FEBS Letters

FEBS PRESS science publishing by scientists

Transiently disordered tails accelerate folding of globular proteins

Saurav Mallik[1,2], Tanaya Ray[3] and Sudip Kundu[1,2]

1 Department of Biophysics, Molecular Biology and Bioinformatics, University of Calcutta, India
2 Center of Excellence in Systems Biology and Biomedical Engineering (TEQIP Phase-II), University of Calcutta, India
3 Harish-Chandra Research Institute, HBNI, Allahabad, India

Correspondence
S. Kundu, Department of Biophysics, Molecular Biology and Bioinformatics, University of Calcutta, 92, Acharya Prafulla Chandra Road, Kolkata 700009, India
Fax: +91 33 2351 0360
Tel: +91 33 2350 8386
E-mail: skbmbg@caluniv.ac.in

(Received 27 December 2016, revised 26 April 2017, accepted 13 June 2017, available online 8 July 2017)

doi:10.1002/1873-3468.12725

Edited by Alfonso Valencia

Numerous biological proteins exhibit intrinsic disorder at their termini, which are associated with multifarious functional roles. Here, we show the surprising result that an increased percentage of terminal short transiently disordered regions with enhanced flexibility (TstDREF) is associated with accelerated folding rates of globular proteins. Evolutionary conservation of predicted disorder at TstDREFs and drastic alteration of folding rates upon point-mutations suggest critical regulatory role(s) of TstDREFs in shaping the folding kinetics. TstDREFs are associated with long-range intramolecular interactions and the percentage of native secondary structural elements physically contacted by TstDREFs exhibit another surprising positive correlation with folding kinetics. These results allow us to infer probable molecular mechanisms behind the TstDREF-mediated regulation of folding kinetics that challenge protein biochemists to assess by direct experimental testing.

Keywords: flexibility; folding rate; protein folding; regulation; secondary structure; transient disorder

Accurate folding of nascent proteins to their native structures is the basis of enzymatic activities in a living cell. However, understanding the mechanistic details of folding at molecular level is a challenging scientific problem [1]. Only recently, combination of advanced experimental techniques and computer simulations are helping to understand the molecular details of folding [1,2].

There are different theoretical models that attempt to portray the structural reconstitutions of a nascent protein from completely unstructured to fully folded state [1–5]. Models that describe the progress of folding through stepwise formation of small folding units cooperatively coalescing into the native structure, fit the best with experimental data [2,3]. In these models, each individual step is hypothesized to be rate-limited by a single energy barrier representing a cooperative relationship between two successive folding units [2]. Misassembled folding units or their lack of cooperativity often results in kinetic blocks/traps hindering the progression of folding [2,5]. A number of studies have depicted the dependency of folding kinetics on different factors (proxy of thermodynamic constraints), such as length [6], secondary structural makeup [7], 3D topology [8] and patterns of residue-level coevolution [9], which establishes an intimate

Abbreviations
ACO, absolute contact order; *ACOr*, absolute contact order ratio; IstDREF, internal short transiently disordered regions with enhanced flexibility; *NSSE*, number of secondary structural elements; *PDR*, percent of short transiently disordered residues; *PDRi*, IstDREFs for *PDR* estimation; *PDRt*, TstDREFs in *PDR* estimation; pIDR, predicted internal disordered regions; *PSSEc*, percent of SSEs contacted by TstDREFs; pTDR, predicted terminal disordered regions; stDREF, short transiently disordered regions with enhanced flexibility; TstDREF, terminal short transiently disordered regions with enhanced flexibility.

relationship between thermodynamic and kinetic aspects of folding.

A growing number of evidences are recently coming forward exploring the surprising fact that cooperative folding of multidomain globular proteins is regulated by segments that are inherently disordered in nature [4,10–13]. The 'disorder' of these segments is often conditional or transient in nature: they remain unstructured in isolation, but during folding the whole segment or a part of it experiences disorder-to-order transition [10–14]. This dormant disorder reawakens during the unfolding events, promoting the stepwise disassembly of folding units [13,14]. For multidomain globular proteins there are some indications that disorder promotes rapid, marginally cooperative folding of individual domains [4]. Intrinsically disordered, transiently disordered, and highly flexible regions of biological proteins participate in concerted structural rearrangements associated with protein folding, assembly, and unfolding, thus regulating the entire cascade of protein turnover [10–12,15,16]. These results suggest that segments with enhanced flexibility and/or disorder-like composition may have some unappreciated relationship with both folding and unfolding kinetics. Indirect evidences such as shorter half-lives of highly flexible proteins rich in disorder-like segments also support this possibility, as half-life itself is directly related to the stability acquired in the folding process [15,16].

Here, based on a computational study on a set of globular proteins with experimentally determined folding kinetics and high-resolution crystal structures, we depict a direct relationship between folding rates and the extent of short transiently disordered segments (with enhanced flexibility) included in the sequence-space that expands the realm of this question amenable to direct experimental testing. Increased short disorder-like segments with elevated flexibility at the two termini compared to the remaining sequence is a common trait for a wide spectrum of biological proteins [17]. Here, we find that the percent of short transiently disordered regions with enhanced flexibility (stDREF) at protein termini has a surprising positive correlation with protein folding rates. Evolutionary conservation of stDREFs at protein termini and drastic alterations of folding kinetics upon point-mutations at these sites further suggests some unappreciated regulatory role of these segments in shaping the folding kinetics. The stDREFs exhibit very long-range intramolecular interactions and the percent of native secondary structural elements in contact with terminal stDREFs also positively correlate with folding rates. These results allow us to infer probable molecular mechanisms behind stDREF-mediated regulation of folding kinetics.

Materials and methods

Protein folding rate, sequence, and 3D structure dataset

We begin with a recently published dataset of 108 globular proteins [18] with experimentally determined folding (log k_f) and unfolding rates (log k_u). Excluding three synthetic proteins, we collect 3D structures of the remaining 105 proteins (Dataset-1) from Protein Data Bank [19] that includes 71 two-state and 34 multi-state proteins. In case of multistate folding since there are multiple transition phases, kinetic data of the slowest phase is used as the protein has to pass through it anyway [7]. Our dataset includes proteins from different structural classes and different phylogenetic kingdoms. We extract a nonredundant subset of these proteins (Dataset-2) under 30% sequence identity cut-off. If there are ≥ 2 proteins having $\geq 30\%$ identity, only one (that with higher resolution crystal structure and/or experimental point-mutation data) is included in Dataset-2, which comprises 82 proteins. However, in Dataset-1, proteins with $\geq 30\%$ identity are often associated with very different (un)folding rates (e.g., mesophilic and thermophilic cold shock proteins, having 52% sequence identity, exhibit 0.61 and -1.74 log k_u values respectively). So, here we analyze both Dataset-1 and Dataset-2.

Inferring disordered protein segments

Disordered sequences generally exhibit a wide-spectrum of attributes in structure-space (from completely unstructured to ordered and featuring SSEs), for which there is no clear conceptual and operational definition of structural disorder [20]. Hence, to investigate the relationship between folding kinetics and disorder present in the sequence-space, we infer the disorder status of every residue of the globular proteins from their sequence data using multiple disorder prediction methods: VSL2B [21], PONDR-FIT [22], and IUPred short [23]. VSL2B is a combination of neural network predictors trained on experimentally identified disordered sequences, IUPred evaluates the energy resulting from inter-residue interactions, assuming disorder is associated with insufficient residue-residue contacts and PONDR-Fit combines different predictors to seek a consensus of the scores.

Elastic Network Model analysis

We exploit the Elastic Network Models implemented in ElNemo package [24] to infer structural flexibility in terms of predicted B-factors based on 100 normal modes. We apply this method to examine whether the predicted disordered segments have high conformational flexibility.

CH-distance estimation

A critical balance of hydropathy and charge content in the sequence-space is required to drive protein self-folding process. This sequence-structure relationship is represented as a diagram-of-states, where positive and negative charge content and the net hydropathy of protein sequences are plotted in a 3D space [25,26]. CH-distance is estimated as the vertical distance from the location of the protein to the boundary separating folded and disordered phases. CH-distance > 0 reflects the proportion of hydrophobic and charged amino acids is inadequate for folding.

Residue-residue contacts and absolute contact order (*ACO*)

The *ACO* of a protein structure is defined as the average amino acid separation of 3D contacts [27]:

$$ACO = \frac{1}{n_c}\sum_{i>j}\Delta(i,j)|s_i - s_j|$$

where n_c is the total number of residue-residue contacts, s_i and s_j are the sequence positions of residues i and j, and $\Delta(i, j)$ is the selection criteria that includes i and j into analysis only if they are in contact and if $|i-j| \geq 4$. This $|i-j| \geq 4$ criterion ensures that contacts included in *ACO* estimation reflect 3D topology of the proteins, rather than secondary structures. If any two atoms from two different amino acids (i and j) are within a cut-off distance of 5 Å (distance cutoff for van der Waals contact), they are considered as connected [28].

Residue contact network

The residue pairs (nodes) in noncovalent contact (edges) with each other are represented as undirected, unweighted residue contact networks. If there are n number of nodes in the network and E number of edges, network density is defined as $2E/n(n-1)$. A connected component is defined as a subgraph, in which any two vertices are connected to each other by paths, and which is connected to no additional vertices in the whole network.

Accessible surface area of amino acids

The accessible surface areas of amino acids are estimated by SURFACE RACER [29] for 1.4 Å probe radius.

Results

Globular protein termini exhibit transiently disordered segments with enhanced flexibility

It is generally known that intrinsic disorder is nonhomogeneously distributed within the protein sequences, with residues located in the protein termini generally being more disordered than residues in the middle [17]. Consistent with this observation, short predicted disordered regions are identified in at least one terminal of all 105 proteins and in both termini of 93 proteins by all three predictors. In addition, there are 30 proteins having predicted internal disordered segments. Based on the distribution of predicted terminal (pTDR) and internal disordered regions (pIDR), both Dataset-1 and Dataset-2 proteins can be classified into six groups (Fig. 1A). A distribution plot of the lengths of N- and C-terminal pTDRs is presented in Fig. 1B. However, pTDRs and pIDRs appear in the crystallographic maps of 104 out of 105 proteins in our dataset. This suggests these short sequences may either be just flexible regions (such as high-B-factor ordered) or they may refer to conditionally or transiently disordered segments [13,14,30].

To understand the physical nature of pTDRs and pIDRs we first compare the sequence composition of these segments with that of experimentally identified datasets of short disordered regions and high-B-factor ordered regions collected by Dunker and coworkers [30]. We estimate the compositional bias of amino acid residues in our pTDR and pIDRs (Appendix S1) and obtain that this bias has a remarkable 0.61 Spearman correlation with that of the short-disordered regions, compared to only 0.12 correlation with that of the high-B-factor ordered regions. This clearly suggests that pTRDs exhibit a strong compositional similarity with experimentally identified short disordered regions.

We exploit the elastic network models implemented in ElNemo package [24] to examine whether both the pTDRs and pIDRs are associated with enhanced flexibility. The predicted B-factors (based on 100 normal modes) of pTDRs and pIDRs are compared with that of the whole protein by two-sample t-tests at 95% significance level under the null hypothesis of equal mean. Signatures of enhanced flexibility are found at (a) pTDRs of 88 out of 93 proteins with both termini disordered, (b) pTDRs of 12 out of 12 proteins with only one terminal disordered and (c) pIDRs of 30 of 30 proteins having internal disordered regions (Appendix S1). In summary, we obtain signatures of enhanced flexibility in pTDRs of 99 of 105 proteins and in pIDRs of 30 of 30 proteins.

Hence, for a set of globular proteins with known folding rates, we obtain some predicted short disordered regions (at protein termini and in some cases in the middle) that (a) exhibit a strong compositional similarity with experimentally identified short disordered regions, collected by Dunker and co-workers [30] and (b) are associated with enhanced structural

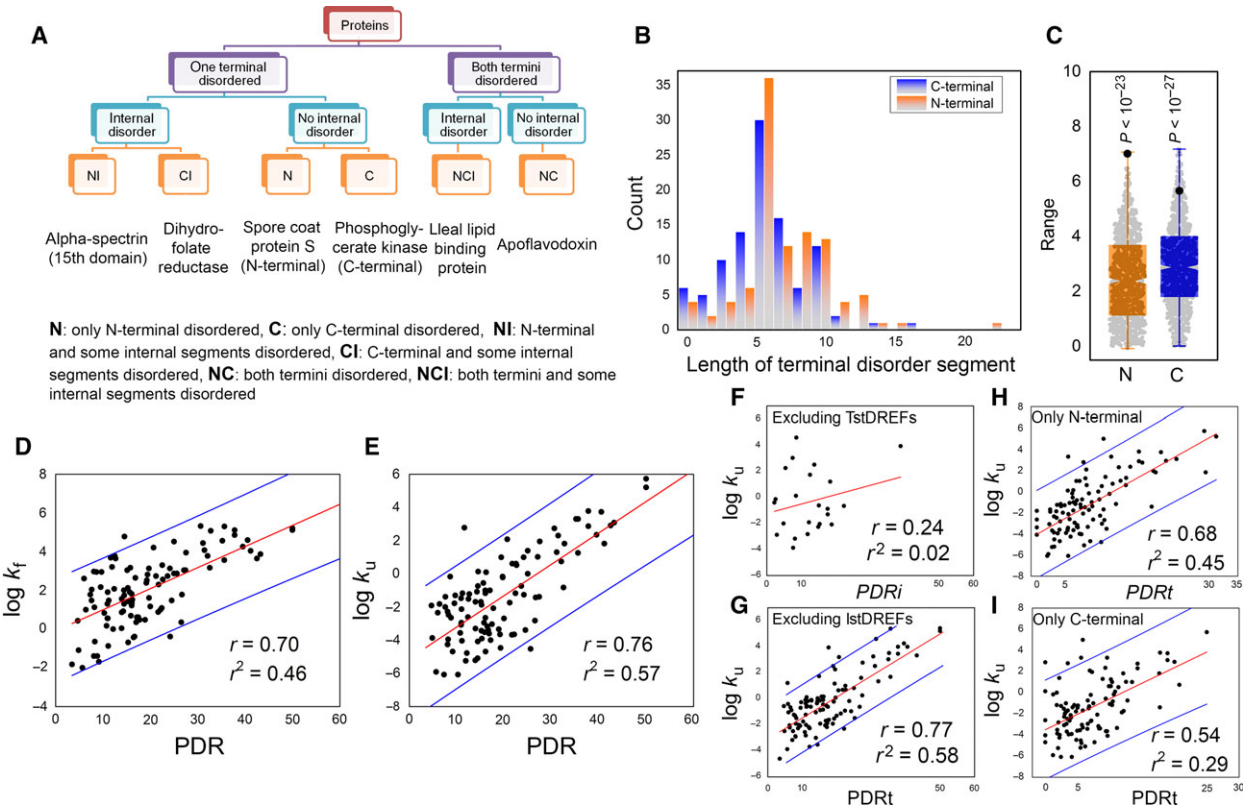

Fig. 1. (A) Protein classification according to the distribution of predicted disordered segments in the sequence. One example of each class is mentioned. (B) Length distribution of N- and C-terminal predicted disordered tails for the 105 proteins in Dataset-1. (C) Boxplot of the length distribution of N- and C-terminal disordered tails for 1000 randomized sequences. Black dots represent the average length of terminal disordered segments obtained for Dataset-1. One sample Wilcoxon signed rank tests are performed to test whether these values significantly differ from the randomized distribution and respective *P*-values are mentioned above. (D) The linear fitting of *PDR* against folding and (E) unfolding rates for Dataset-1; red lines represent the linear fitting, blue lines represent the 95% prediction band. (F–I) The linear fitting of unfolding rate and *PDR* for Dataset-1: in panel-(F) we neglect TstDREF in *PDR* estimation; IstDREF is neglected in panel-G; only N- and C-terminal TstDREFs are used for *PDR* estimation in panel-(H) and (I) respectively.

flexibility evident from elastic network model analysis. Based on these observations, pTDRs and pIDRs are termed as 'terminal short transiently disordered regions with enhanced flexibility' (TstDREF) and 'internal short transiently disordered regions with enhanced flexibility' (IstDREF), respectively, throughout this work.

Folding rates of globular proteins positively correlate with the percent of stDREF residues

In general, presence of longer disordered segments in proteins is known to associate with their lesser globularity, higher structural flexibility, and shorter half-lives [15,26,30]. But surprisingly, the percent of short transiently disordered residues with enhanced flexibility (*PDR*) exhibit strong positive correlations with both folding ($\log k_f$) and unfolding rates ($\log k_u$) (Fig. 1D,E). For example, using IUPred, we find $r^{D1}_{PDR,\log k_f} = 0.70$ ($P < 10^{-16}$) and $r^{D1}_{PDR,\log k_u} = 0.76$ ($P < 10^{-21}$) correlations for Dataset-1 and $r^{D2}_{PDR,\log k_f} = 0.67$ ($P < 10^{-12}$) and $r^{D2}_{PDR,\log k_u} = 0.75$ ($P < 10^{-16}$) correlations for Dataset-2. This suggests stDREF tendency in the sequence-space acts as one of the major regulators of (un)folding rate. Correlations obtained for different prediction methods are provided in Appendix S1. These correlations are much stronger for two-state proteins (e.g., $r^{two-state}_{PDR,\log k_u} = 0.84$) than that for multi-state proteins ($r^{multi-state}_{PDR,\log k_u} = 0.54$).

Next, we test whether this correlation is random or it has some functional relevance. In an iterative process of 1000 steps amino acids are randomly reshuffled for individual protein sequences and at each step disorder inference is made. After 1000 cycles of randomization, we obtain statistically insignificant correlations between (un)folding rates and the randomized average

PDR ($r_{D1} = 0.07$, $r_{D2} = 0.06$ for log k_f; $r_{D1} = 0.11$, $r_{D2} = 0.09$ for log k_u), clearly suggesting that original positive correlations are not random. Further, predicted TstDREF lengths for randomized sequences are significantly smaller (one sample Wilcoxon signed rank test $P_{D1} < 10^{-22}$, $P_{D2} < 10^{-15}$) than that obtained for actual sequences (Fig. 1C). This supports Uversky's hypothesis that this elevated tendency of TstDREFs is more likely an outcome of evolutionary selection [17].

The stDREF tendency in the sequence-space explains the differential (un)folding rates of proteins with high degree of sequence identities as well. Here, we find 15 protein pairs with ≥ 30% identity, among which 11 pairs further exhibit nearly similar lengths and contact orders. For 9 of 11 pairs we find that the protein with higher *PDR* (un)folds faster (Appendix S1).

For any disorder prediction method, *PDR* has a slightly stronger correlation with log k_u than that with log k_f (Fig. 1D,E), similar to what was obtained for *ACO* [18], which is a measure of long range contacts within a protein structure. This stronger correlation meets the plausible expectation that more flexible proteins would unfold faster. This accelerated unfolding by stDREFs can lead to rapid proteolytic degradation and shorter half-lives [15] and thus may be the key to regulating protein turnover in biological cells. But surprisingly, the partial correlation between *PDR* and log k_f, controlling for the effect of log k_u ($r(P)^{D1}_{PDR,\log k_f} = 0.27$), is also statistically significant with $P < 0.01$. This suggests the effect of stDREFs on folding rate is irrespective of its impact on the unfolding rate.

TstDREFs are more important for (un)folding than IstDREFs

We investigate the contribution of TstDREFs and IstDREFs in regulating folding kinetics by excluding one type of segment from *PDR* estimation at a time and examining how the correlation with (un)folding rate changes. First, we exclude N- and C-terminal TstDREFs and only consider IstDREFs for *PDR* estimation (*PDRi*), which results statistically insignificant correlations ($r^{internal}_{PDRi,\log k_u|D1} = 0.35$, $r^{internal}_{PDRi,\log k_u|D2} = 0.29$) with log k_u (Fig. 1F). However, including only TstDREFs in *PDR* estimation (*PDRt*) we obtain strong correlations with log k_u ($r^{D1}_{PDRt,\log k_u} = 0.75$, $r^{D2}_{PDRt,\log k_u} = 0.64$) (Fig. 1G). This result suggests that stDREF-tendency at protein termini probably plays some critical role in accelerating globular protein folding irrespective of their lengths, 3D architecture and folding kinetics. To investigate whether the two termini have differential impacts, we separately include only N- and only C-terminal TstDREFs in *PDRt* estimation respectively. The former gives a stronger correlation ($r^{D1}_{PDRt(N),\log k_u} = 0.68$, $r^{D2}_{PDRt(N),\log k_u} = 0.67$) with log k_u than the latter ($r^{D1}_{PDRt(C),\log k_u} = 0.54$, $r^{D2}_{PDRt(C),\log k_u} = 0.55$) (significant with $P < 10^{-4}$), suggesting N-terminal TstDREFs may have a stronger impact in regulating the (un)folding kinetics (Fig. 1H,I). This trend holds for both two-state ($r^{two-state}_{PDRt(N),\log k_u} = 0.62$, $r^{two-state}_{PDRt(C),\log k_u} = 0.47$) and multi-state proteins ($r^{multi-state}_{PDRt(N),\log k_u} = 0.56$, $r^{multi-state}_{PDRt(C),\log k_u} = 0.38$), while stronger correlations are found in the former case.

Point-mutations at TstDREFs have comparatively more drastic effects on the folding kinetics

Experimental analysis of the effects of point-mutations on folding kinetics is a powerful tool to reveal the importance of different sites in folding process [31,32]. Using our previously used dataset of 617 point-mutations for two-state proteins [33], we compare the percent of folding rate changes upon point-mutations at TstDREFs and globular order regions (Fig. 2A). Mutations at the former generally have stronger impacts on the folding rate than that of the latter ($P_{D1} < 10^{-12}$, $P_{D2} < 10^{-9}$).

We ask whether this result is a consequence of differential solvent exposure and average number of amino acid connectivity of TstDREFs and globular region residues. Estimating the solvent accessible surface area of individual amino acids and the number of residue-residue contacts we find that distributions of these two parameters do not differ significantly between TstDREFs and globular regions (Fig. 2B,C). Hence, drastic alterations of folding kinetics upon point-mutations at TstDREFs likely reflect some critical functional role(s) in the (un)folding process.

TstDREFs are evolutionarily more conserved than IstDREFs

If the TstDREFs have a significant impact on regulating protein folding kinetics, one can expect comparatively higher evolutionary conservation of predicted disorder tendency (not necessarily sequence) at protein termini, compared to that at internal segments. We investigate this by comparing the evolutionary conservation of predicted disorder tendency at TstDREF and IstDREF segments. We estimate the percent score of predicted disorder conservation (if a residue site is predicted as disordered in 50 out of 100 sequences in the alignment, we say that predicted disorder is 50% conserved) for all TstDREF and IstDREF residues. The average percent score of predicted disorder

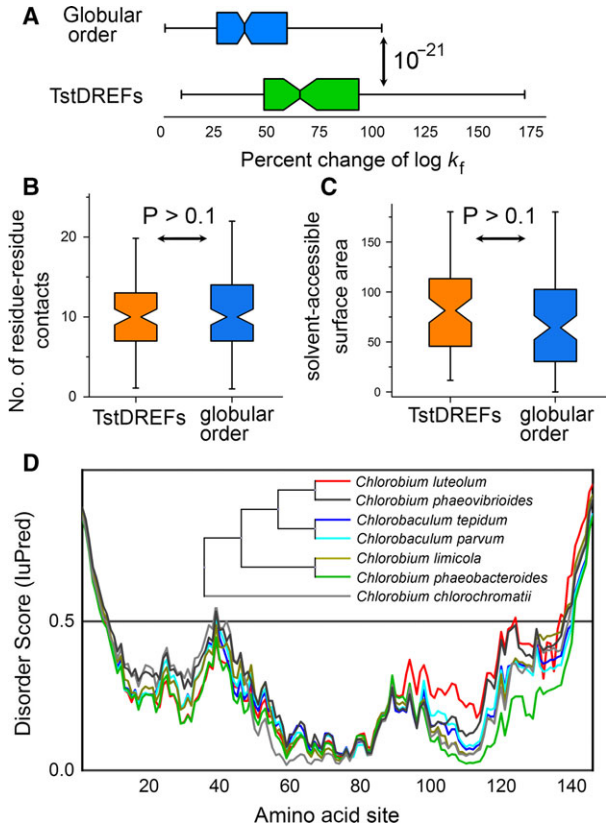

Fig. 2. (A) Distribution plots of the percent of folding rate changes upon point-mutations at residues located at TstDREFs and globular order regions. (B) Distribution of number of residue-residue contacts and (C) solvent accessible surface area for the two regions are shown as boxplots; permutation Mann–Whitney *U*-tests depict these datasets do not differ significantly. (D) Evolutionary variation in predicted disorder at different regions of ribonuclease H1 from *Chlorobium tepidum* among different organisms of Chlorobiaceae family.

conservation for TstDREFs is 87%, while that for IstDREFs is 43%. This difference is statistically significant with $P < 10^{-88}$, suggesting TstDREFs are more evolutionarily conserved than IstDREFs in terms of their predicted disorder tendency. An example for ribonuclease H1 from *Chlorobium tepidum* is demonstrated in Fig. 2D, where we have plotted the IUPred-predicted disorder scores along the sequence for seven different organisms of Chlorobiaceae family.

TstDREFs are associated with very long-range interactions

The physical basis of protein folding is the fact that folding is rate-limited by a free-energy-uphill diffusive configurational search for a transition state that can allow protein chains to go forward in a free-energy-downhill manner [1,2]. It is hypothesized that a major fraction of native residue–residue contacts are also present in the intermediate stages of folding and they are energetically favorable by design, which drives the folding toward the native-like transition state [34,35]. This hypothesis explains why proteins featuring an elevated fraction of native short range contacts fold faster [36,37]. Motivated by these findings, we investigate the extent of short/long range contacts associated with TstDREFs, IstDREFs, and globular order regions.

For a protein of length L, for any two residue sites i and j, we estimate $(i - j)/L$ for each individual residue-residue contact (if $(i - j)/L \to 0$, it's a short-range contact, if $(i - j)/L \to 1$, it's a long-range contact). Distribution plots of this parameter for Dataset-1 for globular order (Fig. 3A), IstDREFs (Fig. 3B) and TstDREFs (Fig. 3C) depict their fundamental differences. The proportion of residue-residue contacts associated with globular order and IstDREFs gradually falls as contacts are more and more long-range, while residue-residue contacts associated with TstDREFs exhibit nearly a homogeneous distribution over the range $0 \le (i - j)/L \le 1$, with two small but distinctive peaks at 0.05 and 0.86. The second peak refers to the fact that in native structure nearly all two-state and some multistate proteins have their TstDREFs in contact [38]. Permutation Mann–Whitney *U*-tests confirm that higher long-range contacts of TstDREFs compared to the other regions is statistically significant with $P_{D1} < 10^{-210}$ and $P_{D2} < 10^{-164}$ (Fig. 3D). We present the contact maps of two representative proteins in Fig. 3E and F.

If different segments of a folded protein chain contact each other randomly, the terminal segments would always exhibit higher long-range contacts compared to the remaining regions. A statistical test is performed to investigate whether the 'long-rangeness' of TstDREFs significantly differs from such randomized contacts. In a 1000-step iterative procedure we randomize the protein contact networks (amino acids as nodes, their contacts as edges) derived in our work while conserving the original network density and the number of connected components. In each iteration, we estimate the $|(i-j)/L|$ value for TstDREFs (Materials and methods). This analysis reveals that long-range contacts of TstDREFs obtained for biological proteins are significantly higher ($P_{D1} < 10^{-157}$, $P_{D2} < 10^{-131}$) than that estimated in this randomization approach (Fig. 3G).

Involvement of TstDREFs in long-range contacts suggests that these structural constraints may be the molecular basis of TstDREF-mediated regulation of folding kinetics. If this hypothesis were true, one can

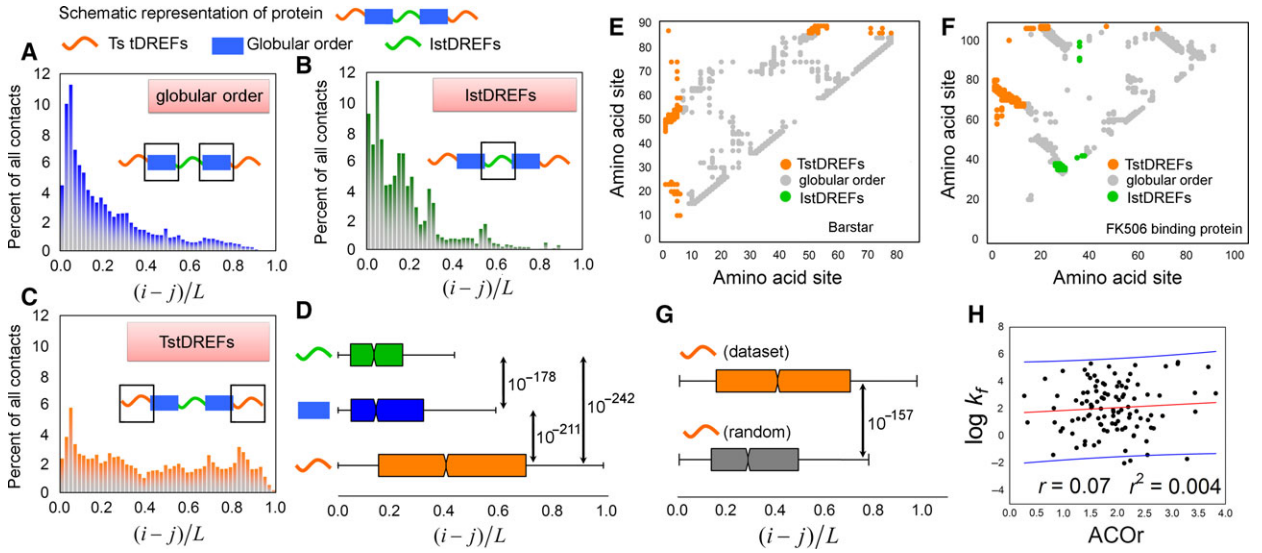

Fig. 3. (A) Distribution plots of short/long range contacts associated with globular order, (B) IstDREFs and (C) TstDREFs for Dataset-1. (D) Boxplot distributions of the residue-residue contacts associated with the three regions are compared using permutation Mann–Whitney *U*-test and *P*-values are mentioned. (E) Contact maps for barstar and (F) FK506 binding proteins; contacts associated with the three regions are highlighted in different colors mentioned in the figure. (G) Permutation Mann–Whitney *U*-test is performed to compare the long-range contacts associated with TstDREFs for Dataset-1 and that obtained by randomization approach, presented as boxplot distributions, and the *P*-value is mentioned. (H) The correlation between absolute contact order ratio and the folding rate.

expect stronger correlations between *PDRt* and (un)-folding rates for proteins with higher *ACO*. Approximating the *ACO* distribution of the 105 proteins as a normal distribution (Appendix S1) we extract the proteins associated with (a) top 10%, (b) top 15% and (c) top 20% *ACO* values and estimate the correlation between *PDRt* and (un)folding rate for the three sets as following: $r^{\text{top }10\%}_{PDRt,\log k_u} = 0.88$, $r^{\text{top }15\%}_{PDRt,\log k_u} = 0.84$ and $r^{\text{top }20\%}_{PDRt,\log k_u} = 0.81$. Same correlations obtained for the background data in each case are $r^{\text{bottom }90\%}_{PDRt,\log k_u} = 0.73$, $r^{\text{bottom }85\%}_{PDRt,\log k_u} = 0.70$ and $r^{\text{bottom }80\%}_{PDRt,\log k_u} = 0.68$. Permutation *t*-tests at 95% significance level demonstrate that stronger correlation with respect to the background set is significant in all three cases with $P < 0.01$, $P < 0.001$ and $P < 0.001$ respectively.

This interesting dependence of $r_{PDRt,\log k_u}$ on *ACO* encourages us to ask whether the partial linear correlation between folding rates and *PDRt*, controlling for the effect of unfolding rate ($r(P)^{\log k_u}_{PDRt,\log k_f}$) would be lower for proteins with lower *ACO*. We gradually remove proteins with higher *ACO* from the dataset to estimate $r(P)^{\log k_u}_{PDRt,\log k_f}$ for bottom 80%, 70%, 60%, 50%, 40%, and 30% *ACO* values. Interestingly, a gradual fall of $r(P)^{\log k_u}_{PDRt,\log k_f}$ (0.31, 0.26, 0.22, 0.17, −0.04, and −0.003, respectively) is observed and statistically significant partial correlations are found only in the former three datasets (those including higher proportions of proteins with high *ACO*; Appendix S1). These two results depict that disorder-mediated folding acceleration is more prominent for proteins that generally fold slower due to higher contact order. These results further demonstrate that structural constraints associated with the long-range interactions of TstDREFs likely play a key role in regulating the folding kinetics of globular proteins.

We define 'absolute contact order ratio' (*ACOr*) as average $|i - j|$ for TstDREFs, divided by that for globular order, such that $ACOr > 1$ means the former exhibits higher long range contacts than the latter. But there is no significant correlation between folding rates and *ACOr* ($r^{\text{D1}}_{ACOr,\log k_f} = 0.09$, $r^{\text{D2}}_{ACOr,\log k_f} = 0.05$), suggesting that the magnitude of difference of long-range contacts between TstDREFs and globular order has little impact on folding kinetics (Fig. 3H).

Folding kinetics is simultaneously coregulated by multiple factors including length and number of SSEs

A careful investigation reveals that different proteins with nearly identical predicted *PDR* often exhibit a wide range of folding rates (Fig. 1D). This suggests folding kinetics is simultaneously coregulated by multiple factors. For any protein pair, if one or more of these factors remain invariant, the others determine the relative folding rate. For example, *PDR* of

hydrogenase maturation protein and ileal lipid binding protein are 10.99 and 11.02, respectively, and the former with comparatively shorter chain length (un)folds faster. Based on such examples, protein length (L) was proposed as another major regulator of folding rate [6] and here we find $r^{D1}_{L,\log k_f} = -0.59$ ($P < 10^{-11}$), $r^{D2}_{L,\log k_f} = -0.64$ ($P < 10^{-11}$) and $r^{D1}_{L,\log k_u} = -0.61$ ($P < 10^{-12}$), $r^{D2}_{L,\log k_u} = -0.60$ ($P < 10^{-9}$) correlations.

Ivankov and Finkelstein [7] proposed that this poor correlation between length and folding rate refers to the fact that folding chain contains some 'independently folding blocks' along with some unstructured loops, for which the effective length of the folding chain (L_{eff}) becomes shorter than L. For a smaller dataset of 64 proteins, considering only the α-helices (predicted from amino acid sequences) as independently folding blocks, they found −0.81 correlation between folding rates and L_{eff}. In spite of the conceptual originality, the main limitation of this approach was its inapplicability for β-only proteins. We seek to generalize this concept by investigating the correlation between folding rates and the number of secondary structural elements (*NSSE*) retrieved from crystal structures. The variations of log k_f and log k_u with *NSSE* can be best fitted into exponential decay curves (Fig. 4A,B) with r-values ($r^{D1}_{\log k_f} = 0.77$, $r^{D2}_{\log k_f} = 0.72$ and $r^{D1}_{\log k_u} = 0.78$, $r^{D2}_{\log k_u} = 0.81$) as strong as that found for L_{eff}.

The remarkable aspect of an exponential decay relationship is that, for smaller proteins with a few SSEs, the *NSSE* would be a very strong regulator of folding kinetics. For example, among C-terminal domain of spore coat protein S (*PDR* = 11.11, *L* = 90, *NSSE* = 10), barstar (*PDR* = 11.24, *L* = 89, *NSSE* = 7) and acyl-coenzyme A binding protein (*PDR* = 11.63, *L* = 86, *NSSE* = 4), faster (un)folding rates are associated with smaller *NSSE*. But after the threshold *NSSE* ≥ 12, the impact of *NSSE* on folding kinetics becomes negligible, leaving other factors to play the dominating role.

Folding rate positively correlates with the percent of SSEs in contact with TstDREFs

Interestingly, even after combining all the parameters (*PDR*, *NSSE*, *L*, *ACO*) known to influence the folding rate, we fail to anticipate relative folding rates of some protein pairs. For example, PI3 SH3 domain and alpha-ketoacid dehydrogenase exhibit very close *PDR* (16.47 and 16.67) and *L* (85 and 84), but they (un)fold at nearly identical rates, contradicting our expectations from their *NSSE* (6 and 8) and *ACO* values (13.98 and

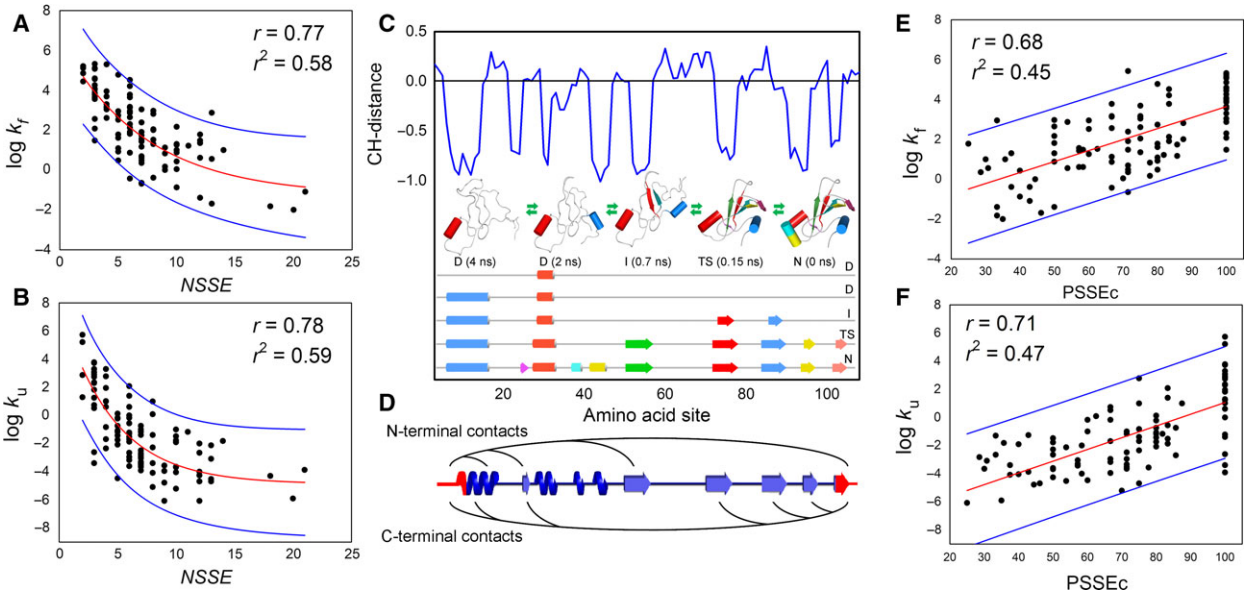

Fig. 4. (A) The exponential decay fitting of folding and (B) unfolding rates and *NSSE* for Dataset-1; red curves represent the exponential fitting, blue curves represent the 95% prediction band. (C) The amino acid site vs. CH distance plot for barnase protein: sharp falls denote segments rich in hydrophobic amino acid that are observed as SSEs in the native state. The folding pathway of barnase and the gradual formation of secondary structural elements in the course of folding are represented schematically. Different states of folding: D, denatured; I, intermediate; TS, transition state; N, native. (D) A schematic representation of how TstDREFs (red) contact different secondary structural elements in the native structure of barnase. (E) The linear regression fitting of folding and (F) unfolding rates and *PSSEc* for Dataset-1.

15.91). In these cases, the only parameter that predicts which protein would fold faster is *ACOr* (protein with higher *ACOr* folds faster) that itself does not correlate with log k_f. This clearly suggests that folding rate is regulated by at least one additional and previously unappreciated factor that has some unknown association with the 'long-rangeness' of TstDREF contacts.

In some chaperone-mediated pathways of *de novo* folding (DnaK, Hsp70, Hsc70 etc.), disordered extensions of the chaperons bind hydrophobic amino acids-enriched segments of the target protein (later appearing as SSEs in native structure), thus delaying any premature hydrophobic collapse until sufficient structural information for productive folding is available [1]. Interestingly, an investigation on all 105 proteins in our dataset demonstrates that their TstDREFs contact an average of 72% SSEs present in the native state. An example of barnase protein is shown in Fig. 4C,D. We partition the sequence into several overlapping blobs of size 5 and for each blob we estimate the CH-distance that crudely represents whether a given segment has the ability of self-folding (CH-distance < 0). In Fig. 4D the secondary structural makeup of barnase is plotted in the same scale as Fig. 4C to show (a) how the hydrophobic amino acid rich segments represent native SSEs and (b) how TstDREFs contacts 7 of 10 SSEs present in the native structure. In fact, the percent of SSEs contacted by TstDREFs (*PSSEc*) exhibits another surprising positive correlation with the (un)folding rate ($r^{D1}_{PSSEc,\log k_f} = 0.68$, $P < 10^{-15}$ and $r^{D2}_{PSSEc,\log k_f} = 0.66$, $P < 10^{-11}$; $r^{D1}_{PSSEc,\log k_u} = 0.71$, $P < 10^{-16}$ and $r^{D2}_{PSSEc,\log k_u} = 0.65$, $P < 10^{-11}$) that is nearly as strong as that obtained for *ACO* (Fig. 4E, F). Further, *PSSEc* perpetually explains the relative folding rates of several protein pairs that were otherwise contradicting the expectations from *PDR*, *NSSE*, *L*, and *ACO* combined (Appendix S1).

PDRt, ACO, and *PSSEc* are the strongest regulators of (un)folding kinetics

Given the scenario that multiple factors simultaneously coregulate the (un)folding rate, it is important to shed some light on which variables independently regulate the (un)folding rate and which variables are correlated. We estimate the partial linear correlations between each parameter and the (un)folding rate, by controlling for the effect of one of the other parameters at a time (Appendix S1). *PDRt*, *ACO* and *PSSEc* turn out to be the three strongest regulators of (un)folding rates, while the impacts of *L* and *NSSE* are highly correlated. A multiple linear regression analysis depicts the former three parameters independently explain 58%, 45% and 41% variations of (un)folding rates respectively (Appendix S1).

Discussion

Deletion experiments depict TstDREFs may have some critical role in protein (un)folding

Correct molecular encounter and interactions are indispensable to proper structural reconstitutions in the molecular world. Flexibility and disorder plays pivotal roles in a wide spectrum of molecular recognition and structural reconstitution events [26]. Here, we find that the extent of stDREF segments at protein termini exhibits a strong positive correlation with folding kinetics. This notion has not been explicitly addressed before; however, there are several experimental evidences where deletion of terminal segments (inferred as stDREFs in our work) causes protein destabilization and unfolding [38], e.g., cytochrome c, ribonuclease A, staphylococcal nuclease, CI2, Titin, TNfn3, bovine pancreatic DNase, botulinum neurotoxin type A light chain and Fyn SH3. Drastic alteration of folding rates upon point-mutations suggests some unappreciated pivotal roles of TstDREFs in regulating the folding process, for which their disorder-like tendency also remains conserved in the course of evolution (Fig. 2).

Stronger impact of N-terminal supports cotranslational folding of biological proteins

We encounter another crucial observation in this aspect that stDREF-mediated regulation of folding kinetics likely involves a stronger impact of N-terminal than that of C-terminal (Fig. 1H,I). Differential roles of the two termini in self-folding pathways are established for large biomolecules such as 16S ribosomal RNA [39] and multidomain proteins [40,41], where folding is cotranscriptional/cotranslational in nature and N-terminal regions that are transcribed/translated first cooperatively guide the folding of C-terminal segments. However, recent cotranslational nascent chain force measurements and fluorescence resonance energy transfer studies on translating ribosomes demonstrate that *in vivo* folding of even small single-domain globular proteins is cotranslational in nature, where the N-terminal segments form a compact, non-native intermediate within the vestibule of exit tunnel, and this intermediate rapidly rearranges itself along with the C-terminal regions into a native-like structure immediately after emerging from the ribosome [42,43]. This cotranslational folding explains the stronger regulatory role of N-terminal stDREFs than that of C-terminal.

Are the correlations between *PDRt* and (un)folding rates a result of evolutionary selection?

Whether evolutionary selection favors accelerated folding of biological proteins has been a matter of scientific debate. While some theoretical studies [44–46] supported this hypothesis, others doubted it, based on the observation that mutations that accelerate the folding rate are as frequent as mutations that decelerate it [47,48]. Bastolla *et al.* [46] examined a theoretically based mechanism for accelerating the folding rate and showed that in fast folding sequences short range contacts are preferentially stronger, thereby producing a positive correlation between contact range and contact energy. They further observed that this correlation is stronger in proteins with larger *ACO*, suggesting the fact that slow-folding proteins are subjected to stronger selective pressure favoring sequence features that accelerate the folding.

In this work, presence of TstDREFs is demonstrated as another sequence feature that promotes fast folding. We make three observations: (a) TstDREFs contribute to majority of the long-range contacts (Fig. 3C); (b) the positive correlation between *PDRt* and (un)folding rates is stronger for slow-folding proteins with larger *ACO*; and (c) there is no statistically significant partial correlation between folding rate and *PDRt* (controlling for the effect of unfolding rate) for fast folding proteins with low *ACO*. Thus, our result not only supports Bastolla's hypothesis [46] but further suggests that correlations between *PDRt* and (un)folding rates may be a consequence of evolutionary selection.

A probable molecular mechanism of TstDREFs-mediated regulation of folding kinetics

The inherent molecular mechanism of stDREF-mediated regulation of folding kinetics is not precisely clear from our study. But integrating our results to the established models of protein folding allows us to infer different possible mechanisms: (a) presence of stDREFs at protein termini limits the configurational diffusion within a reduced conformational space by establishing long-range intramolecular contacts at initial stages of folding and/or (b) TstDREFs probably act as structural templates for stepwise self-assembly of secondary structural elements, minimizing the accumulation of folding intermediates.

Compared to ordered regions, disordered segments exhibit greater capture radius for efficient molecular recognition [49]. N- and C-terminal docking is essential to form an initial kinetic folding intermediate for nearly all two-state and some multistate proteins [38]. If TstDREFs form majority of the long-range contacts at the initial stages, the remaining folding can progress rapidly with ordered regions making only the short-range contacts.

Moreover, long-range contacts by TstDREFs with majority of the native SSEs probably allow the former to act as a structural template guiding the coordinated self-assembly of the latter. We explain this in the light of a model protein (barnase) folding pathway [3] shown in Fig. 4C. Barnase contains a significant proportion of its native SSEs at denatured states. Folding initiates with N- and C-terminal TstDREFs being placed very close to each other. At regions close to the N-terminal, folding is nucleated by fluctuating native helical structures formed in contact with the N-terminal stDREF, while at regions close of C-terminal a native-like β-hairpin is formed in contact with the C-terminal stDREFs. Similar type of SSE nucleation has been observed for other proteins such as Cytochrome c and CI2 [2,3]. However, it is difficult to conclude whether contacts between TDRs and hydrophobic amino acid rich segments promote or disfavor SSE nucleation.

Contacts with TstDREFs at the very first stages of folding likely predispose the nucleated SSEs to self-assemble in a stepwise manner. Such a guided self-assembly reduces the possibility of formation of additional intermediates and define one (or more) dominating folding pathway(s) [2]. Existence of such dominating folding pathway(s) is believed to enhance the efficiency of folding [2]. Our model dictates that cases in which folding propagates through the accumulation of intermediates (multistate folding), significantly lower *PSSEc* values is expected compared to cases where intermediates are either absent or cannot be captured by experimental tools (two-state folding). In our dataset of 74 two-state proteins, the average *PSSEc* estimated is 79%, while for the 31 multistate proteins the average *PSSEc* is only 58% and this difference is statistically significant with $P < 10^{-5}$. This result stands as a strong support for our 'TstDREFs-guided secondary structural element assembly' scenario. The main hydrophobic core of barnase is formed by a spontaneous hydrophobic collapse of N-terminal helical segments (contacting the N-terminal stDREF) onto the C-terminal β-hairpins (contacting the C-terminal stDREF) [3], further supporting a TstDREFs-guided self-assembly model.

In summary, these results show that short transiently disordered segments with enhanced flexibility at globular protein termini acts as one of the major regulators of folding rate. This observation not only

challenges protein biochemists to assess this concept to direct experimental testing, but may further have a broad range of applications in protein engineering and synthetic antibody design, short-listing the targets for antibiotic trials and for anti-prion therapeutics.

Acknowledgments

The authors sincerely acknowledge Kevin W. Plaxco (University of California Santa Barbara), the editor and the anonymous reviewers for their critical assessments to the manuscript and many constructive suggestions. This work is supported by the Center of Excellence in Systems Biology and Biomedical Engineering (TEQIP Phase-II), University of Calcutta, India.

Author contributions

SM, TR, and SK conceptualized and designed the research, SM developed computational methodologies and analyzed the data, SM, TR, and SK discussed and interpreted the results, SM and SK wrote the manuscript.

References

1 Hartl FU and Hayer-Hartl M (2009) Converging concepts of protein folding in vitro and in vivo. *Nat Struct Mol Biol* **16**, 574–581.
2 Englander SW and Mayne L (2014) The nature of protein folding pathways. *Proc Natl Acad Sci USA* **111**, 15873–15880.
3 Daggett V and Fersht A (2003) The present view of the mechanism of protein folding. *Nat Rev Mol Cell Biol* **4**, 497–502.
4 Muñoz V, Campos LA and Sadqi M (2016) Limited cooperativity in protein folding. *Curr Opin Struct Biol* **36**, 58–66.
5 Brockwell DJ and Radford SE (2007) Intermediates: ubiquitous species on folding energy landscapes? *Curr Opin Struct Biol* **17**, 30–37.
6 De Sancho D, Doshi U and Munoz V (2009) Protein folding rates and stability: how much is there beyond size? *J Am Chem Soc* **131**, 2074–2075.
7 Ivankov DN and Finkelstein AV (2004) Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. *Proc Natl Acad Sci USA* **101**, 8942–8944.
8 Plaxco KW, Simons KT and Baker D (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol* **277**, 985–994.
9 Mallik S and Kundu S (2015) Co-evolutionary constraints of globular proteins correlate with their folding rates. *FEBS Lett* **589**, 2179–2185.
10 Moritsugu K, Terada T and Kidera A (2012) Disorder-to-order transition of an intrinsically disordered region of sortase revealed by multiscale enhanced sampling. *J Am Chem Soc* **134**, 7094–7101.
11 Gruszka DT, Mendonça CA, Paci E, Whelan F, Hawkhead J, Potts JR and Clarke J (2016) Disorder drives cooperative folding in a multidomain protein. *Proc Natl Acad Sci USA* **113**, 11841–11846.
12 Gruszka DT, Whelan F, Farrance OE, Fung HK, Paci E, Jeffries CM, Svergun DI, Baldock C, Baumann CG, Brockwell DJ *et al.* (2015) Cooperative folding of intrinsically disordered domains drives assembly of a strong elongated protein. *Nat Commun* **6**, 7271.
13 Uversky VN (2015) Functional roles of transiently and intrinsically disordered regions within proteins. *FEBS J* **282**, 1182–1189.
14 Jakob U, Kriwacki R and Uversky VN (2014) Conditionally and transiently disordered proteins: awakening cryptic disorder to regulate protein function. *Chem Rev* **114**, 6779–6805.
15 van der Lee R, Lang B, Kruse K, Gsponer J, de Groot NS, Huynen MA, Matouschek A, Fuxreiter M and Babu MM (2014) Intrinsically disordered segments affect protein half-life in the cell and during evolution. *Cell Rep* **8**, 1832–1844.
16 Reniere ML, Haley KP and Skaar EP (2011) The flexible loop of *Staphylococcus aureus* IsdG is required for its degradation in the absence of heme. *Biochemistry* **50**, 6730–6737.
17 Uversky VN (2013) The most important thing is the tail: multitudinous functionalities of intrinsically disordered protein termini. *FEBS Lett* **587**, 1891–1901.
18 Broom A, Gosavi S and Meiering EM (2015) Protein unfolding rates correlate as strongly as folding rates with native structure. *Protein Sci* **24**, 580–587.
19 Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and Bourne PE (2000) The protein data bank. *Nucleic Acids Res* **28**, 235–242.
20 Dosztányi Z, Csizmok V, Tompa P and Simon I (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* **347**, 827–839.
21 Peng K, Radivojac P, Vucetic S, Dunker AK and Obradovic Z (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* **7**, 208.
22 Xue B, Dunbrack RL, Williams RW, Dunker AK and Uversky VN (2010) PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim Biophys Acta* **1804**, 996–1010.
23 Dosztányi Z, Csizmok V, Tompa P and Simon I (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 3433–3434.

24 Suhre K and Sanejouand YH (2004) ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res* **32** (Suppl 2), W610–W614.
25 Uversky VN, Gillespie JR and Fink AL (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* **41**, 415–427.
26 Habchi J, Tompa P, Longhi S and Uversky VN (2014) Introducing protein intrinsic disorder. *Chem Rev* **114**, 6561–6588.
27 Grantcharova V, Alm EJ, Baker D and Horwich AL (2001) Mechanisms of protein folding. *Curr Opin Struct Biol* **11**, 70–82.
28 Aftabuddin M and Kundu S (2007) Hydrophobic, hydrophilic, and charged amino acid networks within protein. *Biophys J* **93**, 225–231.
29 Tsodikov OV, Record MT Jr and Sergeev YV (2002) A novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. *J Comput Chem* **23**, 600–609.
30 Radivojac P, Obradovic Z, Smith DK, Zhu G, Vucetic S, Brown CJ, Lawson JD and Dunker AK (2004) Protein flexibility and intrinsic disorder. *Protein Sci* **13**, 71–80.
31 Fersht AR and Sato S (2004) Phi-value analysis and the nature of protein-folding transition states. *Proc Natl Acad Sci USA* **101**, 7976–7981.
32 Sánchez IE and Kiefhaber T (2003) Origin of unusual phi-values in protein folding: evidence against specific nucleation sites. *J Mol Biol* **334**, 1077–1085.
33 Mallik S, Das S and Kundu S (2016) Predicting protein folding rate change upon point mutation using residue-level coevolutionary information. *Proteins* **84**, 3–8.
34 Nymeyer H, García AE and Onuchic JN (1998) Folding funnels and frustration in off-lattice minimalist protein landscapes. *Proc Natl Acad Sci USA* **95**, 5921–5928.
35 Chung HS, Piana-Agostinetti S, Shaw DE and Eaton WA (2015) Structural origin of slow diffusion in protein folding. *Science* **349**, 1504–1510.
36 Plaxco KW, Simons KT, Ruczinski I and Baker D (2000) Topology, stability, sequence, and length: defining the determinants of two-state protein folding kinetics. *Biochemistry* **39**, 11177–11183.
37 Dinner AR and Karplus M (2001) The roles of stability and contact order in determining protein folding rates. *Nat Struct Biol* **8**, 21–22.
38 Krishna MM and Englander SW (2005) The N-terminal to C-terminal motif in protein folding and function. *Proc Natl Acad Sci USA* **102**, 1053–1058.
39 Talkington MW, Siuzdak G and Williamson JR (2005) An assembly landscape for the 30S ribosomal subunit. *Nature* **438**, 628–632.
40 Elcock AH (2006) Molecular simulations of cotranslational protein folding: fragment stabilities, folding cooperativity, and trapping in the ribosome. *PLoS Comput Biol* **2**, e98.
41 Tanaka T, Hori N and Takada S (2015) How co-translational folding of multi-domain protein is affected by elongation schedule: molecular simulations. *PLoS Comput Biol* **11**, e1004356.
42 Holtkamp W, Kokic G, Jäger M, Mittelstaet J, Komar AA and Rodnina MV (2015) Cotranslational protein folding on the ribosome monitored in real time. *Science* **350**, 1104–1107.
43 Nilsson OB, Hedman R, Marino J, Wickles S, Bischoff L, Johansson M, Müller-Lucks A, Trovato F, Puglisi JD, O'Brien EP *et al.* (2015) Cotranslational protein folding inside the ribosome exit tunnel. *Cell Rep* **12**, 1533–1540.
44 Mirny LA, Abkevich VI and Shakhnovich EI (1998) How evolution makes proteins fold quickly. *Proc Natl Acad Sci USA* **95**, 4976–4981.
45 Öztop B, Ejtehadi MR and Plotkin SS (2004) Protein folding rates correlate with heterogeneity of folding mechanism. *Phys Rev Lett* **93**, 208105.
46 Bastolla U, Bruscolini P and Velasco JL (2012) Sequence determinants of protein folding rates: positive correlation between contact energy and contact range indicates selection for fast folding. *Proteins* **80**, 2287–2304.
47 Gu H, Kim D and Baker D (1997) Contrasting roles for symmetrically disposed β-turns in the folding of a small protein. *J Mol Biol* **274**, 588–596.
48 Yi Q, Rajagopal P, Klevit RE and Baker D (2003) Structural and kinetic characterization of the simplified SH3 domain FP1. *Protein Sci* **12**, 776–783.
49 Shoemaker BA, Portman JJ and Wolynes PG (2000) Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proc Natl Acad Sci USA* **97**, 8868–8873.

Supporting information

Additional Supporting Information may be found online in the supporting information tab for this article:

Appendix S1. Extended Results and Discussions.