# Two-State Folding Kinetics of Small Proteins in the Sequential Collapse Model: Dependence of the Folding Rate on Contact Order and Temperature

**Fernando Bergasa-Caceres*,† and Herschel A. Rabitz**

*Department of Chemistry, Princeton University, Princeton, New Jersey 08544*

*Received: March 31, 2003; In Final Form: July 21, 2003*

In this paper, the dynamics of the collapse-like folding transitions of globular proteins with two-state kinetics is studied. The analyses rely on a simple free energy functional for the formation of protein contacts developed for the study of multistate folding pathways in the sequential collapse model (SCM). The resulting model predicts an approximate linear dependence of the folding rate with the contact order of the native structure. It is also consistent with experimental results that show an Arrhenius-like temperature dependence for the collapse rate when corrected for the effect of protein stability.

## 1. Introduction

Many small proteins of less than ∼75−100 amino acids have been observed to undergo a two-state folding transtion without detectable intermediates between the unfolded and the native state.[1−7] Longer proteins tend to fold through complex multistate pathways.[8−14] The ellucidation of the dynamics of these collapse-like transitions for small proteins has attracted considerable experimental[1−7] and theoretical[15−22] attention in recent years, following the discovery that there are regular relationships linking structural features of the native state with the folding rate.[5,6] In particular, it has been observed that there is an linear relationship between the contact order of the native structure and the folding rate.[5,6] Also, an Arrhenius-like dependence of the folding rate with temperature has been observed.[7] The existence of these regular relationships has led to suggestions that the folding code might be simpler than expected,[5,6] in accord with earlier work.[23] The purpose of this paper is to present a simple model that is able to analytically account for these observations. The free energy functional for contact formation employed relies on a similar formulation used in the sequential collapse model (SCM) for the ellucidation of multistate folding pathways.[24−27]

## 2. The Model

In the SCM the free energy of formation of a successful contact is written as

$$\Delta G_{\text{cont}} = \Delta G_{\text{loop}} + \Delta G_{\text{int}} \qquad (1)$$

where $\Delta G_{\text{loop}}$ is the free energy change associated with loop closure and is expected to be positive because loop formation defines a state that has fewer conformational possibilities than an open protein chain, thus inducing a large entropic loss $\Delta S_{\text{loop}}$. The term $\Delta G_{\text{int}}$ represents all the interactions that help stabilize the contact, with $\Delta G_{\text{int}} < 0$.[28] Hydrophobic interactions have been observed to be the central driving force of the folding process.[29,30] Thus, it is reasonable to write eq 1 as

$$\Delta G_{\text{cont}} \approx \Delta G_{\text{loop}} + \Delta G_{\text{hyd}} \qquad (2)$$

where $\Delta G_{\text{hyd}}$ is the free energy change associated with the burial of hydrophobic groups and the gain of entropy in the solvent and is expected to be negative.[28] Two distinct regions constitute a protein loop: $c$ residues will form the contact and $n$ residues form the open loop, such that the loop length is $L = c+n$. Up to a constant, $\Delta G_{\text{conf}}$ can be written approximately as[24]

$$\Delta G_{\text{loop}} = -kT[c \ln(f_c/f_0) + n \ln(f_n/f_0) - 3/2 \ln n] \qquad (3)$$

where $f_i$ represents the conformational freedom of the amino acids in a given region $i$ (i.e., $i = n, c$) of the protein, and $f_0$ is the conformational freedom of the amino acids in the random coil. $\Delta G_{\text{loop}}$ can be shown to have two minima[24] as a function of loop length $n$: a deeper one at $65 < n < 85$ amino acids, called the optimal loop length $n_{\text{op}}$, and a shallower one at $n \approx$ 3−4 amino acids. The shallow minimum represents the shortest length over which the protein chain can reverse its direction. Because a few amino acids are required to form a stable contact, a minimal loop is defined to be generated by a protein contact between two segments of $n_{\text{c,min}}$ ∼3−5 amino acids linked by a turn of ∼4 amino acids. Thus, the minimal loop size is $n_{\text{min}} \approx n + n_{\text{c,min}} = 10-14$ amino acids. Within the SCM, most of the contacts will form at $n_{\text{min}}$, because no more than a few contacts may form at the optimal distance,[24] consistent with experimental evidence showing that short-range contacts predominate over long-range ones in protein structures.[31]

For loops close to minimal length (i.e., short loops) we have $n \ln(f_n/f_0) \gg 3/2 \ln n$[27] and $f_c \approx f_n$, as in a short loop all of the amino acid side chains will have severely restricted conformational freedom after loop definition. Then, eq 3 can be written as

$$\Delta G_{\text{loop}} \approx -kT[(n + c) \ln(f_c/f_0)] \qquad (4)$$

and $\Delta G_{\text{cont}}$ becomes

$$\Delta G_{\text{cont}} \approx -kT\,(n + c) \ln(f_c/f_0) + \Delta G_{\text{hyd}} \qquad (5)$$

Section 3 discusses the physical basis for the two-state folding mechanism within the SCM. Sections 4 and 5 will show that consideration of eq 5 and transition-state theory suffice to reproduce the dependence of folding rate with contact order and temperature observed in small proteins.

[*] Corresponding author.

[†] Current address: ENDESA, C/ Ribera del Loira 60, 28042 Madrid, Spain. Fax: 34 91 213 1908. E-mail: Fbergasa@Endesa.Es

10.1021/jp030384q CCC: $25.00
Published on Web 10/29/2003

## 3. Cooperativity and Two-State Folding Transitions

A two-state folding process in the SCM neccesarily implies that the protein does not form an initial contact at $n_{\mathrm{op}}$ (i.e., a primary contact) leading to a multistate folding pathway.[24] There are two possible reasons why a primary contact would not form in the SCM: (1) the protein is shorter than $n_{\mathrm{op}}$ or (2) the primary contact and tails are not hydrophobic enough to nucleate the formation of a stable long-range intermediate. By hypothesis, the free energy change $\Delta G_{ij}$ for formation of a contact between segments $i$ and $j$ at the minimal distance in general satisfies $\Delta G_{ij} > kT$ in the SCM. Otherwise, because $n_{\mathrm{min}} \ll n_{\mathrm{op}}$, the folding pathway of globular proteins would almost always be initiated by formation of loops of length close to $n_{\mathrm{min}}$. This hypothesis implies that several minimal loops must form at once to initiate the folding process for proteins of length $< n_{\mathrm{op}}$ in order to gain extra stabilization energy from multiple interactions between the segments defining the minimal loops.

It is then suggestive to hypothethize that a two-state nucleation process in which a significant portion of the protein chain enters the folding process simultaneously, initiates the folding pathway of short proteins. The folding nucleus thus created would nucleate the fast folding of the remaining tails. The number of minimal loops that must enter the first nucleation event could vary with the number and location of the hydrophobic amino acids along the sequence, as the burial of these would provide most of the stabilization energy of the native structure. The mechanism proposed here differs from the hypothesis made within the funnel model[32,33] that the existence of multistate and two-state folding mechanisms reflects the different "smoothness" of folding landscapes as a function of the primary sequence.[15] "Rough" folding landscapes with significant activation barriers would lead to multistate folding pathways, while "smooth" folding landscapes would lead to two-state folding transitions.[15]

In the SCM, both multistate and two-state folding reflect the existence of loop-length dependent configurational barriers. The distinction between the two-state and multi-state SCM mechanism arises because in most proteins, long-range contacts initiate the folding pathway leading to a detectable intermediate molten-globule-like-state state (MGLIS).[24] This is probably not a fundamental discrepancy with the funnel picture as a whole, but rather with the specific proposals made to explain the nature of the collapse mechanism within the funnel context.[21] Presently it is not possible to fully establish whether the hypothesis presented here to explain collapse-like folding transitions within the SCM is correct, due to the need to develop a better understanding of the relative contributions of the interactions included in eq 1 to the stability of the folding intermediates.

## 4. Dependence of the Folding Rate with Contact Order

In the SCM, definition of the native-like topology $\Gamma$ of the collapsing region in the SCM must precede establishment of its native structure.[25] Thus, it is natural to assume that the entropic loss associated with the restriction of the conformational space available to the amino acids included in the collapsing region acts as an activation barrier to formation of the native-like set of contacts that constitute $\Gamma$. Definition of the native topology has also been assumed to be the rate-limiting step in the folding of small proteins in recent work.[17,34–37] This assumption allows for a good first principles approximation to the prediction of folding rates of small proteins.[37] The entropic activation barrier $E_{ij}$ to formation of one of the N contacts that define the native topology between amino acids $i$ and $j$ can be written as

$$E_{ij} = -T\,\Delta S_{ij} = \Delta G_{\mathrm{conf},ij} \approx -k\mathrm{T}\,[n_{ij}\ln(f_n/f_0)] \qquad (6)$$

Furthermore, in a collapse-like process all amino acids enter the pathway at the same time. Then eq 6 becomes

$$E_{ij} = -T\,\Delta S_{ij} = \Delta G_{\mathrm{conf},ij} \approx -kT\,n_{ij}\ln(f_{\mathrm{c}}/f_0) \qquad (7)$$

Cooperative collapse implies the simultaneous formation of $N$ contacts. Then, the average time scale for collapse can be written as the $N$th root of the product of the typical time scales for each of the contacts (i.e., the time scale corresponding to the average loop length):

$$\tau_{\mathrm{coll}}(\Gamma) \sim \mathrm{g}^{-1}\exp\,[\sum_{ij}^{N}\Delta G_{\mathrm{conf},ij}/(NkT)] \approx g^{-1}\,[\Pi\,(f_0/f_{\mathrm{c}})^{n_{ij}}]^{(1/N)} \qquad (8)$$

where $g$ is a characteristic transition frequency that can be taken in first approximation as independent of the specific topology considered.[17] The rate of collapse is $\kappa_{\mathrm{coll}} = \tau^{-1}{}_{\mathrm{coll}}$, and we can write

$$\ln\kappa_{\mathrm{coll}}(\Gamma) \approx \ln\mathrm{g} + N^{-1}\sum_{ij}^{N} n_{ij}\ln(f_c/f_0) \qquad (9)$$

The overall stabilization energy of the native folded protein $\Delta G_{\mathrm{Nat}}$, becomes

$$\Delta G_{\mathrm{conf,Nat}} + \Delta G_{\mathrm{int,Nat}} = \Delta G_{\mathrm{Nat}} \approx -\mathrm{k}T\,L\ln\,[f_c/f_0] + \Delta G_{\mathrm{int,Nat}} \qquad (10)$$

Solving for $f$ in eq 10 yields

$$f_c \approx f_0\exp\,[(\Delta G_{\mathrm{int,Nat}} - \Delta G_{\mathrm{Nat}})/(LkT)] \qquad (11)$$

Equation 8 now becomes

$$\ln\kappa_{\mathrm{coll}} \approx \ln g + (NL)^{-1}\sum_{\mathrm{ij}}^{\mathrm{N}} n_{ij}\,[(\Delta G_{\mathrm{int,Nat}} - \Delta G_{\mathrm{Nat}})/(kT)] \qquad (12)$$

where $(\mathrm{NL})^{-1}\,\Sigma_{ij}^{N}\,n_{ij}$ is the contact order (c.o.).[38] $\Delta G_{\mathrm{int,Nat}}$ and $\Delta G_{\mathrm{Nat}}$ are both $< 0$; moreover, $|\Delta G_{\mathrm{int,Nat}}| \gg |\Delta G_{\mathrm{Nat}}|$, as the total stabilization energy of a protein is always much less than the cumulative energy of the attractive interactions defining its structure, due to the large entropic contribution opposing folding.[25] Then, in the SCM, the rate of collapse can be written as

$$\log\kappa_{\mathrm{coll}} \approx \log g - 4.3\;10^{-3}\ \mathrm{c.o.}\ (\,\Gamma)|\Delta G_{\mathrm{int,Nat}}|/(kT) \qquad (13)$$

where c.o. is expressed as a percentage, and the logarithms are in base 10. Thus, the SCM predicts that the logarithm of the collapse rate depends inversely with the contact order of the native topology. This relationship bears a striking resemblance with that observed experimentally.[5,6] Comparison of the logarithm of the two-state folding rate with the contact order (in %) for 24 proteins that undergo a two-state folding transition under at least some conditions[5,38] yields a slope for the linear relationship of $\sim -0.39$ with correlation $r = 0.92$. The expected value $\sim -4.3\;10^{-3}\,|\Delta G_{\mathrm{int,Nat}}|$ for the slope of the SCM linear relationship between the logarithm of the folding rate and contact order and the experimental value of $\sim -0.39$ agree if we set $|\Delta G_{\mathrm{int,Nat}}| \approx 90kT$. This is a reasonable value for a typical protein of ~80 amino acids containing ~1/3 hydrophobic amino acids, each contributing ~3−5 kT to the stabilization energy,[39] but the full validity of eq 12 will only be firmly established when a better understanding of the energetics of the native state is gained.

The approach followed in this paper to derive a relationship between contact order and collapse rates is broadly similar to that followed in the topomer-search model to study the folding rates of small proteins,[17] with being both based on transition-state theory. A significant difference is that in the present work an explicit form for the average equilibrium constant for formation of contacts based on side chain crowding effects is assumed (i.e., eq 10). The present model is not conceptually incompatible with the extended nucleus model (ENM),[18] as both assume that hydrophobic interactions drive the formation of loops, and that there is an entropy cost to the formation of protein loops prior to the establishment of full secondary structure. The SCM and the ENM differ however in the way they account for the entropy loss due to loop closure, as in the SCM there is an extra term, additional to the Jacobson-Stockmeyer (JS) chain entropy, to account for the crowding of the side chains.

As explained previously, the model presented here differs from several proposals intended to explain the dynamics of fast folding proteins within the funnel context.[20,21] The SCM differs from simulations carried out to understand the transition states of fast folding proteins within the capillarity theory of protein folding[20] in the way that the entropy loss upon loop closure is accounted for.[20] Also, the SCM derivation presented above does not include consideration of the protein energetic heterogeneity,[21] that has been proposed to be correlated with the barrier height, such that native sequences having stronger contact energies correlated with the least loop-closure entropic cost (measured in JS form) fold faster. The SCM picture bears some resemblance to simulations of the folding dynamics that assume that contacts form only when the native structure of the intervening chain is already defined,[22] since in the model presented here, establishment of a native-like configuration in the loops defined by the contacts is assumed to be the rate-limiting step for collapse. The model presented here is in a different spirit than approaches that trace the dependence of the folding rate with the topology to network connectivity measures.[16] The SCM attempts to explain the dependence of the folding rate with contact order within a "classical" physical picture that relies only on the concepts of chemical reaction transition state theory. Whether the two approaches are compatible at some level remains to be investigated.

## 5. Dependence of the Folding Rate Upon the Temperature

Experiments carried out with CspB and protein L have shown that, when corrected for the temperature dependence of protein stability[40] (i.e., at constant $\Delta G_{\text{nat}}/T$), plots of $\ln\kappa_{\text{coll}}$ showed an inverse linear dependence upon the temperature.[7] Relatively simple applications of transition state theory have been observed to suffice for a good prediction of protein folding rates.[41,42] Employing transition state theory within the SCM, we can write eq 8 as

$$\tau_{\text{coll}}(\Gamma) \sim g_0^{-1} \exp[(\Delta G_{\text{TS}}/N + \langle E_{ij}\rangle)/(kT)] \qquad (14)$$

where $\Delta G_{\text{TS}}$ is the stabilization energy of the transition state,[7] $\langle E_{ij}\rangle$ is the average activation barrier $\langle E_{ij}\rangle \equiv N^{-1} \sum_{ij}^{N} \Delta G_{\text{conf},ij}$, and $g_0$ is a constant. Because the main interactions that hold the native state together are the same as those that are expected to stabilize the transition state (i.e., hydrophobic interactions compatible with the native topology), fixing $\Delta G_{\text{nat}}/T$ is equivalent to fixing $\Delta G_{\text{TS}}/T$.[7] Then, from eq 14 we can write

$$\ln \kappa_{\text{coll}} \approx \ln g_0 - \Delta G_{\text{TS}}/(NkT) - \langle E_{ij}\rangle/(kT) \qquad (15)$$

which at fixed $\Delta G_{\text{TS}}/T$ shows an Arrhenius-like dependence such that the slope of the $\ln \kappa_{\text{coll}}$ dependence with $1/(kT)$ is, comparing with eqs 7−13, the average entropic activation barrier to collapse (which should not depend on $T$ as long as the protein remains folded):

$$\mathrm{d}^{-1}_{(KT)}(\ln \kappa_{\text{coll}}) = -\langle E_{ij}\rangle = \sum_{N}^{ij} \Delta G_{\text{conf},ij}/N \qquad (16)$$

## 6. Conclusions

In this paper, the SCM was applied for the first time to the study of the two-state folding transitions of small proteins. It was shown that the SCM predicts a dependence of the folding rate with contact order and temperature that is consistent with experimental observations. The fact that the same free energy functional used in the study of multistate folding pathways may be used here in a consistent fashion suggests that common physical principles underlie the dynamics of multistate and two-state folding pathways.[41]

**Acknowledgment.** F.B-C. thanks Prof. Juan J. Saenz at the Universidad Autónoma de Madrid for useful discussions. H.R. thanks the National Science Foundation.

## References and Notes

(1) Jackson S. E.; Fersht A. R. *Biochemistry* **1991**, *30*, 10428.

(2) Viguera, A. R.; Martínez, J. C.; Filimonoc, V. V.; Mateo P. L.; Serrano, L. *Biochemistry* **1994**, *33*, 2142.

(3) Huang, G. S.; Oas, T. G. *Biochemistry* **1995**, *34*, 3884.

(4) Schönbruner, N.; Koller, K.-P.; Kiefhaber, T. *J. Mol. Biol.* **1997**, *268*, 526.

(5) Plaxco, K. W.; Simons, K. T.; Ruczinski, I.; Baker, D. *Biochemistry* **2000**, *39*, 11177.

(6) Baker, D. *Nature* **2000**, *405*, 39.

(7) Scalley, M. L.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 10636.

(8) Jennings, P. A.; Wright, P. E. *Science* **1993**, *262*, 892.

(9) Jacobs, M. D.; Fox, R. O. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 449.

(10) Ballew, R. M.; Sabelko, J.; Gruebele, M. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 5759.

(11) Gladwin, S. T.; Evans, P. A. *Fold. Design* **1996**, *1*, 407.

(12) Eliezer, D.; Yao, J.; Dyson H. J.; Wright, P. E. *Nat. Struct. Biol.* **1998**, *5*, 148.

(13) Nishimura, C.; Prytulla, S.; Dyson H. J.; Wright, P. E. *Nat. Struct. Biol.* **2000**, *7*, 679.

(14) Kuwata, K.; Shastry, R.; Cheng, H.; Hoshino, M.; Batt C. A.; Roder, H. *Nat. Struct. Biol.* **2001**, *8*, 151.

(15) Dill, K. A.; Chan, H. S. *Nat. Struct. Biol.* **1997**, *4*, 10.

(16) Dokholyan, N. V.; Li, L.; Ding, F.; Shaknovich, E. I. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 8637.

(17) Makarov, D. E.; Plaxco, K. W. *Protein Sci.* **2003**, *12*, 17.

(18) Fersht A. R., *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 1525.

(19) Dinner A. R.; Karplus, M. *Nat. Struct. Biol.* **2001**, *8*, 21.

(20) Galzitskaya, O. V.; Finkelstein, A. V. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11299.

(21) Plotkin, S. S.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 6509.

(22) Muñoz, V.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11311.

(23) Hinds, D. A.; Levitt, M. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 2536.

(24) Bergasa-Caceres, F.; Ronneberg, T. A.; Rabitz, H. A. *J. Phys. Chem. B* **1999**, *103*, 9749.

(25) Bergasa-Caceres, F.; Rabitz, H. A. *J. Phys. Chem. B* **2001**, *105*, 2874.

(26) Bergasa-Caceres, F.; Rabitz, H. A. *J. Phys. Chem. B* **2002**, *106*, 4818.

(27) Bergasa-Caceres, F.; Rabitz, H. A. *J. Phys. Chem. B* **2003**, *107*, 3606.

(28) Dill, K. A. *Biochemistry* **1990**, *29*, 7123.

(29) Kauzmann W. *Adv. Protein Chem.* **1959**, *14*, 1.

(30) O'Neil, K. T.; Degrado, W. F. *Science* **1990**, *250*, 646.

(31) Schulz, G. E.; Schirmer, R. H. *Principles of Protein Structure*, Springer-Verlag: New York, 1979.

(32) Sali, A.; Shakhnovich, E.; Karplus, M. *Nature* **1994**, *369*, 248.

(33) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G., *Proteins: Struct. Funct., Genet*. **1995**, *21*, 167.

(34) Sheinerman, F. B.; Brooks, C. L. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 1562.

(35) Socci, N. D.; Onuchic, J. N.; Wolynes, P. G. *Proteins: Struct. Funct., Genet.* **1998**, *32*, 136.

(36) Sosnick, T. R.; Mayne, L.; Englander, S. W. *Nat. Struct. Biol.* **1994**, *1*, 149.

(37) Debe, D. A.; Goddard, W. A. *J. Mol. Biol.* **1999**, *294*, 619.

(38) Plaxco, K. W.; Spitzfaden, C.; Campbell, I. D.; Dobson, C. M. *J. Mol. Biol.* **1997**, *270*, 763.

(39) Fauchère, J. L.; Pliska, V. *Eur. J. Med. Chem.* **1983**, *18*, 369.

(40) Schindler, T.; Schmid, F. *Biochemistry* **1996**, *35*, 16833.

(41) Bergasa-Caceres, F.; Rabitz, H. A. *Chem. Phys. Lett.* **2003**, *376*, 612.

(42) Socci, N. D.; Onuchic, J. N.; Wolynes, P. G. *J. Chem. Phys.* **1996**, *104*, 5860.