Evidence for the principle of minimal frustration in the evolution of protein folding landscapes

Franco O. Tzuib,b, Daniel Vasilchukb,c, and George I. Makhatazdeab,c,

*Department of Biological Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180; Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY 12180; and *Department of Chemistry and Chemical Biology, Rensselaer Polytechnic Institute, Troy, NY 12180

Edited by José N. Onuchic, Rice University, Houston, TX, and approved January 24, 2017 (received for review August 19, 2016)

Theoretical and experimental studies have firmly established that protein folding can be described by a funneled energy landscape. This funneled energy landscape is the result of foldable protein sequences evolving following the principle of minimal frustration, which allows proteins to rapidly fold to their native biologically functional conformations. For a protein family with a given functional fold, the principle of minimal frustration suggests that, independent of sequence, all proteins within this family should fold with similar rates. However, depending on the optimal living temperature of the organism, proteins also need to modulate their thermodynamic stability. Consequently, the difference in thermodynamic stability should be primarily caused by differences in the unfolding rates. To test this hypothesis experimentally, we performed comprehensive thermodynamic and kinetic analyses of 15 different proteins from the thioredoxin family. Eight of these thioredoxins were extant proteins from psychrophilic, mesophilic, or thermophilic organisms. The other seven protein sequences were obtained using ancestral sequence reconstruction and can be dated back over 4 billion years. We found that all studied proteins fold with very similar rates but unfold with rates that differ up to three orders of magnitude. The unfolding rates correlate well with the thermodynamic stability of the proteins. Moreover, proteins that unfold slower are more resistant to proteolysis. These results provide direct experimental support to the principle of minimal frustration hypothesis.

The energy landscape theory provides a conceptual physico-chemical framework for understanding protein folding. This theory is based on the principle of minimal frustration that “...quantifies the dominance of interactions stabilizing the specific native structure over other interactions that would favor nonnative, topologically distinct traps” (1). A consequence of this is that the folding energy landscape of naturally occurring proteins is funnel-shaped (1–22). The shape of this funnel depends on two main factors that can introduce frustration and roughness: topology and the extent of nonnative interactions. Topological frustration can occur when certain native interactions are formed too early and need to be undone to allow for other interactions to form first, leading to backtracking and/or cracking (23–28). Weak nonnative interactions can have complex effects on the folding landscape (29–31): small amounts of weak nonnative interactions can assist folding, whereas larger amounts can create internal friction that will slow folding (32–35). For a given protein fold, both topological and energetic frustrations will depend on the amino acid sequence. Theoretical studies have suggested that naturally occurring proteins have selected sequences that are compatible with the principle of minimal frustration (36).

The goal of this study is to experimentally test the evolutionary validity of the principle of minimal frustration by characterizing protein sequences that span a wide range of stability and evolutionary time. The principle of minimal frustration states that all naturally evolved proteins have optimized folding energy landscapes, with topological or energetic barriers that are often a consequence of forming the biologically functional folded state. We can assume that all proteins within a given functional fold evolved from a common ancestor. This common ancestor was presumably a well-folded protein and had robust folding rates. To acquire such properties, the energy landscape had to follow the principle of minimal frustration. Numerous lines of evidence suggest that the last universal common ancestor (LUCA) was a thermophilic organism (37). Correspondingly, the proteins in this organism were more thermostable than those from modern mesophilic organisms (38–43). During early evolution, different organisms had to colonize different environments that required their proteins to be more or less stable than those of LUCA. If they evolved in such a way that their folding became slower (less robust), this would suggest that the folding energy landscape is not minimally frustrated anymore, and it would violate the principle of minimal frustration. If the proteins evolved to have significantly faster folding rates, this would suggest that the LUCA was in violation of the principle of minimal frustration. Thus, proteins that belong to the same functional protein family are expected to have similar folding energy landscapes, which macroscopically should be manifested via similar folding rates. However, protein stability also needs to evolve to be compatible with the growth temperature of their host organisms. Thus, for a simple two-state system, the modulation of stability within a given family of proteins should occur via modulation of unfolding rates (i.e., more stable proteins should have slower unfolding rates, whereas folding rates remain largely independent of stability). Computational modeling, both on lattice (3) and off lattice (44–46), provides additional support for this hypothesis. Here, we experimentally test this hypothesis using functional homologs from the...
thioredoxin (Trx) protein family from extant organisms or those resurrected using ancestral sequence reconstruction (ASR) as our model system.

**Results and Discussion**

**Model Protein Family.** Trx is a class of essential small (12 kDa) redox proteins found in nearly all known organisms (47) and believed to be present in early life (37). They function as antioxidants by facilitating the reduction of other proteins through cysteine thiol-disulfide exchange. The α/β protein fold of Trx consists of five antiparallel β-sheets sandwiched by four α-helices and is highly conserved (Fig. 1). We studied extant Trx from eight different organisms, including those living under extreme environmental conditions: psychrophilic (optimal growth temperature from 5 to 10 °C) Shewanella benthica (TrxSB) and Colwellella piezophilia (TrxCP); mesophilic (optimal growth temperature of ~37 °C) Escherichia coli (TrxEC) and Homo sapiens (TrxHS); and thermophilic (optimal growth temperature of >70 °C) Thermus thermophilus (TrxTT), Pyrococcus yayanosii (TrxPY), Methanococcus jannaschii (TrxMJ), and Sulfolobus tokodaii (TrxST).

**Extant Trx Stabilities and Folding/Unfolding Kinetics.** Previous studies of stability of Trxs indicated that these proteins are highly stable (41, 48). Indeed, even at pH 2, the transition temperatures range between 40 and 95 °C (Fig. 2A). This variation in the transition temperatures seems to correlate with the optimal growth temperatures of the corresponding organisms. For example, Trxs from thermophilic organisms, such as TrxTT, TrxPY, TrxMJ, and TrxST, have much higher Tm values than their mesophilic homologs TrxEC and TrxHS. The mesophilic proteins, in turn, have higher melting temperatures that the proteins from psychrophilic organisms TrxSB and TrxCP. This correlation of the growth temperature and protein stability is also evident from the denaturant-induced unfolding studies. Fig. 2B shows urea-induced unfolding profiles for the same set of proteins. The midpoint for the urea-induced unfolding transitions ranges from 1.5 to 7 M. For one protein, thermophilic TrxST, the midpoint is probably even higher, because no indication of unfolding was observed at 10 M urea.

How are the differences in thermodynamic stabilities manifested in the kinetic stabilities? Changes in thermodynamic stability can be caused by the differences in the folding rates, unfolding rates, or both. We thus performed characterization of the folding and unfolding of extant Trx variants using standard stopped flow methods. Fig. 2C shows the experimentally measured folding and unfolding rate, k_{obs}, in the form of chevrons for the extant Trx variants. Each data point on the chevron is an average of at least five independently measured traces, and their SDs are reported as the error bars. It is evident from these chevron plots that there is minimal variation in the rates of Trx folding. They all fold in the absence of denaturant in 80 ± 60 ms, and the difference between the fastest (20 ms) and the slowest (140 ms) is only sevenfold.

More notably, inspection of the chevron plots shown in Fig. 2D reveals that the variation in the unfolding rates is dramatic. Extrapolation to a zero denaturant concentration gives unfolding rate constants in the range between 7 and 22,000 s⁻¹ [6 d; i.e., ~3,000-fold difference between the fastest (TrxCP) and the slowest (TrxPY) unfolding Trx proteins]. Importantly, the unfolding rates correlate with stability, whereas folding rates show little correlation with stability within the extant Trx protein family (Fig. 3). This observation suggests that the differences in stabilities of extant Trx
Thermodynamic stability of Trx variants correlates with the unfolding rates and does not correlate with the folding rates. Equilibrium stabilities, $\Delta G_{eq}$, are plotted vs. folding ($\Delta G_{folding}^\#$, black regression line) and unfolding ($\Delta G_{unfolding}^\#$, red regression line) barriers. The folding/unfolding rates are extrapolated to zero denaturant concentration using data for extant and ancestral Trx variants shown in Fig. 2 C and D. Symbol/color coding for different proteins is the same as in Fig. 2.

are largely caused by the differences in the unfolding rates of these proteins.

**Ancestral Trx Stabilities and Folding/Unfolding Kinetics.** Similarity of the folding rates among extant Trx proteins is not an unexpected finding, because we are assessing the kinetics of a specific fold, and there are several reports correlating various topology metrics to folding kinetic rates (49–52). Moreover, the observed robust folding rates suggest that the Trx energy landscapes are unfrustrated (3, 5, 29, 36, 53, 54). This observation, in turn, implies that the folding of extant proteins of the Trx family obeys the principle of minimal frustration and that variations in stability are caused by the unfolding rates. Was this part of the evolutionary considerations?

To answer this question, we turned our attention to the analysis of the thermodynamics and kinetics of the Trx proteins for which sequences were obtained from ASR (41, 55). ASR is a bioinformatics technique that allows sequence extrapolation from extant protein sequences back to their common ancestors. It has been used to resurrect ancestral nodes for a number of proteins and shown to produce catalytically and/or functionally active proteins (40–42, 56–58). The sequences for the ancestral Trx proteins were obtained using maximum likelihood sequence reconstruction, targeting several Precambrian nodes during Trx evolution (41). The seven sequences that were studied here are last animal and fungi common ancestor (TrxLAFC), last eukaryotic common ancestor (TrxLECA), last bacterial common ancestor (TrxLBCA), last $\gamma$-proteobacteria common ancestor (TrxLGPCA), archaea/eukaryota common ancestor (TrxAECA), last cyanobacterial, deinococcus and thermus common ancestor (TrxLPBCA), and last archaeal common ancestor (TrxLACA). These sequences were dated back to over 4 billion years, thus enabling us to probe the thermodynamics and kinetics of folding/unfolding for proteins with sequences that arguably represent a wide evolutionary timespan (41). Furthermore, the crystal structures of these proteins have been determined and support the high conservation of the Trx protein fold (55).

Thermodynamic stabilities of the ancestral Trx proteins at pH 2 are compared in Fig. 2A. As with the extant Trx proteins, ancestral Trx variants showed significant variation in their transition temperatures. Proteins that are dated earlier on the evolutionary tree (e.g., LACA and AECA) have higher $T_m$ values than the more recent proteins, such as LECa and LAFCA, which in turn, are more stable than TrxEC. Equilibrium urea-induced unfolding profiles also support this correlation (Fig. 2B). The differences in transition temperatures at pH 2 agree well with the $T_m$ values obtained at pH 7 (Fig. S1), indicating a generality of the stability trend (41).

The results of folding/unfolding experiments for ancestral Trx variants measured using standard stopped flow methods are presented in the form of chevron plots in Fig. 2D. Again, as in the case of the extant Trx proteins, the folding rates of ancestral Trx are very similar, whereas the unfolding rates vary dramatically. This observation might suggest that the principle of minimal frustration played a role in the sequence selection early on in evolution.

Both ancestral and extant Trx proteins maintain strong correlation of increasing apparent unfolding barrier (i.e., decrease in the unfolding rate) with increase in thermodynamic stability, whereas the apparent folding barrier (as defined by the folding rates) remains independent of the thermodynamic stability. The thermal and kinetic stabilization conferred to the oldest ancestral variants is comparable with the most stable extant Trx proteins (e.g., TrxPY and TrxMJ).

The observation that both extant and ancestral Trx proteins fold with similar rates directly supports the assumption that
protein sequences have evolved with minimal energetic frustration so as to rapidly fold into their stable 3D states (36). It is also clear that the differences in thermodynamic stabilities are primarily caused by differences in unfolding rates for this protein family.

**Correlation of Kinetic Stability with Resistance to Proteolysis.** The variation in the unfolding rate defines kinetic (unfolding rate) stability of proteins, and it was argued that one of the important consequences of the kinetic stability is to protect proteins from proteolytic degradation (59–61). To probe the correlation of the kinetic stability and resistance to proteolysis, we treated Trx samples with pepsin and thermolysin (62). Both proteases have very broad sequence specificity but very different pH activity profiles: pepsin has maximal activity at acidic pH (pH 2.0), whereas thermolysin is maximally active at neutral pH (pH 7.0). We find that both pepsin and thermolysin resistance directly correlate with the unfolding rates at pH 2.0 (Fig. 4A). Furthermore, the thermolysin resistance directly correlates with the unfolding rates measured previously for a subset of Trx variants at pH 7.0 (Fig. 4B). This correlation holds for both extant and ancestral Trx variants, supporting the notion that, independent of the pH, high kinetic stability indeed confers proteins with greater resistance to proteolytic digestion.

**Concluding Remarks**

We have characterized the thermodynamics and kinetics of the Trx functional fold using extant proteins from organisms inhabiting a variety of living temperatures. We also compared the thermodynamic and kinetic properties of eight extant proteins with seven Trx proteins obtained through ASR with an evolutionary timespan of 4 billion years. The main goal was to use this model protein family for studying the evolution of protein folding and unfolding rates. We have shown that, despite a dramatic difference in the thermodynamic stabilities, Trx proteins fold with a very similar rate, and the variation in thermodynamic stability is largely defined by the unfolding rates (i.e., the more thermodynamically stable proteins are kinetically more stable as well). The variation in kinetic stability correlates with the resistance against proteolytic degradation. The observation that differences in thermodynamic stability are largely defined by the differences in unfolding rates based on 15 different proteins from the Trx family seems to be specific not only to this protein family. Carstensen et al. (43) compared stabilities and folding kinetics of modern (ββ8)-barrel protein HisF with its designed ancestral homolog Sym1. The authors found that Sym1 shows the same folding mechanism as HisF and that the higher thermodynamic stability of Sym1 is caused by major changes in the unfolding rates (43). Furthermore, compilation of literature data for other protein families, although not as extensive as this study and limited to extant proteins, shows very similar trends (Fig. 5) (i.e., the stability within a given protein structural family is largely modulated by the unfolding rates). This finding can be mapped to the funneled landscape, in which the transition state for a given protein fold remains minimally frustrated and modulation of the thermodynamic stability is achieved primarily through the optimization of the interactions in the native state. The illustration of the likely scenario of such effect, mapped onto a hypothetical folding energy landscape, is shown in Fig. 6.

Finally, it remains to be seen if other ancestrally reconstructed proteins maintain similar kinetic properties. If they do, ASR may be an efficient way to engineer kinetically stable enzymes.

**Methods**

**Protein Expression, Purification, and Characterization.** The gene sequences of extant Trx proteins were codon-optimized and synthesized (Blue Heron Biotechnology Inc.) for expression in *E. coli*. The DNA sequences of ancestral Trx proteins were optimized in a similar manner as described previously (41). Sequences for the extant and ancestral Trxs are given in Table S1. His6 tags were engineered at the N terminus of all protein sequences.

Plasmids containing genes for the extant and ancestral proteins were transformed into *E. coli* BL21 (DE3), and cells were grown at 37 °C. The expression was induced with 1 mM isopropyl β-D-thiogalactopyranoside when the cell density reached ~0.8 optical units at 600 nm. Cells were harvested after 6 h of induction. All proteins were purified to homogeneity under native conditions using nickel-nitrilotriacetic acid (Novagen; Merck KGaA) affinity resin followed by size exclusion chromatography according to previously published protocols (63–65).

Purities and identities of the recombinant proteins were confirmed by SDS gels and MALDI-TOF MS as previously described (63). In all cases, a single major peak was observed with a mass within 2–5 Da of that expected on the basis of the amino acid sequence (Table S2). Protein concentrations were determined spectrophotometrically using the molar extinction coefficients listed in Table S2.

The oligomerization state of all Trxs was characterized via analytical ultracentrifugation (AUC). Sedimentation equilibrium experiments were performed on a Beckman XLA analytical ultracentrifuge at pH values 3.0 (30 mM glycine-HCl), 5.5, and 7.0 (30 mM sodium cacodylate). Absorbance was monitored at 280 nm using either short- or long-column cells, and samples were allowed to equilibrate at three different rotor speeds at 20 °C. Global

**Fig. 5.** Correlation between thermodynamic stabilities and unfolding rates in eight different protein families: black circle, SH3 domains (78–80); red square, immunity binding domains (81, 82); light-green triangle, protein G (83, 84); yellow inverted diamond, cold shock proteins CspB (85); blue diamond, isopropyl malate dehydrogenases IPMDH (86); pink hexagon, acyl-CoA binding proteins ACPB (87); light-blue circle, rubredoxins (88); gray square, Trxs (this work). In all cases, there is a linear dependence (average slope of 1.00 ± 0.19; maximum of 1.33; minimum of 0.75) with statistically significant correlations ($R^2 = 0.95 ± 0.08$). Table S5 has details.

**Fig. 6.** Hypothetical folding energy landscape illustrating increase in stability (from left to right) because of a decrease of the unfolding rates, while folding rates remain the same. These changes in the energy landscape are consistent with the observed experimental data for the Trx protein fold (i.e., protein stability increase is caused by the increase in the energy barrier between transition-state and folded ensembles, whereas the energy barrier between unfolded and transition-state ensembles remained largely unchanged).
analysis of the centrifugation profiles was done as previously described (66). Table 33 shows molecular masses obtained from AUC experiments.

Differential Scanning Calorimetry Measurements. Differential scanning calorimetry (DSC) experiments were performed on a VP-DSC instrument (MicroCal/GE Healthcare) at a scan rate of 1.5 °C/min using protein concentrations of about 1.0 mg/mL. Buffers used were 30 mM glycine-HCl for pH values 2–3.5 experiments and 30 mM sodium-citrate for pH 7.0 experiments (67). Proteins were extensively dialyzed against the corresponding buffer. Raw DSC data were analyzed using the Origin-DSC (OriginLab) software package. Global fitting of the heat capacity profiles was done using in-house scripts using the nonlinear regression (NLREG) routine (68, 69).

Equilibrium Stability Measurements. Initially, all Trx protein stocks were extensively dialyzed in acidified water and diluted 20-fold into 30 mM glycine-HCl or buffered urea solutions adjusted to pH 2. Dilutions were done to final protein concentrations of 4–10 μM depending on their fluorescence quantum yield. Stock protein samples were diluted into buffered urea with concentrations incrementing up to 9.5 M and incubated at room temperature for 24 h. Trypsinophen emission spectra were collected as a function of urea concentration using a fluorescence plate reader (TECAN infinite M1000 Pro) in a 384-well black plate. Plate and samples were thermostated in the instrument at room temperature (23 °C ± 1 °C) for 5 min before data collection. Excitation was done at 260 nm with emission spectra recorded from 305 to 450 nm in 1-nm increments. Trx MJ does not have a trypsinophen; therefore, tyrosine emission fluorescence was used instead. Trx proteins that did not produce a significant intensity change were additionally characterized via far-UV CD.

Urea unfolding profiles (intensity at 355 nm vs. urea concentration) were obtained globally for all Trx variants according to the linear extrapolation model (70, 71) using in-house written NLREG scripts as previously described (72).

Kinetic Stopped Flow Experiments. All Trx protein samples were described as above. Data for chevron plots were collected by standard stopped flow methods on a JASCO J-815 spectrophotometer operating in fluorescence mode that was equipped with an SFM 300 mixing module (BioLogic Science Instruments) containing a high density (HDO) mixer and a 30-μL FC-15 observation cuvette (73). Buffer containing 30 mM glycine-HCl adjusted to pH 2 was used for all kinetic experiments. Folding and unfolding reactions were initiated by a 20-fold dilution of protein-buffered stock solutions (80–200 μM) with the desired buffered urea concentration. Because of slow kinetics of unfolding, equilibration in buffered urea was required before refolding experiments could be successfully initiated. The reagent mixtures, changing mix, and observation cuvette were thermostated at 20 °C using a circulating water bath. Fluorescence emission intensity from an N-WG 320-nm (N-WG 305 nm for TrxMJ) cutoff filter (BioLogic Science Instruments) was collected after excitation at 295 nm (280 nm for TrxMJ) through a 3-nm slit from a mercury lamp source. Voltages applied to the photomultiplier tube were set constant based on the fluorescence signal intensity at maximum amplitude (~920 V). The kinetic model required multiequilibrium fits with a slow phase related to proline isomerization (74–76). Thus, the pertinent fastest folding and unfolding phases are reported. Rate constants, κ ±obs, were obtained using stretched exponential as previously described (73):

\[ l(t) = pb + t + y_o + A \exp\left(\frac{\kappa_{\text{obs}}}{t}\right) \]  

where \( l(t) \) is fluorescence intensity as a function of time, \( y_o \) is the initial fluorescence intensity, \( A \) is the amplitude of the change between initial and final fluorescence intensities, \( \kappa_{\text{obs}} \) is the observed kinetic rate constant associated with the fluorescence intensity relaxation, and \( pb \) is the sloping baseline correction for the photobleaching effect used in the Bio-Kine32 software. All traces were corrected for instrumental dead time (6 ms) before fitting, because some Trx protein variants showed significantly fast \( k_{\text{obs}} \).

The natural logarithms of the \( k_{\text{obs}} \) are plotted as a function of urea concentration in the form of chevron plots. Each data point on the chevron plot is an average of five individual traces, and errors are taken as the SDs. Extrapolated \( k(H_2O) \) and \( k(\text{urea}) \) values were obtained by globally fitting the chevrons of all Trx variants to Eq. 2 below (77) using in-house scripts for the NLREG data-fitting package:

\[ \ln(k(\text{urea})) = \ln\left(\exp\left(\frac{\ln(k(H_2O)) + m_1(\text{urea})}{RT}\right) + \exp\left(\ln(k(H_2O)) + m_2(\text{urea})\right)\right) \]  

where \( k(H_2O) \) and \( k(\text{urea}) \) are the folding and unfolding rates in the absence of denaturant, respectively, and \( m_1 \) and \( m_2 \) are the kinetic and unfolding m values (measured in kilojoules per mole per molal), respectively (77). The apparent folding/unfolding barriers have been calculated as \( \Delta G_{\text{folding}} = RT\ln(1/k) \), where \( R \) is the universal gas constant (8.14 J/mol K), and \( T \) is the temperature (Kelvin).

Proteolytic Digestion. Extant and resurrected Trx variants were subjected to proteolytic digestion to assess their kinetic stabilities at two pH values using appropriate proteases. For the acidic range, they were digested with the Sigma-Aldrich in 50 mM sodium phosphate, pH 2.0. Proteolysis were carried out at 37 ± 1 °C with a protein:enzyme mass ratio of 10:1 and a final protein concentration of 0.5 mg/mL. At the desired digestion lapse time, an aliquot was removed and quenched with 50% (v/v)vol ammonium hydroxide solution to a final concentration of 1%. For the neutral range, thermolysin (Sigma-Aldrich) digestion was done in 50 mM sodium phosphate and 0.5 mM calcium chloride, pH 7 buffer. Proteolysis was carried out at 50 ± 3 °C, with a protein:enzyme mass ratio of 10:1 and a final protein concentration of 0.5 mg/mL. At the desired digestion lapse time, an aliquot was removed and quenched with 10% (v/v)vol formic acid to a final concentration of 0.5%.

Protein aliquots were immediately flash frozen in liquid nitrogen and stored at −20 °C until SDS/PAGE gels were ran. Samples were diluted in a 1:1 ratio with 2× bromophenol dye containing β-mercaptoethanol and resolved on prerace 4–20% (v/v) Precise Protein SDS/PAGE gels (Thermo Scientific). Gels were stained with Coomassie Blue, and subsequently they were destained and scanned at 1,200 dots per inch resolution using a flatbed scanner. Band intensities were analyzed with ImageJ software. Because some of the proteins were highly resistant to proteolysis, even at elevated temperatures, the \( t_{50} \) could not be determined. Thus, we based our analysis on the fraction of the native band remaining in solution after 2 h.

Acknowledgments. We thank Jose M. Sanchez-Ruiz for providing the plasmids for ancestral Trx and Catherine Roger for the usage of the high-throughput fluorescence plate reader. This work was supported by US National Science Foundation Grant 1506468/1330249.