Single-Chain versus Dimeric Protein Folding: Thermodynamic and Kinetic Consequences of Covalent Linkage

Huan-Xiang Zhou

Department of Physics, Drexel University Philadelphia, Pennsylvania 19104

Received April 10, 2001
Revised Manuscript Received June 4, 2001

The construction of a single-chain protein by linking the C terminal of one subunit with the N terminal of another in an otherwise dimeric protein has been considered as a strategy for increasing protein stability.1,2 In this Communication we investigate the role of the covalent linkage in the stability of the folded protein and the implication for the folding kinetics.

The class of dimeric proteins under consideration have hydrophobic cores formed by side chains from both subunits, such that they become unfolded when the subunits dissociate. In addition, the covalent linkage is assumed to be flexible and not form specific contacts with the rest of the single-chain protein. In this class there are a number of well-characterized proteins, including the gene V dimer of bacteriophage T4,1 the Arc repressor of bacteriophage P22,2 and the coiled-coil region of the yeast transcription factor GCN4.3

Our primary goal is to relate the folding stability of the dimeric protein to that of the single-chain version (see Figure 1). Let the two subunits be denoted as A and B. If the partition functions of subunit α in the unfolded and folded states are $u_A$ and $u_B$, respectively then the equilibrium constant for the dimeric protein is:

$$K^d = f_{d}u_{d} / u_{A}u_{B}$$

where $f_{d}$ arises from the interaction between the two subunits in the folded state. In the simplest case where the two subunits are modeled as spheres interacting via a central symmetric potential $U(r)$, one has $f_{d} = \exp[-U(r)/k_B T] \pi r^2 d r$, where $k_B$ is Boltzmann’s constant and $T$ is the absolute temperature. Note that in defining $f_{d}$ one assumes that the subunit has the structure in the folded state (which by itself, i.e., without the other subunit, may not be stable).

To derive a simple expression for the equilibrium constant for the single-chain version, we assume that the linkage is structureless and modeled as a polymer chain. In the unfolded state the end-to-end distance of the linkage is able to sample all possible values. On the other hand, in the folded state $d$ is restricted to small fluctuations around a fixed value $d_0$. Let the probability density for the end-to-end vector be $p(d_0)$ when $d = d_0$, then the equilibrium constant for the single-chain protein is:

$$K^c = f_{c}u_{c}p(d_0) / u_{A}u_{B}$$

In writing eq 2 we assume that (1) for the unfolded chain the partition function is the product of three factors, $u_{A}$, $u_{B}$, and a corresponding quantity for the linkage (2) in the folded state the linkage behaves the same way as it does in the unfolded state.

![Figure 1. The folding of the dimeric protein (with equilibrium constant $K^d$) and the single-chain version (with equilibrium constant $K^c$). The dotted curve represents the covalent linkage.](image)

except that its end-to-end distance is fixed at $d_0$. In particular the latter assumption means that the interaction between the linkage and the rest of the protein is negligible. Comparison of eqs 1 and 2 leads to:

$$K^d / K^c = p(d_0)$$

Recently Zhou3 found that unstructured loops in proteins can be modeled very well by the worm-like chain4 with a persistence length $l_p = 3$ Å. For this polymer model one has:

$$p(d) = (3/4\pi l_p^4)^{3/2} \exp(-3d^2/4l_p^4)(1 - 5l_p^2/4l_p + 2d^2/l_p^2 - 33d^4/160l_p^6)^{1/2}$$

$$- 37/160l_p^2 - 329d^2/120l_p^4 + 6799d^4/1600l_p^6 - 344d^6/2800l_p^8 + 1089d^8/12800l_p^{10}$$

where $l_p = l_p$ with $L$ the number of residues forming the linkage, and $b = 3.8$ Å, the nearest $C_n-C_n$ distance.

The $K^d / K^c$ ratio calculated from eqs 3 and 4 for the Arc repressor, the gene V protein, and GCN-p1 is shown in Table 1. These results agree with the experimental values to within a factor of 2. Considering that the calculated results do not involve any adjustable parameters and the values of $K^d / K^c$ span 2 orders of magnitude, the agreement is very satisfactory.

We also examined the prediction of eq 3 on two cases in which some of the requirements for the use of eq 3 are not met. Streptomyces subtilisin inhibitor (SSI; pdb code 2sic) is folded only as a homodimer. Tamura and Privalov8 constructed a single-chain version by mutating Asp83, located within the dimer interface, to Cys and cross-linking with a disulfide bond. The link is not between the C terminal of one subunit and the N terminal of the second and the Asp to Cys mutation itself may affect the stability of the folded state. Nonetheless, the fact remains that the $C_n-C_n$ distances of the two Cys83 residues is restricted to around $d_0 = 6.6$ Å in the folded state and will have a much wider range in the unfolded state. Using the distribution of the $C_n-C_n$ distances of disulfide bonded Cys in proteins as a model for $p(d_0)$, we estimated $p(d_0) \approx 1.7$ M. Tamura and Privalov

---

Table 1. Experimental and Calculated Ratio of $K_d$ and $K^d$

<table>
<thead>
<tr>
<th>protein$^a$</th>
<th>condition</th>
<th>$K^d$(M$^{-1}$)</th>
<th>$K^d$</th>
<th>linkage$^b$</th>
<th>$K^d/K^u$(mM)</th>
<th>$p(d_u)$ (mM)$^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arc repressor (1myk)</td>
<td>4.19 M urea; T = 298 K</td>
<td>10$^4$</td>
<td>2.44</td>
<td>A50–B7;</td>
<td>L = 25; $d_o = 29.9$ Å</td>
<td>2.44</td>
</tr>
<tr>
<td>Gene V protein (1gvp)</td>
<td>2.6 M Gdn-HCl; T = 298 K</td>
<td>10$^4$</td>
<td>1.51 $\times$ 10$^4$</td>
<td>A86–B1; L = 8; $d_o = 12.4$ Å</td>
<td>151</td>
<td>74</td>
</tr>
<tr>
<td>GCN4-p1 (2zaa)</td>
<td>4 M Gdn-HCl; T = 283 K</td>
<td>353</td>
<td>85</td>
<td>A2–B2; L = 9; $d_o = 6.3$ Å</td>
<td>241</td>
<td>129</td>
</tr>
</tbody>
</table>

$^a$ Entry in parentheses is the PDB code. $^b$ Ai–Bj refers to the two end residues of the covalent linkage (e.g., residue 50 of subunit A). These were typically selected as the last unstructured residue of subunit A and the first unstructured residue of subunit B. Additional linking residues were introduced in the experiments, giving rise to the listed total residue L for the covalent linkage. The end-to-end distance $d_o$ was calculated between residues Ai and Bj from the PDB structure. $^c$ $p(d_u)$ has the unit of density (i.e., number of ends per unit volume). It is converted to the unit of molarity by dividing by the Avogadro number.

found $K^d = 2.4 \times 10^4$ M$^{-1}$ and $K^u = 4.3 \times 10^3$ at $T = 353$ K, thus $K^d/K^u = 18$ M. While the value of $p(d_o)$ estimated for the present case is about an order of magnitude higher than those for the gene V protein and GCN4-p1, it is still an order of magnitude lower than the experimental value of $K^d/K^u$. The additional stabilization of $\sim 1.6$ kcal/mol may be attributed in part to the elimination of the repulsion between the two negatively charged Asp83 residues in the disulfide bonded mutant.

In the second case we compare the folding of the intact chymotrypsin inhibitor 2 (CI2) and the association/folding of a dimeric version in which the peptide bond between Met40 and Glu41 is broken. A main difference with the proteins listed in Table 1 is that the linkage in CI2, taken to be the reactive site Asp83 residues in the disulfide bonded mutant.

For the proteins listed in Table 1, the specific interactions maintaining the folded state are identical in the dimeric and single-chain versions. The two respective unfolding rate constants $k_d$ and $k_u$, dictated by the breaking of some of these specific interactions, are thus expected to be not too different. This is indeed the case. Specifically, for the Arc repressor $k_u = 18$ s$^{-1}$ and $k_d = 25$ s$^{-1}$ at 7 M urea, for the gene V protein $k_u = 3.5 \times 10^{-3}$ s$^{-1}$ and $k_d = 2.5 \times 10^{-3}$ s$^{-1}$ at 5 M Gdn-HCl, and for GCN4-p1 $k_u = 1.9$ s$^{-1}$ and $k_d = 0.4$ s$^{-1}$ at 4 M Gdn-HCl. A hallmark for the covalent linkage serving as just a tether (as opposed to an active participant in stabilizing the folded structure) should be the tracking of $k_d$ by $k_u$. More recent variants of the covalent linkage for the Arc repressor designed by Robinson and Sauer led to substantial slowing down of the unfolding process (with $k_d = 0.7$ s$^{-1}$, i.e., one 26th of $k_u$, at 7 M urea). This slowing down indicates an active role for the covalent linkage (and consequently the breakdown of eq 3).

As a consequence of the result $k_u \approx k_d$, the folding rate constants $k_d$ and $k_u$ of the dimeric and single-chain proteins can be related via $k_d/k_u = p(d_u)$. This allows us to estimate the limit of the folding rate of a single-chain protein. The folding rate constant of the dimeric protein is limited by the diffusional approach of the two subunits. When the folded dimer is stereospecific and long-range electrostatic interactions are absent, the diffusion-limited rate constant is $~10^6$ M$^{-1}$ s$^{-1}$. If $p(d_u)$ is of the order of magnitude of 0.1 M (see Table 1), then $k_d \leq 10^5$ s$^{-1}$. This estimate complements one previously proposed by Eaton et al. Recent experimental results on the folding of two small proteins approach this limit.

In summary, we explored the connection between the stability of dimeric proteins and that of their single-chain versions. We restricted ourselves to the simplest case where the covalent linkage acts merely as a tether. In this case we can relate the ratio of the two respective equilibrium constants to the probability density for the end-to-end distance of the covalent linkage to be at the value in the folded structure. We further argue that the unfolding rates of the dimer and the single-chain protein should be similar. On the basis of these results we estimate a rate limit of $10^5$ s$^{-1}$ for the folding of single-chain proteins.

Acknowledgment. This work was supported in part by the National Institutes of Health (GM-58187).