**Disorder drives cooperative folding in a multi-domain protein**

Dominika T. Gruszka[a], Carolina A.T.F. Mendonça[a], Emanuele Paci[b], Fiona Whelan[c], Judith Hawkhead[c], Jennifer R. Potts[c] and Jane Clarke[a,1]

[a]Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, UK
[c]Department of Biology, University of York, Wentworth Way, York, YO10 5DD
DTG Current address: The Francis Crick Institute, Clare Hall Laboratory, Blanche Lane, South Mimms, EN6 3LD, UK

[1]To whom correspondence should be addressed. E-mail: jc162@cam.ac.uk; Phone: +44 (0) 1223 336426

Keywords: IDP, protein folding, parallel pathways, protein engineering, cooperativity

**Abstract**

Many human proteins contain intrinsically disordered regions, and disorder in these proteins can be fundamental to their function – for example, facilitating transient but specific binding, promoting allostery, or allowing efficient post-translational modification. SasG, a multi-domain protein implicated in host colonisation and biofilm formation in *Staphylococcus aureus*, provides another example of how disorder can play an important role. Approximately half of the domains in the extracellular repetitive region of SasG are intrinsically unfolded in isolation, but these E domains fold in the context of their neighbouring, folded G5 domains. We have previously shown that the intrinsic disorder of the E domains mediates long-range cooperativity between non-neighbouring G5 domains, allowing SasG to form a long, rod-like, mechanically strong structure. Here we show that the disorder of the E domains coupled with the remarkable stability of the inter-domain interface, results in cooperative folding kinetics across long distances. Formation of a small structural nucleus at one end of the molecule results in rapid structure formation over a distance of 10 nm, which is likely to be important for the maintenance of the structural integrity of SasG. Moreover, if this normal folding nucleus is disrupted by mutation, the inter-domain interface is sufficiently stable to drive the folding of adjacent E and G5 domains, along a parallel folding pathway, thus maintaining cooperative folding.

**Significance statement**

Understanding the role played by disorder in Biology is becoming increasingly important. Disordered proteins are central to signalling, development, initiation of transcription and other vital cellular processes. How and why disordered proteins are used is not entirely clear, but disorder can be important in allostery, facilitate regulatory post-translational modification and allow rapid, specific, yet promiscuous binding. Here, our investigations of the biofilm-promoting protein SasG illustrates that disorder can play another role. We demonstrate that the intrinsic disorder of half the domains is important for imparting long range cooperativity in folding of a large multidomain protein –allowing formation of a very small local element of structure to precipitate cooperative folding of adjacent disordered domains across a length scale of ~10 nm.

**\body**

## Introduction

It has been suggested that as much as 20% of the proteome may be intrinsically disordered (1), mainly manifested as intrinsically disordered regions (IDRs) within multidomain proteins, although a few proteins are apparently entirely disordered. Some proteins function as a consequence of disorder: for example, disordered PEVK regions of titin act as an entropic spring (2), while in the nuclear pore complex disordered nucleoporins provide a thick selective barrier controlling nuclear import (3). Disorder can also play other roles: it facilitates posttranslational modification, and may promote allostery (4, 5). SasG is a cell wall attached protein from *Staphylococcus aureus* that promotes intercellular adhesion during the accumulation phase of biofilm formation via its C-terminal repetitive region (6-8). We previously showed that this part of SasG contains alternating E and G5 domains (Fig. 1*A*) and that E folds when it is N-terminal of a G5 domain. The disorder of E domains in isolation is essential for formation of a long, stiff, mechanically strong, rod-like structure (9) capable of projecting the N-terminal A domain, which is involved in host colonisation (6).

Here we combine biophysical measurements, protein engineering and simulation to show that the disorder in the E domains of SasG also promotes cooperative folding and unfolding pathways. We find that SasG domains have a highly polarized transition state structure, where formation of a small portion of a three-stranded sheet in the far C-terminal region of a SasG G5 domain is sufficient to drive the folding of structure over a distance of 10 nm. Our studies reveal the importance of the E-G5 interface in driving this cooperativity. Furthermore, when the usual folding nucleus is disrupted by mutation in the multidomain protein, then this interface is sufficiently stable to drive folding of the two adjacent domains. Thus we propose that disorder can play a key role in ensuring cooperative folding over long distances in multidomain proteins.

## Results

*SasG domains fold cooperatively at equilibrium:* SasG domains are highly homologous: the sequence identity between G5 domains (except for the first and last) and between E domains is >97%. Here we investigated the first E domain and the second G5 domain ($G5^2$), either alone, or in tandem (E-$G5^2$) (Fig. 1). We have

previously shown that the E domain is fully unfolded in isolation (10). Since SasG domains have no tryptophans, (un)folding was followed by monitoring intrinsic tyrosine fluorescence. We have demonstrated that urea-induced equilibrium denaturation curves of E-G5$^2$ monitored by fluorescence coincide with those recorded by ellipticity at 235 nm (7) and with domain-specific FRET probes (9), demonstrating that equilibrium unfolding of the two-domain construct is fully cooperative: two-state, with concerted disruption of both domains and of secondary and tertiary structure, with no accumulation of intermediates (Fig. 1*C*). The stability of E-G5$^2$ is around 3.5 kcal mol$^{-1}$ greater than that of an isolated G5$^2$ domain (6.3 vs. 2.8 kcal mol$^{-1}$, respectively).

*Kinetic experiments reveal that SasG domains fold and unfold cooperatively.* The refolding kinetics of G5$^2$ and E-G5$^2$ can both be described by a sum of two exponential phases with a fast folding phase (accounting for at least 30% of the amplitude) and a slower phase that represents proline *cis-trans* isomerization-limited folding events (E-G5$^2$ and G5$^2$ have 17 and 8 prolines, respectively). Only the faster phase is discussed here. The rate constant for folding of E-G5$^2$ is the same as that of G5$^2$ at all denaturant concentrations (Fig. 1*D*). Under unfolding conditions, at urea concentrations ≤ 6.5 M, only a single kinetic phase is detected for both G5$^2$ and E-G5$^2$, but E-G5$^2$ unfolds significantly more slowly, and the dependence of the logarithm of the rate constant for unfolding on denaturant concentration ($m_{ku}$) is significantly higher[*]. The unfolding limbs of the chevron plots are curved (Fig. 1*D*). To account for non-linearity in the observed unfolding rate constant, the chevron plot data were fitted to a sequential transition states model (12), in which denaturant induces a switch between two barriers separated by a high-energy intermediate.

At denaturant concentrations below ~6.5 M urea all the evidence suggests that both G5$^2$ and the two-domain construct E-G5$^2$ fold via a two-state pathway where the two domains fold and unfold cooperatively: we observe, for both constructs, that the values of $m_{D\text{-}N}$ obtained by combining kinetic *m*-values are the same within error as the equilibrium values (Supplementary Table 1). Similarly, the values of free energy

[*] The dependence of folding/unfolding rate constants on [urea] (kinetic *m*-values, $m_{kf}$ and $m_{ku}$) is determined by the change in SASA between the denatured state, D, and the transition state, TS, (in folding) and TS and the native state, N (for unfolding) (11). Thus, since E-G5$^2$ and G5$^2$ have the same folding *m*-values we can assume that they fold via the same transition state. The unfolding *m*-value ($m_{ku}$) is higher for E-G5$^2$ than for G52 because the entire E domain, plus a significant proportion of the G5$^2$ domain unfold between N and TS.

of unfolding ($\Delta G_{\mathrm{D-N}}^{\mathrm{H_2O}}$) calculated from the kinetic data match the equilibrium $\Delta G_{\mathrm{D-N}}^{\mathrm{H_2O}}$ values (Supplementary Table 1). Furthermore, double-jump stopped-flow experiments showed no evidence of additional phases that might reveal populated intermediates for either construct.

*Cooperative unfolding breaks down at high denaturant concentrations.* The unfolding of E-G5$^2$ and G5$^2$ results in a decrease in tyrosine fluorescence. However, in the unfolding kinetics of E-G5$^2$ *only*, at urea concentrations >7.0 M, we observed a second, faster rate, associated with an increase in fluorescence that shows very weak denaturant dependence (Fig. 1*D* and Supplementary Fig. 1). A similar extra phase was also observed for the E-G5$^2$ construct labelled with E500W-E532C$^{\mathrm{IAEDANS}}$ FRET pair (Fig. 1*D*), which reports specifically on the (un)folding of E. In contrast, the unfolding kinetics of E-G5$^2$ probed by I555W-E613C$^{\mathrm{IAEDANS}}$ (resulting in FRET only when G5$^2$ is folded) is monophasic (Fig. 1*D*). We infer that the minor rate detected at high urea concentration is related to unfolding of the E domain, perhaps when the stabilizing interface fails at high denaturant concentrations. Note that two other mutations that strongly destabilized the E domain (G524A and G527A) also decoupled the unfolding of E and G5$^2$ (Supplementary Fig. 2).

*G5$^2$ and E-G5$^2$ fold via the same highly polarized transition state.* Since G5$^2$ and E-G5$^2$ fold at the same rate, and the dependence of the refolding rate constant on denaturant concentration is the same (Fig. 1*D*), we infer that they fold via the same rate-limiting transition state. To map out which regions are structured early in the folding of G5$^2$ and E-G5$^2$ a mutational, Φ-value analysis was carried out. SasG domains do not have a compact hydrophobic core and all side chains are exposed to solvent. Mutation of surface residues rarely results in sufficient loss of stability to undertake Φ-value analysis (13). Hence, a series of non-conservative mutations (mainly Pro-to-Ala and Gly-to-Ala) were introduced in both G5$^2$ and E-G5$^2$, and their influence on the thermodynamic stability and kinetics was investigated (Supplementary Tables 2-5). Φ-values were calculated (Supplementary Tables 4,5) for mutants where the destabilization energy ($\Delta\Delta G_{\mathrm{D-N}}^{\mathrm{H_2O}}$) ≥ 0.7 kcal·mol$^{-1}$ (14). In general, non-conservative mutations, such as those we are using here, have to be interpreted with care. But the resultant chevron plots show that here we can be unequivocal (Fig. 2*A*,*B*). Unusually, mutations either alter only the folding kinetics, meaning Φ is close to 1 and the region is fully structured in the TS, *or* alter only the

unfolding kinetics meaning $\Phi \sim 0$, suggesting the region is as unstructured in the TS as in the denatured state. There are no intermediate Φ-values. When mapped onto the structures the Φ-value pattern is clear (Fig. 2*C,D*). It is only in the extreme C-terminal loop/β-sheet region that any structure is formed at all in the transition state ($\Phi \geq 0.8$) in both $G5^2$ *and* in E-$G5^2$. This reveals that the rate-limiting TS for folding is common for the two constructs and strongly polarized to the C-terminal region of the $G5^2$ domain. The rest of the protein folds only after formation of this initial embryonic structure, formation of which establishes the correct register for the β-strands of the $G5^2$ domain.

*Simulations reveal more details about the folding pathway.* After the main, rate-limiting TS our kinetic experiments are relatively "blind" to the subsequent steps. With simulations it is possible to probe the entire pathway. Long equilibrium simulations for $G5^2$ and E-$G5^2$ were carried out using a coarse-grained native-centric model, which allowed us to follow a number of unfolding and folding reactions. In all these simulations, the first step in the folding of both $G5^2$ and E-$G5^2$ is formation of the C-terminal β-sheet/loop motif of $G5^2$ (Fig. 3). In the case of E-$G5^2$, the C-terminal region of E folds concurrently with the N-terminal part of $G5^2$, resulting in formation of the E-G5 interface. This is followed by folding of the N-terminal β-sheet of E which completes the E-$G5^2$ structure (Fig. 3*B*); thus folding of the interface is key to the folding of E (See also Supplementary Fig. 3). At the mid-point temperature, where the proteins are folded 50% of the time (approximately 320K for both $G5^2$ and E-$G5^2$), we observed only a few complete folding events, as the domains are rarely fully unfolded. Hence we performed a large number of shorter simulations starting from completely unfolded structures (from simulations at high temperature) setting the temperature well below the folding temperature. Folding occurs in most of these short simulations and in all cases the sequence of events is that described above. In a few cases, where the E domain folds first, its unfolding is required before the E-$G5^2$ folds.

*The stability of the interface is essential to ensure cooperative unfolding of E-$G5^2$.* We identified two mutations in the E-domain of E-$G5^2$ (G517A and G548A), at the interface between the two domains, that, although the interface was sufficiently stable to promote the folding of the E domain, resulted in unfolding kinetics that were completely uncoupled; two unfolding phases are observed in *all* unfolding traces (Figure 4*A-C*). As was seen in wild-type (WT) E-$G5^2$, the fast unfolding phase,

ascribed to the unfolding of the E domain (which has a low amplitude and is associated with an increase in fluorescence) has a weak dependence on denaturant concentration. Importantly, the slower unfolding phase, associated with the larger fluorescence change, now has the unfolding *m*-value of the $G5^2$ domain alone, further evidence that, for these mutations at the interface, the E and $G5^2$ domains now unfold independently.

We investigated this further using the interface mutant P599A found in the $G5^2$ domain, which has no effect on the thermodynamic stability and kinetics of $G5^2$ in isolation but perturbs E-$G5^2$ (Fig. 4*D,E*). Pro599 is located in the N-terminal loop of $G5^2$. In the isolated domain Pro599 is exposed to solvent, whereas in the context of E-$G5^2$ it contributes to the hydrophobic cluster at the E-G5 inter-domain interface, where it makes contacts with Phe510 and Tyr547 from the E domain (Fig. 4*A*). We introduced the E500W-E532C$^{IAEDANS}$ FRET pair (Fig. 4*A*) in E-$G5^2$-P599A, which results in FRET only when E is folded. The unfolding kinetics were monitored by the decrease in 1,5-IAEDANS fluorescence (Fig. 4*E*), and at high denaturant concentrations that promote unfolding, a single phase was detected, corresponding to the faster unfolding phase found for E-$G5^2$-P599A (similar rate constants and the same urea-dependence) clearly representing unfolding of E uncoupled from $G5^2$. Note that we still observe the same single refolding phase for this mutant (except around the midpoint, Fig 4E), when followed by FRET because the folding of $G5^2$ is the rate-limiting step for folding of the E domain. Thus, again, we found that the interface is key to cooperative folding.

*Mutations reveal an alternative folding pathway for E-$G5^2$*. We found five destabilizing mutations within the $G5^2$ domain that alter the folding pathway in E-$G5^2$. Three of these (G576A, Y625W and G626A) are located in the C-terminal β-sheet/loop region of $G5^2$ (Fig. 5*A*) where, as shown above, folding is nucleated in both $G5^2$ and E-$G5^2$. These mutations destabilize the proteins by >1 kcal mol$^{-1}$ relative to WT $G5^2$ and E-$G5^2$ (Fig. 5*B,C*, Supplementary Tables 2,3). In $G5^2$ alone these three variants all have a Φ-value of ~1 that is, they unfold exactly as WT and all the change in stability is reflected in a change in the rate of folding (Fig. 2*A*, RH panel). Importantly, the dependence of the rate constant for folding on denaturant concentration ($m_{kf}$) is exactly the same as for WT $G5^2$. In E-$G5^2$, however, although these mutants again unfold exactly as WT now the *folding* kinetics are clearly

different (Fig. 5*D*). All still fold more slowly than WT but now the $m_{kf}$ values are significantly increased compared to WT suggesting that these variants are folding via a different, significantly more compact TS, with a $\beta_T = 0.53$ (compared to 0.33 for WT E-G5$^2$)[†].

Two other Gly-to-Ala mutations within the triple-helical region of G5$^2$ (G584A and G587A; Fig. 5*A*) destabilized the domain so significantly that the mutants are largely disordered at 0 M urea (Fig. 5*B* and Supplementary Table 2). In E-G5$^2$, these mutations are also destabilizing, but now both E and G5$^2$ are folded (Fig. 5*C* and Supplementary Table 3). Interestingly, the chevron plots of both E-G5$^2$-G584A and E-G5$^2$-G587A demonstrate the same $m_{k_f}$ value as the mutants that destabilize the extreme C-terminal region of E-G5$^2$ (Fig. 5*D*), suggesting that these variants also fold via a new, more compact, TS (with a $\beta_T$ of 0.53). Note that folding is still cooperative; in a control experiment the kinetics of E-G5$^2$-G584A recorded using the E500W-E532C$^{IAEDANS}$ FRET pair (reporting specifically on folding of E) were characterized by an identical $m_{k_f}$ to the one measured by intrinsic tyrosine fluorescence (Fig. 5*D*).

Thus, if we make mutations that significantly destabilize the folding nucleus at the extreme C-terminal end of the G5$^2$ domain, or mutations that are essential for formation of the triple helix connecting the nucleus to the rest of the protein, we apparently alter the folding pathway – but only when the E-domain is present.

*Formation of the interface is key to driving folding along the alternative pathway.* Crucially, for some of these mutations in the G5$^2$ domain (e.g. Y625W and G576A) the folding pathway of isolated G5$^2$ does not change; the new pathway is only accessible when the E-domain is present and yet we know that E does not fold in isolation. Given the importance of the interface between the two domains in imparting stability and cooperativity, we hypothesized that the alternative TS (characterized by $\beta_T$ of 0.53) involves formation of a structured E-G5$^2$ interface as an early step in this alternative pathway.

---

[†] The Tanford β-value, $\beta T = \left(\frac{m_{kf}}{m_{kf}+m_{ku}}\right)$ is a measure of the position of the transition state (in terms of SASA, or compactness) between D and N (11). An alternative explanation for a switch in $m_{kf}$ is that a mutation results in destabilisation of a TS that fall later on the same single pathway. Several lines of evidence suggest that this is a less reasonable explanation than parallel pathways. Only mutations that destabilise the WT pathway (with $\Phi \sim 1$) are affected; the same mutations in G5$^2$ alone do not result in a change in $m_{kf}$.; a residue with $\Phi \sim 1$ in WT has $\Phi \sim 0$ in Y625W (see below).

If this hypothesis is correct, then residues close to the E-$G5^2$ interface, in the E and $G5^2$ domains which all originally have a Φ-value ~0 should have increased Φ-values in this new pathway and residues in the region with high Φ-values in WT would have low Φ-values in this alternative pathway. We would also predict that a mutation that destabilized the interface could switch the new pathway back to the original polarized TS in E-$G5^2$. Thus we performed a mutational analysis based on Φ-values, in which E-$G5^2$-Y625W was treated as a pseudo-WT (Fig. 5*A*,*E,* Supplementary Table 6). (A crystal structure of the protein at 1.6 Å resolution reveals that this substitution does not disrupt the structure of $G5^2$ (See Supplementary Fig. 4 and Supplementary Table 7.) In that background we introduced a number of Pro-to-Ala mutations, most of which originally had Φ-values of 0 in the background of WT E-$G5^2$. P531A and P540A in E and P618 in $G5^2$ (all Φ~0) were designed to probe the folding of the individual domains, and P512A and P599A (also Φ~0) were designed to weaken the interface. P571A, which originally had Φ~1 is found in the C-terminal loop at the centre of the nucleation site for the WT pathway. Whilst half of the mutants (P512A, P531A and P618A) were insufficiently destabilizing to determine Φ-values in the background of E-$G5^2$-Y625W, three of the mutants gave us information.

*(i) The E domain is partly structured in the transition state of alternative pathway:* The P540A mutation resulted in a fractional Φ (0.7) in the context of E-$G5^2$-Y625W (compared to Φ-values of 0 for Gly-to-Ala mutations in the same region of the WT E domain). Folding is more affected than unfolding, implying that the triple-helix of the E domain is now significantly structured in the TS (Fig. 5*E*).

*(ii) The C-terminal loop of $G5^2$ is not formed in the transition state of the alternative pathway*: the P571A mutation now has no effect on the folding rate. The Φ-value is low in the background of E-$G5^2$-Y625W (Fig. 5*E*) (Φ = 0.1, compared to Φ = 1 in WT).

*(iii) If the interface is destabilized then E-$G5^2$ reverts to the original folding pathway:* the chevron plot of E-$G5^2$-Y625W-P599A shows the same $m_{k_f}$ as E-$G5^2$-P599A and WT E-$G5^2$, indicative of the WT-like folding pathway (Fig. 5*E*). We infer that the mutation P599A at the E-$G5^2$ interface destabilizes the new TS and causes folding to revert to the original, WT pathway. These results confirm that the new TS involves

formation of structure at the interface between the two domains in the alternative folding pathway.

**Discussion**

SasG is a protein that challenges some of our preconceptions of protein structure and folding. First, it has an unusual sequence composition typical of an intrinsically disordered protein (~60% of the residues are charged, Pro or Gly), yet it demonstrably folds cooperatively – albeit to an unusual single-sheet extended structure. Despite this unusual structure, the biophysical parameters for folding (*m*-value, stability) are quite unremarkable for a protein of this size (E-G5 and G5 have 132 and 82 residues, respectively). What *is* remarkable is that G5 domains fold far more rapidly than might be predicted from their relative contact order (15) (Supplementary Fig 5). The interface between the E and $G5^2$ domains provides most of the stability for the protein. This is exemplified when we consider the mutation of two highly conserved Gly residues in the triple helical region of the $G5^2$ domain (G584A and G587A) which both cause $G5^2$ to be unfolded; when we mutate these same residues in E-$G5^2$, the protein folds (Fig. 5*B,C*). Thus we can take an unfolded $G5^2$ domain, add an intrinsically unfolded E domain and produce a folded protein. We have estimated that the interface imparts at least 6 kcal $mol^{-1}$ to the stability of E-$G5^2$ (compared with $\Delta G_{D\text{-}N}$ for WT $G5^2$ and E of 2.8 and $\leq$ -2.5 kcal $mol^{-1}$, respectively) (9). This interface is also key to maintaining cooperative folding and for the long-range cooperativity that imparts stiffness to the SasG structure. Here we have demonstrated that the interface is essential to ensure that the entire E-G5 motif folds and unfolds in a single cooperative step – mutations at the interface disrupt cooperative folding. And yet, to our surprise, our data suggest that the interface between E and $G5^2$ is completely unformed at the transition state for folding (the E-domain and the N-terminal region of the $G5^2$ domain are both unstructured).

Our data show that folding of SasG is initiated at the far C-terminal end of the $G5^2$ domain. At this point there is a turn between the two outer β-strands and the terminal, 'docking' strand is inserted between these, into the loop (Fig. 3). Assembly of this small structural element in one domain is sufficient to drive folding of the entire E-G5 molecule over a distance of more than 10 nm. However, folding at the interface is clearly an option, since destabilization of the C-terminal nucleation site

allows folding via a higher energy transition state where formation of the interface is key. E-G5$^2$ can thus fold via parallel pathways but the lowest energy pathway involves formation of the C-terminal nucleus. It is unclear why this WT pathway should be lower in energy than a pathway involving formation of the interface, the most stable region of the structure and essentially the only region where there is any significant burial of hydrophobic residues. It may be because the entropic cost of forming the interface is larger; it involves bringing together loops from the E and the G5 domain that are distant in sequence (~85 residues apart), although the interactions in the C-terminal nucleus are by no means short range (~ 50 residues between the C-terminal residues of the final strand and the turn). Alternatively, the intrinsic disorder of the E domain may again be key. The formation of the interface involves the folding, at least in part, of the E domain, a process that is inherently costly in terms of free energy. Importantly, however, cooperative folding is a feature of both pathways, because the E domain cannot fold in the absence of G5.

In wild-type protein (except under very destabilising conditions, as described) the protein folds and unfolds as a single unit; no intermediates are populated in folding or in unfolding, or at equilibrium. This is, by definition, cooperative folding. Such tight and robust cooperativity in folding behaviour has not been seen previously in multi-domain proteins. Even where there are significant interfaces between domains, kinetics reveal that the domains fold in a non-two state manner, with each domain behaving as an independent folding unit (16, 17). The obligate cooperativity of SasG arises because E can only fold in the presence of folded G5, but once folded the entire domain is very significantly more stable than the sum of the stability of the two domains individually.

The kind of cooperativity we are observing in the SasG protein ('obligate' folding cooperativity) is reminiscent of the folding of repeat proteins. These comprise tandem arrays of small repeats (20-40 residues) that are unstable on their own, and which fold, apparently cooperatively, through formation of interfaces between the repeats (18-25). But tandem repeats are very different to SasG, where contacts within the domains themselves and between domains are very long-range, whereas contacts in repeat proteins are very local (Supplementary Fig. 5). While there is a dominant folding pathway in SasG, parallel pathways are a key feature of repeat proteins, in particular as the number of repeats increases. However, despite each subunit being intrinsically unstable alone, kinetic cooperativity is not generally maintained beyond

3-4 subunits in repeat proteins, but SasG is able to maintain cooperative folding across a distance of ~12 nm.

**Conclusion**

The importance of intrinsic disorder in Biology is becoming increasingly apparent, but why would Nature choose disordered domains to form a multi-domain protein? We had previously shown that disorder-mediated thermodynamic cooperativity allows SasG to adopt long, mechanically strong, rod-like structures (9). Now we have shown how this disorder, coupled with the remarkable stability of the inter-domain interface, can result in cooperative folding kinetics, with no populated intermediates, across long distances. The folding of classic multi-domain proteins is highly cooperative, but only within the relatively local confines of a single domain. In repeat proteins short-range cooperativity is apparent between 3-4 individually unstable repeats. SasG provides a paradigm for much longer-range cooperative folding – by the obligatory folding of alternate intrinsically disordered domains with their folded neighbors.

**Materials and Methods**

All experimental procedures are described in detail in the Supplementary Information.

*Analysis of kinetic data:* For some mutants kinetic data were fitted to a model allowing for parallel pathways (see Supplementary Fig. 6 for details).

*Simulations:* Simulations were performed using a coarse-grained model where only $C_\alpha$ atoms are represented and interactions depend on the native reference structure and on the residue type (26). Details are given in the Supplementary Information.

*Determination of the structure of E-G5$^2$-Y625W:* Details of the crystallization and structure determination of E-G5$^2$ Y625W can be found in the Supplementary Information. The coordinates and structure factors have been deposited in the protein data bank with accession code 5DBL.

**Acknowledgements**

This research was supported by Biotechnology and Biological Research Council Grants BB/J006459/1 (D.T.G. and J.C.), BB/J005029/1 (F.W. and J.R.P), C.A.T.F.M is supported by the Cambridge Trust and CAPES Science without Borders Cambridge Scholarship. J.R.P holds a British Heart Foundation Senior Basic Science Fellowship (FS/12/36/29588). J.C. is a Wellcome Trust Senior Research Fellow (WT/095195).

**Figure Legends**

**Fig. 1. Structure and biophysical data for wild-type (WT) SasG $G5^2$ and E-$G5^2$.** (*A*) Schematic representation of SasG from *S. aureus* NCTC 8325. The A domain promotes adhesion to host cells. The core region comprises tandemly arrayed G5 (red) and E (blue) domains (10). The E-$G5^2$ fragment of SasG is indicated with a bar. (*B*) Structure of E-$G5^2$ (PDB accession: 3TIP) illustrating the topology of E and $G5^2$ domains: two single-layer triple-stranded β-sheets connected by a central collagen-like triple-helical region. The tyrosines and positions of engineered FRET pairs are shown. FRET pair E500W-E532C$^{IAEDANS}$ (cyan) results in FRET only when E is folded; I555W-E613C$^{IAEDANS}$ (green) results in FRET when $G5^2$ is folded. (*C*) Equilibrium denaturation curves. Data for WT $G5^2$ and E-$G5^2$, and E-$G5^2$-E500W-E532C$^{IAEDANS}$ taken from (9). (*D*) Urea dependence of the natural logarithm of the observed rate constants (in $s^{-1}$) for proteins shown in *C*. Circles and squares represent major and minor unfolding rate constants, respectively.

**Fig. 2. Mapping the structure of transition states of WT folding pathway for $G5^2$ and E-$G5^2$.** (*A*) Chevron plots for $G5^2$: WT (black) and mutants. (*B*) Chevron plots for E-$G5^2$: WT (black) and mutants. (*A,B*) Left panels: mutants that unfold faster than WT while the folding rate is largely unaffected. Right panels: mutants that fold slower than WT while the unfolding rate is unaffected. (*C,D*) Φ−values of (*C*) $G5^2$, and (*D*) E-$G5^2$ mapped onto the crystal structures. Blue, high Φ-values (> 0.8); Red, low Φ-values (< 0.2); Grey: where $\Delta\Delta G$ was not high enough to obtain reliable Φ-values.

**Fig. 3. Probing the folding pathways of SasG using simulations.** Simulations of (*A*) $G5^2$ and (*B*) E-$G5^2$ by coarse-grained native-centric model simulations at 320 K. Top panel shows the root mean square deviation (RMSD) as a function of simulation time for a typical refolding event. For $G5^2$ (*A*), RMSD values were calculated for all atoms (black), the C-terminal β-sheet/loop region (cyan) and the N-terminal β-sheet/loop region (red). For E-$G5^2$ (*B*), RMSD values were calculated for all atoms (black), the C-terminal β-sheet/loop region of $G5^2$ (cyan), the N-terminal β-sheet/loop region of $G5^2$ together with the C-terminal β-sheet/loop region of E (red) and the N-terminal β-sheet/loop region of E (orange). The bottom panel illustrates corresponding sequential snapshots from the refolding trajectory and the related schematic topology representation. The $G5^2$ domain is shown in red, except for the C-terminal β-

sheet/loop region (cyan) and its central C-terminal 'docking' strand (green). The E domain is shown in blue. Further details from the same trajectory are illustrated in Supplementary Fig. 3.

**Fig. 4. Mutations at the interface break the cooperative unfolding of E-G5$^2$.** (*A*) Structure of E-G5$^2$ showing the location of mutated residues within the E domain (light blue) and G5$^2$ domain (Pro599; orange). Phe510 and Tyr547 (grey) contact Pro599. (*B,C*) Mutations in the E domain: (*B*) Equilibrium denaturation curves and (*C*) urea dependence of the natural logarithm of the observed rate constants for WT and mutants. (*D,E*) Mutations in the G5$^2$ domain: (*B*) Equilibrium denaturation curves and (*C*) urea dependence of the natural logarithm of the observed rate constants for WT G5$^2$ and mutants. Circles and squares in *C* and *E* represent major and minor rate constants, respectively. Mutations at the interface result in the breakdown of the cooperative unfolding of the E and G5$^2$ domains, manifested in the presence of a second unfolding rate constant at all denaturant concentrations, and a decrease in the dependence of ln$k_u$ on [urea].

**Fig. 5. E-G5$^2$ can fold by an alternative folding pathway.** Mutations in the G5 domain that destabilize the folding nucleus cause a switch in pathway in E-G5$^2$, manifested by a change in the dependence of ln$k_f$ on [urea]. (*A*) Structure of E-G5$^2$ showing the location of residues mutated or used to engineer the FRET pair. (*B, C*) Equilibrium denaturation curves for G5$^2$ and E-G5$^2$ respectively (*D*) Chevron plots for WT E-G5$^2$ and mutants. Note the change in slope of the folding limb of the chevron plot for all of these mutants. (*E*) Mutations using Y625W as a pseudo-WT. Chevron plots for WT E-G5$^2$ (black), E-G5$^2$-Y625W (green) and Pro-to-Ala mutants of E-G5$^2$ in the background of Y625W. Note that the interface mutant (P599A) causes the slope to revert to WT. The other mutants have Φ-values that differ from those in the WT background (see text).

## References


1. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, & Jones DT (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 337(3):635-645.
2. Tskhovrebova L & Trinick J (2003) Titin: properties and family relationships. *Nat Rev Mol Cell Biol* 4(9):679-689.

3. Milles S*, et al.* (2015) Plasticity of an ultrafast interaction between nucleoporins and nuclear transport receptors. *Cell* 163(3):734-745.
4. Shammas SL, Travis AJ, & Clarke J (2014) Allostery within a transcription coactivator is predominantly mediated through dissociation rate constants. *Proc Natl Acad Sci U S A* 111(33):12055-12060.
5. Law SM, Gagnon JK, Mapp AK, & Brooks CL, 3rd (2014) Prepaying the entropic cost for allosteric regulation in KIX. *Proc Natl Acad Sci U S A* 111(33):12067-12072.
6. Corrigan RM, Rigby D, Handley P, & Foster TJ (2007) The role of *Staphylococcus aureus* surface protein SasG in adherence and biofilm formation. *Microbiology* 153(Pt 8):2435-2446.
7. Geoghegan JA*, et al.* (2010) Role of surface protein SasG in biofilm formation by Staphylococcus aureus. *J Bacteriol* 192(21):5663-5673.
8. Formosa-Dague C, Speziale P, Foster TJ, Geoghegan JA, & Dufrêne YF (2016) Zinc-dependent mechanical properties of Staphylococcus aureus biofilm-forming surface protein SasG. *Proc Natl Acad Sci U S A* 113:410-415.
9. Gruszka DT*, et al.* (2015) Cooperative folding of intrinsically disordered domains drives assembly of a strong elongated protein. *Nat Commun* 6:7271.
10. Gruszka DT*, et al.* (2012) Staphylococcal biofilm-forming protein has a contiguous rod-like structure. *Proc. Natl Acad. Sci. USA* 109(17):E1011-E1018.
11. Fersht AR (1999) *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding* (W. H. Freeman and Company, New York).
12. Bachmann A & Kiefhaber T (2001) Apparent two-state tendamistat folding is a sequential process along a defined route. *J. Mol. Biol.* 306(2):375-386.
13. Fersht AR, Matouschek A, & Serrano L (1992) The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J Mol Biol* 224(3):771-782.
14. Fersht AR & Sato S (2004) Phi-value analysis and the nature of protein-folding transition states. *Proc Natl Acad Sci U S A* 101(21):7976-7981.
15. Plaxco KW, Simons KT, & Baker D (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol* 277(4):985-994.
16. Batey S, Nickson AA, & Clarke J (2008) Studying the folding of multidomain proteins. *HFSP J* 2(6):365-377.
17. Batey S & Clarke J (2006) Apparent cooperativity in the folding of multidomain proteins depends on the relative rates of folding of the constituent domains. *Proc Natl Acad Sci U S A* 103(48):18113-18118.
18. Werbeck ND, Rowling PJ, Chellamuthu VR, & Itzhaki LS (2008) Shifting transition states in the unfolding of a large ankyrin repeat protein. *Proc Natl Acad Sci U S A* 105(29):9982-9987.
19. Werbeck ND & Itzhaki LS (2007) Probing a moving target with a plastic unfolding intermediate of an ankyrin-repeat protein. *Proc Natl Acad Sci U S A* 104(19):7863-7868.

20. Lowe AR & Itzhaki LS (2007) Biophysical characterisation of the small ankyrin repeat protein myotrophin. *J Mol Biol* 365(4):1245-1255.
21. Tang KS, Fersht AR, & Itzhaki LS (2003) Sequential unfolding of ankyrin repeats in tumor suppressor p16. *Structure* 11(1):67-73.
22. Tripp KW & Barrick D (2008) Rerouting the folding pathway of the Notch ankyrin domain by reshaping the energy landscape. *J Am Chem Soc* 130(17):5681-5688.
23. Barrick D, Ferreiro DU, & Komives EA (2008) Folding landscapes of ankyrin repeat proteins: experiments meet theory. *Curr Opin Struct Biol* 18(1):27-34.
24. Bradley CM & Barrick D (2006) The notch ankyrin domain folds via a discrete, centralized pathway. *Structure* 14(8):1303-1312.
25. Mello CC, Bradley CM, Tripp KW, & Barrick D (2005) Experimental characterization of the folding kinetics of the notch ankyrin domain. *J Mol Biol* 352(2):266-281.

Figures

Figure 1

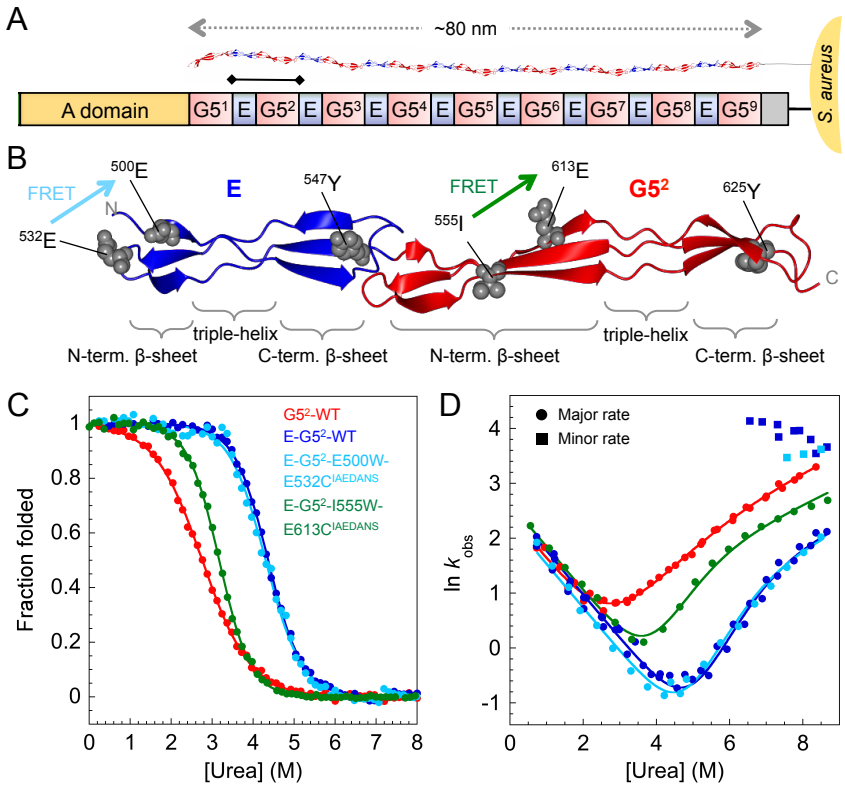


Figure 2

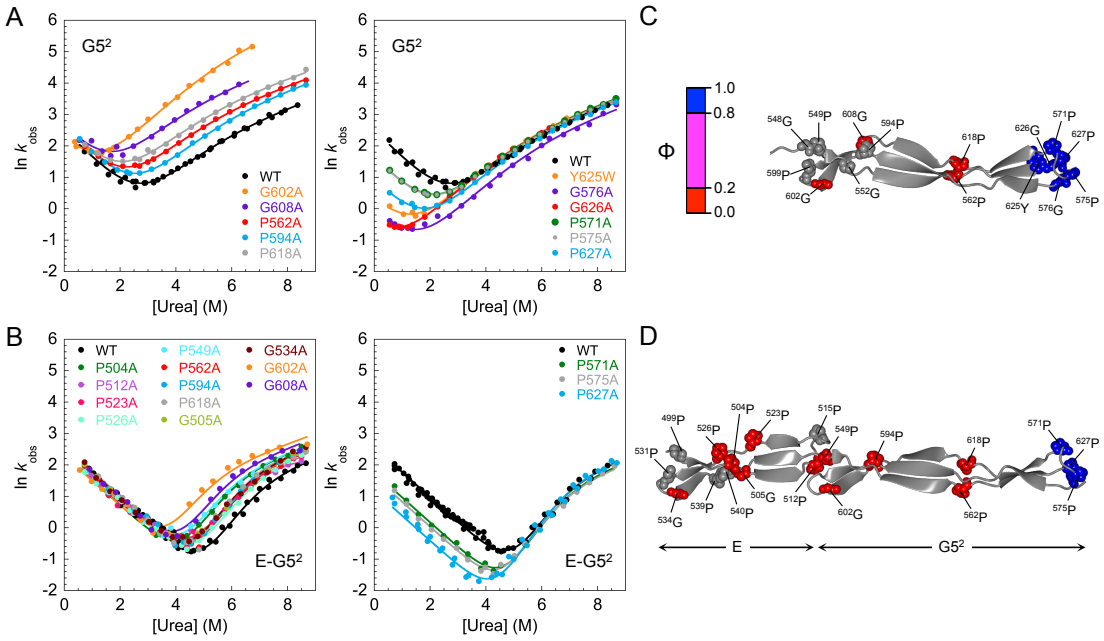

A
G5²
ln k_obs
[Urea] (M)
WT
G602A
G608A
P562A
P594A
P618A
G5²
ln k_obs
[Urea] (M)
WT
Y625W
G576A
G626A
P571A
P575A
P627A
C
Φ
1.0
0.8
0.2
0.0
548G
549P
608G
594P
618P
626G
571P
627P
599P
602G
552G
562P
625Y
576G
575P
B
WT
P504A
P512A
P523A
P526A
P549A
P562A
P594A
P618A
G505A
G534A
G602A
G608A
ln k_obs
E-G5²
[Urea] (M)
WT
P571A
P575A
P627A
ln k_obs
E-G5²
[Urea] (M)
D
499P
526P
504P
523P
515P
549P
531P
534G
539P
540P
505G
512P
602G
594P
618P
571P
627P
562P
575P
E
G5²

Figure 3

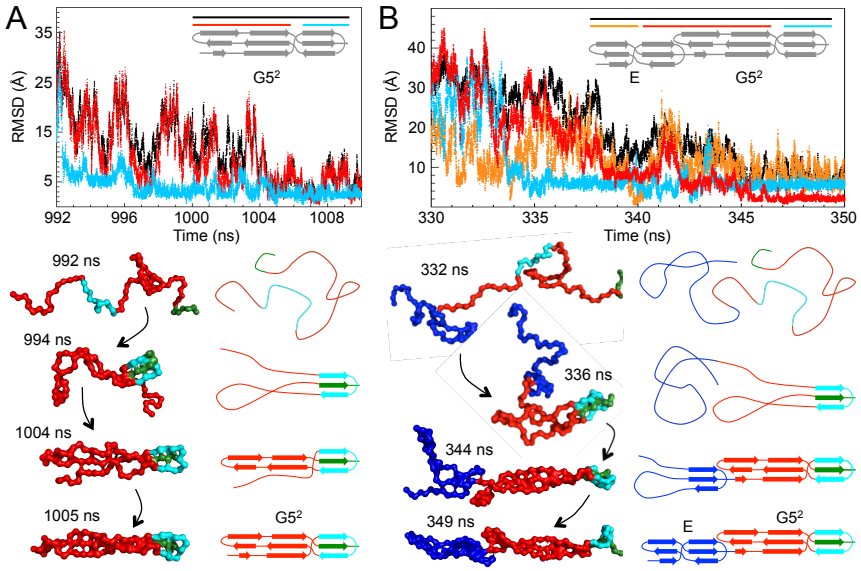

A
RMSD (Å)
35
25
15
5
992
996
1000
1004
1008
Time (ns)
G5²
992 ns
994 ns
1004 ns
1005 ns
G5²
B
RMSD (Å)
40
30
20
10
330
335
340
345
350
Time (ns)
E
G5²
332 ns
336 ns
344 ns
349 ns
E
G5²

Figure 4

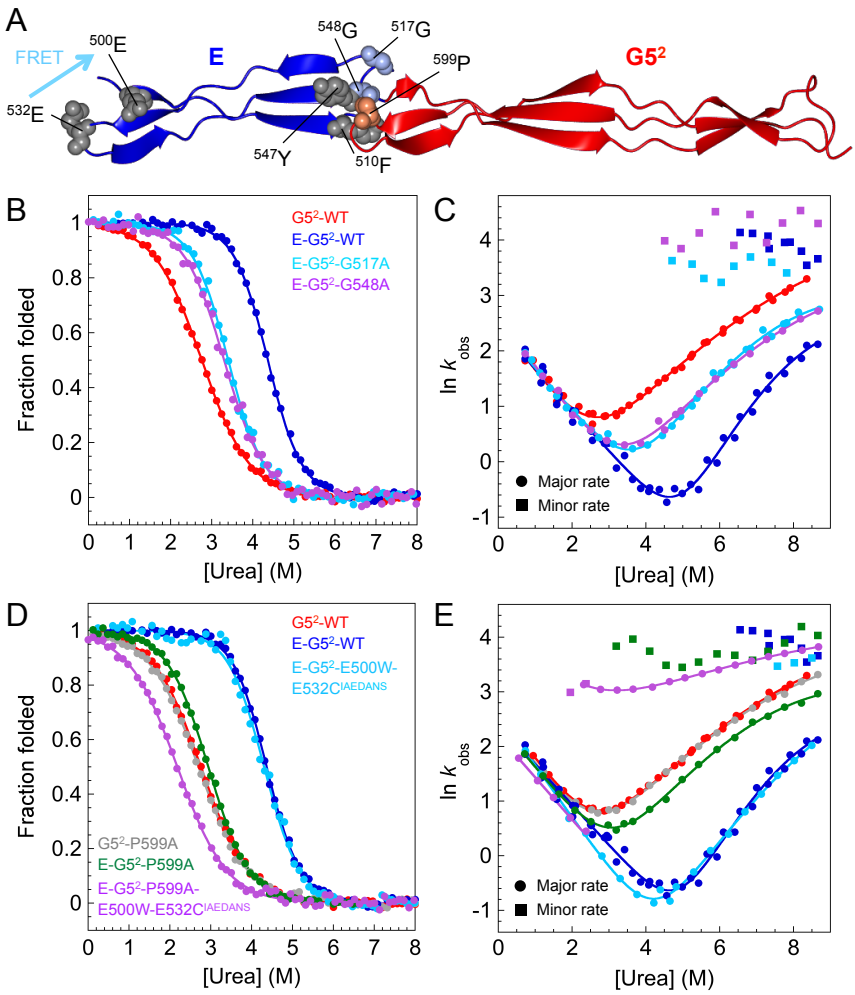
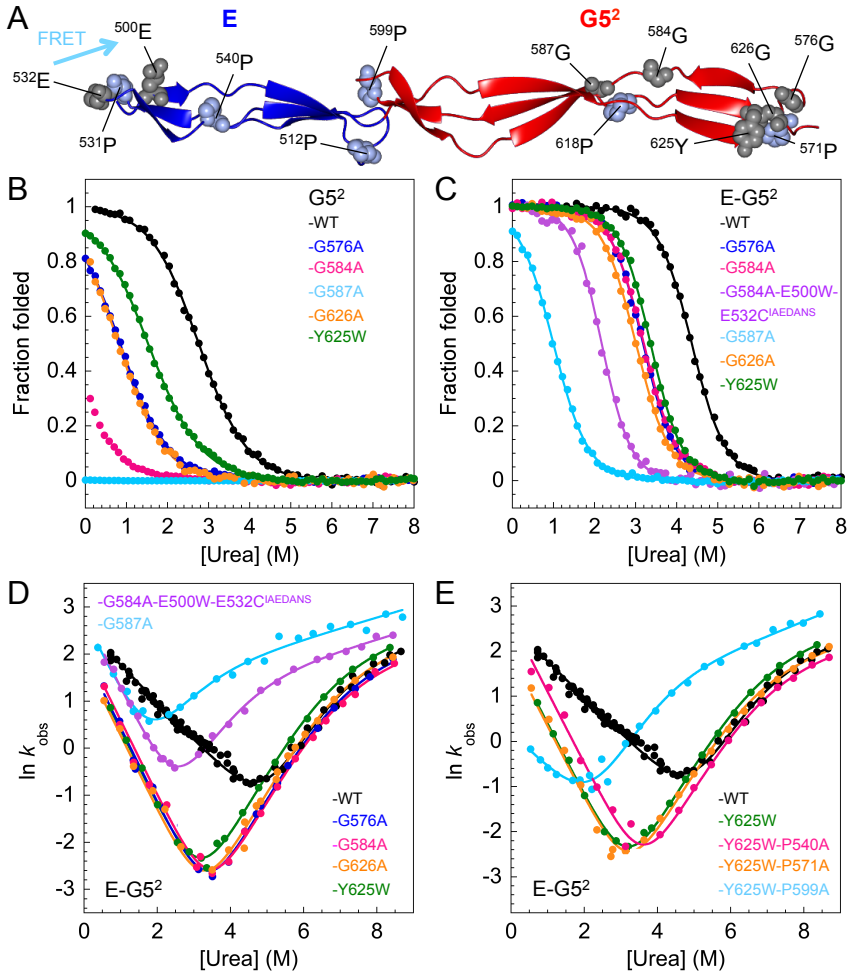

A
FRET
500E
532E
E
548G
517G
599P
547Y
510F
G5²
B
G5²-WT
E-G5²-WT
E-G5²-G517A
E-G5²-G548A
Fraction folded
[Urea] (M)
C
ln k obs
Major rate
Minor rate
[Urea] (M)
D
G5²-WT
E-G5²-WT
E-G5²-E500W-E532C IAEDANS
G5²-P599A
E-G5²-P599A
E-G5²-P599A-E500W-E532C IAEDANS
Fraction folded
[Urea] (M)
E
ln k obs
Major rate
Minor rate
[Urea] (M)

Figure 5


A
FRET
E
G5²
500E
532E
531P
540P
512P
599P
587G
618P
584G
626G
625Y
576G
571P
B
G5²
-WT
-G576A
-G584A
-G587A
-G626A
-Y625W
Fraction folded
[Urea] (M)
C
E-G5²
-WT
-G576A
-G584A
-G584A-E500W-E532C^IAEDANS
-G587A
-G626A
-Y625W
Fraction folded
[Urea] (M)
D
-G584A-E500W-E532C^IAEDANS
-G587A
-WT
-G576A
-G584A
-G626A
-Y625W
E-G5²
ln k_obs
[Urea] (M)
E
-WT
-Y625W
-Y625W-P540A
-Y625W-P571A
-Y625W-P599A
E-G5²
ln k_obs
[Urea] (M)

**Supplementary Information**

**Supplementary Methods:**

*Protein production and purification*: SasG $G5^2$ and E-$G5^2$ (WT and mutants) expression and purification procedures were as previously described (1, 2).

*FRET labels:* Tryptophan (E500W and I555W) and cysteine (E532C and E613C) residues were introduced into SasG $G5^2$ and E-$G5^2$ constructs by site-directed mutagenesis. Both, E-$G5^2$-E500W-E532C and $G5^2$-I555W-E613C were labelled with 5-((((2-iodoacetyl)amino)ethyl)amino)naphthalene-1-sulfonic acid (1,5-IAEDANS; Life Technologies) following the manufacturer's instructions as described previously (1).

*Equilibrium studies* Equilibrium unfolding of the proteins was studied by urea denaturation under standard conditions (phosphate-buffered saline, 25°C). Folding was followed by intrinsic tyrosine (WT, proline- and glycine-to-alanine mutants, excitation wavelength 276 nm; emission 305 nm) and tryptophan (Y265W; excitation 280 nm; emission 350 nm) fluorescence and FRET measurements (excitation 280 nm: emission 490 nm) on a fluorescence spectrometer (Perkin Elmer LS55). The data were analyzed as previously described (1).

*Kinetic studies*: Kinetic experiments following the change in the fluorescence signal at different urea concentrations were carried out using a stopped-flow fluorimeter (Applied Photophysics SX.20) at 25°C constant temperature, as described previously (1). The data were fitted to equations describing single- or double-exponential phases (see text). To account for non-linearity in the observed unfolding rate constant, the chevron plot data were fitted to a sequential transition states model as described previously (3), in which denaturant induces a switch between two barriers separated by a high-energy intermediate.

**Φ**-values were determined using the following equation:

$$\Phi = \frac{\Delta\Delta G_{\text{D-}\ddagger}}{\Delta\Delta G_{\text{D-N}}}$$

Where $\Delta\Delta G_{\text{D-N}}$ was determined using equilibrium experiments, and $\Delta\Delta G_{\text{D-}\ddagger} = RT \ln\left(\frac{k_{\text{wt}}^{\text{H2O}}}{k_{\text{mut}}^{\text{H2O}}}\right)$.

Native reference structures were the crystal structure of *S. aureus* SasG E-$G5^2$ (PDB accession: 3TIP) for both E-$G5^2$ and the $G5^2$ domain alone.

*Simulations:* Simulations were performed using a coarse-grained model where only $C_\alpha$ atoms are represented and interactions depend on the native reference structure and on the residue type (4).

Equilibrium simulations were performed at a broad range of temperatures between 270 and 330 K lasting at least 30 μs. Temperature was controlled using Langevin dynamics, and the timestep for integration of the equations of motion was 15 fs. For both systems the mid-point temperature was approximately 320K. At this temperature E-G5$^2$ completely unfolds only a few times, hence we performed 62 simulations starting from random conformations sampled at 350K and setting the thermostat to temperatures between 270 and 315K at which E-G5$^2$ is expected to be folded. For all simulations in which full folding occurs, the pathway is identical to those observed during the equilibrium simulation reported in Fig. 3.

*Crystallisation of E-G5$^2$-Y625W:* E-G5$^2$ Y625W was purified as described previously (5) and concentrated to 47.6 mg.ml$^{-1}$ in 20 mM Tris, 150 mM NaCl, pH 8. Crystallisation screening with JCSG+ (Molecular Dimensions; (6)) resulted in growth of large single crystals in conditions comprising 100 mM citrate pH 5 and 20% PEG 6000. Crystals were flash cooled in liquid $N_2$ prior to data collection on Diamond beamline I02. Data were indexed, integrated and scaled using XDS (7) and merged using Aimless (8). Phases were determined by molecular replacement with PhaserMR (9) using WT E-G5$^2$ (PDB accession: 3TIP (5)). E-G5$^2$-Y625W crystallised in spacegroup *C*2 with one molecule in the asymmetric unit. The model was improved using Coot (10) and refined to 1.6 Å (Supplementary Table 7) with nine translation/libration/screw (TLS) groups by Phenix (11). The coordinates and structure factors have been deposited in the protein data bank with accession code 5DBL. The structure was aligned by secondary structure matching with WT E-G5$^2$ using Superpose (12) and cartoons were rendered with CCP4mg (13).

**Supplementary Figures:**

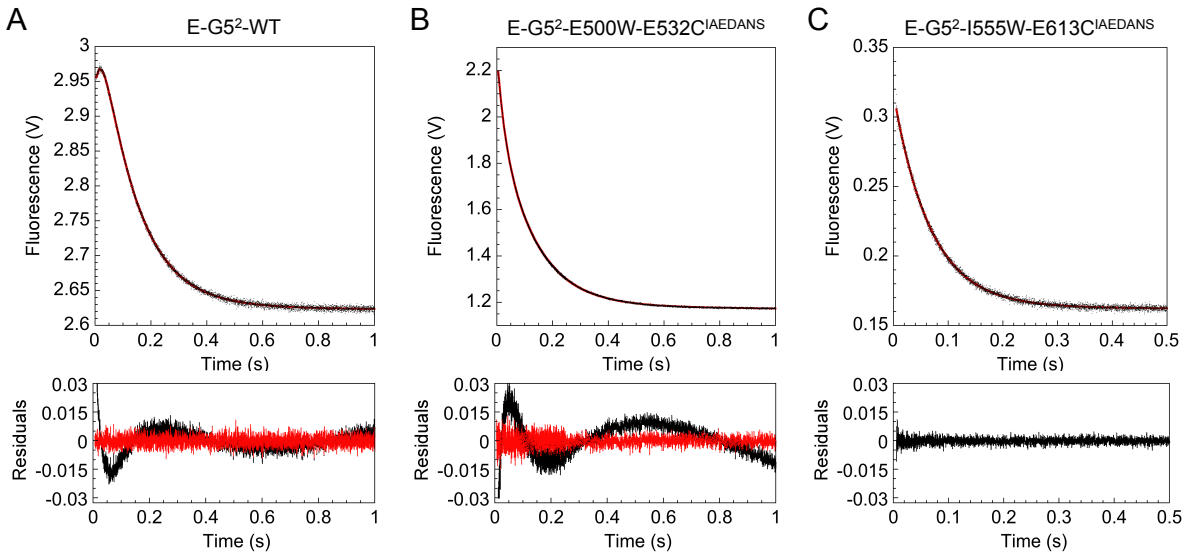



**Fig. S1. At high denaturant concentrations two unfolding phases are observed in E-G5$^2$, as unfolding of the domains becomes uncoupled.** Kinetics of unfolding into 9.5 M urea for E-G5$^2$-WT (*A*), E-G5$^2$-E500W-E532C$^{IAEDANS}$ (*B*) and E-G5$^2$-I555W-E613C$^{IAEDANS}$ (*C*). Traces were collected by monitoring the change in intrinsic tyrosine or 1,5-IAEDANS fluorescence. Unfolding traces of E-G5$^2$-WT (*A*) and E-G5$^2$-E500W-E532C$^{IAEDANS}$ (*B*) were fitted to the sum of two exponentials, which describes the data better than the single exponential. Residuals for the fit to the single exponential and the sum of two exponentials are shown below the data in black and red, respectively. Unfolding traces of E-G5$^2$-I555W-E613C$^{IAEDANS}$ (*C*) (that monitors the unfolding of the G5 domain only) were fitted to a single exponential, which describes the data well. The residuals are shown below the data.

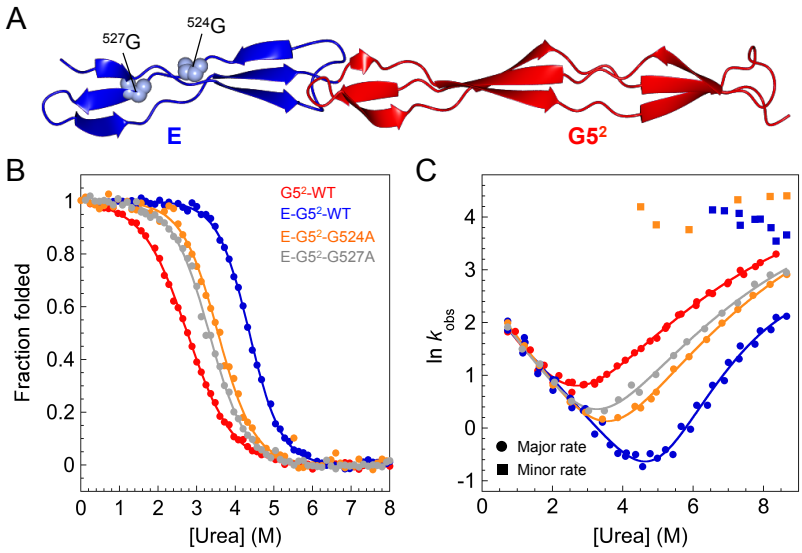



**Fig. S2. Highly destabilizing mutations in the E-domain break the cooperative unfolding of E-G5$^2$.** (*A*) Structure of E-G5$^2$ showing the location of mutated residues within the E domain (Gly524, Gly527 light blue spheres) (*B*) Equilibrium denaturation curves and (*C*) urea dependence of the natural logarithm of the observed rate constants for wild type and mutant proteins. Circles and squares represent major and minor rate constants, respectively. Note that the unfolding *m*-value of these two mutants reverts to that of wild-type G5$^2$ showing that the E and G5$^2$ domains are now unfolding independently.

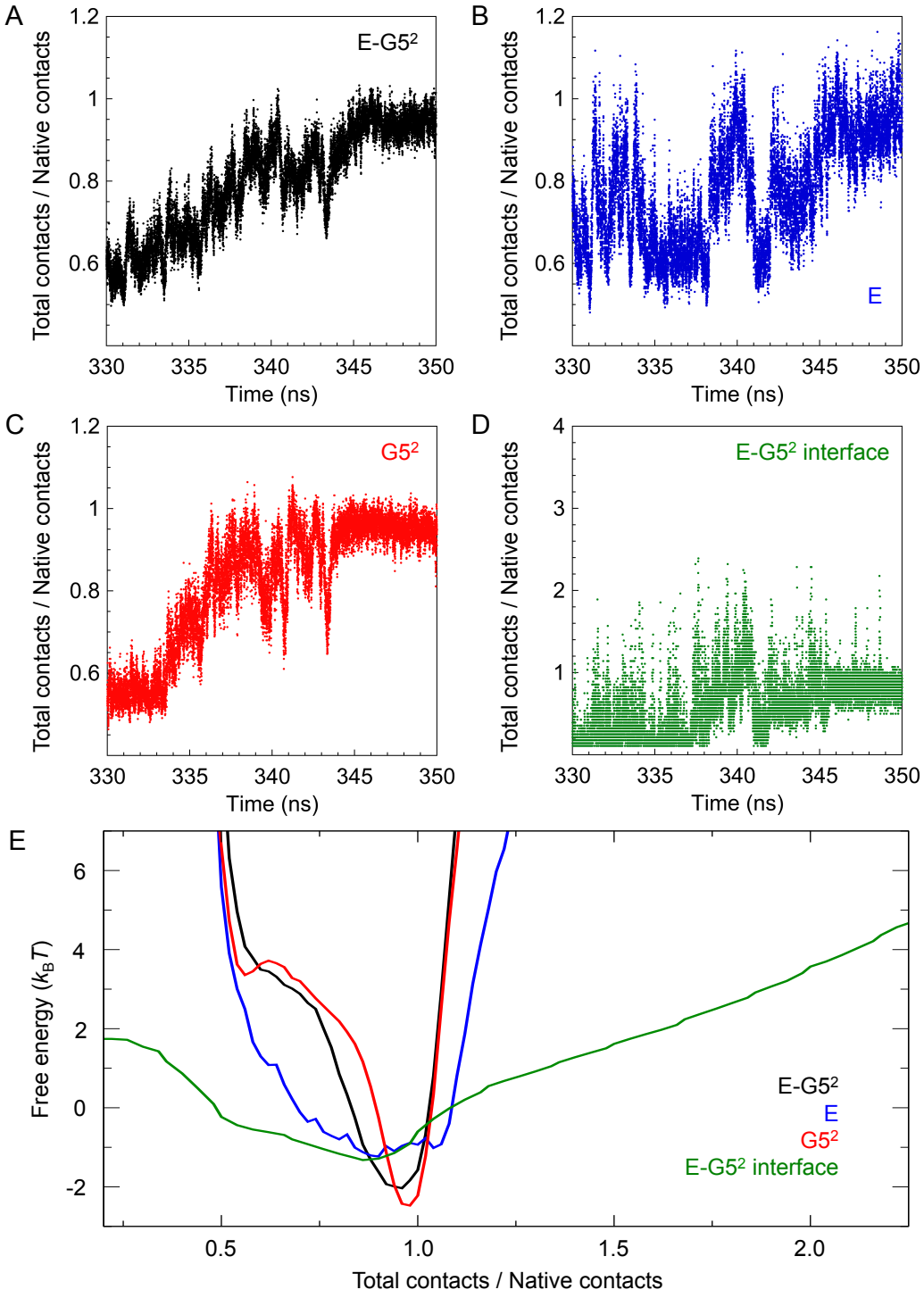



**Fig. S3. Simulations of E-G5$^2$ at 320 K.** Trajectories of the total contacts normalized by the number of native contacts for E-G5$^2$ (*A*), E (*B*), G5$^2$ (*C*) and the E-G5$^2$ interface (*D*), for the same folding event as presented in Fig. 3. Panel *E* shows the free energy change as a function of the ratio of total contacts to native contacts. Domain E is characterized by a broad basin that encompasses both folded and unfolded states whereas the G5$^2$ domain shows a barrier between the unfolded and folded state (at 320 K the native state is much more populated that the unfolded state).

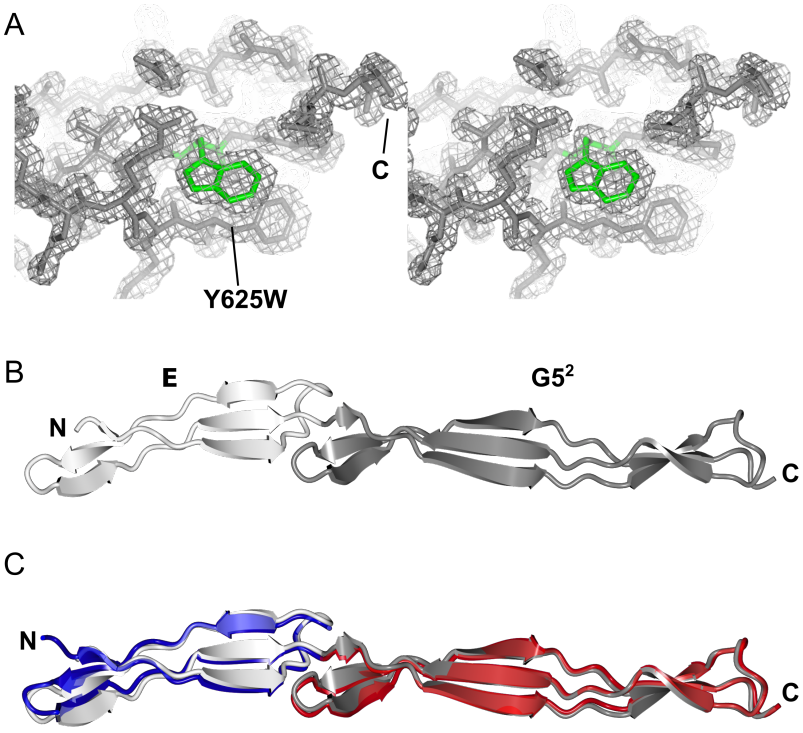



**Fig. S4. The structure of E-G5$^2$-Y625W is highly similar to the wild type protein fold.** An X-ray crystal structure of E-G5$^2$-Y625W was determined at 1.6 Å resolution (PDB: 5DBL). (*A*) Stereo image of the *2mFo-DFc* electron density map (grey) contoured at 1 electron/Å$^3$ at the C-terminus of G5$^2$; the Y625W side-chain is shown in green. (*B*) The X-ray crystal structure of E-G5$^2$ Y625W (E, white and G5$^2$, grey) is highly similar to the wild type (PDB accession: 3TIP E, blue and G5$^2$, red). Alignment by secondary structure matching revealed a Cα root mean square deviation of 1 Å (*C*), confirming the Y625W mutation does not affect the overall structure of E-G5$^2$.

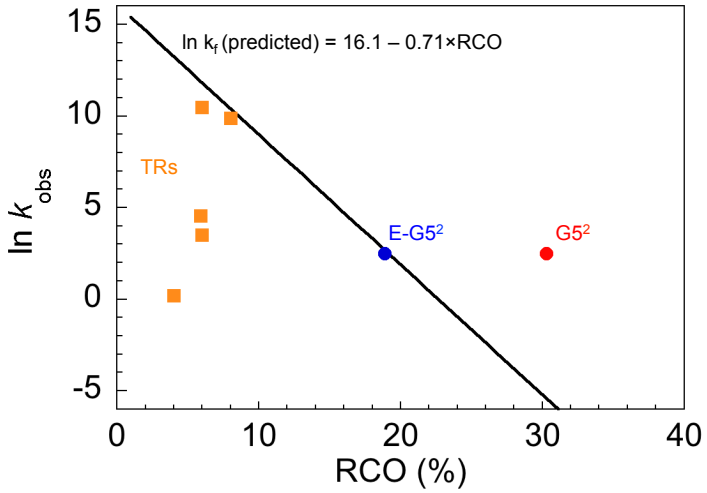



**Fig. S5. Plot of rate of folding vs contact order.** $G5^2$ folds significantly more rapidly than would be predicted from its relative contact order. (Data from Plaxco et al (14) shown by the straight line). Although it has some properties of a repeat protein SasG clearly lacks the short-range interactions that characterize all true tandem repeat proteins (TRs; examples include leucine-rich repeats, ankyrin repeats, and tetratricopeptide repeats, orange).

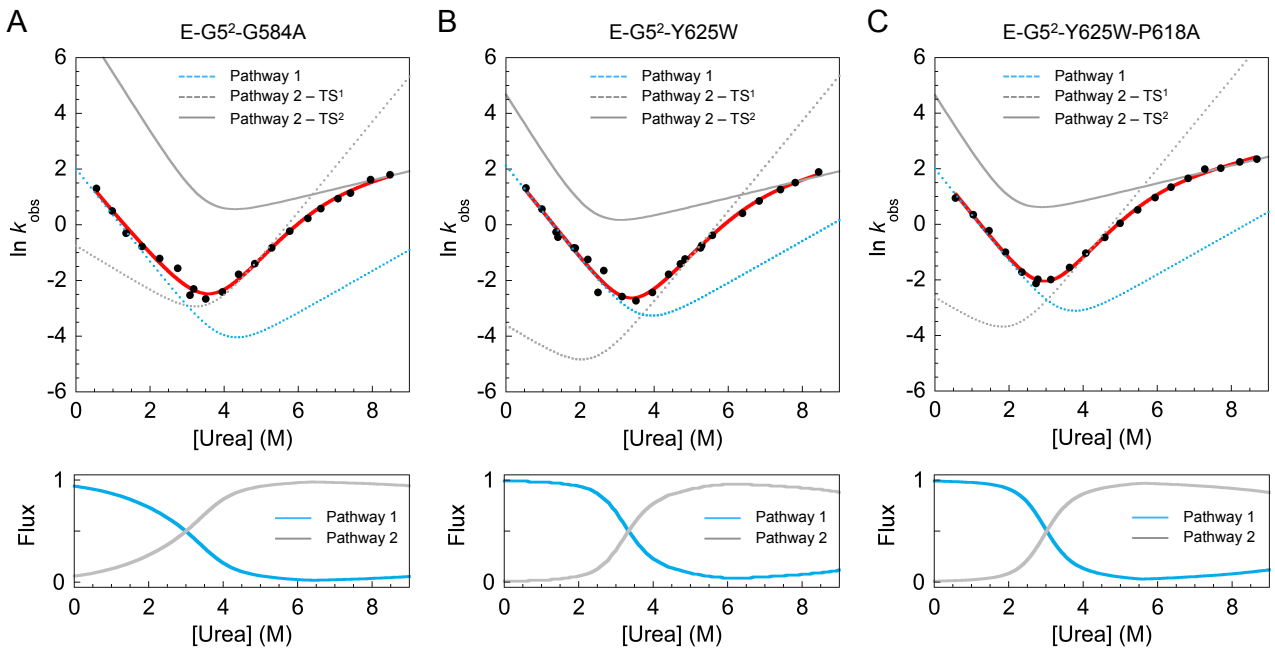



**Fig. S6. Fitting the data to parallel pathways models.** Chevron and flux plots for representative E-G5$^2$ mutants that fold via an alternative pathway: E-G5$^2$-G584A (*A*), E-G5$^2$-Y625W (*B*) and E-G5$^2$-Y625W-P618A (*C*). The chevron plots were fitted globally to a model assuming two parallel pathways, shown in red, in which the observed rate constant is equal to the sum of the rate constants for each pathway (for details see Table S6). The hypothetical chevron corresponding to the alternative pathway (pathway 1) is shown as a dashed blue line. The other (wild-type) pathway (pathway 2) was assumed to follow the sequential transition states model (3) as the wild-type pathway. The hypothetical chevrons corresponding to the wild-type pathway transition state 1 (TS$^1$) and transition state 2 (TS$^2$) are shown as dashed orange and purple lines, respectively. The bottom plots illustrate the fractional fluxes through the alternative pathway (pathway 1, blue) and wild-type pathway (pathway 2, grey).

**Supplementary Tables:**

**Table S1.** Thermodynamic and kinetic parameters for $G5^2$ and E-$G5^2$.

| Protein | Equilibrium* | | Kinetic† | | | |
|---|---|---|---|---|---|---|
| | $m_{D-N}$* (kcal·mol$^{-1}$·M$^{-1}$) | $\Delta G_{D-N}^{H_2O}$* (kcal·mol$^{-1}$) | $m_{D-N}$† (kcal·mol$^{-1}$·M$^{-1}$) | $\Delta G_{D-N}^{H_2O}$† (kcal·mol$^{-1}$) | $k_f^{H_2O}$† (s$^{-1}$) | $k_u^{H_2O}$† (s$^{-1}$) |
| $G5^2$-WT | 1.0 ± 0.1 | 2.8 ± 0.2 | 1.1 ± 0.1 | 2.8 ± 0.2 | 12.2 ± 0.3 | 0.112 ± 0.025 |
| E-$G5^2$-WT | 1.4 ± 0.1 | 6.3 ± 0.2 | 1.4 ± 0.1 | 7.0 ± 0.3 | 13.0 ± 0.3 | (9.4 ± 3.2)×10$^{-5}$ |
| E-$G5^2$-Y547 | 1.4 ± 0.1 | 5.5 ± 0.1 | - | - | - | - |
| E-$G5^2$-Y625 | 1.4 ± 0.1 | 5.6 ± 0.1 | 1.4 ± 0.1 | 6.5 ± 0.4 | 10.6 ± 0.5 | (19.5 ± 6.9)×10$^{-5}$ |
| $EG5^2$-T501C$^{A488}$ -E613C$^{A594}$ | 1.4 ± 0.1 | 6.1 ± 0.3 | 1.4 ± 0.1 | 6.7 ± 0.4 | 10.7 ± 0.5 | (14.0 ± 5.0)×10$^{-5}$ |
| $EG5^2$-E500W-E532C$^{IAEDANS}$ | 1.4 ± 0.1 | 6.1 ± 0.4 | 1.4 ± 0.1 | 6.7 ± 0.4 | 10.3 ± 0.5 | (11.8 ± 4.2)×10$^{-5}$ |
| $EG5^2$-E555I-E613C$^{IAEDANS}$ | 1.4 ± 0.1 | 4.5 ± 0.3 | 1.4 ± 0.1 | 5.4 ± 0.3 | 14.0 ± 0.7 | (1.5 ± 0.5)×10$^{-3}$ |

* Equilibrium parameters were obtained by fitting the data to a two-state equation.

† Kinetic parameters were calculated from fitting the data globally to a sequential transition states model.

**Table S2.** Apparent equilibrium parameters obtained for wild-type G5[2] and its mutants at 25°C.

| Protein | $m_{\mathrm{D-N}}$ (kcal·mol$^{-1}$·M$^{-1}$) | $[D]_{50\%}$ (M$^{-1}$) | $\Delta G_{\mathrm{D-N}}^{\mathrm{H_2O}}$ (kcal·mol$^{-1}$) | $\Delta\Delta G_{\mathrm{D-N}}^{\mathrm{H_2O}}$ (kcal·mol$^{-1}$) |
|---|---|---|---|---|
| G5[2]-WT | 1.00 ± 0.05 | 2.80 ± 0.07 | 2.80 ± 0.16 | - |
| G5[2]-P549A | 0.91 ± 0.02 | 2.42 ± 0.03 | 2.20 ± 0.05 | 0.60 ± 0.16 |
| G5[2]-P562A | 1.02 ± 0.02 | 2.22 ± 0.01 | 2.26 ± 0.05 | 0.54 ± 0.17 |
| G5[2]-P571A | 0.98 ± 0.02 | 1.86 ± 0.05 | 1.83 ± 0.06 | 0.97 ± 0.17 |
| G5[2]-P575A | 1.03 ± 0.02 | 1.94 ± 0.01 | 1.99 ± 0.04 | 0.81 ± 0.16 |
| G5[2]-P594A | 0.99 ± 0.02 | 2.40 ± 0.01 | 2.37 ± 0.04 | 0.43 ± 0.16 |
| G5[2]-P599A | 0.99 ± 0.02 | 2.70 ± 0.01 | 2.68 ± 0.04 | 0.12 ± 0.16 |
| G5[2]-P618A | 1.03 ± 0.02 | 2.18 ± 0.01 | 2.24 ± 0.05 | 0.56 ± 0.16 |
| G5[2]-P627A | 0.96 ± 0.03 | 1.67 ± 0.02 | 1.62 ± 0.05 | 1.18 ± 0.16 |
| G5[2]-G548A | 0.92 ± 0.04 | 2.94 ± 0.06 | 2.71 ± 0.14 | 0.09 ± 0.21 |
| G5[2]-G552A | 1.05 ± 0.05 | 2.90 ± 0.05 | 3.03 ± 0.16 | -0.23 ± 0.22 |
| G5[2]-G576A | 1.01 ± 0.04 | 0.85 ± 0.19 | 0.85 ± 0.19 | 1.95 ± 0.25 |
| G5[2]-G584A | 1.00 | -1.60 ± 0.61 | -1.6 ± 0.61 | 4.40 ± 0.63 |
| G5[2]-G587A | - | - | - | - |
| G5[2]-G602A | 1.06 ± 0.04 | 1.28 ± 0.03 | 1.35 ± 0.07 | 1.45 ± 0.17 |
| G5[2]-G608A | 1.02 ± 0.03 | 1.58 ± 0.02 | 1.61 ± 0.05 | 1.19 ± 0.17 |
| G5[2]-G626A | 0.98 ± 0.07 | 1.11 ± 0.25 | 1.08 ± 0.25 | 1.72 ± 0.30 |
| G5[2]-Y625W | 0.92 ± 0.02 | 1.53 ± 0.01 | 1.41 ± 0.03 | 1.39 ± 0.16 |

The parameters were calculated by fitting the equilibrium denaturation curves to a two-state model. The errors quoted for $m_{\mathrm{D-N}}$ and $[D]_{50\%}$ of G5[2]-WT represent the experimental errors (based on four independent experiments). The errors quoted for the G5[2] mutants are the errors of the fits of the data. In the case of G5[2]-G584A, the data were fit to a two-state equation with the $m_{\mathrm{D-N}}$ value fixed at 1 kcal·mol$^{-1}$·M$^{-1}$. G5[2]-G587A is inherently unstable in water.

**Table S3.** Apparent equilibrium parameters obtained for wild-type E-G5[2] and its mutants at 25°C.

| Protein | $m_{\mathrm{D-N}}$ (kcal·mol$^{-1}$·M$^{-1}$) | $[D]_{50\%}$ (M$^{-1}$) | $\Delta G_{\mathrm{D-N}}^{\mathrm{H_2O}}$ (kcal·mol$^{-1}$) | $\Delta\Delta G_{\mathrm{D-N}}^{\mathrm{H_2O}}$ (kcal·mol$^{-1}$) |
|---|---|---|---|---|
| E-G5[2]-WT | 1.42 ± 0.04 | 4.40 ± 0.02 | 6.27 ± 0.18 | - |
| E-G5[2]-Y547 | 1.40 ± 0.02 | 3.92 ± 0.03 | 5.49 ± 0.09 | 0.78 ± 0.20 |
| E-G5[2]-Y625 | 1.37 ± 0.03 | 4.07 ± 0.01 | 5.58 ± 0.10 | 0.69 ± 0.21 |
| EG5[2]-T501C$^{A488}$-E613C$^{A594}$ | 1.40 ± 0.06 | 4.35 ± 0.02 | 6.09 ± 0.28 | 0.18 ± 0.33 |
| EG5[2]-E500W-E532C$^{IAEDANS}$ | 1.39 ± 0.06 | 4.38 ± 0.02 | 6.10 ± 0.27 | 0.17 ± 0.33 |
| EG5[2]-E555I-E613C$^{IAEDANS}$ | 1.44 ± 0.06 | 3.13 ± 0.02 | 4.50 ± 0.18 | 1.77 ± 0.26 |
| E-G5[2]-P499A | 1.38 ± 0.06 | 4.41 ± 0.02 | 6.10 ± 0.26 | 0.17 ± 0.31 |
| E-G5[2]-P504A | 1.34 ± 0.04 | 3.99 ± 0.01 | 5.36 ± 0.14 | 0.91 ± 0.23 |
| E-G5[2]-P512A | 1.28 ± 0.03 | 4.08 ± 0.01 | 5.22 ± 0.12 | 1.05 ± 0.22 |
| E-G5[2]-P515A | 1.40 ± 0.04 | 4.26 ± 0.01 | 5.97 ± 0.16 | 0.29 ± 0.24 |
| E-G5[2]-P523A | 1.32 ± 0.03 | 4.10 ± 0.01 | 5.40 ± 0.14 | 0.86 ± 0.23 |
| E-G5[2]-P526A | 1.30 ± 0.04 | 4.13 ± 0.02 | 5.37 ± 0.18 | 0.90 ± 0.26 |
| E-G5[2]-P531A | 1.35 ± 0.04 | 4.16 ± 0.02 | 5.63 ± 0.17 | 0.64 ± 0.25 |
| E-G5[2]-P539A | 1.37 ± 0.06 | 4.38 ± 0.03 | 6.02 ± 0.28 | 0.25 ± 0.33 |
| E-G5[2]-P540A | 1.40 ± 0.10 | 4.55 ± 0.04 | 6.37 ± 0.45 | -0.10 ± 0.48 |
| E-G5[2]-P549A | 1.32 ± 0.04 | 3.68 ± 0.02 | 4.85 ± 0.15 | 1.42 ± 0.24 |
| E-G5[2]-P562A | 1.44 ± 0.04 | 3.90 ± 0.01 | 5.61 ± 0.14 | 0.65 ± 0.23 |
| E-G5[2]-P571A | 1.33 ± 0.04 | 3.92 ± 0.02 | 5.19 ± 0.15 | 1.07 ± 0.24 |
| E-G5[2]-P575A | 1.41 ± 0.04 | 3.99 ± 0.02 | 5.63 ± 0.18 | 0.63 ± 0.26 |
| E-G5[2]-P594A | 1.28 ± 0.04 | 4.19 ± 0.02 | 5.39 ± 0.17 | 0.88 ± 0.25 |
| E-G5[2]-P599A | 1.17 ± 0.03 | 2.94 ± 0.02 | 3.44 ± 0.10 | 2.83 ± 0.21 |
| E-G5[2]-P599A-E500W-E532C$^{IAEDANS}$ | 0.93 ± 0.02 | 2.03 ± 0.01 | 1.88 ± 0.04 | 4.39 ± 0.19 |
| E-G5[2]-P618A | 1.32 ± 0.06 | 3.84 ± 0.02 | 5.07 ± 0.21 | 1.20 ± 0.28 |
| E-G5[2]-P627A | 1.28 ± 0.04 | 3.77 ± 0.02 | 4.83 ± 0.14 | 1.44 ± 0.23 |
| E-G5[2]-G505A | 1.34 ± 0.07 | 4.07 ± 0.03 | 5.45 ± 0.30 | 0.82 ± 0.36 |
| E-G5[2]-G517A | 1.35 ± 0.05 | 3.38 ± 0.02 | 4.58 ± 0.19 | 1.69 ± 0.26 |
| E-G5[2]-G524A | 1.23 ± 0.06 | 3.58 ± 0.03 | 4.40 ± 0.22 | 1.87 ± 0.28 |
| E-G5[2]-G527A | 1.23 ± 0.04 | 3.37 ± 0.02 | 4.16 ± 0.15 | 2.10 ± 0.24 |

| | | | | |
|---|---|---|---|---|
| E-G5$^2$-G534A | 1.23 ± 0.03 | 3.34 ± 0.01 | 4.11 ± 0.09 | 2.16 ± 0.21 |
| E-G5$^2$-G548A | 1.20 ± 0.06 | 3.30 ± 0.03 | 3.97 ± 0.19 | 2.30 ± 0.26 |
| E-G5$^2$-G552A | 1.30 ± 0.04 | 3.66 ± 0.02 | 4.75 ± 0.15 | 1.52 ± 0.24 |
| E-G5$^2$-G576A | 1.48 ± 0.05 | 3.18 ± 0.02 | 4.72 ± 0.16 | 1.55 ± 0.24 |
| E-G5$^2$-G584A | 1.45 ± 0.05 | 3.20 ± 0.02 | 4.62 ± 0.15 | 1.64 ± 0.23 |
| E-G5$^2$-G587A | 1.51 ± 0.05 | 1.25 ± 0.02 | 1.88 ± 0.07 | 4.39 ± 0.19 |
| E-G5$^2$-G602A | 1.41 ± 0.03 | 3.04 ± 0.01 | 4.29 ± 0.09 | 1.97 ± 0.21 |
| E-G5$^2$-G608A | 1.41 ± 0.04 | 3.47 ± 0.02 | 4.90 ± 0.14 | 1.37 ± 0.23 |
| E-G5$^2$-G626A | 1.40 ± 0.05 | 3.01 ± 0.02 | 4.23 ± 0.14 | 2.04 ± 0.23 |
| E-G5$^2$-Y625W | 1.53 ± 0.02 | 3.37 ± 0.01 | 5.14 ± 0.08 | 1.13 ± 0.20 |
| E-G5$^2$-Y625W-P512A | 1.48 ± 0.03 | 3.16 ± 0.01 | 4.67 ± 0.11 | 1.60 ± 0.21 |
| E-G5$^2$-Y625W-P531A | 1.51 ± 0.02 | 3.41 ± 0.01 | 5.13 ± 0.05 | 1.14 ± 0.19 |
| E-G5$^2$-Y625W-P540A | 1.63 ± 0.04 | 3.54 ± 0.01 | 5.78 ± 0.14 | 0.49 ± 0.23 |
| E-G5$^2$-Y625W-P571A | 1.51 ± 0.04 | 3.10 ± 0.02 | 4.67 ± 0.14 | 1.60 ± 0.23 |
| E-G5$^2$-Y625W-P599A | 1.42 ± 0.02 | 1.96 ± 0.01 | 2.79 ± 0.04 | 3.48 ± 0.19 |
| E-G5$^2$-Y625W-P618A | 1.57 ± 0.02 | 3.03 ± 0.01 | 4.75 ± 0.07 | 1.52 ± 0.19 |
| E-G5$^2$-G584A-E500W-E532C$^{IAEDANS}$ | 1.58 ± 0.09 | 2.01 ± 0.04 | 3.17 ± 0.18 | 3.10 ± 0.26 |

The parameters were calculated by fitting the equilibrium denaturation curves to a two-state model. The errors quoted for the G5$^2$ mutants are the errors of the fits of the data.

**Table S4.** Kinetic parameters obtained for the $G5^2$ pathway based on the single mutants at 25°C.

| Protein | Equilibrium $\Delta\Delta G_{D-N}^{H_2O}$ (kcal·mol$^{-1}$) | Kinetic $\Delta\Delta G_{D-N}^{H_2O}$ (kcal·mol$^{-1}$) | $k_f^{H_2O}$ (s$^{-1}$) | $k_u^{H_2O}$ (s$^{-1}$) | Φ |
|---|---|---|---|---|---|
| $G5^2$-WT | - | - | 12.2 ± 0.3 | 0.11 ± 0.02 | - |
| $G5^2$-P549A | 0.6 ± 0.2 | -0.2 ± 0.3 | 13.9 ± 0.5 | 0.09 ± 0.02 | - |
| $G5^2$-P562A | 0.5 ± 0.2 | 0.6 ± 0.3 | 13.0 ± 0.5 | 0.30 ± 0.07 | -0.07 |
| $G5^2$-P571A | 1.0 ± 0.2 | 0.6 ± 0.3 | 5.1 ± 0.2 | 0.13 ± 0.03 | 0.91 |
| $G5^2$-P575A | 0.8 ± 0.2 | 0.6 ± 0.3 | 5.0 ± 0.2 | 0.13 ± 0.03 | 0.90 |
| $G5^2$-P594A | 0.4 ± 0.2 | 0.2 ± 0.3 | 13.9 ± 0.5 | 0.18 ± 0.04 | - |
| $G5^2$-P599A | 0.1 ± 0.2 | 0.0 ± 0.3 | 12.4 ± 0.5 | 0.11 ± 0.03 | - |
| $G5^2$-P618A | 0.6 ± 0.2 | 0.7 ± 0.3 | 13.5 ± 0.6 | 0.42 ± 0.09 | -0.11 |
| $G5^2$-P627A | 1.2 ± 0.2 | 0.9 ± 0.3 | 2.5 ± 0.1 | 0.11 ± 0.02 | 1.02 |
| $G5^2$-G548A | 0.1 ± 0.2 | -0.4 ± 0.3 | 13.8 ± 0.5 | 0.07 ± 0.02 | - |
| $G5^2$-G552A | -0.2 ± 0.2 | -0.3 ± 0.3 | 14.1 ± 0.5 | 0.08 ± 0.02 | - |
| $G5^2$-G576A | 2.0 ± 0.2 | 1.3 ± 0.3 | 0.9 ± 0.1 | 0.07 ± 0.02 | 0.80 |
| $G5^2$-G602A | 1.5 ± 0.2 | 1.6 ± 0.3 | 8.5 ± 0.4 | 1.16 ± 0.30 | 0.05 |
| $G5^2$-G608A | 1.2 ± 0.2 | 1.1 ± 0.3 | 13.3 ± 0.6 | 0.82 ± 0.18 | 0.01 |
| $G5^2$-G626A | 1.7 ± 0.3 | 2.0 ± 0.3 | 0.6 ± 0.1 | 0.15 ± 0.03 | 1.06 |
| $G5^2$-Y625W | 1.4 ± 0.2 | 1.5 ± 0.3 | 1.3 ± 0.1 | 0.14 ± 0.03 | 0.95 |

The chevron plots were fitted globally to the sequential transition states model, with the values of $k_{I^*-D}$ and $m_{I^*-D}$ fixed at $1\times10^4$ s$^{-1}$ and 0 M$^{-1}$, respectively, and the values of $m_{D-I^*}$, $m_{I^*-N}$ and $m_{N-I^*}$ shared between the data sets (0.88 ± 0.02 M$^{-1}$, 0.64 ± 0.02 M$^{-1}$ and 0.29 ± 0.02 M$^{-1}$, respectively; kinetic $m_{D-N}$ was 1.07 ± 0.03 kcal·mol$^{-1}$·M$^{-1}$). All other microscopic rate constants were allowed to vary freely. Φ values were calculated using equilibrium rather than kinetic $\Delta\Delta G_{D-N}^{H_2O}$, due to lower associated errors. The rate constants and Φ values presented in the table are for TS[1]. The errors on the Φ values are 5-10%.

**Table S5.** Kinetic parameters obtained for the main (wild-type) E-G5[2] pathway at 25°C.

| Protein | Equilibrium $\Delta\Delta G_{\mathrm{D-N}}^{\mathrm{H_2O}}$ (kcal·mol$^{-1}$) | Kinetic $\Delta\Delta G_{\mathrm{D-N}}^{\mathrm{H_2O}}$ (kcal·mol$^{-1}$) | $k_{\mathrm{f}}^{\mathrm{H_2O}}$ (s$^{-1}$) | $k_{\mathrm{u}}^{\mathrm{H_2O}}$ (s$^{-1}$) | Φ |
|---|---|---|---|---|---|
| E-G5[2]-WT | - | - | 13.0 ± 0.3 | (9.4 ± 3.2)×10$^{-5}$ | - |
| E-G5[2]-Y625 | 0.7 ± 0.2 | 0.6 ± 0.5 | 10.6 ± 0.5 | (2.0 ± 0.7)×10$^{-4}$ | - |
| EG5[2]-T501C$^{A488}$-E613C$^{A594}$ | 0.2 ± 0.3 | 0.4 ± 0.5 | 10.7 ± 0.5 | (1.4 ± 0.5)×10$^{-4}$ | - |
| EG5[2]-E500W-E532C$^{IAEDANS}$ | 0.2 ± 0.3 | 0.3 ± 0.5 | 10.3 ± 0.5 | (1.2 ± 0.4)×10$^{-4}$ | - |
| EG5[2]-E555I-E613C$^{IAEDANS}$ | 1.8 ± 0.3 | 1.6 ± 0.5 | 14.0 ± 0.7 | (1.5 ± 0.5)×10$^{-3}$ | - |
| E-G5[2]-P499A | 0.2 ± 0.3 | 0.3 ± 0.5 | 11.1 ± 0.5 | (1.3 ± 0.5)×10$^{-4}$ | - |
| E-G5[2]-P504A | 0.9 ± 0.2 | 1.1 ± 0.5 | 10.1 ± 0.5 | (4.3 ± 1.5)×10$^{-4}$ | 0.17 |
| E-G5[2]-P512A | 1.0 ± 0.2 | 0.7 ± 0.5 | 12.6 ± 0.5 | (2.8 ± 1.0)×10$^{-4}$ | 0.02 |
| E-G5[2]-P515A | 0.3 ± 0.2 | 0.4 ± 0.5 | 11.6 ± 0.5 | (1.7 ± 0.6)×10$^{-4}$ | - |
| E-G5[2]-P523A | 0.9 ± 0.2 | 0.8 ± 0.5 | 11.2 ± 0.5 | (2.9 ± 1.0)×10$^{-4}$ | 0.10 |
| E-G5[2]-P526A | 0.9 ± 0.3 | 0.5 ± 0.5 | 11.1 ± 0.5 | (1.8 ± 0.6)×10$^{-4}$ | 0.10 |
| E-G5[2]-P531A | 0.6 ± 0.3 | 0.5 ± 0.5 | 11.9 ± 0.5 | (1.9 ± 0.7)×10$^{-4}$ | - |
| E-G5[2]-P539A | 0.3 ± 0.3 | 0.0 ± 0.5 | 12.2 ± 0.5 | (0.9 ± 0.4)×10$^{-4}$ | - |
| E-G5[2]-P540A | -0.1 ± 0.5 | 0.0 ± 0.5 | 11.5 ± 0.5 | (0.9 ± 0.3)×10$^{-4}$ | - |
| E-G5[2]-P549A | 1.4 ± 0.2 | 1.1 ± 0.5 | 12.1 ± 0.6 | (5.1 ± 1.8)×10$^{-4}$ | 0.03 |
| E-G5[2]-P562A | 0.7 ± 0.2 | 0.5 ± 0.5 | 12.8 ± 0.6 | (5.1 ± 1.8)×10$^{-4}$ | 0.02 |
| E-G5[2]-P571A | 1.1 ± 0.2 | 0.4 ± 0.5 | 5.8 ± 0.3 | (9.0 ± 3.2)×10$^{-5}$ | 1.03 |
| E-G5[2]-P575A | 0.6 ± 0.3 | 0.6 ± 0.5 | 5.1 ± 0.2 | (9.7 ± 3.4)×10$^{-5}$ | 0.97 |
| E-G5[2]-P594A | 0.9 ± 0.3 | 0.5 ± 0.5 | 13.3 ± 0.6 | (2.1 ± 0.7)×10$^{-4}$ | -0.01 |
| E-G5[2]-P618A | 1.2 ± 0.3 | 0.5 ± 0.5 | 13.0 ± 0.6 | (2.2 ± 0.8)×10$^{-4}$ | 0.00 |
| E-G5[2]-P627A | 1.4 ± 0.2 | 0.8 ± 0.5 | 3.3 ± 0.1 | (9.8 ± 3.3)×10$^{-5}$ | 0.98 |
| E-G5[2]-G505A | 0.8 ± 0.4 | 0.6 ± 0.5 | 13.1 ± 0.6 | (2.4 ± 0.8)×10$^{-4}$ | 0.00 |
| E-G5[2]-G534A | 2.2 ± 0.2 | 0.5 ± 0.5 | 13.0 ± 0.6 | (2.4 ± 0.8)×10$^{-4}$ | 0.00 |
| E-G5[2]-G602A | 2.0 ± 0.2 | 1.9 ± 0.5 | 9.9 ± 0.5 | (1.7 ± 0.6)×10$^{-3}$ | 0.08 |
| E-G5[2]-G608A | 1.4 ± 0.2 | 1.2 ± 0.5 | 12.1 ± 0.6 | (6.9 ± 2.4)×10$^{-4}$ | 0.03 |
| E-G5[2]-Y625W-P599A | 3.5 ± 0.2 | 3.9 ± 0.5 | 1.2 ± 0.1 | (6.6 ± 1.9)×10$^{-3}$ | - |

The chevron plots were fitted globally to the sequential transition states model, with the values of $k_{\mathrm{I^*-D}}$ and $m_{\mathrm{I^*-D}}$ fixed at 1×10$^4$ s$^{-1}$ and 0 M$^{-1}$, respectively, and the values of $m_{\mathrm{D-I^*}}$, $m_{\mathrm{I^*-N}}$ and $m_{\mathrm{N-I^*}}$ shared between the data sets (0.80 ± 0.01 M$^{-1}$, 1.30 ± 0.04 M$^{-1}$ and 0.32 ± 0.03 M$^{-1}$, respectively; kinetic $m_{\mathrm{D-N}}$ was 1.43 ± 0.05 kcal·mol$^{-1}$·M$^{-1}$). All other microscopic rate constants were allowed to vary freely. Φ values were calculated using equilibrium rather than kinetic $\Delta\Delta G_{\mathrm{D-N}}^{\mathrm{H_2O}}$, due to lower associated errors. The rate constants and Φ values presented in the table are for TS[1]. The errors on the Φ values are 5-10%. Due to little confidence in the Φ values calculated for TS[2] (errors of 5-40%, owing to large errors in the rate constants associated with TS[2]), they are not listed.

**Table S6.** Kinetic parameters for the alternative folding pathway of E-G5[2] at 25°C obtained from the parallel pathways model fitting.

| Protein | Pathway 1 | | Pathway 2 | | | |
|---|---|---|---|---|---|---|
| | | | TS[1] | | TS[2] | |
| | $k_f^{H_2O}$ ($s^{-1}$) | $k_u^{H_2O}$ ($s^{-1}$) | $k_f^{H_2O}$ ($s^{-1}$) | $k_u^{H_2O}$ ($s^{-1}$) | $k_f^{H_2O}$ ($s^{-1}$) | $k_u^{H_2O}$ ($s^{-1}$) |
| E-G5[2]-G576A | 8.4 ± 1.3 | (1.4 ± 1.7)×10$^{-3}$ | 0.03 ± 0.27 | (9.8 ± 3.4)×10$^{-5}$ | (0.1 ± 1.0)×10$^{3}$ | 0.38 ± 0.07 |
| E-G5[2]-G584A | 7.5 ± 1.2 | (0.5 ± 1.4)×10$^{-3}$ | 0.47 ± 0.25 | (9.6 ± 2.8)×10$^{-5}$ | (1.9 ± 1.1)×10$^{3}$ | 0.39 ± 0.05 |
| E-G5[2]-Y625W | 7.9 ± 0.6 | (1.4 ± 0.4)×10$^{-3}$ | 0.066 ± 0.088 | (1.9 ± 0.2)×10$^{-4}$ | (1.8 ± 2.5)×10$^{2}$ | 0.54 ± 0.03 |
| E-G5[2]-Y625W P512A | 6.0 ± 0.7 | (0.8 ± 2.3)×10$^{-3}$ | 0.14 ± 0.20 | (3.5 ± 0.7)×10$^{-4}$ | (2.1 ± 2.9)×10$^{2}$ | 0.51 ± 0.07 |
| E-G5[2]-Y625W P531A | 7.5 ± 0.6 | (6.7 ± 8.7)×10$^{-4}$ | 0.87 ± 0.16 | (1.2 ± 0.2)×10$^{-4}$ | (2.8 ± 0.6)×10$^{3}$ | 0.39 ± 0.03 |
| E-G5[2]-Y625W P540A | 15.3 ± 2.1 | (6.7 ± 8.7)×10$^{-4}$ | 0.27 ± 0.36 | (1.1 ± 0.3)×10$^{-4}$ | (0.1 ± 1.3)×10$^{3}$ | 0.38 ± 0.06 |
| E-G5[2]-Y625W -P618A | 7.4 ± 0.7 | (1.8 ± 1.8)×10$^{-3}$ | 0.07 ± 0.17 | (4.4 ± 0.6)×10$^{-4}$ | (1.0 ± 2.4)×10$^{2}$ | 0.64 ± 0.06 |

The chevron plots were fitted globally to a model assuming two parallel pathways, in which the observed rate constant is equal to the sum of the rate constants for each pathway (see Fig. S5 for representative examples). For the alternative folding pathway (pathway 1) we assumed the simplest two-state model, with the values of $m_{k_f}$ and $m_{k_u}$ fixed at 1.66 $M^{-1}$ and 0.75 $M^{-1}$, respectively. To account for the curvature in the unfolding arm of the chevron plots, the other pathway (pathway 2) was assumed to follow the sequential transition states model, as in the described above wild-type pathway, with the values of $k_{I^*-D}$, $m_{D-I^*}$, $m_{I^*-D}$, $m_{I^*-N}$ and $m_{N-I^*}$ fixed at 1×10$^4$ $s^{-1}$, 0.80 $M^{-1}$, 0 $M^{-1}$, 1.30 ± 0.04 $M^{-1}$ and 0.32 ± 0.03 $M^{-1}$, respectively. All other microscopic rate constants were allowed to vary freely. Data for E-G5[2]-G587A, E-G5[2]-G584A-E500W-E532C$^{IAEDANS}$, E-G5[2]-G626A and E-G5[2]-Y625W-P571A were included in the global fitting, but did not converge.

**Table S7.** Data collection and refinement statistics.

| | E-G5$^2$-Y625W |
|---|---|
| Data collection * | |
| PDB deposition code | 5DBL |
| Space group | *C*2 |
| Cell dimensions | |
| *a*, *b*, *c*; Å | 69.1, 35.0, 69.2 |
| *β*; ° | 104.9 |
| Resolution, Å | 33.4—1.6 (1.63—1.60) |
| $R_{pim}$, % | 4.9 (60.0) |
| $CC_{1/2}$§, % | 99.9 (77.1) |
| *I*/σ*I* | 11.1 (1.6) |
| Completeness, % | 99.0 (98.5) |
| Redundancy | 3.2 (3.2) |
| Refinement | |
| Resolution, Å | 33.4—1.6 |
| No. of reflections | |
| Working set | 20011 |
| Test set | 1,074 |
| $R_{work}/R_{free}$ | 17.4/20.7 |
| No. of atoms | |
| Protein | 1060 |
| Water | 298 |
| B-factors | |
| Protein | 22 |
| Water | 34 |
| rmsd from ideality | |
| Bond lengths, Å | 0.006 |
| Bond angles, ° | 0.992 |
| Ramachandran angles | |
| Favored regions, % | 100 |
| Outliers, % | 0 |

* Values in parentheses are for the highest resolution shell.

§ $CC_{1/2}$ is the half-data-set correlation coefficient.

We acknowledge Johan Turkenburg and Sam Hart for assistance with crystal testing and data collection. The authors would also like to thank Diamond Light Source for beamtime (proposal mx7864), and the staff of beamline I02 for assistance with crystal testing and data collection.

**Supplementary References**


1. Gruszka DT, *et al.* (2015) Cooperative folding of intrinsically disordered domains drives assembly of a strong elongated protein. *Nat Commun* 6:7271.
2. Gruszka DT, *et al.* (2012) Staphylococcal biofilm-forming protein has a contiguous rod-like structure. *Proc. Natl Acad. Sci. USA* 109(17):E1011-E1018.
3. Bachmann A & Kiefhaber T (2001) Apparent two-state tendamistat folding is a sequential process along a defined route. *J. Mol. Biol.* 306(2):375-386.
4. Karanicolas J & Brooks CL, 3rd (2003) Improved Go-like models demonstrate the robustness of protein folding mechanisms towards non-native interactions. *J Mol Biol* 334(2):309-325.
5. Gruszka DT, *et al.* (2012) Staphylococcal biofilm-forming protein has a contiguous rod-like structure. *Proc. Natl Acad. Sci. USA* 109(17):E1011-E1018.
6. Newman J, *et al.* (2005) Towards rationalization of crystallization screening for small- to medium-sized academic laboratories: the PACT/JCSG+ strategy. *Acta crystallographica. Section D, Biological crystallography* 61(Pt 10):1426-1431.
7. Kabsch W (2010) XDS. *Acta Crystallogr. D* 66(Pt 2):125-132.
8. Evans PR & Murshudov GN (2013) How good are my data and what is the resolution? *Acta crystallographica. Section D, Biol Crystallogr* 69(Pt 7):1204-1214.
9. McCoy AJ, *et al.* (2007) Phaser crystallographic software. *J. Appl. Crystallogr.* 40(Pt 4):658-674.
10. Emsley P, Lohkamp B, Scott WG, & Cowtan K (2010) Features and development of Coot. *Acta crystallographica. Section D, Biol Crystallogr* 66(Pt 4):486-501.
11. Adams PD, *et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta crystallographica. Section D, Biol Crystallogr* 66(Pt 2):213-221.
12. Krissinel E & Henrick K (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D* 60(Pt 12 Pt 1):2256-2268.
13. McNicholas S, Potterton E, Wilson KS, & Noble ME (2011) Presenting your structures: the CCP4mg molecular-graphics software. *Acta crystallographica. D, Biological Crystallogr* 67(Pt 4):386-394.
14. Plaxco KW, Simons KT, & Baker D (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol* 277(4):985-994.