Loops, Linkages, Rings, Catenanes, Cages, and Crowders: Entropy-Based Strategies for Stabilizing Proteins

HUAN-XIANG ZHOU*

Department of Physics and Institute of Molecular Biophysics, Florida State University, Tallahassee, Florida 32306

Received September 9, 2003

ABSTRACT

A protein molecule exists in either a compact folded state or a variable and open unfolded state. Since the unfolded state is favored by chain entropy, restricting its entropy is an attractive mechanism for shifting the equilibrium toward the folded state. A number of entropy-based strategies have been engineered or used by natural proteins to increase the folding stability: (a) shortening of loop lengths, (b) covalent linkage of dimeric proteins, (c) backbone cyclization, (d) catenation, (e) spatial confinement, and (f) macromolecular crowding. Theoretical analyses demonstrate the importance of accounting for consequences on the folded as well as the unfolded state and provide guidance for further exploitation of these stabilization strategies.

Introduction

The folding stability of a protein is measured by the free energy difference between the folded and unfolded states. Most discussions of protein stability have focused on the specific interactions (e.g., hydrogen bonding, packing, and burial of nonpolar groups) in the folded state. However, it is now recognized that chain entropy of a protein can be significantly perturbed by changes in loop length[1−5] and by spatial confinement and macromolecular crowding of the protein molecule.[6−10] Attention has also turned to variations on the typical linear chain topology of proteins. A number of proteins that fold only upon dimerization have been studied against single-chain variants obtained by covalently linking the monomers.[11−14] Circular proteins have been obtained by connecting the N and C termini by peptide linkers and, indeed, have been found in nature.[15−17] Most recently catenated proteins, consisting of two interlocked circular chains, have been discovered and designed.[18,19] The stabilization effects of loop length, covalent linkage, circularization, catenation, confinement, and crowding (Figure 1) are all amenable to modeling by polymer theory.[20−26] The theoretical models give indications on the magnitudes of these effects and provide guidance for further exploitation of the stabilization strategies.

Entropy Restrictions on a Polymer

Let us begin with a brief introduction to polymer theory, with a view toward later applications to protein molecules. A polymer chain samples different configurations, thus its end-to-end vector **r** is not fixed. Instead **r** has a broad distribution (Figure 2A). The normalized distribution function will be denoted as $p(\mathbf{r})$. Typically $p(\mathbf{r})$ only depends on the magnitude, i.e., the end-to-end distance r, not the direction of **r**. The simplest polymer chain has a Gaussian distribution function:

$$p_G(r) = (3/2\pi\langle r^2\rangle)^{3/2} \exp(-3r^2/2\langle r^2\rangle) \qquad (1)$$

where the mean square distance is given by $\langle r^2\rangle = b^2 n$, where b is an effective bond length and n is the number of bonds in the polymer chain.

If **r** is restricted to a certain region Ω (Figure 2A), then the fraction of allowed configurations is

$$f = \int_\Omega \mathrm{d}^3\mathbf{r}\, p(\mathbf{r}) \qquad (2)$$

This restriction increases the free energy of the polymer chain by

$$\delta G_1 = -k_B T \ln f \qquad (3)$$

where k_B is the Boltzmann constant and T is the temperature. This free energy reduction is entropic in nature.[27] In particular, Jacobson and Stockmayer[28] modeled ring formation by restricting **r** to the volume within a short contact distance. Other ways to change the fraction of allowed configurations include spatial confinement and macromolecular crowding[25,26] (see below).

Suppose that the vector **r** is restrained to a distribution function $P(\mathbf{r})$ without the polymer chain; then under the simultaneous action of the restraint and the polymer chain, **r** has a distribution function

$$q(\mathrm{r}) = \mathcal{N}^{-1} p(\mathrm{r}) P(\mathrm{r}) \qquad (4)$$

where the normalization factor $\mathcal{N}$ is $\int \mathrm{d}^3\mathbf{r}\, p(\mathbf{r})P(\mathbf{r})$. If the restraint changes from $P_u(\mathbf{r})$ to $P_f(\mathbf{r})$, then the free energy of the system changes by (Figure 2B)

$$\delta G_2 = -k_B T \ln g \qquad (5a)$$

$$g = \int \mathrm{d}^3\mathbf{r}\, p(\mathbf{r}) P_f(\mathbf{r}) / \int \mathrm{d}^3\,\mathbf{r}\, p(\mathbf{r}) P_u(\mathbf{r}) \qquad (5b)$$

Loop Length

Now consider a loop in the folded structure of a protein (Figure 1A). The loop is modeled as a polymer chain in both the folded and the unfolded states. The folded structure restricts the end-to-end vector of the loop to

Huan-Xiang Zhou received his B.S. from Wuhan University (China) and his Ph.D. from Drexel University. He did postdoctoral work at the NIH with Attila Szabo. After faculty appointments at HKUST and Drexel, he moved to Florida State University in 2002. His group does theoretical, computational, and experimental research on protein stability, folding, and interactions.

* Corresponding author: phone (850) 645-1336; fax (850) 644-7244; e-mail zhou@sb.fsu.edu.

10.1021/ar0302282 CCC: $27.50
Published on Web 01/07/2004

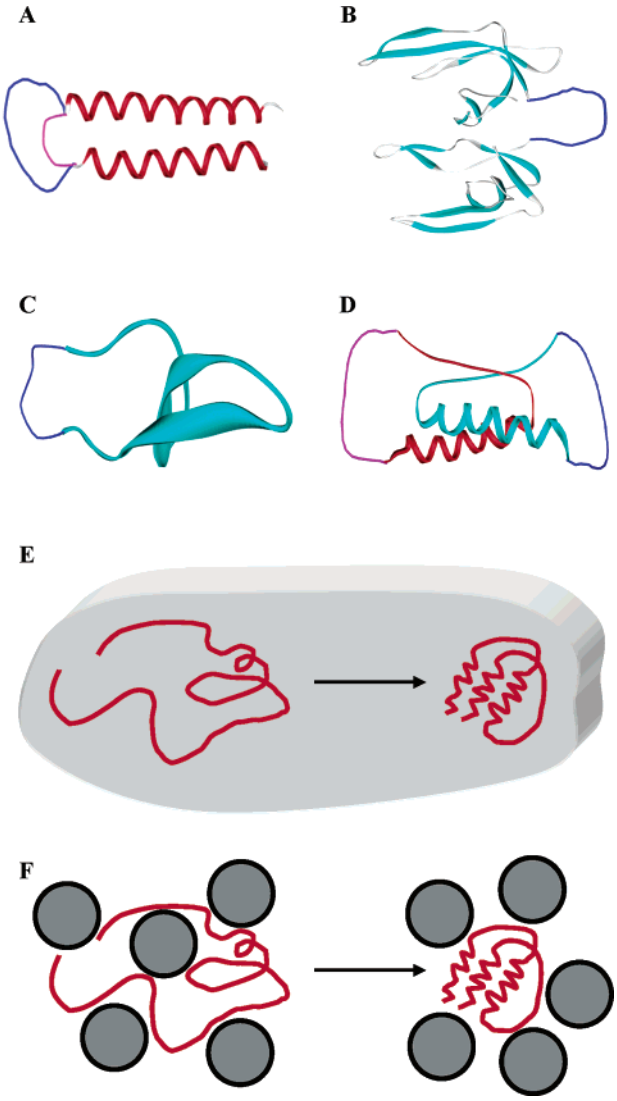

FIGURE 1. Different strategies for stabilizing proteins: change in loop lengths (A); covalent linkage of dimeric proteins (B); backbone cyclization (C); catenation (D); confinement (E); and crowding (F). Loops and linkers are shown in purple or blue. The shadowed box in panel E represents a confining cage, and black spheres in panel F represent crowding macromolecules.

small fluctuations around a fixed displacement, **d**. Thus the restraint $P_f(\mathbf{r})$ in the folded state can be modeled as a δ function:

$$P_f(\mathbf{r}) = \delta(\mathbf{r} - \mathbf{d}) \tag{6a}$$

On the other hand, the unfolded protein chain does not impose any restriction on **r** (if the excluded-volume effect is neglected). Thus the restraint here can be written as

$$P_u(\mathbf{r}) = 1/V \tag{6b}$$

where V is the volume containing the protein molecule. According to eq 5, the loop affects the folding free energy by

$$\delta G_{loop} = -k_B T \ln Vp(\mathbf{d}) \tag{7}$$

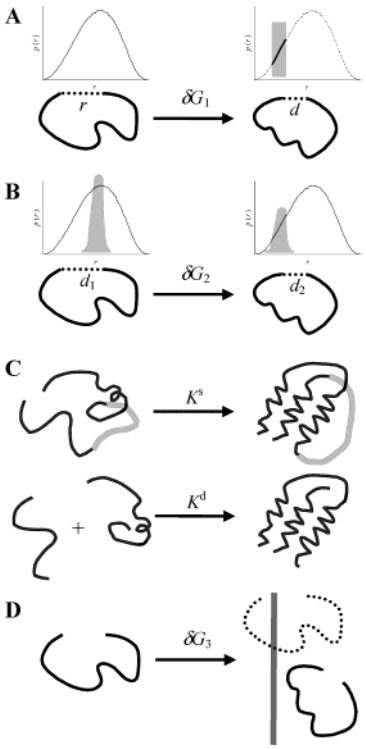

FIGURE 2. Changes in the entropy of a polymer chain by restriction of the end-to-end vector to a particular volume (A), by restraints (B), and by the presence of a physical boundary (D). A specific case of restraints (B) is shown in panel C, where a loop of a protein is either unrestrained (in the unfolded state) or restrained by the folded structure.

If the length of the loop is changed from l_0 to l peptide bonds, then the change in folding free energy is

$$\Delta\delta G_{loop} = -k_B T[\ln p(\mathbf{d};l) - \ln p(\mathbf{d};l_0)] \tag{8}$$

where the dependence of $p(\mathbf{d})$ on loop lengths l and l_0 is explicitly indicated.

For short loops, a model more realistic than the Gaussian chain is the wormlike chain.[20] Accurate values for $p(\mathbf{r};l)$ can be obtained from computer generations of the wormlike chain.[23] An approximate expression for the distribution function of the wormlike chain is given by[29]

$$p(\mathbf{r};l) = (3/4\pi l_p l_c)^{3/2} \exp(-3r^2/4l_p l_c)[1 - w(r;l)] \tag{9a}$$

$$w(r;l) = 5l_p/4l_c - 2r^2/l_c^2 + 33r^4/80l_p l_c^3 + 79l_p^2/160l_c^2 + 329r^2 l_p/120l_c^3 - 6799r^4/1600l_c^4 + 3441r^6/2800l_p l_c^5 - 1089r^8/12800l_p^2 l_c^6 \tag{9b}$$

where $l_c = 3.8l$ Å is the contour length and l_p is the persistence length. Equation 9a deviates from a Gaussian distribution by the presence of a "correction" term $w(r;l)$. The persistence length appropriate for short unstructured peptides has been determined by fitting the distributions of end-to-end distances of loops in protein structures to the predictions of the wormlike chain, with the result $l_p = 3.04$ Å.[20] With the approximate expression of eq 9, eq 8 becomes

$$\Delta\delta G_{loop}/k_BT = (3/2)\ln l + 3d^2/4l_p(3.8l) - \ln[1 - w(d;l)] - A_0 \quad (10)$$

where A_0 is the result of the first three terms at $l = l_0$. The first term, $(3/2)\ln l$, is the classical Jacobson−Stockmayer expression[28] obtained by using a Gaussian distribution and setting $d = 0$. Use of a more realistic polymer model and explicit account of the finite distance spanned by the loop should add to the accuracy of eq 8.

Equation 8 and the use of the wormlike chain model have been tested against experimental results of Nagi and Regan[1] for the effect of loop length on the four-helix-bundle dimeric protein Rop (Figure 1A).[20] Nagi and Regan replaced the native two-residue loop of each subunit by two to 10 glycines. The end-to-end distance for this loop, measured between the C_α atoms of Asp30 and Asp32, is $d = 5.6$ Å (calculated on the Protein Data Bank entry 1rop). With the 10-glycine loop as reference (i.e., $l_0 = 10$), the experimental and theoretical results for $\Delta\delta G_{loop}$ are compared in Figure 3A. Note that since Rop is a dimeric protein with two identical glycine-substituted loops, the prediction of eq 8 is doubled in the comparison with experiment. While the Jacobson−Stockmayer theory does reasonably well in reproducing experimental data, improvement is made by eq 8 (except for the shortest loop at $l = 2$).

The loop model presented above assumes that the ends of the loop are rigidly held by the protein structure but the loop itself is flexible. For the glycine-substituted loops in Rop, the choice of the two helix-terminal residues packed against each other in the wild-type protein as the ends of the loops seems obvious. The choice of the end residues of a loop to be modeled as flexible may not always be clear. Viguera and Serrano[2] studied the effects of the length of the linker connecting the N and C termini in circular permutants of the α-spectrin SH3 domain. The amino acid sequence of the linker and neighboring residues is L61-D62-**S-G**$_n$-T4-G5-K6, where n ranges from 1 to 10. If we choose D62 and T4 as the ends of the loop, which are separated by ∼10.0 Å (according to PDB entry 1g2b), then for $n = 2$, 4, 6, and 10, the loop lengths are 4, 6, 8, and 12. With $l = 4$ as reference, eq 8 predicts $\Delta\delta G_{loop} = 0.25$, 0.39, and 0.62 kcal/mol for the other three loop lengths. These can be compared with the experimental results of 0.27 ± 0.05, 0.49 ± 0.05, and 0.76 ± 0.08 kcal/mol. As observed by Viguera and Serrano, these results are also consistent with the Jacobson−Stockmayer prediction, $(3/2)\ln l/l_0$, if four residues bordering the inserted linkers are considered part of the flexible loop.

The Jacobson−Stockmayer theory predicts that loop entropy decreases with loop length, whereas eq 8 predicts that the end-to-end distance affects the dependence on loop length. It is entropically unfavorable both for a long loop to have close contact between the ends and for a short loop to extend to a long distance between the ends. For larger values of d, the loop length that minimizes $\Delta\delta G_{loop}$ is no longer the minimal number of residues required to span the end-to-end distance. For example, at $d = 10$ Å, the optimal loop length is four peptide bonds, instead of the minimally required three. Most loops in proteins are longer than the predicted optimal lengths;[20] a decrease in $\Delta\delta G_{loop}$ is thus generally expected upon loop shortening. Shortening of loops has been suggested as a mechanism for thermostability and indeed has been observed in some thermophilic proteins in comparison to mesophilic counterparts.[5] In addition, in a combinatorial mutagenesis experiment on the SH3 domain, enrichment of shortened n-src loops was obtained by phage-display selection.[30]

Covalent Linkage

Imagine that the loop considered above is deleted, leading to two separate subunits (Figure 2C). Then in the unfolded state the two subunits are free to move relative to each other, while in the folded state the displacement (**r**) from the end of the first subunit to the beginning of the second is distributed according to $P_f(\mathbf{r})$ instead of eq 4. The effect of covalently linking the two subunits is precisely given by eq 7. If the folding equilibrium constants of the dimeric protein and the single-chain variant are K^d and K^s, respectively, then $K^sV/K^d = \exp(-\beta\delta G_{loop}) = Vp(\mathbf{d})$, where $\beta = (k_BT)^{-1}$. Thus the effective concentration, C_{cl}, for covalent linkage is[21]

$$C_{cl} \equiv K^s/K^d = p(\mathbf{d}) \quad (11)$$

For a dimeric protein, the folded fraction increases with protein concentration. The concentration dependence disappears in the covalently linked single-chain variant. The folded fraction of a homodimeric protein is less than that of the single-chain variant when the total protein concentration is less than $C_{cl}\ (1 + K^s)/2$ (in monomer units). $p(\mathbf{d})$ can be expressed in molar concentration by dividing by Avogadro's number. Specifically, results in millimolar are obtained when a numerical factor $10^7/6.022$ is applied to $p(\mathbf{d})$ in units of angstroms^{-3}.

Folding thermodynamics of single-chain variants of a number of dimeric proteins have been studied,[11−14] allowing for direct test of eq 11.[21,22] One such single-chain variant, of the gene V protein, is illustrated in Figure 1B. As Table 1 shows, without adjustment of parameters, eq 11 is able to predict, to within a factor of 2, experimental results for C_{cl} spanning 2 orders of magnitude. The increase in the effective concentration of covalent linkage can be rationalized by the decreases in the end-to-end distance and the linker length. In contrast to the other four cases in Table 1, the native protein of chymotrypsin inhibitor 2 (CI2) is a single chain, but a dimeric version was created by cleaving the active-site loop at Met40.[31,32]

Traditionally, covalent linkage has been introduced by disulfide bonds. Disulfide bonds are relatively rigid and typically buried in the folded state. Therefore the effect of a disulfide bond will be more complex than merely providing a restraint. The flexible linker model perhaps could still give some indication on the magnitude of the effect. Since the C_α−C_α distances of disulfide-bonded residues are narrowly distributed, the resulting distribution function $p(\mathbf{d})$ can reach much higher values than its

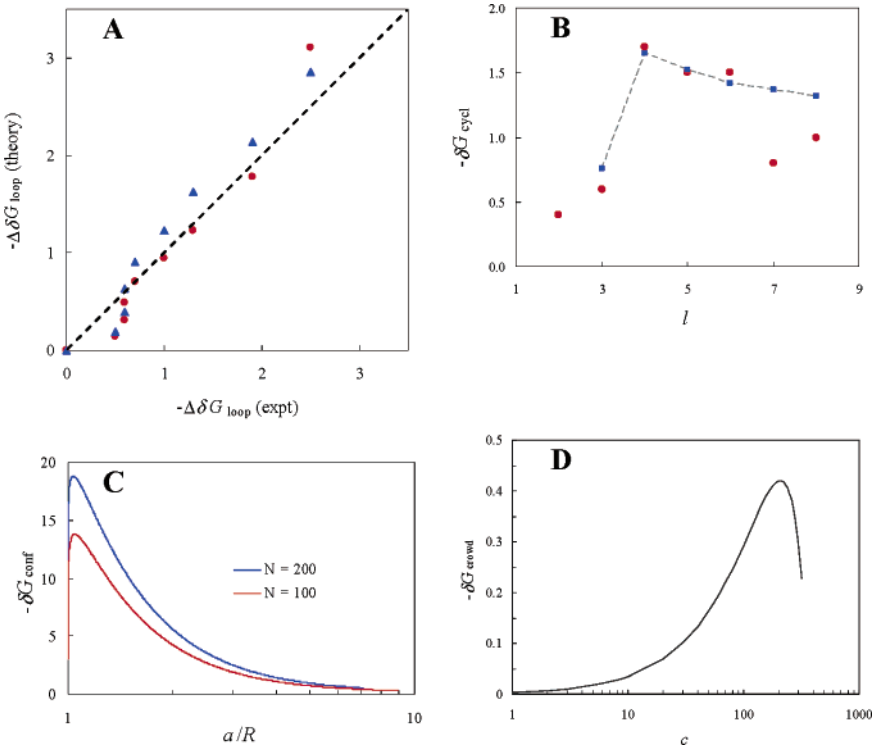

FIGURE 3. Effects on the folding free energy by loop length (A), backbone cyclization (B), spatial confinement (C), and macromolecular crowding (D). In panel A, circles and triangles show predictions of eq 8 and the Jacobson−Stockmayer theory, respectively; a diagonal line is drawn to aid in the comparison with experiment. In panel B, squares (connected by dashed lines) and circles show predictions of eq 12 and experimental results, respectively. In panel C, the radius a of the folded protein is 17.3 Å at $N = 100$ and 21.8 Å at $N = 200$. In panel D, the parameters used are $y = R_g/a_c = 1.95$, $z = a/a_c = 1.12$, and $\phi = (5.4 \times 10^{-4})c$, where c is the crowder concentration in grams per liter. Calculations are done for $T = 298$ K. All energies are in units of kilocalories per mole.

Table 1. Effect of Covalent Linkage on Protein Folding Equilibrium

protein[a]	conditions	K^d (mM^{-1})	K^s	linkage[b]	d^c (Å)	l^d	K^s/K^d (mM)	$p(d)^e$ (mM)
Arc repressor[12] (1myk)	4.19 M urea, 298 K	1	2.44	A R50−B M7	29.9	25	2.44	3.99
Cro repressor[13] (5cro)	326 K	0.19	1	A N61−C E2	24.9	15−23	5.3	6.0−8.1
CI2[31,32] (2ci2)	298 K	2.44×10^4	3.74×10^5	V34−D45	20.7	11	15	12
gene V protein[11] (1gvp)	2.6 M Gdn·HCl, 298 K	10^2	1.51×10^4	A A86−B M1	12.4	7	151	74
GCN4-p1[14] (2zta)	4 M Gdn·HCl, 283 K	0.353	85	A M2−B M2	6.3	9	241	129

[a] PDB entry of protein is given in parentheses. [b] End residues (e.g., R50 of subunit A and M7 of subunit B), [c] End-to-end distance. [d] Number of peptide bonds considered to be flexible. [e] Calculated according to eqs 11 and 9. No significant difference was found when $p(d)$ was obtained from computer generations of the wormlike chain.

counterparts for flexible peptide linkers.[21] The potentially large effective concentrations expected are in line with some experimental data. In particular, Jana et al.[13] also studied a single-chain Cro repressor variant V55C in which the two subunits are connected by the disulfide bond between the two mutated residues. The effective concentration, estimated from the melting temperatures of the dimeric repressor and the disulfide-bonded variant, was as high as 14 M. Similarly, the effective concentration for the disulfide-bonded variant D83C of *Streptomyces* subtilisin inhibitor (SSI) was 18 M.[33] However, potential conflict between the stereochemical requirement for disulfide bonding and the packing environment makes the effect of this type of cross-linking unpredictable.

The simple theory summarized by eq 11 helps resolve a debate between Karplus and Janin[34] and Privalov and Tamura[33,35] regarding the interpretation and theoretical implications of the latter group's experimental result on SSI. An entropy corresponding to the measured $C_{cl} = 18$ M of effective concentration for covalently linking SSI is $-k_B \ln (C_{cl}/1\ M) = -6$ cal/(mol·K). The magnitude of this entropy is much smaller than theoretical estimates for the loss of translational/rotational entropy upon formation of a protein−protein complex. Karplus and Janin attributed this discrepancy to the fact that the SSI subunits are unfolded when dissociated, while Privalov and Tamura pointed to deficiency of theory. According to the simple theory outlined here, C_{cl} primarily captures the effect of the linker. The linker, while constraining the relative translation of the subunits in the unbound state, has its end-to-end vector restricted to a displacement **d**. Other changes accompanying the complex formation, such as folding of the subunits and gain in vibrational entropy, are (or are assumed to be) unaffected by the covalent linking. As a side issue, I personally do not view that the often used 1 M "standard state" holds any physical

significance and the entropy $-k_B \ln (C_{cl}/1\ M)$ based on the use of this practice, in absolute (as opposed to relative) terms as particularly meaningful. In any event, it is hoped that eq 11 satisfies the desire of Privalov and Tamura[35] for a theory "to take into account the contribution of the long flexible linker into the measured entropic effect."

Backbone Cyclization

The connection of the N and C termini of a protein by a peptide linker results in a macrocycle (Figure 1C). The peptide linker, modeled as a polymer chain, is now restrained in both the folded and the unfolded states. Equation 6a again defines the restraint in the folded state. In the unfolded state, the original linear protein restrains the linker because they must have the same end-to-end vector. The restraint is just the distribution function of the unfolded linear protein, which is modeled as a Gaussian chain. Then $P_u(\mathbf{r}) = p_G(\mathbf{r};N)$,where N is the number of peptide bonds in the linear protein. According to eq 5, the effect of backbone cyclization on the folding free energy is given by[23]

$$q_{cycl} = \exp(-\beta\delta G_{cycl}) = p(\mathbf{d})/\int d^3\mathbf{r}\, p(\mathbf{r})p_G(\mathbf{r};N) \quad (12)$$

The effect of circularizing the 34-residue PIN1 WW domain has been carefully studied recently by Deechongkit and Kelly.[16] The N and C termini, at a distance of ~10 Å, were connected by linkers with 1−7 residues (i.e., 2−8 peptide bonds). The experimental results for δG_{cycl} can be quantitatively reproduced by eq 12,[23] as shown in Figure 3B. In particular, a maximal stabilization of 1.7 kcal/mol by a linker with four peptide bonds is correctly predicted.

Greater stabilization might have been expected on the basis of consideration of the entropy reduction of the unfolded chain by the circularization. However, eq 12 makes it clear that both the folded and the unfolded state suffer chain entropy losses. In the unfolded state, both the original linear protein chain and the peptide linker are restrained since they must share the same end-to-end vector. This vector is still able to sample all possible values but with a distribution that is more restrictive (eq 4). On the other hand, in the folded state, only the peptide linker is restrained, but this restraint is rigid—the end-to-end vector is restricted to the fixed displacement between the N and C termini as determined by the folded structure of the linear protein. Entropy losses of the folded and unfolded states also differ in another respect, namely, with respect to the change in chain length of the linear protein. The loss is not affected by chain length in the folded state but becomes more severe in the unfolded state as chain length increases. Hence larger stabilization effects are predicted for circularizing longer protein chains.

Catenation

When the chains of a dimeric protein cross each other, backbone cyclization of the chains creates a catenane, consisting of two interlocked rings (Figure 1D). The folding equilibrium of a catenane is independent of protein concentration. If the original dimeric protein and the catenated variant have folding equilibrium constants K^d and K^c, respectively, the effective concentration for catenation is

$$C_{cat} \equiv K^c/K^d \quad (13)$$

Blankenship and Dawson[19] designed a catenated variant of a dimer mutant of the p53 tetramerization domain. In p53cat^{dim}, the N and C termini (at Glu326 and Lys357, respectively) are extended by four residues each and then ligated. The effective concentration for catenation was found to be 1.7 M.

Catenation introduces two effects: (a) backbone cyclization of the two chains and (b) interlocking of the two unfolded chains, which restrains the relative motion of the chains.[24] The first effect results in a stabilization factor q_{cycl} for each subunit (eq 12). If the effect of interlocking is described by an effective concentration C_{il}, then

$$C_{cat} \equiv q_{cycl,A}q_{cycl,B}C_{il} \quad (14)$$

where A and B in the subscripts refer to the two subunits. For p53cat^{dim}, eq 12 predicts $q_{cycl,A} = q_{cycl,B} = 2.13$. By modeling each unfolded circular chain as a rigid ring, C_{il} was estimated to be 0.12 M. Equation 14 then gives $C_{cat} = 2.13^2 \times 0.12\ M = 0.54\ M$, in reasonable agreement with the experimental result of 1.8 M.

The capsid of bacteriophage HK97 consists of 60 hexameric and 12 pentameric rings.[18] The rings are covalently ligated by isopeptide bonds formed by lysine and asparagine side chains and symmetry-related rings are interlocked. Like p53cat^{dim}, ligation occurs after the assembly of the subunits, and the interlocking provides stabilization.

Spatial Confinement

Spatial confinement brings out a different entropy-based approach to stabilization. Near a boundary of a confined space, many conformations otherwise sampled by an unfolded protein molecule will be disallowed because the protein chain cannot cross the boundary (Figures 1E and 2D). For an unfolded protein modeled as a sufficiently long Gaussian chain starting at $\mathbf{x}_0$, the probability density for residue n to be at $\mathbf{x}$ satisfies the diffusion equation[36]

$$\frac{\partial G(\mathbf{x},n|\mathbf{x}_0)}{\partial n} = \frac{b^2}{6}\nabla^2 G(\mathbf{x},n|\mathbf{x}_0) \quad (15)$$

where the residue number n plays the role of time and the effective bond length b determines the diffusion constant. The solution of eq 15 gives the Gaussian distribution of eq 1 with $\mathbf{r} = \mathbf{x} - \mathbf{x}_0$. Configurations of a Gaussian chain are thus equivalent to random-walk trajectories of a Brownian particle. The confined space is absorbing for $G(\mathbf{x}, n|\mathbf{x}_0)$. The fraction of allowed configurations is

$$f_u = \int d^3\mathbf{x}\, d^3\mathbf{x}_0\, G(\mathbf{x},N|\mathbf{x}_0)/V \quad (16)$$

where N is the total number of peptide bonds in the protein chain and V is now the volume of the confined space. The confined space also restricts the folded protein, which by its finite size can only access a fraction, f_f, of the volume V. Confinement thus changes the folding free energy by (eq 3)

$$\delta G_{conf} = -k_B T(\ln f_f - \ln f_u) \tag{17}$$

Calculations for simple confined geometries have been made to illustrate the significant effect of confinement on the folding equilibrium.[25] For a spherical cage with radius R, the allowed fraction of unfolded chain configurations is $f_u = (6/\pi^2)\sum_{k=0}^{\infty} k^{-2} \exp[-(\pi k R_g/R)^2]$, where $R_g = (N/6)^{1/2} b$ is the radius of gyration of the unfolded chain. If the folded protein is modeled as a sphere with radius a, then the fraction of accessible volume is $f_f = (1 - a/R)^3$. Figure 3C shows that as the spherical cage shrinks down to the size of the folded protein, stabilization of 15 kcal/mol or more is predicted.

The model presented above is highly simplified and only provides qualitative estimates on the effects of confinement.[37] Direct comparison with experiments is therefore not warranted. Nonetheless, large stabilization effects by confinement have been observed by Eggers and Valentine,[6] who encapsulated α-lactalbumin in the pores of silica glass and found the melting temperature to rise by 32 °C. The Anfinsen cage of chaperonins and compartments in cellular organelles may impart similar stabilization effects to proteins.

Macromolecular Crowding

The cytoplasm is crowded with proteins and other macromolecules. Macromolecular crowding may exert a stabilization effect similar to spatial confinement.[8] The model presented above for the entropy restriction of the unfolded chain by confinement has been extended to treat crowding.[26] Continuing the equivalence of a Gaussian chain and a Brownian particle, the problem of entropy restriction by crowding macromolecules can be mapped to the problem of trapping, with the macromolecules serving as traps. The fraction of allowed configurations, f_u, is equivalent to the survival probability of the Brownian particle:

$$f_u = S(t) \tag{18}$$

where it is understood that time t is equivalent to chain length N and the diffusion constant D of the Brownian particle is equivalent to $b^2/6$. At short times, the Smoluchowski theory gives[38,39]

$$-\ln S(t) = c\int_0^t d\tau\, k(\tau) = 4\pi D a_c c t[1 + 2a_c(\pi D t)^{-1/2}] \tag{19a}$$

where a_c is the radius of the crowding macromolecules, assumed to be spherical, and c is their concentration. Mapping into the crowding problem, we have

$$-\ln f_u = 3\phi y^2(1 + 2/\pi^{1/2} y) \tag{19b}$$

where $\phi = 4\pi a_c^3 c/3$ is the volume fraction of the crowders and $y = R_g/a_c$ is the radius of gyration of the unfolded chain scaled by a_c.

The crowders also restrict the folded protein because their excluded-volume effects block many attempts of its placement in the solution. The fraction of successful placement can be obtained from the scaled particle theory, giving[40]

$$-\ln f_f = -\ln(1 - \phi) + (3z + 3z^2 + z^3)\phi/(1 - \phi) + (9z^2/2 + 3z^3)[\phi/(1 - \phi)]^2 + 3z^3[\phi/(1 - \phi)]^3 \tag{20}$$

where $z = a/a_c$, the radius of the folded protein scaled by a_c. The effect of macromolecular crowding, calculated from the difference of eqs 20 and 19b, on the folding free energy, is shown in Figure 3D. Only a modest stabilization, <0.5 kcal/mol, is seen. Given the simplicity of the model, the numerical value of the predicted effect should not be taken literally. However, the disparity in predicted effects for confinement and crowding (see different scales of Figure 3 panels C and D) seems unmistakable, especially since they are obtained from essentially the same model.

Why are the predicted effects of confinement and crowding so different? In confinement, the cage fully encloses the unfolded chain. However, in crowding, there are always interstitial voids that allow the unfolded chain to escape (Figure 1F). At very high crowder concentrations, these voids may become too small in serving as routes of translocation for the compact folded protein. In this situation, crowding is like confinement in a cage with holes, which allow the unfolded chain to leak but cannot let out the folded protein. The folded state may even become disfavored, leading to the decrease of $-\delta G_{crowd}$ with crowder concentration in Figure 3D.

Experimentally, the effects of macromolecular crowding on protein folding stability have indeed been found to be modest in several studies.[7,9,10] For example, Sasahara et al.[10] recently studied the thermal unfolding of hen lysozyme in the presence of dextran, observing a mere 2.5 °C increase in melting temperature even at 300 g/L dextran. Given the qualitative predictions of the simple model, the disparity between these modest effects for crowding and the large effect observed for confinement by Eggers and Valentine[6] is perhaps not coincidental. Further understanding will come from detailed simulations on the effects of confinement and crowding.[37,41]

Connections among the Entropy-Based Stabilization Strategies

In the analyses of all the entropy-based stabilization strategies, it has been important to consider constraints imposed on both the folded and the unfolded states. For the effects of loop length, covalent linkage, backbone cyclization, and catenation, the folded state is treated in the same way, namely, by restricting the end-to-end vector of each loop or linker to a fixed displacement, leading to the appearance of $p(\mathbf{d})$ in eqs 8, 11, and 12. While $p(\mathbf{d})$ essentially accounts for the effects of loop length and covalent linkage, backbone cyclization and catenation

impose additional restraints on the unfolded state. A catenated protein is affected both by the backbone cyclization of its two subunits and by the restrained relative motion of the subunits within the interlocked topology. The second effect was estimated on the basis of rigid rings. A more realistic treatment would allow the chains to sample different configurations. Some of these configurations would violate the topological constraint and should be eliminated. In this sense the topological constraint serves a role similar to that of the boundary of a confined space or a crowding macromolecule.

The lengths of linkers used to create single-chain, circularized, and catenated variants, like loop length, can be optimized for maximal stabilization effects. The optimal linker length was indeed observed in the circularized PIN1 WW domain. Both covalent linkage and catenation change the folding process from bimolecular to unimolecular. However, the approaches are different. In covalent linkage, the linker restrains the relative motion of the two subunits, whereas in catenation the interlocking restrains the relative motion of the two subunits. The latter approach appears to be more restrictive to the unfolded state, leading to a higher effective concentration. Spatial confinement also owes its extraordinary stabilization effect to its ability to severely restrict the unfolded state.

Concluding Remarks

A protein molecule exists in either a compact folded state or a variable and open state, which is favored by chain entropy. Thus an attractive mechanism for increasing the folding stability is restricting the entropy of the unfolded state. The various entropy-based stabilization strategies have now been analyzed by use of polymer theory-based models. Tests of these theoretical models demonstrated their success but also exposed their limitations. In particular, unequivocal applications of these models require information on the conformations of residues making up loops and linkers. Some aspects of the models are clearly too simplistic. The entropy-based stabilization strategies all have wide applicability. The theoretical models can provide guidance for their further exploitation, which in turn will motivate further improvements of the models.

In the literature the term "topology" has generally referred to the arrangement of secondary structures in a protein. This misuse did not pose much problem because as a norm proteins are linear chains. However, now circular proteins and catenated proteins have presented important exceptions to this norm. It may be sensible to refer to the arrangement of secondary structures in a protein as its architecture and reserve the term "topology" for its proper use. In addition to rings and catenanes, other topological varieties (formed by backbone−backbone, backbone−side chain, and side chain−side chain ligations) are being discovered. The theoretical models developed so far may serve as a guide for considering the entropic consequences of these topological links.

I thank anonymous reviewers for several valuable comments. This work was supported in part by NIH Grant GM58187.

References

(1) Nagi, A. D.; Regan, L. An Inverse Correlation between Loop Length and Stability in a Four-helix-bundle Protein. *Folding Des.* **1997**, *2*, 67−75.
(2) Viguera, A.-R.; Serrano, L. Loop Length, Intramolecular Diffusion and Protein Folding. *Nat. Struct. Biol.* **1997**, *4*, 939−946.
(3) Ladurner, A. G.; Fersht, A. R. Glutamine, Alanine or Glycine Repeats Inserted into the Loop of a Protein Have Minimal Effects on Stability and Folding Rates. *J. Mol. Biol.* **1997**, *273*, 330−337.
(4) Grantcharova, V. P.; Riddle, D. S.; Baker, D. Long-range Order in the Src SH3 Folding Transition State. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 7084−7089.
(5) Wallon, G.; Kryger, G.; Lovett, S. T.; Oshima, T.; Ringe, D.; Petsko, G. A. Crystal Structures of *Escherichia coli* and *Salmonella typhimurium* 3-Isopropylmalate Dehydrogenase and Comparison with Their Thermophilic Counterpart from *Thermus thermophilus*. *J. Mol. Biol.* **1997**, *266*, 1016−1031.
(6) Eggers, D. K.; Valentine, J. S. Molecular Confinement Influences Protein Structure and Enhances Thermal Protein Stability. *Protein Sci.* **2001**, *10*, 250−261.
(7) van den Berg, B.; Ellis, R. J.; Dobson, C. M. Effects of Macromolecular Crowding on Protein Folding and Aggregation. *EMBO J.* **1999**, *18*, 6927−6933.
(8) Minton, A. P. Effect of a Concentrated "Inert" Macromolecular Cosolute on the Stability of a Globular Protein with Respect to Denaturation by Heat and by Chaotropes: a Statistical-thermodynamic Model. *Biophys. J.* **2000**, *78*, 101−109.
(9) Qu, Y.; Bolen, D. W. Efficacy of Macromolecular Crowding in Forcing Proteins to Fold. *Biophys. Chem.* **2002**, *101*−*102*, 155−165.
(10) Sasahara, K.; McPhie, P.; Minton, A. P. Effect of Dextran on Protein Stability and Conformation Attributed to Macromolecular Crowding. *J. Mol. Biol.* **2003**, *326*, 1227−1237.
(11) Liang, H.; Sandberg, W. S.; Terwilliger, T. C. Genetic Fusion of Subunits of a Dimeric Protein Substantially Enhances Its Stability and Rate of Folding. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 7010−7014.
(12) Robinson, C. R., & Sauer, R. T. Equilibrium Stability and Submillisecond Refolding of a Designed Single-chain Arc Repressor. *Biochemistry* **1996**, *35*, 13878−13884.
(13) Jana, R.; Hazbun, T. R.; Fields, J. D.; Mossing, M. C. Single-chain Lambda Cro Repressors Confirm High Intrinsic Dimer−DNA Affinity. *Biochemistry* **1998**, *37*, 6446−6455.
(14) Moran, L. B.; Schneider, J. P.; Kentsis, A.; Reddy, G. A.; Sosnick, T. R. Transition State Heterogeneity in GCN4 Coiled Coil Folding Studied by Using Multisite Mutations and Cross-linking. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 10699−10704.
(15) Goldenberg, D. P.; Creighton, T. E. Folding Pathway of a Circular Form of Bovine Pancreatic Trypsin Inhibitor. *J. Mol. Biol.* **1984**, *179*, 527−545.
(16) Deechongkit, S.; Kelly, J. W. The Effect of Backbone Cyclization on the Thermodynamics of β-sheet Unfolding: Stability Optimization of the PIN WW Domain. *J. Am. Chem. Soc.* **2002**, *124*, 4980−4986.
(17) Trabi, M.; Craik, D. J. Circular Proteins−No End in Sight. *Trends Biochem. Sci.* **2002**, *27*, 132−138.
(18) Wikoff, W. R.; Liljas, L.; Duda, R. L.; Tsuruta, H.; Hendrix, R. W.; Johnson, J. E. Topologically Linked Protein Rings in the Bacteriophage HK97 Capsid. *Science* **2000**, *289*, 2129−2133.
(19) Blankenship, J. W.; Dawson, P. E. Thermodynamics of a Designed Protein Catenane. *J. Mol. Biol.* **2003**, *327*, 537−548.
(20) Zhou, H.-X. Loops in Proteins Can Be Modeled as Wormlike Chains. *J. Phys. Chem. B* **2001**, *105*, 6763−6766.
(21) Zhou, H.-X. Single-chain versus Dimeric Protein Folding: Thermodynamic and Kinetic Consequences of Covalent Linkage. *J. Am. Chem. Soc.* **2001**, *123*, 6730−6731.
(22) Zhou, H.-X. The Affinity-enhancing Roles of Flexible Linkers in Two-domain DNA-binding Proteins. *Biochemistry* **2001**, *40*, 15069−15073.
(23) Zhou, H.-X. Effect of Backbone Cyclization on Protein Folding Stability: Chain Entropies of Both the Unfolded and the Folded States Are Restricted. *J. Mol. Biol.* **2003**, *332*, 257−264.
(24) Zhou, H.-X. Effect of Catenation on Protein Folding Stability. *J. Am. Chem. Soc.* **2003**, *125*, 9280−9281.
(25) Zhou, H.-X.; Dill, K. A. Stabilization of Proteins in Confined Spaces. *Biochemistry* **2001**, *40*, 11289−11293.
(26) Zhou, H.-X. Protein Folding and Binding in Confined Spaces and Crowded Solutions. *J. Mol. Regul.* (submitted for publication).
(27) There may actually be an enthalpic contribution if the distribution function $p(\mathbf{r})$ is temperature-dependent.
(28) Jacobson, H.; Stockmayer, W. H. Intramolecular Reaction in Polycondensations. I. The Theory of Linear Systems. *J. Chem. Phys.* **1950**, *18*, 1600−1606.

(29) Gobush. W.; Stockmayer, W. H.; Yamakawa, H.; Magee, W. S. Statistical Mechanics of Wormlike Chains. 1. Asymptotic Behavior. *J. Chem. Phys.* **1972**, *57*, 2839−2843.
(30) Riddle, D. S.; Santiago, J. V.; Bray-Hall, S. T.; Doshi, N.; Grantcharova, V. P.; Yi, Q.; Baker, D. Functional Rapidly Folding Proteins from Simplified Amino Acid Sequences. *Nat. Struct. Biol.* **1997**, *4*, 805−809.
(31) Ladurner, A. G.; Itzhaki, L. S.; de Prat Gay, G.; Fersht, A. R. Complementation of Peptide Fragments of the Single Domain Protein Chymotrypsin Inhibitor 2. *J. Mol. Biol.* **1997**, *273*, 317−329.
(32) Itzhaki, L. S.; Otzen, D. E.; Fersht, A. R. The Structure of the Transition State for Folding of Chymotrypsin Inhibitor 2 Analysed by Protein Engineering Methods: Evidence for a Nucleation-condensation Mechanism for Protein Folding. *J. Mol. Biol.* **1995**, *254*, 260−288.
(33) Tamura, A.; Privalov, P. L. The Entropy Cost of Protein Association. *J. Mol. Biol.* **1997**, *273*, 1048−1060.
(34) Karplus, M.; Janin, J. Comment on "The Entropy Cost of Protein Association." *Protein Eng.* **1999**, *12*, 185−186.
(35) Privalov, P. L.; Tamura, A. Comments on the Comments. *Protein Eng.* **1999**, *12*, 187.
(36) Doi, M.; Edwards, S. F. *The Theory of Polymer Dynamics*; Clarendon Press: Oxford, U.K., 1986.
(37) Recently the effects of confinement have been studied by several detailed simulations (Klimov, D. K.; Newfield, D.; Thirumalai, D. Simulations of β-hairpin Folding Confined to Spherical Pores Using Distributed Computing. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 8019−8024, and Takagi, F.; Koga, N.; Takada, S. How Protein Thermodynamics and Folding Mechanisms Are Altered by the Chaperonin Cage: Molecular Simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 11367−11372). Results of these studies were qualitatively similar to the prediction of eqs 16 and 17.
(38) Smoluchowski, M. V. Versuch einer mathematischen Theorie der Koagulationskinetik kolloider Loeschungen. *Z. Phys. Chem.* **1917**, *92*, 129−168.
(39) Szabo, A.; Zwanzig, R.; Agmon, N. Diffusion-controlled Reactions with Mobile Traps. *Phys. Rev Lett.* **1988**, *61*, 2496−2499.
(40) Lebowitz, J. L.; Rowlinson, J. S. Thermodynamic Properties of Mixtures of Hard Spheres. *J. Chem. Phys.* **1964**, *41*, 133−138.
(41) Elcock, A. H. Atomic-level Observation of Macromolecular Crowding Effects: Escape of a Protein from the GroEL Cage. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 2340−2344.

AR0302282