Loops in Proteins Can Be Modeled as Worm-Like Chains

Huan-Xiang Zhou*

Department of Physics, Drexel University, Philadelphia, Pennsylvania 19104

Received: April 11, 2001; In Final Form: June 1, 2001

The end−end distances of loops in proteins are found to be distributed according to the worm-like chain model with a persistence length $l_p = 3.04$ Å. For a protein with a loop at a certain short end−end distance, increasing the loop length is expected to decrease the protein stability since the entropic cost increases for constraining the loop ends at the given distance. The predicted decrease in stability is tested against experimental results on the four-helix-bundle protein Rop, in which the native two-residue loop is replaced by two to ten glycinces. Without adjustable parameters, the prediction agrees with experiment with a correlation coefficient of 0.99.

A protein structure can be viewed as a stable packing of secondary structure elements (SSEs) (i.e., α-helices and β-strands), with loops providing the necessary links. Loops have no specific conformations and thus have considerably fewer favorable interactions with the rest of the protein. The lengths of the loops are an important factor in protein stability and may play a critical role in the folding kinetics. In this letter, we show that the distribution of end−end distances of loops follows the worm-like chain (WLC) model,[1−3] and the predicted entropic cost for constraining a loop with various lengths at a given distance agrees with experimental results.[4]

We collected a total of 25 975 loops from the 1907 proteins in the FSSP library,[5] which is a nonredundant representation of the Protein Data Bank.[6] A loop consists of all the residues between two regular SSEs (as defined by DSSP[7]). If the last position of one SSE is residue number i and the first position of the next SSE is residue number j, the length of the loop is $L = j - i$. The total number $N(L)$ of L-residue loops is shown as a function of L in Figure 1. It can be very well fitted to an exponentially decreasing function $f(L) = 3469.2 \exp[-(L - 5)/4.488]$ for $L \geq 5$.

The end−end distances of all L-residue loops were pooled to calculate $G(r|L)$, the probability density for the end−end distance at r. For a WLC with a persistence length l_p and contour length l_c, the probability density for $l_c/l_p > \sim 10$ is accurately given by[2,3]

$$G_0(r|l_c) = 4\pi r^2 (3/4\pi l_p l_c)^{3/2} \exp(-3r^2/4l_p l_c)(1 - 5l_p/4l_c + 2r^2/l_c^2 - 33r^4/80l_p l_c^3 - 79l_p^2/160l_c^2 - 329r^2 l_p/120l_c^3 + 6799r^4/1600l_c^4 - 3441r^6/2800l_p l_c^5 + 1089r^8/12800l_p^2 l_c^6) \quad (1)$$

For protein loops, $l_c = Lb$ where $b = 3.8$ Å is the $C_\alpha - C_\alpha$ distance. We find that for $L \geq 7$, $G(r|L)$ agrees very well with

$$G(r|L) = g(r) G_0(r|Lb) \quad (2)$$

where $g(r)$ is the radial distribution function of a hard-sphere liquid with a diameter of $\sigma = 4.56$ Å and a reduced density of $\rho\sigma^3 = 0.83$ (to be discussed later). In Figure 2, we compare with eq 2 the distribution of the end−end distances of loops from the proteins for $L = 7$, 10, and 13. The same persistence length $l_p = 3.04$ Å can be used to reproduce the distributions of end−end distances at all the loop lengths. In particular $l_c/l_p = 8.75$ at $L = 7$.

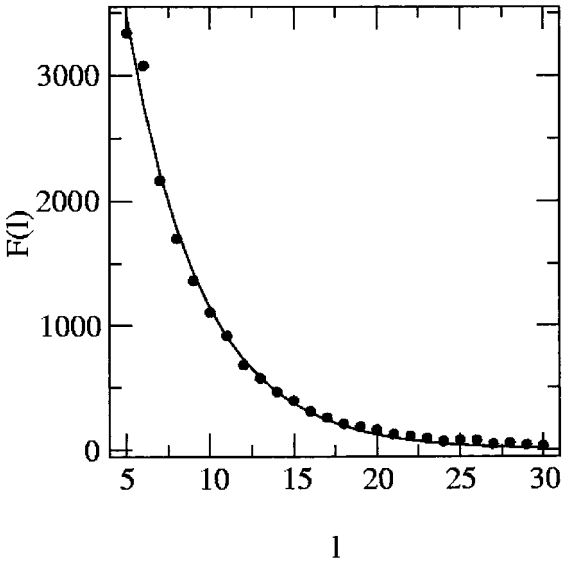

Figure 1. Total number of loops with a given length (circles) fitted to an exponential function (line).

When $l_c/l_p < \sim 2$ the probability density of the end−end distance for a WLC can be calculated from[3]

$$G_0(r|l_c) = Kr^2[1/l_p(l_c - r)]^{3/2} \sum_{k=1}^{\infty} H[(2k-1)^2 l_c^2/4l_p(l_c - r)] \quad (3)$$

where $H(x) = (4x - 2)\exp(-x)$ and K is a normalization constant. For $L = 5$ and 6, one has $l_c/l_p = 6.25$ and 7.5, respectively, which unfortunately fall in a range where neither eq 1 nor eq 3 is very accurate. We find empirically that the average of the two equations provides a reasonable approximation for $G_0(r|l_c)$. This average function when used in eq 2 well reproduces the distributions of end−end distances for $L = 5$ and 6.

* Tel.: (215) 895-2716; fax: (215) 895-5934; e-mail: hxzhou@einstein.drexel.edu.

10.1021/jp011355n CCC: $20.00
Published on Web 06/26/2001

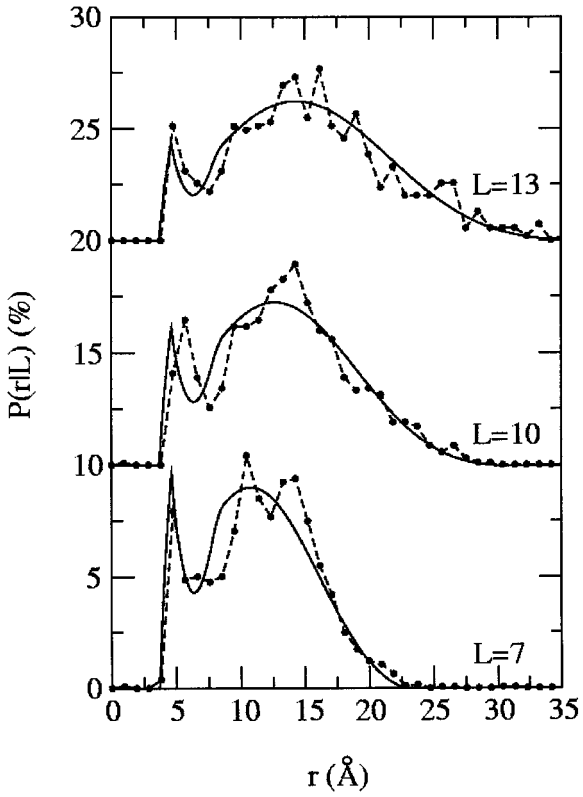

Figure 2. Probability density for the end−end distance at three loop lengths. The circles connected by dashed lines are calculated from loops in proteins and the solid lines are predictions of the WLC model (eq 2).

In Figure 3a we compare the distribution $p(r|L \geq 5)$ of the end−end distances of the 17 866 loops with $L \geq 5$ to

$$p_0(r|L \geq 5) = \sum_{L \geq 5} f(L) G_0(r|Lb) \tag{4}$$

When $r > 10$ Å, $p(r|L \geq 5)$ is very well reproduced by $p_0(r|L \geq 5)$. At $r = \sigma = 4.56$ Å, $p(r|L \geq 5)$ clearly shows a peak, which is absent in $p_0(r|L \geq 5)$. The ratio $p(r|L \geq 5)/p_0(r|L \geq 5)$ is shown in Figure 3b. The peak at $r = \sigma$ followed by a valley around $r = 1.5\sigma$ is reminiscent of the radial distribution function of a hard-sphere liquid, explaining the appearance of $g(r)$ in eq 2. As in a hard-sphere liquid, a packing configuration of a protein in which two C_α atoms are at a distance of 1.5σ is entropically unfavorable because the extra space between the C_α atoms cannot be filled. The finite size of residues can thus simply be accounted for by $g(r)$. With $\sigma = 4.56$ Å, a reduced density $\rho\sigma^3 = 0.83$ corresponds to a volume of $v_r = 114$ Å^3 for individual residues in proteins. This value is consistent with volumes of proteins calculated from their X-ray coordinates.[8]

An alternative way of assessing the applicability of the WLC model is to compare the distribution $p(L|r)$ for loop lengths at a given end−end distance. This is given by

$$\begin{aligned} p(L|r) &= f(L)G(r|Lb)/p(r|L \geq 5) \\ &= f(L)G_0(r|Lb)/p_0(r|L \geq 5) \end{aligned} \tag{5}$$

On going to the second step, the radial distribution function $g(r)$ is removed from both the denominator and the numerator, thus comparison of the distribution in loop lengths to eq 5 is not complicated by $g(r)$. Figure 4 shows the comparison at $r =$ 9.5, 13.5, and 17.5 Å. Note that a peak in the length distribution appears at $r = 17.5$ Å, reflecting the fact that both the ring closure conformation and the fully extended conformation are not probable for a loop. At $r = 9.5$ and 13.5 Å, the peak would occur below $L = 5$.

We now switch our focus to an individual loop in a protein with a end−end distance constrained at $r = r_0$ (by the rest of the protein). In the unfolded state, such a constraint disappears and the end−end distribution is expected to follow eq 2 [perhaps with a radial distribution function $g_u(r)$ appropriate for the unfolded state]. Relative to the unfolded state, the probability for observing the two end residues at a distance of r_0 is

$$P(L,r_0) = G(r_0|L)\Delta V/4\pi r_0^{\,2} = [\Delta V g(r_0)/4\pi r_0^{\,2}] G_0(r_0|Lb) \tag{6}$$

where ΔV is the volume in which one end residue fluctuates relative to the other end residue in the folded state. The free-

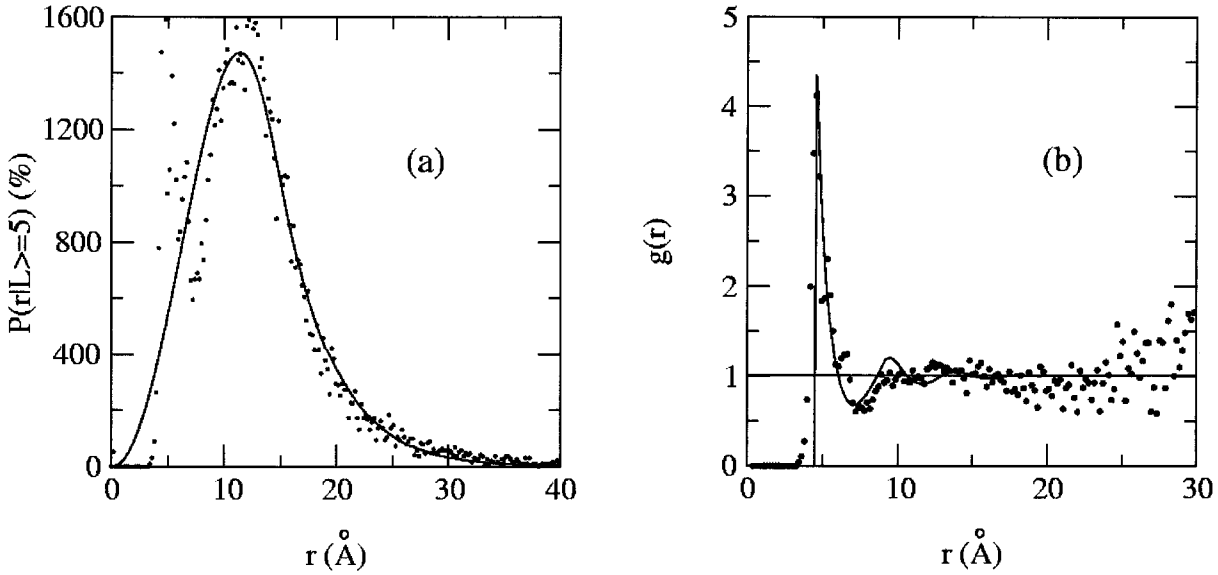

Figure 3. (a) Distribution $p(r|L \geq 5)$ of the end−end distances for all loops with five or more residues (scattered dots) and the prediction $p_0(r|L \geq 5)$ of the WLC model (line; eq 4). (b) Ratio $p(r|L \geq 5)/p_0(r|L \geq 5)$ (scattered dots) compared to the radial distribution function $g(r)$ of a hard-sphere liquid (line).

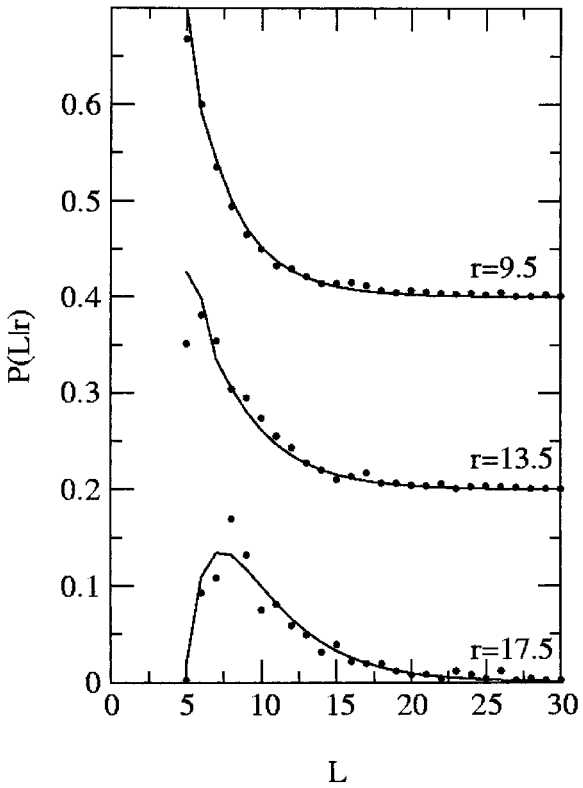

Figure 4. Distribution of the loop length at three given end−end distances. The circles are from loops in proteins and the solid lines are predictions of the WLC model (eq 5).

energy change due to the reduced entropy is thus

$$\Delta G_{\text{loop}}(L) = -k_{\text{B}}T\ln\{[\Delta V g(r_0)/4\pi r_0^2]G_0(r_0|Lb)\} \quad (7)$$

where k_{B} is the Boltzmann constant and T is the absolute temperature. If eq 1 is used for $G_0(r|L_{\text{B}})$, we have

$$\Delta G_{\text{loop}}(L)/k_{\text{B}}T = (3/2)\ln L + 3r_0^2/4bl_{\text{p}}L - \ln(1 - 5l_{\text{p}}/4bL + 2r_0^2/bL^2 - 33r_0^4/80l_{\text{p}}b^3L^3 - 79l_{\text{p}}^2/160b^2L^2 - 329r_0^2l_{\text{p}}/120b^3L^3 + 6799r_0^4/1600b^4L^4 - 3441r_0^6/2800l_{\text{p}}b^5L^5 + 1089r_0^8/12800l_{\text{p}}^2b^6L^6) + C \quad (8)$$

where C is the remaining term independent of the loop length L. If $Lb \gg r_0$ and l_{p}, then all L-dependent terms, except for the first one on the right-hand side, become negligible and one obtains the classical result $(3/2)\ln L + C$.[9,10] In general, a ΔG_{loop} will depend on the constrained value r_0 of the end−end distance.

Equation 8 can be used directly to analyze the experimental results of Nagi and Regan[4] on the four-helix-bundle protein Rop, in which the native two-residue loop was replaced by two to ten glycinces. The end−end distance (between Asp30 and Asp32) for this loop is $r_0 = 5.61$ Å (calculated from PDB entry 1rop[11]). Since Rop is a homodimer, in comparing eq 8 with experiment, the theoretical value should be multiplied by two.[12] In Figure 5, we compare theoretical and experimental results for $\Delta\Delta G = -\Delta G_{\text{loop}}(L) + \Delta G_{\text{loop}}(L{=}10)$. A linear regression analysis gives a correlation coefficient of 0.99.

Nagi and Regan fitted their data to a function $\Delta\Delta G/2k_{\text{B}}T = c\ln(10/L)$ and found $c = 1.2$ gave the best fit (the factor of 2, arising from Rop being a homodimer, was overlooked in their work). This coefficient is smaller than the classical value of 3/2. It should be noted that studies on lattice models have indicated that considerations of excluded volume effects and

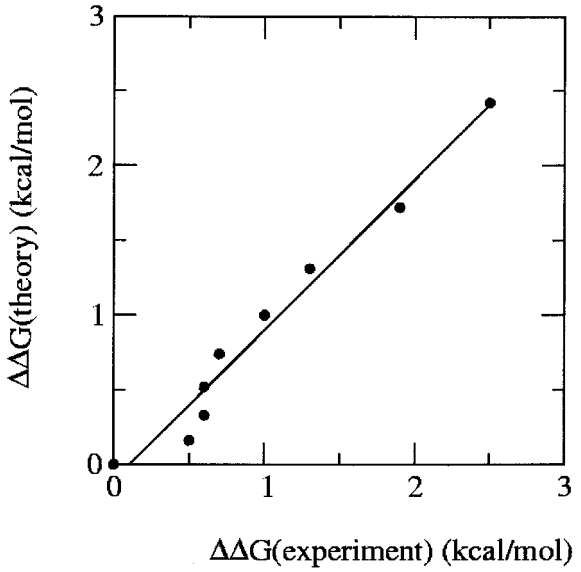

Figure 5. Comparison of experimental[4] and theoretical (eq 8) results for the free energy cost of loop formation. The line, representing the equation $y = -0.106 + 1.006x$, is from a linear regression analysis (correlation coefficient 0.99).

end effects will increase the coefficient c from the classical value of 3/2.[13] The smaller value of c required to fit the experimental data can be easily explained by eq 8. While the first term (the classical result) is a increasing function of L, the second term is a decreasing function and effectively reduces the value of c.

The value of $l_{\text{p}} = 3.04$ Å for the persistence length of loops is to be compared with persistence lengths of 4 to 8 Å for unfolded proteins found in recent single-molecule experiments.[14−17] It is well known that in long polymer chains the excluded volume effect tends to swell the chain and thus increases the effective persistence length.[2,18] Thus, the persistence length of loops found here is consistent with those found for unfolded proteins.[19]

It is remarkable that loops (and, by extension, peptides) of only a few residues long can be accurately modeled as WLCs. This will make it very convenient to analyze other thermodynamic and kinetic data on loop formation[20,21] and gain further information (such as conformational and dynamic parameters involved) on these processes.

References and Notes

(1) Kratky, O.; Porod, G. *Rec. Trav. Chim.* **1949**, *68*, 1106.

(2) Gobush, W.; Yamawaka, H.; Stockmayer, W. H.; Magee, W. S. *J. Chem. Phys.* **1972**, *57*, 2839. Yamawaka, H.; Stockmayer, W. H. *J. Chem. Phys.* **1972**, *57*, 2843.

(3) Wilhelm, J.; Frey, E. *Phys. Rev. Lett.* **1996**, *77*, 2581.

(4) Nagi, A. D.; Regan, L. *Folding Design* **1997**, *2*, 67.

(5) Holm, L.; Sander, C. *Nucl. Acids Res.* **1997**, *25*, 3389.

(6) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, N.; Bourne, P. E. *Nucl. Acids Res.* **2000**, *28*, 235.

(7) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577.

(8) Zhou, H.-X. *Biophys. J.* **1995**, *69*, 2298.

(9) Jacobson, H.; Stockmayer, W. H. *J. Chem. Phys.* **1950**, *18*, 1600.

(10) Pace, C. N.; Grimsley, G. R.; Thomson, J. A.; Barnett, B. J. *J. Biol. Chem.* **1988**, *263*, 11820.

(11) Banner, W.; Kokkinidis, M.; Tsernoglou, D. *J. Mol. Biol.* **1987**, *196*, 657.

(12) Upon folding, each subunit of the homodimer suffers an increase of $\Delta G_{\text{loop}}(L)$ in free energy due to reduced loop entropy. Thus the homodimer as a whole suffers an increase of $2\Delta G_{\text{loop}}(L)$. Note that in the experiment of Nagi and Regan,[4] what is measured is the free-energy difference between the homodimer and the two unfolded subunits.

(13) Chan H. S.; Dill, K. A. *J. Chem. Phys.* **1989**, *90*, 492.
(14) Rief, M. M.; Gautel, F.; Oesterhelt, F.; Fernandez, J. M.; Gaub, H. E. *Science* **1997**, *276*, 1109.
(15) Oberhauser, A. F.; Marszalek, P. E.; Erickson, H. P.; Fernandez, J. M. *Nature* **1998**, *393*, 181.
(16) Rief, M.; Pascual, J.; Saraste, M.; Gaub, H. E. *J. Mol. Biol.* **1999**, *286*, 553.
(17) Yang, G.; Cecconi, C.; Baase, W. A.; Vetter, I. R.; Breyer, W. A.; Haack, J. A.; Matthews, B. W.; Dahlquist, F. W.; Bustamante, C. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 139.
(18) Doi, M.; Edwards, S. F. *The Theory of Polymer Dynamics*; Clarendon Press: Oxford, 1986.
(19) The secondary structure elements (to which the two ends of the loop are attached) may have some influence on the persistence length. It is thus possible that the persistence length of 3.04 Å obtained on the loops collected from the protein structures may indeed be somewhat smaller than what would be obtained if the proteins are unfolded.
(20) Ladurner, A. G.; Fersht, A. R. *J. Mol. Biol.* **1997**, *273*, 330.
(21) Lapidus, L. J.; Eaton, W. A.; Hofrichter, J. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 7220.