J|A|C|S
ARTICLES

Published on Web 10/04/2008

Folding of Elongated Proteins: Conventional or Anomalous?

Tzachi Hagai and Yaakov Levy*

Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel

Received June 6, 2008; E-mail: koby.levy@weizmann.ac.il

Abstract: The complexity of the mechanisms by which proteins fold has been shown by many studies to be governed by their native-state topologies. This was manifested in the ability of the native topology-based model to capture folding mechanisms and the success of folding rate predictions based on various topological measures, such as the contact order. However, while the finer details of topological complexity have been thoroughly examined and related to folding kinetics, simpler characteristics of the protein, such as its overall shape, have been largely disregarded. In this study, we investigated the folding of proteins with an unusual elongated geometry that differs substantially from the common globular structure. To study the effect of the elongation degree on the folding kinetics, we used repeat proteins, which become more elongated as they include more repeating units. Some of these have apparently anomalous experimental folding kinetics, with rates that are often less than expected on the basis of rates for globular proteins possessing similar topological complexity. Using experimental folding rates and a larger set of rates obtained from simulations, we have shown that as the protein becomes increasingly elongated, its folding kinetics becomes slower and deviates more from the rate expected on the basis of topology measures fitted for globular proteins. The observed slow kinetics is a result of a more complex pathway in which stable intermediates composed of several consecutive repeats can appear. We thus propose a novel measure, an elongation-sensitive contact order, that takes into account both the extent of elongation and the topological complexity of the protein. This new measure resolves the apparent discrimination between the folding of globular and elongated repeat proteins. Our study extends the current capabilities of folding-rate predictions by unifying the kinetics of repeat and globular proteins.

Introduction

Proteins fold by navigating through an energy landscape that is globally funneled toward a structurally defined native state. The funneled nature of energy landscapes, which is an outcome of the evolutionary minimization of conflicting non-native interactions, explains why native-state topology critically determines protein-folding kinetics.[1−4] The rates at which proteins fold have been shown to vary by more than 8 orders of magnitude.[5,6] Attaining a theoretical framework for protein-folding kinetics that incorporates the key determinants of these complex processes and predicts folding rates and mechanisms for a variety of proteins having different sizes, structures, and topologies has been the focus of a long, ongoing effort and is seen by many as a major constituent of the "folding problem".[6,7]

The folding rate has been shown in many studies to be correlated with protein topological complexity. While rapidly folding proteins tend to have a substantial number of local structures, slow folders often have structures in which distant parts of the protein come in contact to form the folded state.[8] These observations are implicit in the measure known as contact order (CO), a parameter that indicates the complexity of the protein topology.[9] The CO, which is widely used, is calculated as the average sequence distance between pairs of interacting residues in the native state. Its initial variant, the relative contact order (Rel-CO), defined as Rel-CO $= (CN)^{-1}\sum^{C}|i-j|$, where N is the number of residues, C is the number of contacts, and $|i-j|$ is the sequence separation between residues in contact, has been successful in predicting the folding rates of small single-domain proteins that fold via a two-state mechanism.[10] However, extending the set in question to larger proteins (more than 110 residues) or to multiple-state folders has reduced its predictive power and called for the use of two other size-sensitive variants, the absolute contact order (Abs-CO)[11] and the size-modified contact order (SM-CO),[12] which are obtained by multiplying the Rel-CO by N (the protein's size) and $N^{2/3}$, respectively.

Although CO has been successful in predicting folding rates in many cases, in others the experimental folding rates were

(1) Onuchic, J. N.; Wolynes, P. G. *Curr. Opin. Struct. Biol.* **2004**, *14*, 70–75.
(2) Abkevich, V. I.; Gutin, A. M.; Shakhnovich, E. I. *Biochemistry* **1994**, *33*, 10026–10036.
(3) Clementi, C.; Nymeyer, H.; Onuchic, J. N. *J. Mol. Biol.* **2000**, *298*, 937–953.
(4) Ferrara, P.; Caflisch, A. *J. Mol. Biol.* **2001**, *306*, 837–850.
(5) Jackson, S. E. *Folding Des.* **1998**, *3*, R81–R91.
(6) Dill, K. A.; Ozkan, S. B.; Weikl, T. R.; Chodera, J. D.; Voelz, V. A. *Curr. Opin. Struct. Biol.* **2007**, *17*, 342–346.
(7) Onuchic, J. N.; Luthey Schulten, Z.; Wolynes, P. G. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545–600.
(8) Baker, D. *Nature* **2000**, *405*, 39–42.
(9) Plaxco, K. W.; Simons, K. T.; Baker, D. *J. Mol. Biol.* **1998**, *277*, 985–994.
(10) Plaxco, K. W.; Simons, K. T.; Ruczinski, I.; Baker, D. *Biochemistry* **2000**, *39*, 11177–11183.
(11) Ivankov, D. N.; Garbuzynskiy, S. O.; Alm, E.; Plaxco, K. W.; Baker, D.; Finkelstein, A. V. *Protein Sci.* **2003**, *12*, 2057–2062.
(12) Koga, N.; Takada, S. *J. Mol. Biol.* **2001**, *313*, 171–180.

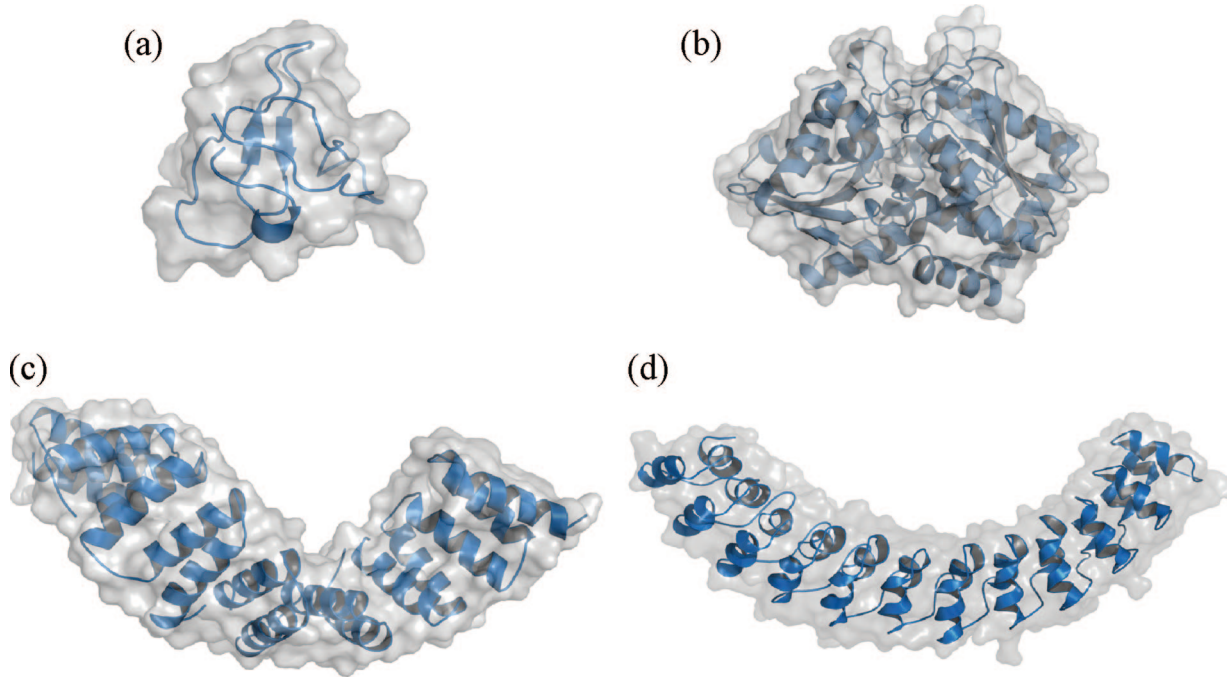

Figure 1. Structures of globular and elongated repeat proteins. (a) src SH3 domain (PDB entry 1SRL), a small globular domain ($N = 57$ aa). (b) Trp synthase β2 subunit (PDB entry 1QOP), a large globular domain ($N = 390$ aa). (c) CTPR8 (based on PDB entry 2AVP), a consensus-designed TPR protein with eight units. (d) ANK12 (PDB entry 1N11), the D34 fragment of AnkyrinR with 12 ANK units. All of the images were created using Pymol.

reported to substantially deviate from the CO-based prediction. Some of the proteins for which a CO-based approach is unsuccessful tend to have a rather elongated shape instead of the classical globular structure. Globularity seems to be a common motif by which single-domain proteins fold. However, elongated proteins, and in particular repeat proteins, which are composed of successive homologous structural units, are also widespread in nature, as they have been evolutionarily selected for functional reasons, such as the need to increase the outer surface for protein−protein interactions.

There are relatively few studies that have investigated folding of elongated proteins. However, in recent years, repeat proteins (composed of up to seven repeating units) have been the target of several folding studies because the regularity and simplicity of their structures allows them to be used to address fundamental questions of protein folding.[13−17] Repeat proteins can be viewed as simple cases of multidomain proteins in which the interface between the repeats (which act as the domains) is significantly larger than interfaces in common multidomain proteins. Accordingly, a repeat protein is a one-dimensional multidomain protein in which small domains are densely packed[18] and internal repeating units can be affected by their neighboring units, resulting in extensive cross-talk.[15] Repeat proteins become more elongated as the number of repeating units increases, so the effect of elongation on protein folding can be studied using proteins with different numbers of repeats. The different structural appearances of globular and elongated repeat proteins can be seen in Figure 1 and in Figure S1 in the Supporting Information.

Several kinetics studies on repeat proteins have provided ambiguous findings with respect to the topology-based prediction. Folding rates of shorter repeat proteins having a small number of repeating units [such as the consensus-designed tetratricopeptide repeat (CTPR) series members CTPR2 and CTPR3] were reported to correlate well with the CO prediction.[19] However, more elongated repeat proteins (such as Myotrophin, $p16^{INK4a}$, and $p19^{INK4d}$ from the ankyrin repeats family) fold several orders of magnitude more slowly than anticipated on the basis of their topological complexity.[20−22] These deviations from predictions based on topology-based measures seem to become larger as the proteins become more elongated. Two of the longer repeat proteins studied, the *Drosophila* Notch receptor ANK domain and the Internalin B LRR domain, have both been reported to fold much more slowly than anticipated on the basis of various CO measures,[23−25] deviating significantly from the kinetics for globular and shorter repeat proteins. The most-elongated repeat protein studied, D34 (a fragment of AnkyrinR composed of 12 ANK repeats), has recently been shown to exhibit complex folding where multiple intermediates are significantly populated. The slowest refolding phase observed deviated by ∼4 orders of magnitude from the

(13) Main, E. R.; Jackson, S. E.; Regan, L. *Curr. Opin. Struct. Biol.* **2003**, *13*, 482–489.
(14) Main, E. R.; Lowe, A. R.; Mochrie, S. G.; Jackson, S. E.; Regan, L. *Curr. Opin. Struct. Biol.* **2005**, *15*, 464–471.
(15) Kloss, E.; Courtemanche, N.; Barrick, D. *Arch. Biochem. Biophys.* **2008**, *469*, 83–99.
(16) Ferreiro, D. U.; Walczak, A. M.; Komives, E. A.; Wolynes, P. G. *PLoS Comput. Biol.* **2008**, *4*, e1000070.
(17) Interlandi, G.; Wetzel, S. K.; Settanni, G.; Pluckthun, A.; Caflisch, A. *J. Mol. Biol.* **2008**, *375*, 837–854.
(18) Han, J. H.; Batey, S.; Nickson, A. A.; Teichmann, S. A.; Clarke, J. *Nat. Rev. Mol. Cell. Biol.* **2007**, *8*, 319–330.
(19) Main, E. R.; Stott, K.; Jackson, S. E.; Regan, L. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 5721–5726.
(20) Lowe, A. R.; Itzhaki, L. S. *J. Mol. Biol.* **2007**, *365*, 1245–1255.
(21) Tang, K. S.; Guralnick, B. J.; Wang, W. K.; Fersht, A. R.; Itzhaki, L. S. *J. Mol. Biol.* **1999**, *285*, 1869–1886.
(22) Low, C.; Weininger, U.; Zeeb, M.; Zhang, W.; Laue, E. D.; Schmid, F. X.; Balbach, J. *J. Mol. Biol.* **2007**, *373*, 219–231.
(23) Bradley, C. M.; Barrick, D. *J. Mol. Biol.* **2005**, *352*, 253–265.
(24) Mello, C. C.; Bradley, C. M.; Tripp, K. W.; Barrick, D. *J. Mol. Biol.* **2005**, *352*, 266–281.
(25) Courtemanche, N.; Barrick, D. *Protein Sci.* **2008**, *17*, 43–53.

rate predicted by CO measures and is thought to represent a partially folded protein. The folding rate of the entire D34 12-repeat protein is estimated to be smaller, indicating a larger deviation from the CO rate prediction.[26]

This apparent trend in which the deviation from predictions given by topology-based measures seems to increase as the protein becomes more elongated encouraged us to examine the folding kinetics of various repeat proteins, with the aim of correlating their unique structural features with their folding kinetics and understanding the origins of their significant deviations from the folding of globular proteins. We applied a coarse-grained model based on native topology to shed light on the discrepancy between the folding of elongated and globular proteins. This model was previously applied to study the kinetics of globular proteins and showed a significant degree of correlation with experimental results.[12,27] Recently, this model was successfully employed to investigate the folding mechanism of several ankyrin repeat proteins.[28,29] For our simulations, we used sets of globular and repeat proteins whose folding kinetics had been studied experimentally. In addition to all of the individual repeat proteins that were studied experimentally, we used the ANK and CTPR series of consensus-based-design repeat proteins. These allowed us to consistently test how the extent of elongation affects folding behavior without regard to the structural variations caused by differences in the sequences (as the consensus repeats are nearly identical[30−36]) and to address the apparent anomalous folding kinetics of repeat proteins.

Results and Discussion

Different Topological Measures Vary in Their Rate Predictions and Predictive Power. Since different topology-based measures were used in studies of folding rates of various repeat proteins, we examined the success of three different variants of CO (Rel-CO, Abs-CO, and SM-CO) in predicting folding rates of both globular and elongated proteins. The Abs-CO and SM-CO measures were used to distinguish between the effects of protein size (number of residues) and protein length (geometrical shape), as they are appropriate to use with large proteins as well. We also used the topomer-search-model measure, Q_D, and its size-sensitive version,[37] as it has been argued that this measure provides the physical rationale for the capabilities of CO to predict folding rates.

Rel-CO has previously been shown to correlate strongly with the folding rates of small two-state folders, though the correlation significantly weakens when large proteins (>110 amino acids) or multistate-folding proteins are included.[11] This observation is seen in our set of globular proteins (Figure 2a), for which there is no apparent correlation between the experimental folding rates and the Rel-CO values across the entire globular set. Separating the set of globular proteins, as was previously described,[11] into two-state and multistate folders (marked in Figure 2a as open and filled circles, respectively) reveals two distinct and significant correlations with Rel-CO ($R = -0.82$ and $R = 0.88$, respectively). Comparison of the globular protein set to the repeat protein set shows an apparent trend of deviation from the Rel-CO predictions. The two CTPR proteins (composed of two and three repeating units) seem to behave similarly to the two-state globular proteins. This is consistent with their geometrical features, as they seem to have globular rather than elongated structures (see Figure S3 in the Supporting Information and later discussion on exact quantification of elongation degree). However, more elongated repeat proteins deviate from the line of the two-state proteins.

The Abs-CO and SM-CO measures are size-sensitive and more suitable for use with large proteins and multistate folders. These measures significantly correlate with the experimental folding rates of the entire set of globular proteins ($R = -0.87$ and $R = -0.83$, respectively), as shown for Abs-CO in Figure 2c (data for SM-CO are not shown). Therefore, the use of these measures may be more appropriate for investigating the kinetics of repeat proteins, as some of these proteins are very large. The deviation of the two CTPR proteins from the globular trend line seems relatively small. However, the four ANK proteins (Myotrophin, p16, p19, and Notch, having four, five, seven, and seven repeats, respectively) and the LRR protein InternalinB (composed of seven repeats) deviate markedly from this line, and their experimental folding rates are smaller than the rates that can be read off the globular trend line (Figure 2c). This deviation becomes larger as the proteins become more elongated (i.e., as more repeating units are added), yet this deviation is not due to the greater size of the elongated proteins, since the folding rates of the large globular proteins correlate well with these measures. A similar trend is seen when using the SM-CO measure (data not shown).

Deviation of the Folding Kinetics of Repeat Proteins from That Expected on the Basis of Its Relationship with Topological Measures Increases As the Proteins Become More Elongated. The available experimental folding rates of repeat proteins implied that there might exist a consistent behavior associated with repeat proteins that differs from that of globular proteins. Since the data on the kinetics of repeat proteins are relatively scarce, especially with respect to the more elongated representatives, we used simulations of consensus-designed series of repeat proteins for both TPR and ANK repeat proteins to complement the experimental data, in addition to simulating globular and repeat proteins for which experimental data is known. Comparing the folding rates of globular proteins with those of increasingly more elongated repeat proteins, on the basis of their simulated kinetics, reveals some striking features.

Although Rel-CO cannot be taken as a proper measure, given its weak correlation with the folding rates for the entire globular set (Figure 2a,b), an interesting tendency is observed in the two repeat protein series when the simulated folding rate is plotted as a function of Rel-CO (Figure 2b). Rel-CO values for the

(26) Werbeck, N. D.; Rowling, P. J.; Chellamuthu, V. R.; Itzhaki, L. S. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 9982–9987.
(27) Chavez, L. L.; Onuchic, J. N.; Clementi, C. *J. Am. Chem. Soc.* **2004**, *126*, 8426–8432.
(28) Ferreiro, D. U.; Cho, S. S.; Komives, E. A.; Wolynes, P. G. *J. Mol. Biol.* **2005**, *354*, 679–692.
(29) Barrick, D.; Ferreiro, D. U.; Komives, E. A. *Curr. Opin. Struct. Biol.* **2008**, *18*, 27–34.
(30) Main, E. R.; Xiong, Y.; Cocco, M. J.; D'Andrea, L.; Regan, L. *Structure* **2003**, *11*, 497–508.
(31) Wetzel, S. K.; Settanni, G.; Kenig, M.; Binz, H. K.; Plückthun, A. *J. Mol. Biol.* **2008**, *376*, 241–257.
(32) Tripp, K. W.; Barrick, D. *J. Mol. Biol.* **2007**, *365*, 1187–1200.
(33) Devi, V. S.; Binz, H. K.; Stumpp, M. T.; Plückthun, A.; Bosshard, H. R.; Jelesarov, I. *Protein Sci.* **2004**, *13*, 2864–2870.
(34) Ferreiro, D. U.; Cervantes, C. F.; Truhlar, S. M.; Cho, S. S.; Wolynes, P. G.; Komives, E. A. *J. Mol. Biol.* **2007**, *365*, 1201–1216.
(35) Mosavi, L. K.; Minor, D. L., Jr.; Peng, Z. Y. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 16029–16034.
(36) Stumpp, M. T.; Forrer, P.; Binz, H. K.; Plückthun, A. *J. Mol. Biol.* **2003**, *332*, 471–487.
(37) Makarov, D. E.; Keller, C. A.; Plaxco, K. W.; Metiu, H. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 3535–3539.

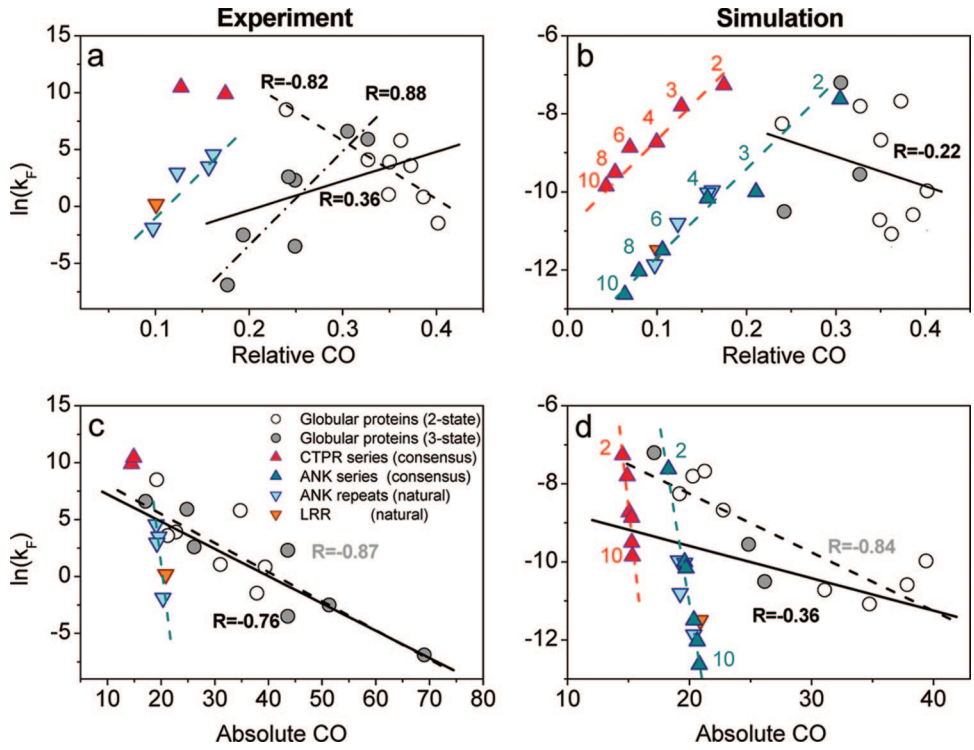

Figure 2. Correlation between folding rates and topology-based measures: experimental and simulation-based folding rates [expressed as $\ln(k_F)$, where k_F is the rate constant for folding] for various proteins are shown as functions of various topological measures. (a) Correlation of experimental rates with Rel-CO, calculated for all globular proteins (all types of circles), two-state folders only (open circles) and multiple-state folders only (filled circles). The correlation for all globular proteins (dotted line, $R = 0.36$) is much weaker than that for two-state folders only (dashed-dotted line, $R = -0.82$) and for multistate folders only (dashed line, $R = 0.88$). Repeat proteins (shown as triangles) were not included in any of these correlations. (b) Correlation of simulation-based rates and Rel-CO. The correlation was calculated for all globular proteins and shown to be insignificant ($R = -0.22$); thus, no further correlations were calculated. Dashed red and blue lines are drawn through the members of the CTPR and ANK series, respectively (natural ANK proteins and the LRR protein InternalinB fall close to the ANK series line). (c) Correlation of experimental rates and Abs-CO. The correlation for globular proteins (dashed gray line, $R = -0.87$) weakens when CTPR and ANK proteins are included in the correlation (solid black line, $R = -0.76$). Dashed red and blue lines are drawn through the members of the consensus-designed CTPR series (CTPR2 and CTPR3) and through the natural ANK and LRR proteins (Myotrophin, p16, p19, Notch, and InternalinB, listed in order of increasing number of repeating units), respectively. (d) Correlation of simulation-based rates and Abs-CO. The correlation for globular proteins (dashed gray line, $R = -0.84$) is significantly weaker when the CTPR and ANK proteins are included in the correlation (solid black line, $R = -0.36$).

CTPR and ANK series decrease as more repeating units are included in the protein, thus predicting a greater rate for more elongated proteins. This is contrary to our simulation-based results, which show a decreasing folding rate as the number of repeating units in the protein increases (e.g., the Rel-CO value for CTPR10 is smaller than that for CTPR2, but the folding rate for CTPR10 is 5 orders of magnitude smaller than that of CTPR2). In addition, the deviation from the globular trend line is larger for the more elongated proteins in both series, similar to the results with natural repeat proteins (see Figure 2b).

This deviation is seen when correlating the folding rates with the Abs-CO measure, despite the fact that it is a more appropriate measure to use in this case because it is sensitive to size. The Abs-CO measure predicts an almost identical rate for all of the members of the CTPR and ANK repeat series (reflected by the almost vertical line of the rates as a function of Abs-CO), irrespective of the increasing number of repeating units (see Figure 2d). Incorporating the repeat proteins along with the set of globular proteins causes a significant weakening of the correlation, from $R = -0.84$ for the globular proteins alone to $R = -0.36$ when both the globular and repeat protein sets are included. The similarity of the Abs-CO values for all of the members of the CTPR and ANK series is a mathematical outcome of the similar internal connectivity of each of the repeats and the similarity of the contacts between successive units. This demonstrates the inability of Abs-CO to discriminate between various repeat proteins in the series and to provide a credible rate prediction. A similar trend is seen when using the SM-CO measure (data not shown).

The deviation of the CTPR and ANK repeat protein series from the globular protein trend line also occurs when Q_D-based rate predictions are used. Smaller members of the CTPR and ANK series fit the globular trend line, while larger members of these series deviate substantially and fold faster than expected (Figure S5b in the Supporting Information). The deviation from the predicted rate given by Q_D is contrary to the deviation observed for CO-based measures, in which the more elongated members of the series fold more slowly than expected on the basis of the linear fit of the globular proteins. Using the size-sensitive variant of Q_D gives a consistent trend in terms of deviation, with the more elongated proteins of both series folding more slowly than would be expected from this measure (data not shown).

The discrepancy between the folding rates of globular and elongated proteins as reflected by the different topological-measure analyses is also seen in the experimental rates (Figure 2a,c). However, since the experimental data set includes fewer elongated proteins, the consistent deviation is less clearly observed.

In summary, all of the measures tested thus far have given predictions that consistently deviate from the behavior expected on the basis of that of globular proteins. The smaller members

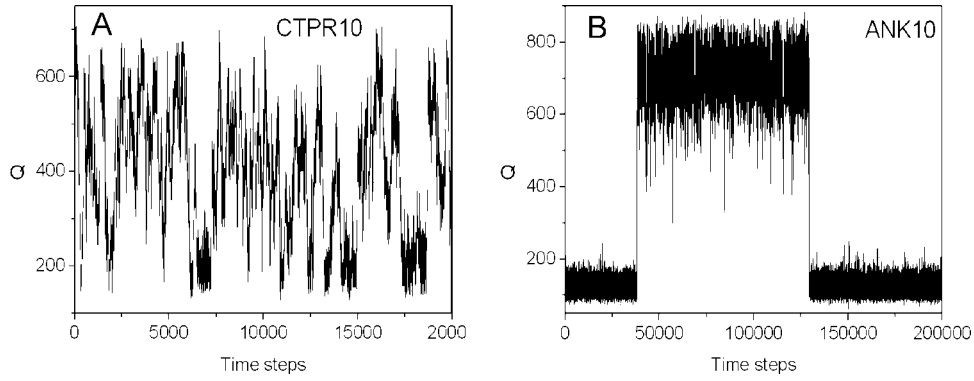

Figure 3. Folding events of elongated repeat proteins: time evolution of the total number of native contacts (Q) along representative folding trajectories at T_F for (A) CTPR10 and (B) ANK10.

of the repeat protein series fit the line of the globular proteins, but the larger members deviate from the expected kinetics. This deviation increases as the members of the repeat protein series become more elongated, and this effect cannot be accounted for solely by their increased size because the deviation is observed with size-sensitive measures as well.

Possible Origins for the Deviation of Repeat Proteins from the Kinetics of Globular Proteins. While the mathematical origin of the apparent deviation seen in the rate predictions given by the various topology-based measures can be easily understood, the physical origins of the apparently different folding characteristics of globular and elongated repeat proteins are not as clear or as simple to explain. Moreover, several studies aimed at quantifying the role of native topology in dominating the folding rate have been conducted,[38−42] yet the determinants that govern the success of the original CO measure in rate predictions are not well-understood.[37,43−45]

The success of topology-based measures in folding-rate predictions for globular proteins is based on the assumption that creating contacts between residues located far away from each other in the primary sequence is more difficult entropically. This assumption holds in many cases, despite its simplifications, because the contacts between pairs are formed relatively in parallel during the process of collapse to the three-dimensional folded structure and the magnitude of the entropy−enthalpy compensation is comparable. Thus, these measures assume an extreme case of an all-or-none behavior in which all of the contacts are either formed or still absent. In the case of globular proteins, which fold via a multistate reaction, topology-based measures are still predictive, since the rate-limiting formation of their intermediates involves formation of contacts between residues that are distant in the primary sequence.[46] In repeat proteins, however, this cannot be the case,[47] as only residues that are relatively close in the primary sequence form contacts (see Figure S1 in the Supporting Information), and their sequence separation is not the primary determinant in the overall folding rate.

The folding pathway for repeat proteins is sequential, as the folding event must propagate one-dimensionally because non-neighboring repeats have no interface between them. Even in cases where the folding of repeat proteins was observed to follow two-state behavior,[21,48,49] a relatively rapid sequence of folding events must occur, and intermediates are kinetically detected.[20,21,24] The deceleration seen in the folding reaction of longer repeat proteins can thus be attributed to either the nucleation stage, in which a stable folded nucleus evolves, or to the propagation stage, in which units adjacent to the folding nucleus sequentially fold. In the nucleation scenario, the formation of the nucleus (of the same size or with a size correlated with the protein length) can be slower for more elongated proteins because of either kinetic or thermodynamic factors. In the propagation scenario, the reaction will be slower as the number of units that need to become folded increases, if one assumes a constant rate of folding of additional units.

The folding nuclei of elongated proteins might be polarized (as opposed to diffusive),[50,51] with structured residues that are spatially clustered while the rest of the residues show little definite order. Such nuclei are similar to the capillarity approximation in nucleation, in which the free energy of a stable phase droplet is separated from the metastable phase by a sharp interface. Folding can be described as the expansion of the edge between the folding nucleus and unfolded parts of a protein until the entire molecule is folded.[52] The propagation of the growth of the nucleus, which proceeds through broadening of the interface surrounding it by "wetting" of residues at the interface, is directional and can be affected by not only the protein's size but also its shape. The capillarity scheme supports the notion that the folding of an elongated protein can be affected by the nucleation rate and by the rate of the one-dimensional propagation of the front between the folded and unfolded parts, which is orthogonal to the long axis of the protein.

In order to examine these folding scenarios for repeat proteins, we analyzed folding trajectories of consensus-based designed series of ANK and CTPR proteins at a temperature (T_F) at which

(38) Plotkin, S. S.; Wang, J.; Wolynes, P. *J. Chem. Phys.* **1997**, *106*, 2932–2948.
(39) Plotkin, S. S.; Wang, J.; Wolynes, P. G. *Phys. Rev. E* **1996**, *53*, 6271–6296.
(40) Wang, J.; Plotkin, S.; Wolynes, P. *J. Phys. I* **1997**, *7*, 395–421.
(41) Shea, J. E.; Onuchic, J. N.; Brooks, C. L. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 12512–12517.
(42) Plotkin, S. S.; Onuchic, J. N. *Q. Rev. Biophys.* **2002**, *35*, 205–286.
(43) Wallin, S.; Chan, H. S. *Protein Sci.* **2005**, *14*, 1643–1660.
(44) Weikl, T. R.; Dill, K. A. *J. Mol. Biol.* **2003**, *329*, 585–598.
(45) Weikl, T. R. *Arch. Biochem. Biophys.* **2008**, *469*, 67–75.
(46) Maity, H.; Maity, M.; Krishna, M. M.; Mayne, L.; Englander, S. W. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 4741–4746.
(47) Cortajarena, A. L.; Mochrie, S. G.; Regan, L. *J. Mol. Biol.* **2008**, *379*, 617–626.
(48) Mosavi, L. K.; Williams, S.; Peng, Z.-y. *J. Mol. Biol.* **2002**, *320*, 165–170.
(49) Zweifel, M. E.; Barrick, D. *Biochemistry* **2001**, *40*, 14357–14367.
(50) Lindberg, M.; Tangrot, J.; Oliveberg, M. *Nat. Struct. Biol.* **2002**, *9*, 818–822.
(51) Oliveberg, M.; Wolynes, P. G. *Q. Rev. Biophys.* **2005**, *38*, 245–288.
(52) Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 6170–6175.

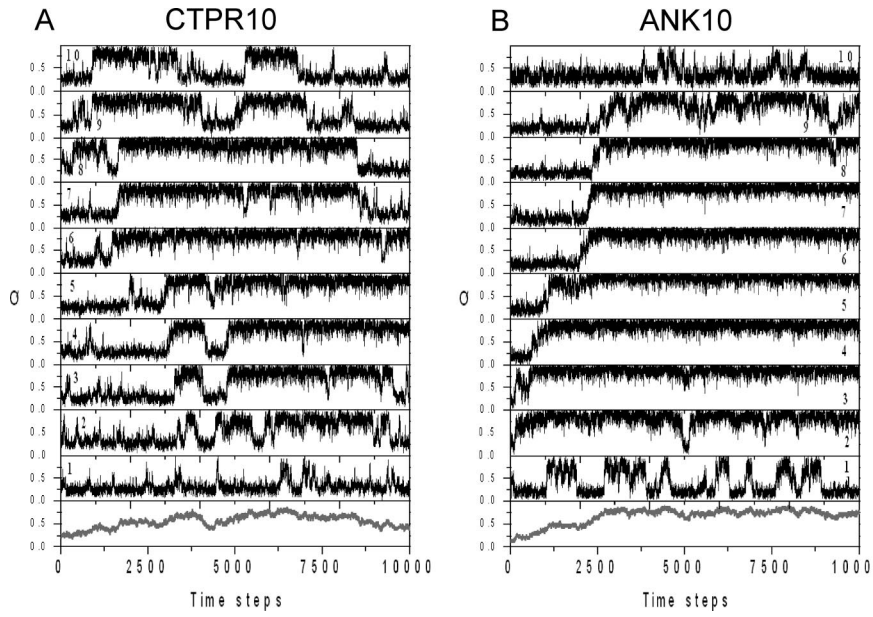

Figure 4. Folding transitions of elongated repeat proteins: time evolution of the number of intrarepeat contacts of the 10 repeating units constituting (A) CTPR10 and (B) ANK10 along a typical folding transition. For each protein, the number of native contacts (Q) of the entire protein is shown in the bottom panel. The numbers of native contacts of the individual repeating units are shown separately in the upper panels (which are numbered accordingly).

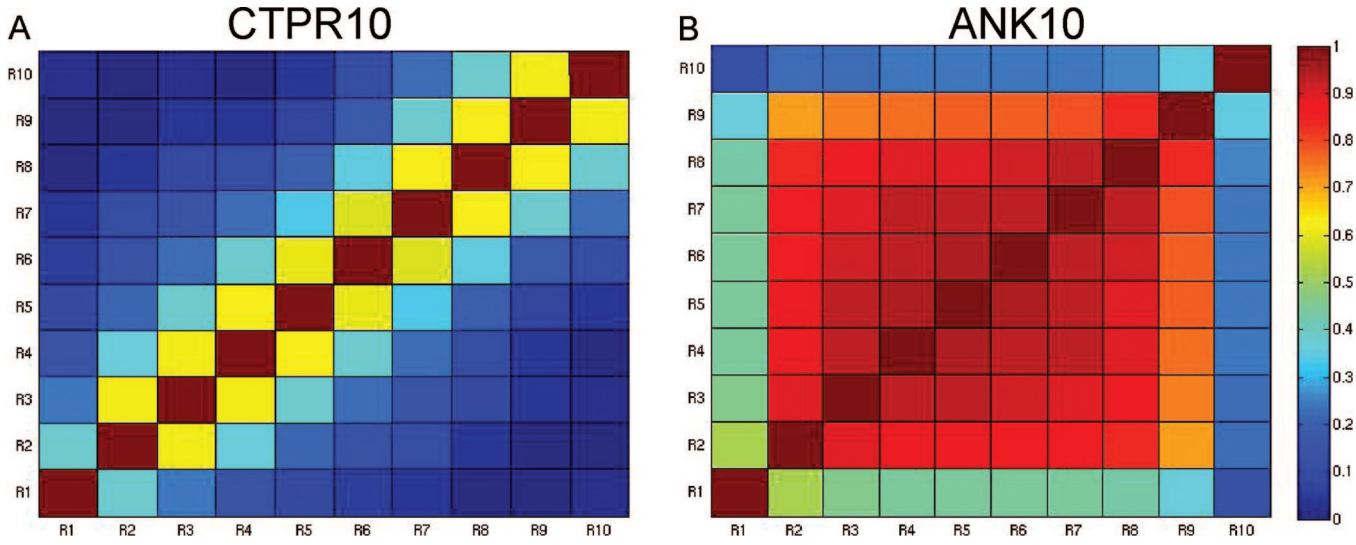

Figure 5. Coupling between the folding of individual repeats: correlation between the degree of folding of all of the pairs of repeating units in (A) CTPR10 and (B) ANK10. The correlation coefficients between the repeats are based on the Q values of each repeat throughout the trajectories of these proteins at T_F.

the folded and unfolded states are equally stable. We explored the folding of the entire protein as well as the coupling between its repeating units. Figure 3 shows representative folding events of ANK and CTPR proteins composed of 10 repeats and illustrates that an ANK protein folds significantly more slowly than a CTPR protein with the same number of repeats (also see Figure 2d). The different folding rates most likely stem from the different topological properties of the intra- and inter-repeat interactions of the ANK and CTPR families. Exploring the coupling between the folding of different repeating units during a complete folding event of the entire protein (Figure 4) illustrates the behaviors observed in these two families. While the folding transition of ANK is more cooperative than that of CTPR, it is still composed of sequential events involving assembly of several repeats at a time until the folding is completed.

When the correlations between the folding of the repeating units constituting a given repeat protein are examined, an interesting pattern is seen (see Figure 5). In the case of the CTPR family, the folding behavior of a given repeat is most correlated with its neighboring units, and this correlation weakens as the units become more separated from each other (Figure 5a). The results for the ANK family present a very different picture, in which the central units are all well-correlated with one another, leaving only the terminal units to fold independently from the central bulk (Figure 5b). These observations are in agreement with experimental observations that at equilibrium, many ANK proteins fold in a cooperative two-state manner[21,48,49] while CTPR proteins fold and unfold in a sequential way.[47]

Thus, in the CTPR series, there are relatively fast transitions between unfolded and folded regions, and the units rapidly fold and unfold in a sequential and weakly cooperative manner.

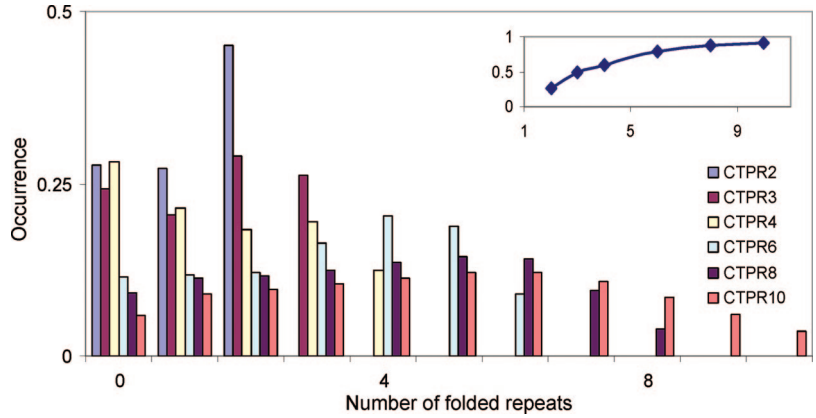

Figure 6. Population of individual folded repeating units in the folding reactions of several members of the CTPR series. The occurrence of intermediates can be seen in the appearance of 1 to $(n - 1)$ folded repeating units in the folding reaction of CTPRn [i.e., for a two-state protein, one would expect a bimodal distribution where no folded repeats are found (the unfolded state) or all of the repeats are folded (the folded state)]. The population of partially folded intermediates increases as the number of repeating units rises. This can be seen in the inset, where the accumulated occurrence of 1 to $(n - 1)$ folded repeats is shown as a function of n in CTPRn.

During this sequential process, intermediates that include a number of consecutive repeats are accumulated. Comparison of the results for the members of the CTPR series leads to the observation of a consistent trend in which the occurrence of these intermediates increases as the number of repeating units rises. These stable intermediates are mostly composed of structured repeating units that fold independently, leading to a breaking of cooperativity (Figure 6). It seems that the major difference between the CTPR family members occurs in the intermediate accumulation stages, in which the protein sequentially folds. The CTPR family seems likely to fold more slowly as the proteins become more elongated, as a result of the slower propagation stage, which requires assembly of a larger number of folded units

The folding in the ANK family is similar to that in the CTPR family but shows some discrepancies (see Figures 4 and 5). While CTPR proteins undergo sequential folding and unfolding in a rapid manner, it seems that the ANK proteins significantly populate the unfolded and folded states and that the transition between them is relatively fast. In order to examine whether differences in the nucleation stages result in the observed deceleration, we analyzed the unfolded region of the ANK series. The rate of forming two or three consecutive folded repeats (which can be treated as the folding nucleus[53]) in the unfolded state is similar in all of the larger members of the ANK series (data not shown), suggesting that the propagation of the 1D folding is the origin of the rate decrease as the length of the ANK protein increases. The high cooperativity in ANK folding is obviously another origin of slower folding. In a theoretical study,[28] it was suggested that in small ANK repeat proteins, the coupling energy between the units (i.e., the strength of the interface) is higher than the stability of each individual repeat (each repeat was shown to be unstable by itself[53,54]), leading to a cooperative folding behavior. As the number of ANK repeating units increases, stable intermediates composed of several structured repeats may be formed, resulting in a smaller folding rate. This is seen in experimental studies of some elongated repeat proteins.[22,55,56] Others[24] have speculated that the barrier responsible for slowing down the folding reaction of more elongated ANK proteins, such as the Notch receptor ANK domain, stems from the necessity of forming an intermediate that includes a larger number of structured consecutive repeats. Recently, the slow folding of an ANK protein composed of 12 repeating units (the D34 fragment of AnkyrinR[26,55]) was explained using the capillarity picture of folding of an elongated protein. It was suggested that impeding of the 1D propagation of the folding front by an intermediate dictates the folding rate.[57]

In the CTPR and the ANK families, similar trends of deceleration are observed as the proteins become more elongated. While the folding deceleration seems to originate from slower propagation of the assembly of the repeats in a 1D geometry, it is probable that other factors (e.g., the intrinsic stability of the internal units as well as the size of the interface) also contribute to this phenomenon. Further studies are required in order to decipher the molecular origins of the differences between the folding kinetics and mechanisms of the CTPR and ANK repeats. In particular, the exact theory of folding by the capillarity approximation can quantify the interplay between nucleation and propagation of elongated proteins. This theory has recently been formalized for globular proteins,[58] classifying the growth of the nucleus into different patterns of core and interface condensation. It would be interesting to explore the folding of elongated proteins as well using a similar analysis. Furthermore, various Ising models, which have provided an insightful perspective on the folding thermodynamics of repeat proteins,[16,59−61] may be valuable in exploring the dynamics and kinetics of these systems and the effect of their molecular details on the folding rate.

(53) Zhang, B.; Peng, Z. *J. Mol. Biol.* **2000**, *299*, 1121–1132.
(54) Forrer, P.; Stumpp, M. T.; Binz, H. K.; Plückthun, A. *FEBS Lett.* **2003**, *539*, 2–6.
(55) Werbeck, N. D.; Itzhaki, L. S. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 7863–7868.
(56) Low, C.; Weininger, U.; Neumann, P.; Klepsch, M.; Lilie, H.; Stubbs, M. T.; Balbach, J. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 3779–3784.
(57) Ferreiro, D. U.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 9853–9854.
(58) Qi, X.; Portman, J. J. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 11164–11169.
(59) Kajander, T.; Cortajarena, A. L.; Main, E. R.; Mochrie, S. G.; Regan, L. *J. Am. Chem. Soc.* **2005**, *127*, 10188–10190.
(60) Mello, C. C.; Barrick, D. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14102–14107.
(61) Wetzel, S. K.; Settanni, G.; Kenig, M.; Binz, H. K.; Plückthun, A. *J. Mol. Biol.* **2008**, *376*, 241–257.

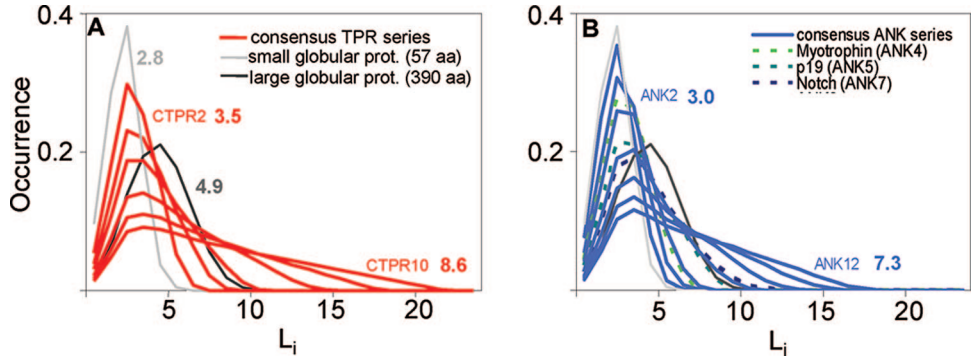

Figure 7. Distributions of the shortest path distances in various proteins: (a) CTPR series composed of 2−12 repeating units; (b) consensus-designed ANK repeat series composed of 2−12 repeating units and three natural ANK proteins (Myotrophin, p19, and Notch, composed of four, five, and seven repeating units, respectively). For comparison, the distribution for a small globular protein [src SH3 domain (PDB entry 1SRL, 57 aa)] appears in light gray and that for a larger globular protein [Trp synthase β2 subunit (PDB entry 1QOP, 390aa)] appears in dark gray in both panels. The average values of the shortest path distances, $\langle L \rangle$, for several proteins are indicated. It is apparent from the distributions that the structural features of small repeat proteins resemble those of globular proteins, while longer repeat proteins deviate by an amount that increases with length.

An Elongation-Sensitive Contact Order Gives Better Estimates of Folding Rates for Elongated Repeat Proteins. As discussed above, although different topological measures (i.e., Rel-CO, Abs-CO, SM-CO, and Q_D) have been shown to correlate with folding rates for many globular proteins, various elongated repeat proteins tend to deviate from that correlation. Since deviation from the expected rate correlates with deviation from globularity, we speculated that modifying the original topological CO measures in such a way that the degree of protein elongation would also be taken into account could yield a measure that would correlate well with folding rates for both globular and elongated repeat proteins.

For this reason, we developed several "elongation scales" that allow the quantification of the degree of a protein's elongation. This quantification was achieved using five different measures, among which were the radius of gyration (R_g), the protein's surface-to-volume ratio, and the average shortest path length connecting all of the nonlocal pairs in the protein ($\langle L \rangle$). All of the elongation scales allow one to distinguish among classical globular proteins, small repeat proteins (which are globular in nature), and more elongated repeat proteins. The details of these quantification methods are outlined in Figure S4 and associated text in the Supporting Information.

Using these elongation scales, we suggest an elongation-sensitive measure that is simply calculated by multiplying Abs-CO (representing topological complexity) by $\langle L \rangle$ (representing the degree of elongation). We selected the Abs-CO measure because of its good correlation with the folding rates of many proteins, irrespective of their folding mechanisms and sizes (see Figure 2c,d). We elected to represent the degree of elongation by $\langle L \rangle$ because it impressively distinguishes between globular and elongated proteins (Figures 7 and Figure S4 in the Supporting Information), yet other elongation scales might alternatively be employed. Various scaling-law relationships involving Abs-CO and $\langle L \rangle$ were examined with the aim of finding the optimal correlation with folding rates. Specifically, k_F was plotted versus Abs-CO $\times \langle L \rangle^{\alpha}$, where α ranged from -5 to $+5$; the best correlation was observed for $\alpha \approx 1$ (i.e., for a linear dependence on $\langle L \rangle$) (see Figure S6 in the Supporting Information). However, the physical interpretation of this relation is nontrivial. It is probable that $\langle L \rangle$ successfully represents the decrease in the ability of different parts of the protein to fold simultaneously as well as the sequential propagation processes in elongated proteins. Since in repeat protein series the geometry and topology remain very similar, they represent a simple case in which the folding time scale is directly related to the protein's length.

Plotting the experimental folding rates versus our Abs-CO $\times \langle L \rangle$ measure yields a trend line for the globular proteins near which most of the elongated proteins also fall without damaging the strength of the correlation for the globular proteins (Figure 8a). This novel elongation-sensitive CO measure can improve the prediction of folding rates for elongated proteins by several orders of magnitude. For example, the rate prediction for the Notch receptor ANK domain is improved 300-fold, and the correlation of the experimental folding rates of elongated and globular proteins with Abs-CO $\times \langle L \rangle$ is improved relative to that with Abs-CO while the correlation for the globular proteins with these two measures is unaffected (Figures 2c and 8a). This improvement can be better observed when the simulated folding rates, which include a larger number of elongated proteins, are compared using the Abs-CO $\times \langle L \rangle$ measure. For the simulated rates, the correlation coefficient for all of the tested proteins based on Abs-CO $\times \langle L \rangle$ (-0.73) is improved compared with that based on Abs-CO (-0.36), and the correlation of the globular proteins also shows an improvement when it is fitted using the elongation-sensitive CO (Figures 2d and 5b). In regard to the two repeat series, the smaller proteins CTPR2−4 and ANK2 sit almost on the correlation line, indicating that their folding is similar to that of globular proteins. Their more elongated counterparts, the CTPR and ANK proteins with 8 and 10 repeats, deviate from their expected rates as predicted using Abs-CO $\times \langle L \rangle$ by up to 1 order of magnitude, compared with deviations of 4 and 8 orders of magnitude for the Abs-CO predictions for CTPR10 and ANK10, respectively (Figure 2d). Thus, the suggested new measure substantially improves the correlation and allows a better prediction of folding rates for repeat proteins as well.

The elongation-sensitive CO measure predicts that as the protein incorporates more repeating units, its folding rate will become slower. This is consistent with the experimental folding rates of Myotrophin, p16, p19, and Notch (four different ANK repeat proteins that include four, five, and seven repeating units, respectively), which show that each series member folds more slowly than its shorter predecessors. The folding rates of these natural ANK proteins are similar to the simulated folding rates of consensus-designed analogues of comparable size, strengthening the concept that folding rate progressively decreases as

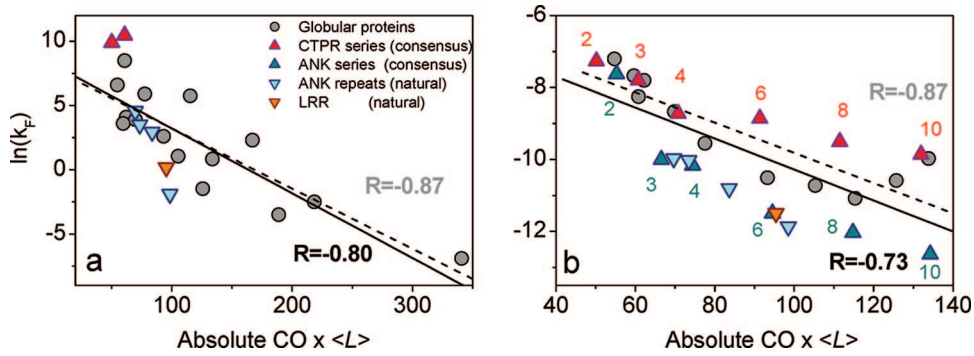

Figure 8. Correction of folding prediction using elongation-sensitive contact order: folding rates [$\ln(k_F)$] of globular and repeat proteins as a function of the elongation-sensitive CO measure Abs-CO × $\langle L \rangle$. (a) Correlation of experimental folding rates for globular proteins alone (dashed gray line, $R = -0.87$) and that when repeat proteins are included (solid black line, $R = -0.80$). (b) Correlation of simulation-based folding rates for globular proteins alone (dashed gray line, $R = -0.87$) and that when repeat proteins are included (solid black line, $R = -0.73$). In both cases, the correlation of the elongation-sensitive CO with globular and repeat proteins is better than the correlation with the Abs-CO measure (see Figure 2c,d).

the protein incorporates more repeating units. However, we note that because of the plasticity of the energy landscape of repeat proteins[29,62−64] it is possible that variations in their sequence and structure, which are unavoidable in natural sequences of repeat proteins, may affect their folding rates. Accordingly, these local perturbations may break the symmetry of the repeat protein and may compete with the topological and shape factors that often dictate the folding rate and mechanism.

Conclusions

Many studies have shown that the folding rates of proteins are largely dependent on the topological complexity of their native state, which allows prediction of folding kinetics and mechanism based solely on the protein's structure in the native state. However, the gross geometrical properties of proteins other than size were ignored by these measures. In this study, we have shown that the overall geometry of the protein may significantly affect its folding kinetics. Unique geometrically shaped proteins, such as elongated proteins, may have folding rates that deviate from those expected on the basis of a topologically based prediction, as their geometries are very different from that of classical globular proteins.

Using both experimental kinetics studies and simulations of repeat proteins, we found that the deviation from the expected kinetics behavior correlated with the extent of the protein's deviation from globularity. This deviation, manifested in the much smaller rate than expected as the proteins become more elongated, is a probable outcome of the increased complexity of the folding pathway. Periodic repeat proteins with a quasi-1D geometry can fold via more intermediates, resulting in slower kinetics, as the ability to fold in a simple two-state manner is impeded when the structure becomes more elongated.

Using a simple elongation scale that quantifies the protein's degree of elongation, we have suggested a modification to the known topology-based measures. This measure takes into account the elongation scale but does not damage the correlation for globular proteins. The modification, a simple multiplication of Abs-CO by the average shortest path length of a protein, $\langle L \rangle$, encompasses topological complexity and size as well as the protein's gross geometrical structure and allows a better prediction of the kinetics of various repeat proteins. Thus, while previously used measures failed to discriminate among various repeat proteins with different numbers of repeating units, the suggested elongation-sensitive contact order measure, which takes their degree of elongation into consideration, yields good predictions for both classical globular proteins and elongated repeat proteins.

Methods

Proteins Used in the Study. In this work, we used dozens of globular and repeat proteins for various analyses. All of the proteins are listed in Table S1 in the Supporting Information, which provides their Protein Data Bank (PDB) entry number, size (N), and folding rate [as $\ln(k_F)$].

Simulation Procedures and Analyses. Selected proteins were studied using molecular dynamics, simulated by the Langevin equation. We used a simple native-topology-based model (also called the Gō model) that assumes a perfectly funneled energy landscape.[65] This model has been previously shown to reproduce experimental kinetics rates as well as to capture other processes involved in folding, such as folding intermediates and protein dimerization and assembly[3,12,27,66−72] (see Figures S2 and S3 in the Supporting Information). Determination of protein-folding rates was done using the mean passage time (MPT) method, in which the folding rate of the protein is correlated with the length of time for which the protein remains unfolded in transition on average. The length and number of simulations for each of the proteins were big enough to provide a large sampling. A more detailed description of the model and the rate calculations is given in the Supporting Information.

Topology-Based Rate Measures. Folding-rate values from both experimental and simulation studies were compared to various topology-based measures: Rel-CO, Abs-CO, SM-CO, and Q_D and its size-sensitive analogue.

The CSU method[73] was used to determine the contacting residues required to calculate the various measures, as it is considered to be

(62) Ferreiro, D. U.; Komives, E. A. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 7735–7736.
(63) Lowe, A. R.; Itzhaki, L. S. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 2679–2684.
(64) Tripp, K. W.; Barrick, D. *J. Am. Chem. Soc.* **2008**, *130*, 5681–5688.
(65) Go, N. *Annu. Rev. Biophys. Bioeng.* **1983**, *12*, 183–210.
(66) Takagi, F.; Koga, N.; Takada, S. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 11367–11372.
(67) Clementi, C.; Jennings, P. A.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5871–5876.
(68) Levy, Y.; Cho, S. S.; Onuchic, J. N.; Wolynes, P. G. *J. Mol. Biol.* **2005**, *346*, 1121–1145.
(69) Wang, J.; Xu, L.; Wang, E. *Biophys. J.* **2007**, *92*, L109–L111.
(70) Levy, Y.; Caflisch, A.; Onuchic, J. N.; Wolynes, P. G. *J. Mol. Biol.* **2004**, *340*, 67–79.
(71) Levy, Y.; Onuchic, J. N. *Acc. Chem. Res.* **2006**, *39*, 135–142.
(72) Lu, Q.; Lu, P.; Wang, J. *Phys. Rev. Lett.* **2007**, *98*, 128105.
(73) Sobolev, V.; Wade, R. C.; Vriend, G.; Edelman, M. *Proteins* **1996**, *25*, 120–129.

more exact than a simple distance cutoff definition of native interactions in determining which pairs are in contact. From the resulting contacting list, CO values were calculated using the following formulas: Rel-CO = $(CN)^{-1}\sum^{C}|i - j|$, Abs-CO = Rel-CO × N, and SM-CO = Rel-CO × $N^{2/3}$, where N is the number of residues, C is the number of contacts, and $|i - j|$ is the distance in sequence between residues in contact. Our suggested elongation-sensitive contact order measure was calculated as Abs-CO × $\langle L \rangle$. A detailed description of the calculation of $\langle L \rangle$ as well as of other elongation measures is provided in the Supporting Information.

Acknowledgment. We thank Lynne Regan for sharing unpublished results and Joel Sussman, Tzviya Zeev Ben-Mordehai, Debora Fass, and Vladimir Sobolev for helpful talks on geometrical characterization. We also thank Paul Whitford, Amit Mor, Ariel Azia Amitai, Amir Marcovitz, Dalit Shental-Bechor, Asaf Carmi, Yuval Shifriss, and Yoel Kahanka for technical assistance and Reut Avni for initial work on the ANK repeat proteins. This work was supported in part by the Kimmelman Center for Macromolecular Assemblies, the Clore Center for Biological Physics, the Center for Complexity Science (Y.L.), and the Foundation Fernande et Jean Gaj. Y.L. is the incumbent of the Lilian and George Lyttle Career Development Chair.

Supporting Information Available: Figures S1−S6 and Table S1. This material is available free of charge via the Internet at http://pubs.acs.org.

JA804280P