect. Additionally, we observed the third refolding kinetic phase for wt and the amplitude of which is about 15% of that of the major kinetic phase (red crosses in the first column of Figure 6). In the presence of intermediate denaturant concentrations (1.5-3M GdnHCl), the rates associated with the third kinetic phase can be superimposed with the fastest refolding kinetic phase when refolded from intermediate state, suggesting that these two kinetic phases correspond to the same refolding pathway (green crosses in the first column of Figure 6). In the case of tandem HP0242, we observed two refolding arms when refolded from denatured and intermediates, but only of one of which could be superimposed (last column in Figure 6).

Furthermore, when refolded from denatured state, the faster refolding arms dominated over the slower one with the associated amplitudes being more than 10-fold of that of the slower kinetic phase. In contrast, when refolding was initiated from intermediate state, a faster kinetic phase was observed and its rate marginally increased in the presence of a lower amount of denaturant as expected from productive refolding events, whereas the slower kinetic phase overlapped with the major refolding phase when refolded from denatured state: the corresponding rates decreased monotonously on decreasing denaturant concentration, indicative of non-productive off-pathway folding reactions. Importantly, the amplitudes associated with the two refolding kinetic phases when initiated from intermediate state are very

similar and are about half of that of the major refolding kinetic phase when initiated from denatured state. According to the kinetic observations for both wt and tandem HP0242, it is likely that productive folding intermediate can only be populated when incubated under intermediate conditions. Otherwise non-productive, off-pathway folding intermediates are more likely to be populated when refolding is initiated from fully denatured states. Although the presence of a rolled-over refolding arm could be explain by either an on-pathway or an off-pathway folding intermediate, our observation is very similar to the folding kinetics of a PDZ domain, which also exhibits a very strongly rolled-over refolding arm that has been unambiguously established to be associated with the formation of a misfolded intermediate [38].

Although the folding pathways of wt and tandem HP0242 are very complex, rendering it difficult to assign individual kinetic phases to specific steps of the folding pathways, we tentatively associate the fast and linear refolding arm derived from intermediate-to-native refolding (green cross in Figure 6) with the unfolding arm (filled and open blue circles for wt and tandem, respectively) to fit to a two-state folding model to extract the kinetic and thermodynamic parameters (Table 4) in order to compare with those derived from equilibrium unfolding (Tables 1 and 3). Indeed, for both wt and tandem HP0242, the transition points, $[D]_{50\%}$, derived from the chevron plot analysis (2.66±0.24 and 2.00±0.67 M for wt and tandem HP0242, respectively) are consistent with the $[D]_{50\%,N\leftrightarrow I}$ values derived from equilibrium unfolding monitored by intrinsic fluorescence (2.49±0.04 and 2.02±0.02 M for wt and tandem HP0242, respectively, Table 1), implying that the fast unfolding and refolding arms correspond to pathway between the native and productive folding intermediate states. Furthermore, the free energy of unfolding derived from the chevron plot analysis is quite consistent with the intrinsic fluorescence-based equilibrium

unfolding (4.42±0.28 versus 5.33±17 (Table 1) or 3.84±0.06 (Table 3) kcal mol^{-1}), given the uncertainty of the fitting of the equilibrium data. For tandem HP0242 however, the free energy of unfolding derived from the chevron plot analysis (1.52±0.38 kcal mol^{-1}) is significantly smaller than the equilibrium-based value (3.41±0.23 kcal mol^{-1}), primarily due to the difference between their respective m-values.

The analyses of the rolled-over refolding arms of HP0242 variants using a three-state folding model with a misfolded, off-pathway intermediate did not converge with the exception for the m-value of the unfolding arms. In the case of wt, the unfolding arm of the three-state analysis was the same as the two-state analysis and hence the m-value is essentially the same. For tandem HP0242, the slower unfolding arm (filled blue circles, last column of Figure 6) was linked with middle refolding arm (filled red triangle and open green diamond, last column of Figure 6). The resulting m-value of the unfolding arm is effectively the same as the faster one (0.42 versus 0.41±0.08). Furthermore, the amplitudes associated with the two unfolding arms are virtually the same and the sum of which is similar to that of wt. The same m-values of the two unfolding arms therefore suggest the presence of parallel folding pathways, which require additional kinetic experiments to establish (see next section).

We next examined the folding kinetics of F57W/W18F and F60W/W18F. Both variants yielded similar results as that of tandem HP0242, but the lack of tryptophan side-chain at position 18 rendered it impossible to observe significant spectral changes when refolding was initiated from intermediate states (middle columns in Figure 6). In terms of amplitude analysis, however, the refolding arms of F57W/W18F were similar in amplitude while those of F60W/W18F differed

substantially: the amplitudes of the faster refolding arm were approximately twice as large as those of the slower refolding arm. Both double mutants exhibited strongly rolled-over refolding arms similar to those of wt and tandem HP0242 when refolded from denatured states, but the extent of the rolled-over was much more pronounced, which is a clear indicator of misfolded, off-pathway folding intermediate [38]. The chevron plot of F57W/W18F was also subjected to data fitting using a three-state off-pathway model. The only converging parameter was the m-value associated with the linear unfolding arm (m_u=0.42±0.01), which is within uncertainty the same as those of wt and tandem HP0242. This is suggesting that all HP0242 variants are prone to form misfolded, non-productive folding intermediates, which could involve a parallel reaction scheme, and that the misfolded conformation encompasses structural elements beyond the knotted region, which involves the endogenous tryptophan, W18. Direct comparison between the double mutants, wt and tandem repeat should be made with caution as the native states of both double mutants are substantially destabilised with respect to the intermediates (Figure 3).

Comparing the folding kinetics of wt in different pH values showed that the unfolding kinetics was significantly accelerated at pH 8.0 compared to that at pH 6.5 (Figure 7). Unfolding kinetics could not be accurately determined at pH 5.0 due to increased aggregation propensity. The observed refolding rates of the productive refolding arm associated with the I-N refolding were slightly faster at pH 8.0 compared to those observed at pH 6.5 and were about 10-fold faster than those observed at pH 5.0. When initiated from denatured states, the productive refolding arm could only be observed at pH 6.5 while the corresponding refolding arm became strongly rolled-over and was not detectable at pH 5.0 (Figure 7). Furthermore, the amplitudes associated with the major refolding arm of the D-N refolding were

significantly larger at pH 8.0 compared to those at pH 5.0, while those at pH 6.5 were in between the two extremes. The strong pH-dependence may be associated with the charge distribution of the native structure of HP0242, which exhibits significant positively charged surface around the knotted region (Figure 1).

3.2.2 Double-jump stopped-flow fluorescence analysis

In order to further characterize the folding kinetics of HP0242 variants, we carried out double-jump stopped-flow fluorescence measurements. Standard interrupted refolding (D-N-D) experiments identified an additional unfolding phase that is about one order of magnitude faster than the original unfolding rate observed in single-jump experiments (Figure 8 left panel). According to the amplitude (population) build-up rates, both unfolding phases originated from the faster D-N refolding arm. In fact, these additional unfolding rates coincide with those observed in the single-jump experiments of tandem HP0242 (Figure 6). Considering that two unfolding arms of tandem HP0242 share similar m-values, which are indicative of parallel folding pathways, the same may also apply to wt HP0242.

We next carried out an unconventional double-jump refolding experiment whereby the denatured protein was first refolded to intermediate state to build up intermediate populations before the next refolding step to native conditions (D-I-N; Figure 8 middle panel). Indeed, similar to the single-jump experiments that started from an equilibrated intermediate state, we observed three refolding arms with the fastest one corresponding to the productive kinetic phase while the two slower ones were strongly rolled-over. Importantly, the amplitude build-up of the fastest kinetic phase was accompanied by decaying amplitudes of the two slower kinetic phases of which the decaying rates were similar to the build-up rate, implying that the three

kinetic pathways correspond to three distinct folding intermediates that interconvert during the aging period of the double-jump measurements.

Finally, we carried out interrupted unfolding (N-I-N; Figure 8 right panel) measurements by which native protein was unfolded and aged under intermediate conditions before refolding under native conditions. In this case, two well-defined refolding arms could be observed with the faster one, which corresponds to the productive refolding arm that was observed only through intermediate-to-native single-jump refolding, being the major pathway with the corresponding amplitudes being *ca.* four times greater than that of the slower refolding phase.

In the case of tandem HP0242, we consistently observed two kinetic phases in all three types of double-jump kinetic measurements. In the interrupted refolding experiments (D-N-D; Figure 9 left panel), two distinct unfolding kinetic phases are associated with the same rollover refolding arm that exhibit intermediate reaction rates between 10 and 100 s^{-1}, suggesting the presence of two distinct folding intermediates that can be populated under near-native conditions. In the sequential refolding experiments (D-I-N; middle panel of Figure 9), accumulation of folding intermediate leads to the formation of two distinct species, one of which is observed in the interrupted refolding experiment as mentioned (*ca.* 10 s^{-1}, filled red square). This folding pathway appears to be a kinetically more favorable one since the initial populations at short aging times are markedly higher than the other two species that exhibits a much faster refolding rate (*ca.* 100 s^{-1}, filled blue circle). These two kinetic phases exhibit similar population conversion as a function of aging time under the intermediate condition, which is similar to the observation for wt that exhibits an additional interconverting kinetic phase (filled green triangle, Figure 8 middle panel). Finally, in the case of interrupted unfolding experiments (N-I-N; Figure 9 right panel),

the two refolding arms originated from the same unfolding kinetic phase, which in fact correspond to the slower one.

3.3. Kinetic folding pathway of HP0242 variants

According to the three different double-jump stopped-flow measurements, it seems plausible that there exist at least three different folding intermediates along the folding pathways of wt. When refolding starts from a fully denatured state, HP0242 tends to form misfolded folding intermediates, which correspond to the two rollover refolding arms (Figure 8), the rates of which are protein concentration-dependent, indicating that the misfolded populations are also dimeric (data not shown). Furthermore, the productive folding intermediate can only be significantly populated when incubated under intermediate conditions to allow redistribution of the three distinct intermediate populations, potentially due to the high kinetic barriers that separate the three different intermediates (Figure 6 and Figure 8). Given the disparity in the number of observed unfolding kinetic phase and refolding phases, we have not been able to assign the observed kinetic rates to specific (un)folding reactions. While quantitative derivation of the reaction rates associated with the folding kinetics are currently unavailable, we propose two models to account for the observed folding kinetics of HP0242 variants (Figure 10). The first model consists of three distinct folding intermediates of which two are misfolded and off-pathway (I' and I"), which readily populate upon refolding and require unfolding to the denatured state for the formation of a productive folding intermediate (I) to complete the folding process into the native state. Indeed, refolding kinetics of wt showed three refolding arms with the fastest, productive folding rate being the least populated (Figure 6). Refolding of wt HP0242 from intermediate conditions renders redistribution of the

three intermediate populations with the fastest refolding rate being the most populated (Figure 6).

In the case of tandem HP0242, however, the productive refolding phase was not observed when refolding was initiated from the denatured state. Instead, two strongly rolled-over refolding arms were observed while the amplitude of the slower one was almost negligible (Figure 6). The productive refolding phase of tandem HP0242 was only observed when refolding was initiated from the intermediate state and its population (amplitude) was about the same as that of the faster rollover kinetic phase (Figure 6). One explanation for the lack of productive, linear arm in the single jump refolding of tandem HP0242 is that the fluorescence associated with this reaction is too weak to be detected in comparison with those of the two rollover phases. However, an alternative explanation may be that the kinetic folding pathway of tandem HP0242 is different from that of wt. Indeed, tandem HP0242 exhibits an additional linear unfolding arm that is approximately one order of magnitude faster than that of the common unfolding arm shared between wt and tandem HP0242 (Figures 6, 8 and 9). We therefore propose the second model in which the misfolded populations are on-pathway kinetic intermediates and the folding step from the misfolded conformation into the correctly folded one is the rate-limiting step (Figure 10). The misfolded folded intermediates may correspond to wrongly knotted populations that require back-tracking to overcome the kinetic traps as observed in the simulations [25].

According to the recently reported simulation analyses, multiple highly populated folding intermediates were present in the free energy landscape of tandem HP0242 near the native basin, *i.e.*, exhibiting high Q values [26]. Li *et al*. identified at six intermediate substates between the native and denatured state; the

three most populated substates exhibit high Q-values (> 0.7). Importantly, multiple folding routes were identified to branch from the second substate, which has a Q-value of *ca.* 0.5, after the formation of the transition state [25, 27]. Furthermore, it was found that protein knots were mostly formed after the completion of the transition state although some knotting events were also identified in the early stage of the folding process [25]. While the simulations by Sulkowska *et al.* indicated that the folding of tandem HP0242 and an unknotted but disulfide-linked HP0242 variant is two-state, unlike the more complex free energy landscape reported by Li *et al.,* [27], they indeed observed rollover refolding arms for both HP0242 variants as a function of simulation temperature, rather than chemical denaturant [25]. While the observed folding events could be fit to double exponentials, the authors suggested that a compressed exponential fit, *i.e.*, $A(t) = A_0 e^{-(k_{obs}t)^{\beta}}$, with β=1.4, may be a better model to describe the observed folding kinetics. Such a compressed exponential model is used to describe the relaxation process of jammed soft materials, which may reflect the properties of stressed polypeptide chains due to misfolded and knotted configurations. Our bulk fluorescence measurements have not been able to dissect the contributions of heterogeneous folding intermediates of more than three distinct populations. As suggested by the simulations, most of the native-like kinetic folding intermediates exhibit similar Q-values while some of the knotted but non-native-like intermediates are only transiently populated and hence difficult to be identified using our current experimental setup. Comprehensive description of the kinetic folding pathways of HP0242 variants may require single molecule fluorescence spectroscopy with the capacity to detect kinetic events that are much faster than the ms timescale.

4. Conclusion

In this study, we systematically characterized the folding equilibria and kinetics of HP0242 variants using a variety of spectroscopic tools. By constructing a number of single- and double-tryptophan HP0242 variants, we show that the highly populated folding intermediates under equilibrium conditions contain a significant amount of residual structure that mostly encompasses helixes H2, H3 and H4 that are distant from the knotted region, and that the intermediate is dimeric and heterogeneous. Indeed, simulation analyses suggest that dimerisation of H2 upon hydrophobic collapse is rapid and precedes knot formation [25, 27]. Using stopped-flow fluorescence measurements, we further demonstrate that the highly populated intermediate state is structurally heterogeneous and that these folding intermediates can interconvert under equilibrium condition to favor productive refolding into native conformations provided that the system is incubated under intermediate conditions for a sufficient period of time to populate productive folding intermediate that is less attainable when direct refolding from the denatured state is initiated. Overall, the folding kinetics of HP0242 variants are several orders of magnitudes faster than those of YibK [14], YbeA [15], MJ0366 [18], UCH-L1 [16] and UCH-L3 [17], possibility due to the all α-helical structural content. Nevertheless, the complex rollover refolding kinetics has not been reported for other knotted proteins. Further experimental analyses are required to piece out the contributions of substates along the complex folding pathways in order to verify the proposed folding pathways based on the atomistic simulations [25, 27].

Acknowledgement

The project is supported by a Career Development Award of the International Human Frontier Science Program, and funding from the Ministry of Science and Technology, National Tsing Hua University and Academia Sinica, Taiwan. Yu-Nan Liu was a recipient of a short-term EMBO fellowship to carry out preliminary experiments in Dr Sophie Jackson's laboratory at the Department of Chemistry, University of Cambridge.

Reference

[1] Mansfield M L, 1994 *Nat. Struct. Biol.* **1** 213-214.

[2] Taylor W R, 2000 *Nature* **406** 916-919.

[3] Sulkowska J I, Rawdon E J, Millett K C, Onuchic J N, Stasiak A, 2012 *Proc Natl Acad Sci U S A* **109** E1715-1723.

[4] Millett K C, Rawdon E J, Stasiak A, Sulkowska J I, 2013 *Biochem Soc Trans* **41** 533-537.

[5] Khatib F, Weirauch M T, Rohl C A, 2006 *Bioinformatics* **22** e252-259.

[6] Jamroz M, Niemyska W, Rawdon E J, Stasiak A, Millett K C, Sulkowski P, Sulkowska J I, 2015 *Nucleic Acids Res*. **43**, D306-D314

[7] Virnau P, Mirny L A, Kardar M, 2006 *PLoS Comp. Biol.* **2** e122.

[8] Lai Y L, Chen C C, Hwang J K, 2012 *Nucleic Acids Res* **40** W228-231.

[9] Lai Y L, Yen S C, Yu S H, Hwang J K, 2007 *Nucleic Acids Res* **35** W420-424.

[10] Virnau P, Mallam A, Jackson S, 2011 *J. Phys. Condensed Matter* **23** 033101.

[11] Mallam A L, 2009 *FEBS J* **276** 365-375.

[12] Sulkowska J I, Noel J K, Ramirez-Sarmiento C A, Rawdon E J, Millett K C, Onuchic J N, 2013 *Biochem Soc Trans* **41** 523-527.

[13] Mallam A L, Jackson S E, 2012 *Nat. Chem. Biol.* **8** 147-153.

[14] Mallam A L, Jackson S E, 2005 *J Mol Biol* **346** 1409-1421.

[15] Mallam A L, Jackson S E, 2007 *J Mol Biol* **366** 650-665.

[16] Andersson F I, Werrell E F, McMorran L, Crone W J, Das C, Hsu S T, Jackson S E, 2011 *J Mol Biol* **407** 261-272.

[17] Andersson F I, Pina D G, Mallam A L, Blaser G, Jackson S E, 2009 *FEBS J* **276** 2625-2635.

[18] Wang I, Chen SY, Hsu S T D, 2015 *J Phys Chem B*, in press (DOI: 10.1021/jp511029s)

[19] Mallam A L, Rogers J M, Jackson S E, 2010 *Proc Natl Acad Sci U S A* **107** 8189-8194.

[20] Hsieh S J, Mallam A L, Jackson S E, Hsu S T, 2014 *Biomol NMR Assign* **8** 439-442.

[21] Hsieh S J, Mallam A L, Jackson S E, Hsu S T, 2014 *Biomol NMR Assign* **8** 283-285.

[22] Wallin S, Zeldovich K B, Shakhnovich E I, 2007 *J Mol Biol* **368** 884-893.

[23] Sulkowska J I, Sulkowski P, Onuchic J, 2009 *Proc Natl Acad Sci U S A* **106** 3119-3124.

[24] Noel J K, Sulkowska J I, Onuchic J N, 2010 *Proc Natl Acad Sci U S A* **107** 15403-15408.

[25] Sulkowska J I, Noel J K, Onuchic J N, 2012 *Proc Natl Acad Sci U S A* **109** 17783-17788.

[26] Noel J K, Onuchic J N, Sulkowska J I, 2013 *J. Phys. Chem. Lett.* **4** 3570-3573.

[27] Li W, Terakawa T, Wang W, Takada S, 2012 *Proc Natl Acad Sci U S A* **109** 17789-17794.

[28] Lui C S, 2004 Master thesis National Tsing Hua University

[29] Tsai J Y, Chen B T, Cheng H C, Chen H Y, Hsaio N W, Lyu P C, Sun Y J, 2006 *Proteins* **62** 1138-1143.

[30] Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin A G, 2014 *Nucleic Acids Res* **42** D310-314.

[31] King N P, Jacobitz A W, Sawaya M R, Goldschmidt L, Yeates T O, 2010 *Proc Natl Acad Sci U S A* **107** 20732-20737.

[32] Liu W, Feng X, Zheng Y, Huang C H, Nakano C, Hoshino T, Bogue S, Ko T P, Chen C C, Cui Y, Li J, Wang I, Hsu S T, Oldfield E, Guo R T, 2014 *Sci. Rep.* **4** 6214.

[33] Laue T M, Shah B D, Ridgeway T M, Pelletier S L, in: S.E. Harding, A.J. Rowe, J.C. Horton (Eds.) Analytical Ultracentrifugation in Biochemistry and Polymer Science, Royal Society of Chemistry, 1992.

[34] Andersson F I, Werrell E F, McMorran L, Crone W J, Das C, Hsu S T, Jackson S E, 2011 *J. Mol. Biol.* **407** 261-272.

[35] Fersht A R. Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding; W. H. Freeman: New York, USA, 1998.

[36] Buchner J, Kiefhaber T, Protein Folding Handbook, Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2005.

[37] Horng J C, Tracz S M, Lumb K J, Raleigh D P, 2005 Biochemistry 2005, 44, 627−634.

[38] Gianni S, Ivarsson Y, De Simone A, Travaglini-Allocatelli C, Brunori M, Vendruscolo M, 2010 Nat. Struc. Mol. Biol. **17**, 1431-1437

Table 1. Thermodynamics parameters of equilibrium unfolding induced by GdnHCl

wt	Intrinsic fluorescence					Far-UV CD					ANS[b]
[Protein] (μM)	50	25	12.5	6.25	3.125	50	25	12.5	6.25	3.125	11.25
$m_{N_2 \leftrightarrow I_2}$ (kcal mol^{-1} M^{-1})			2.14±0.06					1.93±0.15			N.C.[c]
$[D]_{50\%,N_2 \leftrightarrow I_2}$ (M)			2.49±0.04					2.39±0.03			
$m_{I_2 \leftrightarrow 2D}$ (kcal mol^{-1} M^{-1})			0.38±0.07					1.26±0.13			
$[D]_{50\%,I_2 \leftrightarrow 2D}$ (M)	3.80±0.26	3.07±0.31	2.98±0.32	2.45±0.38	2.64±0.36	5.45±0.16	5.10±0.14	4.94±0.12	4.84±0.09	4.60±0.10	
F57W[d]											
[Protein] (μM)	50	25	12.5	6.25	3.125	50	25	12.5	6.25	3.125	11.25
$m_{N_2 \leftrightarrow I_2}$ (kcal mol^{-1} M^{-1})			2.00±0.04					1.98±0.14			2.07±0.34
$[D]_{50\%,N_2 \leftrightarrow I_2}$ (M)			1.86±0.01					2.05±0.03			1.36±0.14
$m_{I_2 \leftrightarrow 2D}$ (kcal mol^{-1} M^{-1})			1.07±0.06					1.06±0.09			0.55±0.08
$[D]_{50\%,I_2 \leftrightarrow 2D}$ (M)	4.68±0.01	4.56±0.01	4.36±0.01	4.17±0.01	3.94±0.01	5.09±0.14	4.75±0.12	4.66±0.11	4.65±0.10	4.32±0.10	7.45±0.73
F57W/W18F											
[Protein] (μM)	50	25	12.5	6.25	3.125	50	25	12.5	6.25	3.125	11.25
$m_{N_2 \leftrightarrow I_2}$ (kcal mol^{-1} M^{-1})			1.41±0.11					0.99±0.12			N.C.
$[D]_{50\%,N_2 \leftrightarrow I_2}$ (M)			0.56±0.11					0.56[a]			
$m_{I_2 \leftrightarrow 2D}$ (kcal mol^{-1} M^{-1})			0.10±0.07					1.19±0.13			
$[D]_{50\%,I_2 \leftrightarrow 2D}$ (M)	4.28±0.10	4.13±0.08	4.03±0.20	3.81±0.16	3.54±0.08	4.98±0.16	4.66±0.13	4.54±0.13	4.31±0.13	4.08±0.13	
F60W/W18F											
[Protein] (μM)	20	10		5		20	10		5		11.25
$m_{N_2 \leftrightarrow I_2}$ (kcal mol^{-1} M^{-1})			0.77±0.21					0.89±0.10			N.C.
$[D]_{50\%,N_2 \leftrightarrow I_2}$ (M)			0.90±0.33					0.90[a]			
$m_{I_2 \leftrightarrow 2D}$ (kcal mol^{-1} M^{-1})			1.49±0.38					1.08±0.09			
$[D]_{50\%,I_2 \leftrightarrow 2D}$ (M)	5.68±1.55	5.24±1.14		4.80±1.15		6.21±0.28	5.58±0.17		5.55±0.13		
Tandem											11.25
$m_{N \leftrightarrow I}$ (kcal mol^{-1} M^{-1})			1.69±0.10					0.88±0.09			1.12±0.21
$[D]_{50\%,N \leftrightarrow I}$ (M)			2.02±0.02					1.38±0.26			2.86±0.14
$m_{I \leftrightarrow D}$ (kcal mol^{-1} M^{-1})			2.01±0.30					1.84±0.31			0.77±0.11
$[D]_{50\%,I \leftrightarrow D}$ (M)			5.68±1.55					6.35±0.19			5.01±0.20

a. The fitting results did not converge and hence the corresponding values derived from intrinsic fluorescence analysis were used and fixed for the global fitting of the remaining parameters.
b. For ANS fluorescence, only the m-values were shared for fitting and the reported $[D]_{50\%}$ values correspond to the first and largest SVD component
c. Not converged (N.C.) The fitting results did not converge
d. For F57W, in total 10 protein concentrations ranging from 50 to 1 μM were used for intrinsic fluorescence measurements as shown in Figure 4. For simplicity, only those that have corresponding values in other variants are tabulated herein.

Table 2. Thermodynamics parameters of equilibrium unfolding induced by GdnHCl monitored by maximum emission wavelength (λ_{max}) of intrinsic fluorescence

wt					
[Protein] (μM)	50	25	12.5	6.25	3.125
$m_{N_2 \leftrightarrow I_2}$ (kcal mol^{-1} M^{-1})			2.04±0.06		
$[D]_{50\%,N_2 \leftrightarrow I_2}$ (M)			2.74±0.01		
$m_{I_2 \leftrightarrow 2D}$ (kcal mol^{-1} M^{-1})			1.36±0.20		
$[D]_{50\%,I_2 \leftrightarrow 2D}$ (M)	4.84±0.15	4.93±0.14	4.78±0.13	4.72±0.16	4.95±0.12
F57W					
[Protein] (μM)	50	25	12.5	6.25	3.125
$m_{N_2 \leftrightarrow I_2}$ (kcal mol^{-1} M^{-1})			1.64±0.13		
$[D]_{50\%,N_2 \leftrightarrow I_2}$ (M)			2.11±0.03		
$m_{I_2 \leftrightarrow 2D}$ (kcal mol^{-1} M^{-1})			1.21±0.13		
$[D]_{50\%,I_2 \leftrightarrow 2D}$ (M)	4.51±0.11	4.50±0.10	4.29±0.09	4.15±0.10	4.21±0.10
F57W/W18F[a]					
[Protein] (μM)	50	25	12.5	6.25	3.125
$m_{I_2 \leftrightarrow 2D}$ (kcal mol^{-1} M^{-1})			1.14±0.04		
$[D]_{50\%,I_2 \leftrightarrow 2D}$ (M)	4.44±0.06	4.31±0.05	4.04±0.05	4.06±0.05	3.55±0.05
F60W/W18F[a]					
[Protein] (μM)	20		10		5
$m_{I_2 \leftrightarrow 2D}$ (kcal mol^{-1} M^{-1})			1.73±014		
$[D]_{50\%,I_2 \leftrightarrow 2D}$ (M)	5.92±0.09		5.60±0.07		5.31±0.07
Tandem[b]					
$m_{N \leftrightarrow I}$ (kcal mol^{-1} M^{-1})			1.19±0.12		
$[D]_{50\%,N \leftrightarrow I}$ (M)			2.06±0.11		

a. The results can only be fit to a two-state model which corresponds to the I-D transition

b. The results can only be fit to a two-state model which corresponds to the N-I transition

Table 3. Thermodynamic parameter of equilibrium unfolding of HP0242 variants derived from a dimeric intermediate model

	wt	F57W	F57W/W18F	F60W/F18W
Intrinsic fluorescence				
$\Delta G^{H_2O}_{N_2 \leftrightarrow I_2}$ (kcal mol^{-1})	3.84±0.06	3.34±0.03	2.48±0.16	1.36±0.02
$m_{N_2 \leftrightarrow I_2}$ (kcal mol^{-1} M^{-1})	1.53±0.02	1.70±0.02	0.49±0.02	0.99±0.02
$\Delta G^{H_2O}_{I_2 \leftrightarrow 2D}$ (kcal mol^{-1})	N.C.[a]	5.16±0.10	2.25±0.32	9.64±0.07
$m_{I_2 \leftrightarrow 2D}$ (kcal mol^{-1} M^{-1})		1.40±0.02	0.99±0.06	1.94±0.01
Y_I	0.98±0.002	0.68±0.003	0.60±0.11	0.26±0.002
Far-UV CD				
$\Delta G^{H_2O}_{N_2 \leftrightarrow I_2}$ (kcal mol^{-1})	4.88±0.23	3.43±0.07	1.43±0.04	1.49±0.02
$m_{N_2 \leftrightarrow I_2}$ (kcal mol^{-1} M^{-1})	1.90±0.10	1.63±0.04	0.75±0.03	1.05±0.02
$\Delta G^{H_2O}_{I_2 \leftrightarrow 2D}$ (kcal mol^{-1})	8.11±0.38	7.18±0.18	8.57±0.37	9.70±0.16
$m_{I_2 \leftrightarrow 2D}$ (kcal mol^{-1} M^{-1})	1.81±0.07	1.72±0.03	1.89±0.06	1.95±0.03
Y_I	0.48±0.009	0.52±0.004	0.54±0.01	0.45±0.002

a. The fitting result did not converge to meaningful values.

Table 4. Chevron plot analysis of HP0242 variants

Two-state folding model[a]	wt	Tandem
$k_f^{H_2O}$ (s^{-1})	1695±349	184±29
m_f (kcal mol^{-1} M^{-1})	-1.25±0.09	-0.35±0.15
$k_u^{H_2O}$ (s^{-1})	0.96±0.41	14.1±8.7
m_u (kcal mol^{-1} M^{-1})	0.41±0.05	0.41±0.08
$\Delta G_{N-D}^{H_2O}$ (kcal mol^{-1})	4.42±0.28	1.52±0.38
$[D]_{50\%}$ (M)	2.66±0.24	2.00±0.67

a. The fitting results are shown in solid blue lines in Figure 6.

Figure legends

Figure 1. Structure and knotted topology of HP0242. Left panel. Wild-type HP0242 forms a homodimeric structure with intertwining helices between helices 1 and 4 around the only tryptophan residue, W18 (shown in red spheres), to form a pseudoknot. The two phenylalanine residues, F57 and F60, located on the opposite side of W18 are shown in green and blue spheres, respectively. Given the two-fold symmetry, the labeling of the secondary structures and residue side-chains of the second monomer are indicated with a prime symbol. Middle panel. Topological representation of tandem HP0242. The individual helices are shown in thick line connected by thin lines that represent the loop structures. The topological representation is colour-ramped from blue to red to indicate the sequential connections from the N- to C-terminal region. The position of the linker that connects the C-terminus of the first HP0242 molecule to the N-terminus of the second HP0242 molecule is indicated by an arrow. Right panel. Surface charge distribution of HP0242 homodimer. The positively and negatively charged protein surfaces are coloured in blue and red, respectively. A threshold of ±3 eV is used. The electrostatics calculation is made by the PDB2PQR server and rendered by APBS (http://www.poissonboltzmann.org/docs/calculating/) using the molecular visualization software PyMol (http://www.pymol.org/).

Figure 2. SEC-MALS analysis of HP0242 in solution. Monomeric HP0242 has a molecular weight (MW) of 11.1 kDa. The calculated MW of the elution peaks in the presence and absence of 4M GdnHCl are 20.4 and 23.6 kDa, respectively, indicating that HP0242 remains dimeric in the presence of 4M GdnHCl.

Figure 3. Multi-parametric analysis of the folding equilibria of HP0242 variants. Intrinsic fluorescence, fluorescence anisotropy, far-UV CD and ANS fluorescence of wt, F57W, F57W/W18F, F60W/W18F and tandem repeat variants of HP0242 as a function of GdnHCl concentration are shown in descending order as indicated and are colour-ramped from red to blue on increasing GdnHCl concentration from 0 to 8M. The fractions of unfolding of HP0242 variants of various protein concentrations derived from SVD analyses of intrinsic fluorescence and far-UV CD spectroscopy are shown in red and green, respectively, and the light to dark data points correspond to low to high protein concentrations, respectively. The singular values of the first five SVD components of the intrinsic fluorescence analyses are shown in bar charts with the corresponding correlation coefficient indicated above. The fluorescence anisotropy was determined with a protein concentration of *ca.* 20 μM. The corresponding wavelength is indicated at the lower left corner of each panel. For ANS fluorescence, the amplitudes of the first two SVD components are shown in blue and red dots and are fit to a three-state equilibrium unfolding model, except for F57W/W18F. The populations of native (N), intermediate (I) and denatured (D) states derived from far-UV CD data are shown in the last column in blue, green and red lines, respectively. The positions of the N-I and I-D transitions of wt are indicated by two vertical dashed lines across the population plots of all variants to highlight the changes in folding stabilities as a result of the point mutations or covalent linkage in the case for tandem HP0242.

Figure 4. Protein concentration-dependent I-D transition of HP0242 variants. The I-D transition points derived from far-UV CD and intrinsic fluorescence are shown in green and red symbols. The results derived from changes of a single-wavelength and SVD analysis are shown in open and filled symbols, respectively.

Figure 5. Changes of maximum intrinsic fluorescence emission wavelength of HP0242 variants as a function of denaturant concentration. All transition curves are fit to a two-state folding model except for that of F57W, which is fit to a three-state folding model.

Figure 6. Folding kinetics of HP0242 variants as a function of denaturant concentration. The N-D unfolding, D-N refolding and I-N refolding rates are shown in blue, red and green symbols, respectively, in the top panels. The corresponding amplitudes are shown in the bottom panels. The N-I and I-D transition points derived from equilibrium unfolding analyses are shown in transparent vertical lines as references. The fitting results of a two-state and three-state folding model are shown in blue and red line, respectively.

Figure 7. pH-dependent folding kinetics of HP0242. The observed folding rates are shown in the same styles as in Figure 6.

Figure 8. Double-jump folding kinetics analysis of wt HP0242. Interrupted refolding (D-N-D), sequential refolding (D-I-N) and interrupted unfolding (N-I-N) results are shown from left to right, accordingly. The observed folding rates are derived from global-fitting of the kinetic traces of different delays between the first and second mixing steps to a double-exponential decaying function (bottom panels). The fast and slow rates are shown in filled blue and red symbols, respectively (top panels). The folding rates derived from single-jump experiment are shown in grey as reference. The amplitudes corresponding to the two kinetic phases with different delays are subjected to single exponential fitting (middle panels) and the corresponding build-up rates are shown in open symbols (top panels). For

clarification, the data points and residuals are smoothed by a running average of four points.

Figure 9. Double-jump folding kinetics analysis of tandem HP0242. The definitions of the symbols are the same as those used in Figure 8.

Figure 10. Proposed off-folding pathway (A) and on-pathway (B) folding models of HP0242 variants. Schematic representation of the populations of native dimer (N_2), denatured monomer (2D) and three different dimeric intermediates (I, I' and I") as a function of positional root-mean-squared-deviation (RMSD) versus the reaction coordinate expressed as the native contact formed, Q. I, I' and I" are circled in blue, red and green, corresponding to the three refolding kinetic phases observed in wt (Figure 8). The thickness of the arrowheads corresponds to the likelihood of reactions. Grey arrowheads between I'_2 and $I"_2$ are introduced to suggest the potential interconversion between the two intermediates, which has not yet been confirmed experimentally.

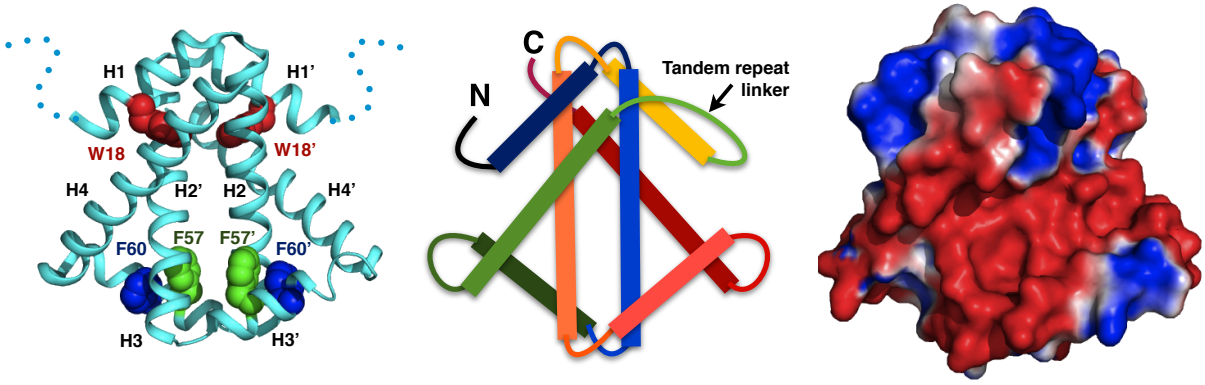
H1
H1'
W18
W18'
H4
H2'
H2
H4'
F60
F57
F57'
F60'
H3
H3'
C
N
Tandem repeat
linker

Figure 1

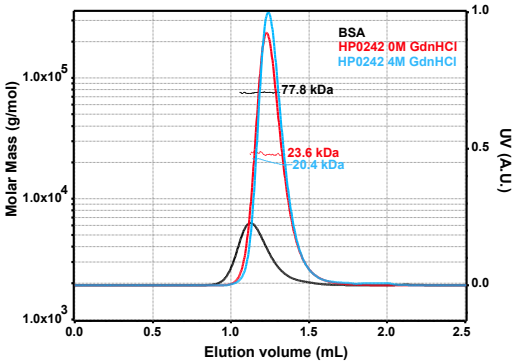
BSA
HP0242 0M GdnHCl
HP0242 4M GdnHCl
77.8 kDa
23.6 kDa
20.4 kDa
Molar Mass (g/mol)
1.0x10^5
1.0x10^4
1.0x10^3
UV (A.U.)
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
2.5
Elution volume (mL)

Figure 2

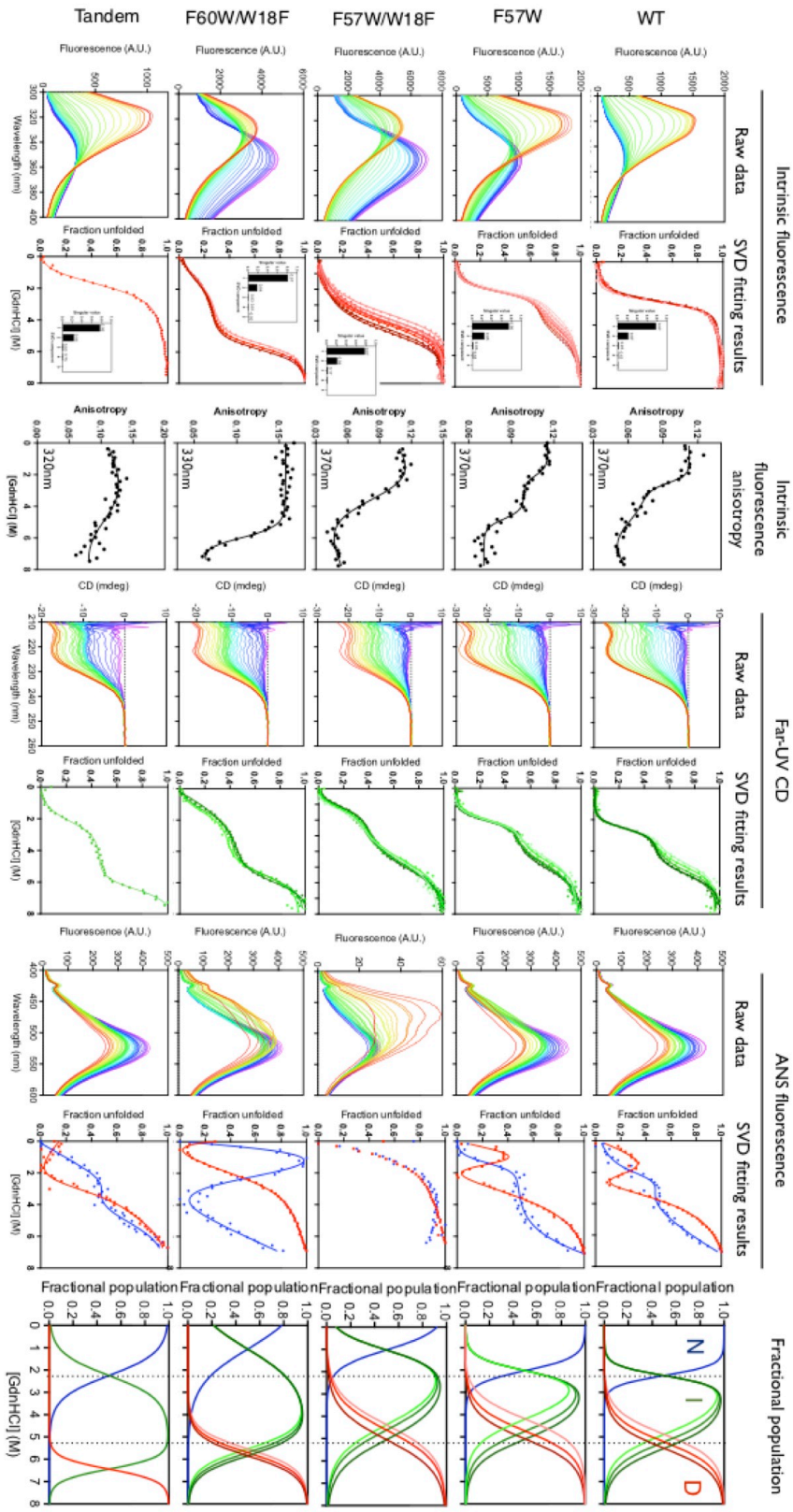
Tandem
F60W/W18F
F57W/W18F
F57W
WT
Intrinsic fluorescence
Raw data
SVD fitting results
Intrinsic fluorescence anisotropy
Far-UV CD
ANS fluorescence
Fractional population
Fluorescence (A.U.)
Wavelength (nm)
Fraction unfolded
[GdnHCl] (M)
Anisotropy
320nm
330nm
370nm
CD (mdeg)
N
I
D

Figure 3

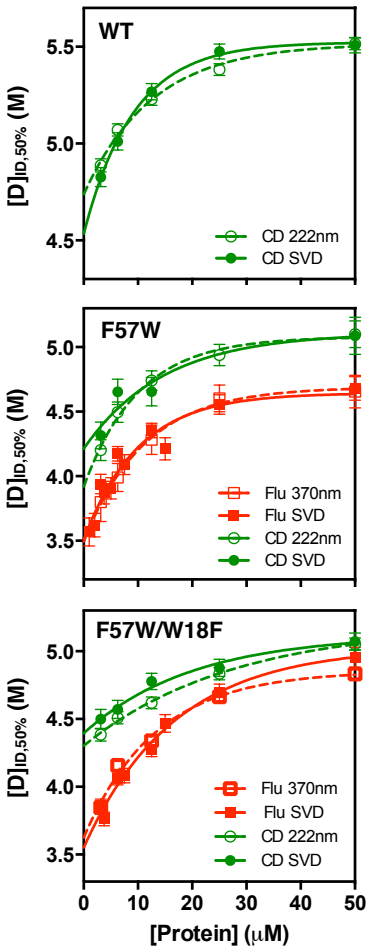
WT
5.5
5.0
4.5
[D]ID,50% (M)
CD 222nm
CD SVD
F57W
5.0
4.5
4.0
3.5
[D]ID,50% (M)
Flu 370nm
Flu SVD
CD 222nm
CD SVD
F57W/W18F
5.0
4.5
4.0
3.5
[D]ID,50% (M)
Flu 370nm
Flu SVD
CD 222nm
CD SVD
0
10
20
30
40
50
[Protein] (μM)

Figure 4

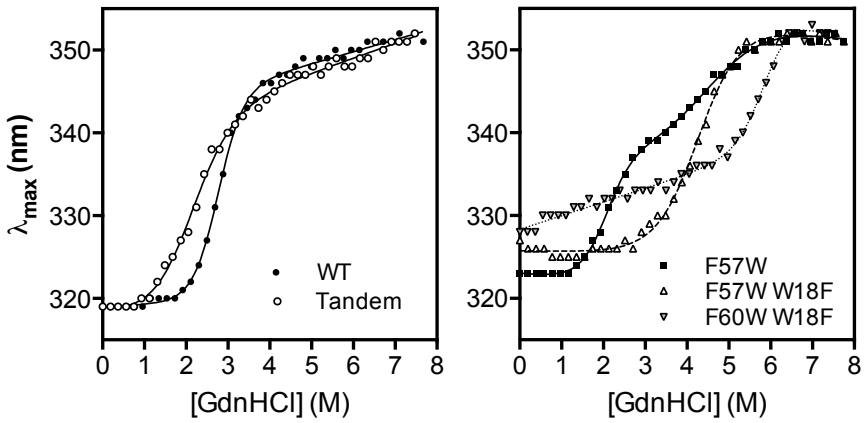
λ_{max} (nm)
350
340
330
320
0 1 2 3 4 5 6 7 8
[GdnHCl] (M)
WT
Tandem
F57W
F57W W18F
F60W W18F

Figure 5

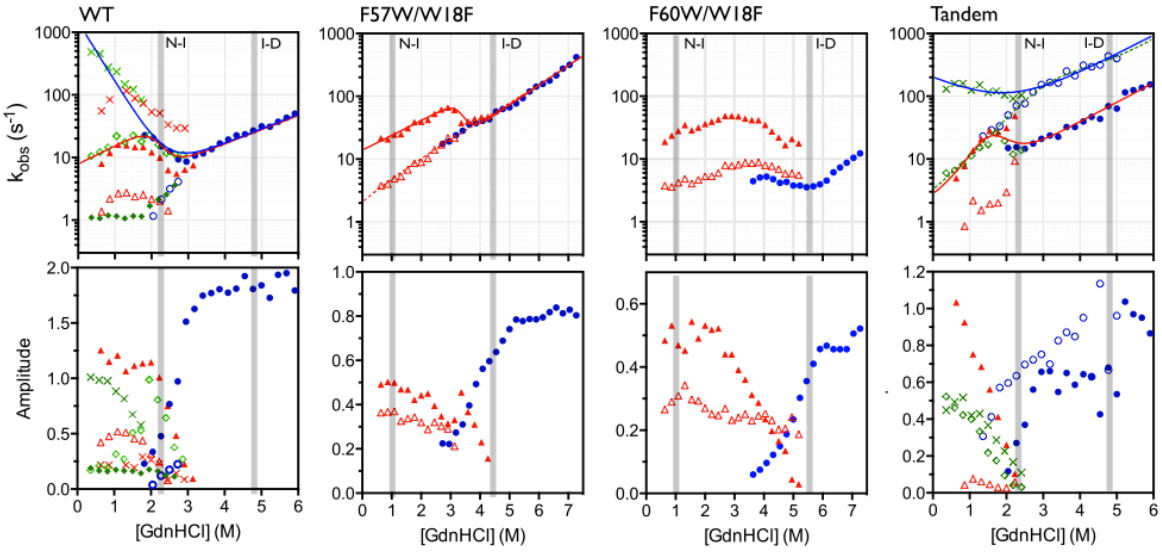
WT
F57W/W18F
F60W/W18F
Tandem
N-I
I-D
k_{obs} (s^{-1})
Amplitude
[GdnHCl] (M)

Figure 6

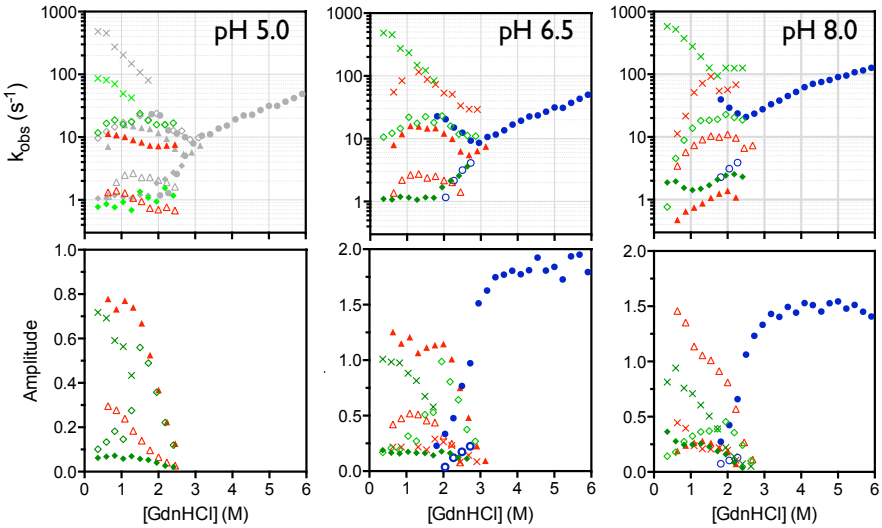

Figure 7

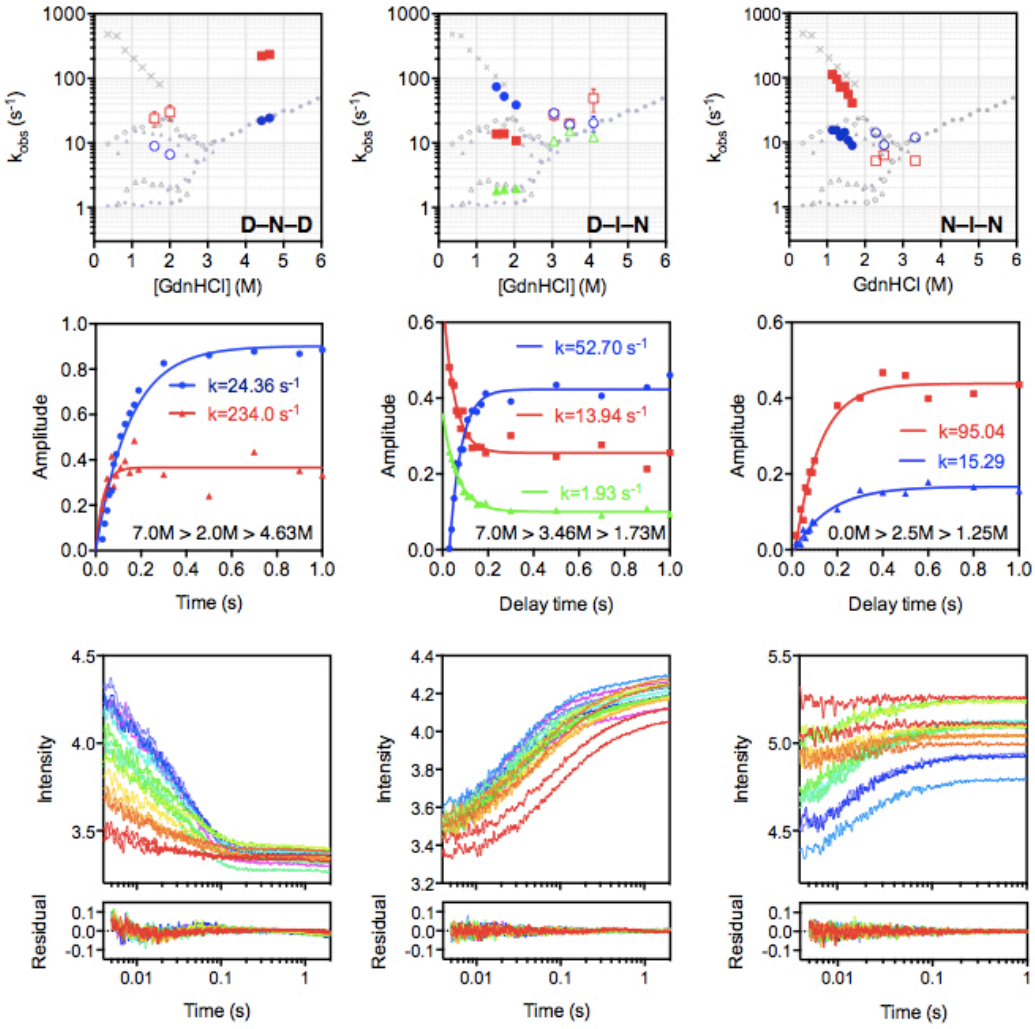
D–N–D
D–I–N
N–I–N
[GdnHCl] (M)
GdnHCl (M)
k_obs (s^-1)
k=24.36 s^-1
k=234.0 s^-1
7.0M > 2.0M > 4.63M
k=52.70 s^-1
k=13.94 s^-1
k=1.93 s^-1
7.0M > 3.46M > 1.73M
k=95.04
k=15.29
0.0M > 2.5M > 1.25M
Amplitude
Time (s)
Delay time (s)
Intensity
Residual

Figure 8

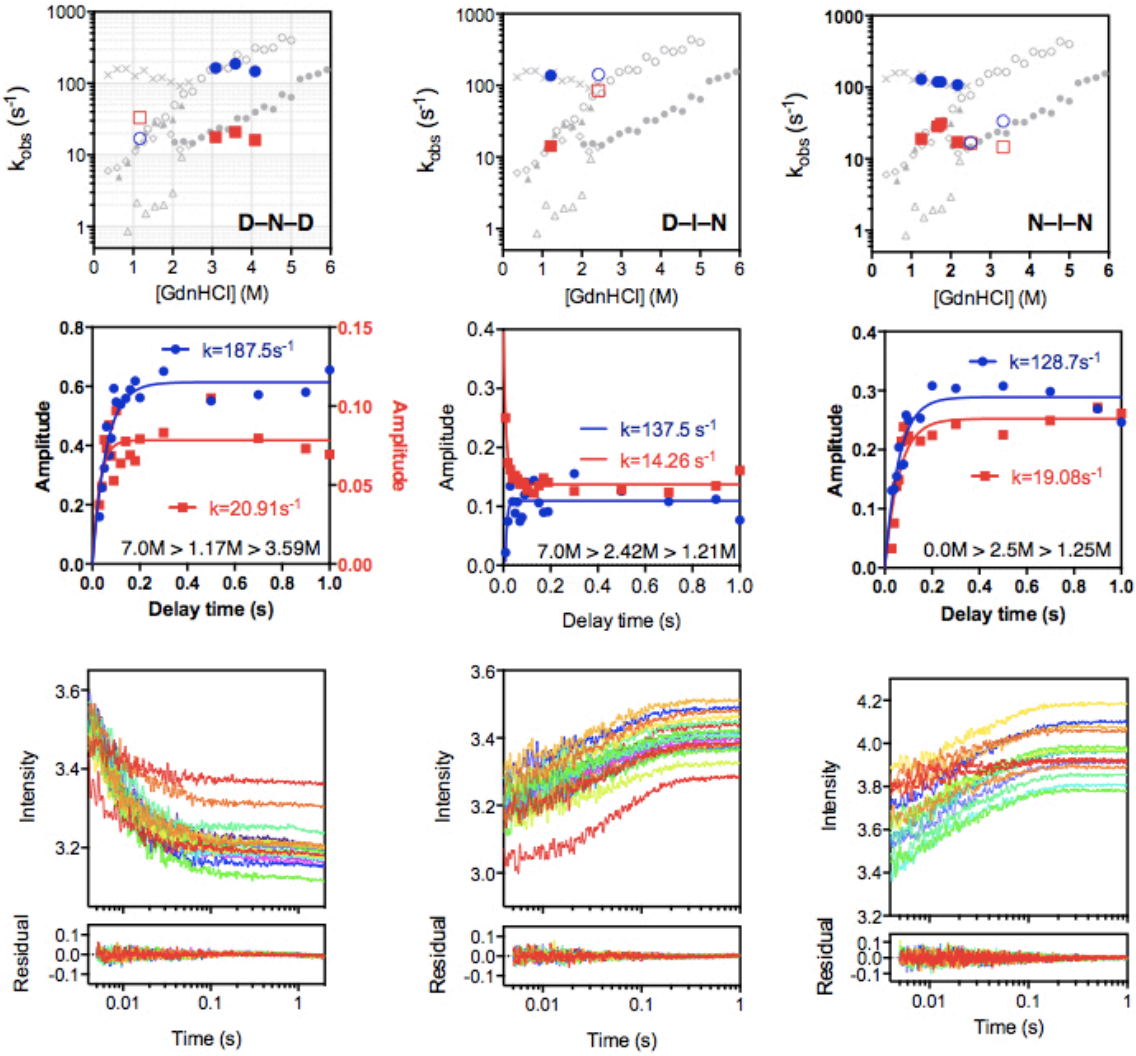
D–N–D
D–I–N
N–I–N
[GdnHCl] (M)
k=187.5s-1
k=20.91s-1
7.0M > 1.17M > 3.59M
k=137.5 s-1
k=14.26 s-1
7.0M > 2.42M > 1.21M
k=128.7s-1
k=19.08s-1
0.0M > 2.5M > 1.25M
Amplitude
Delay time (s)
Intensity
Residual
Time (s)

Figure 9

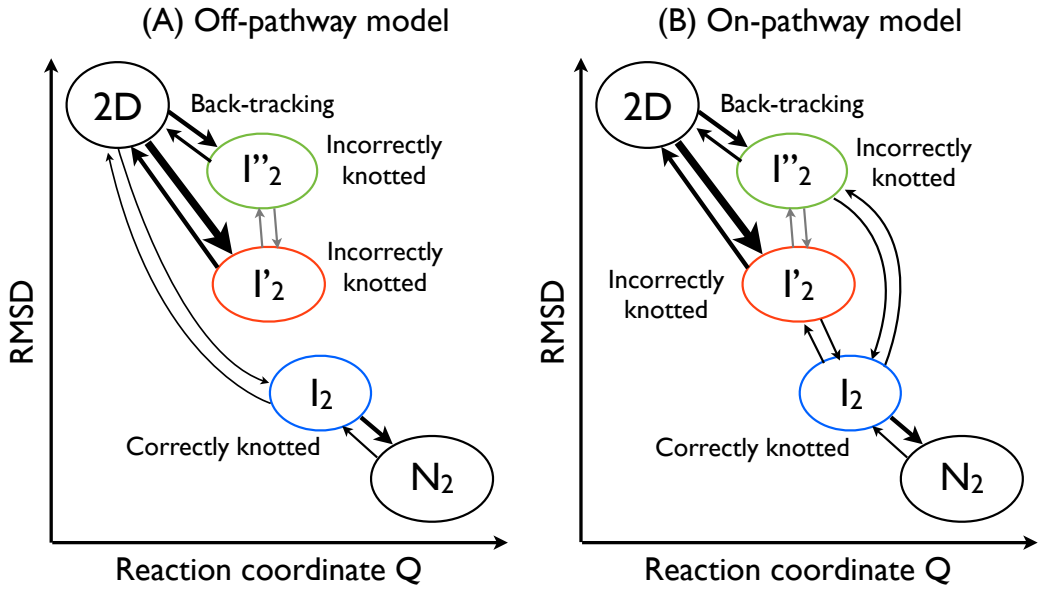
(A) Off-pathway model
2D
Back-tracking
I''2
Incorrectly knotted
I'2
Incorrectly knotted
I2
Correctly knotted
N2
RMSD
Reaction coordinate Q
(B) On-pathway model
2D
Back-tracking
I''2
Incorrectly knotted
Incorrectly knotted
I'2
I2
Correctly knotted
N2
RMSD
Reaction coordinate Q

Figure 10