Time Scales for the Formation of the Most Probable Tertiary Contacts in Proteins with Applications to Cytochrome *c*

D. Thirumalai

Department of Chemistry and Biochemistry and Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742

Received: May 26, 1998; In Final Form: October 9, 1998

A simple theory (practically without any adjustable parameter) for calculating the diffusion-limited rates of forming loops of arbitrary length in polypeptide chains is given. The time needed to form a tertiary contact between met80 and his18 in Cytochrome *c* is 40 μs, which is in excellent agreement with experimental determinations.

Introduction

A plausible elementary early event in the acquisition of the three-dimensional topology of proteins is a contact between two residues that are distant in sequence. If we can estimate the rates of formation of tertiary contacts, then one can get an idea on how fast a protein can fold starting from an ensemble of denatured states. Recently, Eaton and co-workers[1−3] have attempted to answer this question using a combination of experimental methods and some theoretical arguments. The result of this work suggests that the fastest time in which a single-domain globular protein can fold is 1 μs. This estimate is made on the premise that a protein cannot fold any faster than the rate at which a single most probable tertiary contact between two nonbonded residues can form. If, we denote the time scale for formation of the most probable contact as $\tau_{\rm MP}$, then it follows that the fastest a protein can fold $\tau_{\rm F} \approx \tau_{\rm MP}$, where $\tau_{\rm F}$ is the folding time. Thus, to estimate $\tau_{\rm F}$ it suffices to directly measure or infer an approximate value for $\tau_{\rm MP}$, which apparently is possible.

It has not been possible to measure $\tau_{\rm MP}$ directly. Related experiments that monitor the kinetics of coil to globule transition in homopolymers have been extremely difficult to perform.[4] Therefore, it is not surprising that measurements of rates of contact formation in proteins have been rare. Since, in a protein each amino acid residue acts as a marker, it may be possible to take advantage of specific residues with characteristic spectroscopic properties. Hagen, Hofrichter, and Eaton (HHE)[2] used time-resolved nanosecond absorption spectroscopy to decipher (indirectly) that the time for forming a contact between met80 and his18 in Cytochrome *c* is about 40 μs (1 μs = 10^{-6} s). Statistical mechanics of stiff chains shows that a loop with the largest probability of forming has $n \approx 10$, where n is the number of residues in the loop.[5] If the loop formation probability is used to scale the time constant for a contact between met80 and his18 ($n = 62$) to $n = 10$, one obtains that $\tau_{\rm MP} \approx 1$ μs. This led HHE to conclude that a generic single domain cannot fold on a time scale faster than 1 μs.

The purpose of this paper is to show that $\tau_{\rm MP}$ (and hence τ_F according to HHE arguments) can be computed accurately and simply without using experimental parameters. Such simple theories are useful because there are subtleties in formulating the problem.[6] Our theory uses simple dimensional analysis to estimate rates of forming loops of various sizes. In the process, we correct an earlier misleading application of this theory in ref 2 to estimate the time constants for forming a contact between met80 and his18 in Cytochrome *c*.

Theory

In general, the time for forming a loop of length n (measured along the contour of the chain) by a diffusion-limited process can be estimated as[7]

$$\tau(n) \approx \frac{\langle R_n^2 \rangle}{P(n) D_0} \qquad (1)$$

where $\langle R_n{}^2 \rangle$ is the mean distance between the two residues that are separated by n monomer units, $P(n)$ is the loop formation probability,[5] and D_0 is an effective monomer diffusion constant. The result in eq 1 follows because in the simplest case the time scale for forming a loop (solely by diffusion process alone) is set by the mean-squared distance between the residues and the mutual diffusion constant. The denominator in the above equation can be interpreted as the mutual diffusion constant that clearly depends on n.

If we assume, to a first approximation, that the polypeptide chain in the denatured state is semiflexible with a moderate persistence length $a \approx bn_{\rm per}$ with b as the average distance between α-carbon atoms, one can show that $P(n)$ for $n > n_{\rm per}$ should have the following structure[5]

$$P(n) = \frac{P_0}{n^{\theta_3}}(1 - \exp(-n/n_{\rm per})) \qquad (2)$$

where the exponent θ_3 in three dimensions is 2.2 and P_0 is an appropriate normalization constant. The value of $n_{\rm per}$ is dependent on chain stiffness, which is a function of the solvent conditions. The loop-formation probability can be derived using a model for semiflexible chains. By expressing the Hamiltonian as arising from bending rigidity, characterized by a persistence length a, $P(n)$ is computed using a variational method introduced elsewhere.[8] The physical meaning of eq 2 is clear. If $n_{\rm per}$ is large, then the chain is essentially stiff, and hence, only loops with $n \gg n_{\rm per}$ are allowed. Similarly, if $n_{\rm per}$ is small, then the chain is flexible, and here loops on all scales down to sizes on the order of few multiples of $n_{\rm per}$ are allowed. Thus, $P(n)$ given

10.1021/jp982362n CCC: $18.00
Published on Web 01/12/1999

in eq 2 has appropriate physical limits and may be computed using a mean-field approach to stiff chains.

For proteins, under nominal denaturing conditions (say 6 M urea), Camacho and Thirumalai (CT) estimated n_{min} to be around 7.[5] This parameter is not needed at all in applying our theory to Cytochrome *c*. The distribution function expresses the fact that loops of length much less than n_{per} are unlikely to form due to intrinsic chain stiffness. When $n \gg n_{\mathrm{per}}$ it is appropriate to think of the chain as a sequence of self-avoiding segments each of length n_{per}. For this case, $P(n)$ should rigorously scale as $n^{-\theta_3}$.[9] Thus, $P(n)$ extrapolates naturally from the flexible limit (n_{per} very small) to semiflexible limit when n_{per} is large.

If eq 1 is to be used to obtain absolute times for forming a loop of size *n*, then the normalization constant P_0 in eq 2 needs to be determined. To determine the value of P_0, it is useful to keep the physical process in mind. Consider a chain with two “stickers” that have a favorable interaction between them. The position between them, namely, *n*, is varied. When conditions are arranged so that the contact between them can form, there would be a distribution of products each with a contact of loop length. We believe that this situation applies to the experiments described by HHE. Because the drive to form a contact overwhelms all other possibilities, the equilibrium is largely shifted toward conformations with loops. The weights associated with other possible conformations are negligible. Among the products, each with a definite loop length, the most probable one will have the largest value of $P(n)$. For the situations described here it suffices to determine P_0 from the condition

$$\int_{n_{\mathrm{min}}}^{n_{\mathrm{max}}} P(n)\mathrm{d}n = 1 \tag{3}$$

where n_{min} is the minimum size of the loop allowed and n_{max} is the largest-sized loop possible.[9] Clearly n_{max} is obtained by bringing the chain ends together, and, hence, is equal to the number of monomers in the chain. Since $P(n)$ converges rapidly we can, without introducing any error in determining P_0, let $n_{\mathrm{max}} = \infty$ even for moderate-sized chains. The value of the normalization constant is

$$P_0 = (\theta_3 - 1){n_{\mathrm{min}}}^{\theta_3 - 1} - \left[\frac{1}{{n_{\mathrm{per}}}^{\theta_3+1}\Gamma(\theta_3 + 1,\ \frac{n_{\mathrm{min}}}{n_{\mathrm{per}}})}\right] \tag{4}$$

where $\Gamma(a,x)$ is the incomplete γ function. For the parametrization of $P(n)$ deemed reliable for polypeptide chains,[5] the contribution from the second term in eq 4 is small, and hence in most cases P_0 simplifies to $(\theta_3 - 1){n_{\mathrm{min}}}^{\theta_3-1}$. In particular, for a polypeptide chain $n_{\mathrm{min}} \approx 7$[5] and $n_{\mathrm{per}} \approx 2$.[11] For these values the second term in eq 4 is less than unity whereas the first term is about 12.4.

With the approximate expression for P_0 (a numerically accurate value of the normalization constant for the loop-formation probability proposed by CT is given in the next section), we can estimate the absolute time of forming an intramolecular contact as a function of *n*. This is given by

$$\tau(n) = \frac{\langle {R_n}^2\rangle n^{\theta_3}}{D_0 P_0\left(1 - \exp\left(-\frac{n}{n_{\mathrm{per}}}\right)\right)} \tag{5}$$

If, under denaturing conditions, the chain is flexible, then n_{per} is small and $\langle {R_n}^2\rangle \simeq a^2 n^{2\nu}$ ($\nu \approx 0.6$ or 0.5 depending on whether the denaturant acts as a good solvent or poor solvent), $a = dn_{\mathrm{per}}$ with *d* being the distance between monomer units. The monomer diffusion constant can be approximated by the Stokes relation $D_0 = k_{\mathrm{B}}T/(6\pi\eta a)$ (η is the viscosity, k_{B} is the Boltzmann constant, and *T* is the temperature) so that the determination of $\tau(n)$ only requires knowing the value of *a*. This is an easy quantity to estimate and can be measured explicitly by computing the radius of gyration by suitable scattering measurements. If we take $a \approx 3.8$ Å (the value used by HHE), $\eta \approx 0.01$ Poise and $T = 25°$ C then $D_0 = 5.7 \times 10^{-6}$ cm^2 /s. For $n = 10$, which roughly corresponds to the most probable value of the loop length given by CT, we get $\tau(10) \approx 0.2-0.4$ μs, depending on whether $\nu = 0.5$ or 0.6. In obtaining these estimates we have used $\langle {R_n}^2\rangle = \mathrm{C}_n n^{2\nu} a^2$ with $\mathrm{C}_n = 8$.[2]

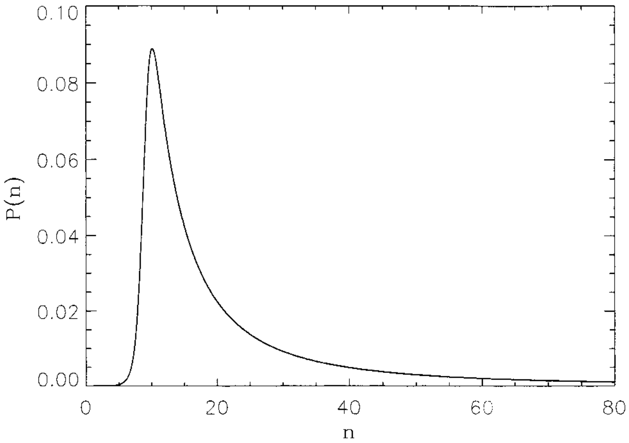

Figure 1. Plot of the normalized loop formation probability $P(n)$ (see eq 6) as a function of the loop length *n* which is measured along the contour of the chain.

Application to Cytochrome *c*

Hagen et al.[2] have estimated that the time scale for forming a loop between met80 and his18 is about 35−40 μs. They used time-resolved spectroscopy, a model for loop formation, and a number of assumptions regarding polymer statistics for cytochrome *c* under fully denatured conditions to arrive at this estimate. The only parameter needed to calculate loop-formation times in our theory is the monomer diffusion coefficient, D_0. Even this is well-described by the Stokes relation. However, to make a direct comparison with the experiments of HHE, we take $D_0 = 10{-}5$ cm^2/s. To calculate τ(met80−his18) (*n* in eq 2 is 62), we use the CT formula for $P(n)$,which is appropriate for polypeptide chains in the random coil state. The precise form of $P(n)$, which has the same structure as eq 2, is[5]

$$P(n) = \frac{18.42\Omega_0}{n^{2.2}[1 + (1.8(9 - n)]} \tag{6}$$

where Ω_0 is a normalization constant. The above formula reduces to that given by eq 2 for *n* greater than about 10. This form for $P(n)$ provides an estimate for loop-formation probability for all values of *n* and correctly takes into account the fact that very short loops occur with smaller probability. The normalization constant Ω_0 is obtained from the condition $\int_1^{n_{\mathrm{max}}} P(n)\mathrm{d}n = 1$. As before, we can safely let $n_{\mathrm{max}} = \infty$ without any errors. Because $P(n)$ rapidly decreases for large *n*, the integral can be easily evaluated numerically and we find that $\Omega_0 = 0.89$. The normalized $P(n)$ is shown in Figure 1. HHE estimated, using random walk statistics, that the mean-squared distance $\langle \mathrm{R}_{62}^2\rangle$ between met80 and his12 is 7225 Å.[2] If we use the estimate for D_0 of HHE, then the time for forming a loop between met80 and his18 according to our theory is

$$\tau_T \approx \frac{\langle R_{62}^2 \rangle}{D_0 P(62)} \approx 38\ \mu\text{s} \tag{7}$$

This value is in exact agreement with the estimates of HHE.

In ref 38 of HHE it is claimed that the prediction, partially made using eqs 1 and 6, compares poorly with experiment. This discrepancy arises solely because the differences in the calculated value of $18.42\Omega_0$ (denoted as P_0 by HHE) and the supposed experimental estimate.[2] HHE presumed that Ω_0 could be measured accurately using their experiment. Their estimate of Ω_0 requires a number of assumptions, the validity of which have by no means been established firmly. Furthermore, unless measurements of $P(n)$ for a number of values of n are made (this does not appear easy), even internal consistency of the errors in determining Ω_0 cannot be made. In the absence of these precise measurements, it is prudent to use the very simple theory given here (requiring essentially one parameter) and compare the predictions directly with the experimental estimates. The only input in our theory is D_0, which, as shown above, can be readily calculated using the Stokes relation. We do not consider the parameter a (the effective persistence length) as adjustable because its value has to be in the range 3.0−5.0 Å. Thus, a fair assessment of theory would be to compute $\tau(n)$ using eqs 1 and 6 with minimal experimental input and compare the results with what the experimentalists (HHE) believe to be accurate estimates. When this is done, as demonstrated here, the agreement between theory and experiment is far better that it should be!

Since the estimate of the mean distance between met80 and his18 involves a few assumptions and appeals to other experiments, it is advisable to check using our theory if the time for loop between met80 and his18 changes drastically by assuming that 6 M GuHCl is a θ solvent. If $\langle R_{62}^2 \rangle \approx a^2 n^{2\nu}$ ($\nu \approx 0.6$ for good solvents), then, with $a = 3.8$ Å, we get τ (met80−his12) $\approx$ 11 μs. This number, while smaller than that obtained by assuming that mean distance between met80 and his18, 85 Å, is still in the range estimated by HHE. Thus, relaxing the assumption that 6 M GuHCl is a θ solvent will not qualitatively alter the conclusions of HHE.

The most important argument advanced by HHE is that protein folding cannot occur any faster than 1 μs. They arrived at this result by noting that the upper limit for the folding rate is roughly equal to the rate at which a loop with the length with the highest probability to form. According to estimates based on the CT theory the most probable length is 10. If the Stokes relation is used for D_0 with $a = 3.8$ Å, we get, using eq 1, $\tau(10) \approx 0.2$ μs. On the other hand, if the experimental estimate of D_0 is used, we find that $\tau(10) \approx 0.1$ μs. Thus, it appears that the fastest a protein can fold is somewhere between 0.1 and 1.0 μs. The completely independent and near parameter free estimates made here support the claims of HHE on the upper limit to the rate of protein folding. Not coincidentally, these numbers roughly correspond to the time needed for forming simple structural motifs such as α helices and β sheets.

Conclusions

In this paper we have given a simple estimate for the time scales for forming loops by a diffusion-limited process. The only parameter needed is an estimate of the effective persistence length of the polypeptide chain. The theoretical value of the rate of formation of a contact between met80 and his18 in cytochrome *c* using the parameters provided by Hagen et al.[2] is in quantitative agreement with the experimental estimate. The estimates for $\tau(n)$ are predicated on the validity of eqs 1 and 2. We can provide additional justification for the validity that our theory should be appropriate by considering the case of a Rouse chain ($\nu = 1/2$) for which simulations for loop formation times have been reported. The exponent θ_3 (see eq 2) is, in general, given by $\nu(d + \theta_2)$, where d is the spatial dimensionality. The value of θ_2 is zero for Rouse chains because there is no self-avoidance. The value of θ_3 for Rouse chains coincides with the Jacobson−Stockmayer result.[12] In this case, we find that $\tau(n) \approx n^{\beta}$ with $\beta \approx 2.5$, which is somewhat larger than the numerical estimates. Comparison of the numerical values estimated using our theory with the simulated results in Table 1 shows that we overestimate the times by about a factor of 5. This simple estimate, therefore, provides a relatively good estimate for loop-formation times.

The estimate for the fastest time in which a protein can fold lies in the range of 10^{-6}−10^{-7} s. The range of this estimate, which is obtained by assuming that contact occurs by a diffusive process, should be considered tentative. This is because in a recent paper by Pitard and Orland,[13] it was shown using an approximate theory that collapse of a moderate-sized chain (75 monomers) in a homopolymer driven by internal forces (two- and three-body solvent-mediated interactions) can take place on the order of 0.1 μs using estimates for the effective two-body interactions and the Kuhn length. Since, in a protein the drive toward forming compact states could be larger due to the presence of a number of hydrophobic residues, it could be conjectured that large corrections to the pure diffusion-limited time scales may result. Additional experimental and theoretical work will be necessary to fully answer this and related questions.

Acknowledgment. I am grateful to John Straub for discussions and encouragement. I also thank W. Eaton with whom I have engaged in many discussions over the last 5 years about loop formation in proteins. This work was supported in part by a grant from the National Science Foundation (Grant No. CHE 96-29845).

References and Notes

(1) Hagen, S. J.; Hofrichter, J.; Szabo, A.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 11615.

(2) Hagen, S. J.; Hofrichter, J.; Eaton, W. A. *J. Phys. Chem. B* **1997**, *101*, 2352.

(3) Eaton, W. A.; Munoz, V.; Thompson, P. A.; Henry, E. R.; Hofrichter. *Acc. Chem. Res.* **1998**, in press.

(4) Chu, B.; Ying, Q.; Grosberg, A. Y. *Macromolecules* **1995**, *28*, 180.

(5) Camacho, C. J.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *93*, 1277.

(6) Pastor, R. W.; Zwanzig, R.; Sazabo, A. *J. Chem. Phys.* **1996**, *105*, 3878.

(7) Guo, Z.; Thirumalai, D. *Biopolymers* **1995**, *36*, 83.

(8) Ha, B.-Y.; Thirumalai, D. *J. Chem. Phys.* **1995**, *103*, 9408.

(9) des Cloizeaux, J. *Phys. (Paris)* **1980**, *41*, 223.

(10) We assume that the ambient conditions are favorable and that a single loop of some arbitrary length forms. If this process is viewed as a chemical equilibrium, then the products consist of at least one structure with a tertiary contact between residues that are not directly bonded to each other.

(11) See refs 27 and 28 of HHE.

(12) Jacobson, H.; Stockmayer, W. H. *J. Chem. Phys.* **1950**, *50*, 1610.

(13) Pitard, E.; Orland, H. *Europhys. Lett.* **1998**, *41*, 467.