Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 10422–10425, October 1994
Biochemistry

Structure of the transition state for the folding/unfolding of the barley chymotrypsin inhibitor 2 and its implications for mechanisms of protein folding

(protein engineering)

DANIEL E. OTZEN, LAURA S. ITZHAKI, NADIA F. ELMASRY, SOPHIE E. JACKSON, AND ALAN R. FERSHT

Medical Research Council Unit for Protein Function and Design and Cambridge Centre for Protein Engineering, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, United Kingdom

Contributed by Alan R. Fersht, July 12, 1994

ABSTRACT **The equilibrium and kinetics of folding of the single-domain protein chymotrypsin inhibitor 2 conform to the simple two-state model. The structure of the rate-determining transition state has been mapped out at the resolution of individual side chains by using the protein engineering method on 74 mutants that have been constructed at 37 of the 64 residues. The structure contains no elements of secondary structure that are fully formed. The majority of interactions are weakened by >50% in the transition state, although most regions do have some very weak structure. The structure of the transition state appears to be an expanded form of the native state in which secondary and tertiary elements have been partly formed concurrently. This is consistent with a "global collapse" model of folding rather than a framework model in which folding is initiated from fully preformed local secondary structural elements. This may be a general feature for the folding of proteins lacking a folding intermediate and is perhaps representative of the early stages of folding for multidomain or multimodule proteins. The major transition state for the folding of barnase, for example, has some fully formed secondary and tertiary structural elements in the major transition state, and barnase appears to form by a framework process. However, the fully formed framework may be preceded by a global collapse, and a unified folding scheme is presented.**

The folding of a protein from its unfolded state to its specific biologically active conformation has been generally assumed to follow a specific pathway or set of pathways in order to occur in a finite time (1). Several different models have been suggested to describe the reaction (2–8). According to the hydrophobic collapse model, the driving force for folding is visualized as the squeezing out of water from a rapidly formed hydrophobic core within which secondary and tertiary structure is subsequently formed, as employed by Dill *et al.* (9) in their "hydrophobic zipper" model. Framework models envisage preformed secondary structural elements initiating folding (2, 10–12). These elements may diffuse together, collide and stabilize each other by local rearrangements (docking) (8, 11), or propagate (13, 14). The existence of defined pathways has been challenged, however. According to the jigsaw puzzle model (15), there is no preferential starting point for folding, and each folding attempt may follow a different path. This mechanism is supported by very recent computer simulations (16).

The pathway of protein folding is now amenable to analysis at the level of individual residues because of developments in protein engineering and NMR (see ref. 16 for a brief review). The structures of unfolded states and stable intermediates may be analyzed by NMR, whereas transition states can be studied only by kinetics, such as the protein engineering method (17–23). In the latter method, described in detail by Fersht *et al.* (20), site-directed mutagenesis is used to remove interactions made by a particular side chain. The change in the Gibbs free energy of folding, $\Delta\Delta G_{\mathrm{F-U}}$, is measured from equilibrium denaturation experiments (F, folded; U, unfolded). The change in the free energy of any other state X, $\Delta\Delta G_{\mathrm{X-U}}$, is measured by kinetics with respect to the unfolded state. A quantity Φ_{F} is obtained, defined by $\Phi_{\mathrm{F}} = \Delta\Delta G_{\mathrm{X-U}}/\Delta\Delta G_{\mathrm{F-U}}$. If $\Phi_{\mathrm{F}} = 1$, then the state X is destabilized by exactly the same amount as the fully folded state, and so the target side chain is assumed to be in a fully native environment in state X. If $\Phi_{\mathrm{F}} = 0$, then the state X is unaffected by the mutation and so the target side chain is assumed to be in a fully denatured environment in X, since $\Delta\Delta G_{\mathrm{X-U}} = 0$. There is not a linear relationship between Φ_{F} and the extent of structure formation for fractional values. Each mutated side chain acts as a reporter group, and a series of such measurements has defined in detail the structures of the major transition state for the unfolding of barnase and a folding intermediate (24). A subsequent computer simulation of the unfolding of barnase gave results in excellent agreement with those from the protein engineering method (25).

We now describe an extensive application of the method to chymotrypsin inhibitor 2 (CI2), a serine protease inhibitor from barley seeds. A truncated, 64-residue form of the protein—lacking the N-terminal, unstructured 19 amino acids—is used in these studies. CI2 is an ideal protein with which to tackle certain aspects of the protein folding problem. It is small and monomeric. Its three-dimensional structure has been solved in both the crystal (26, 27) and the solution states (28–31). Despite its small size, there is considerable secondary and tertiary structure present and a small, compact hydrophobic core. It has no cis peptidyl-prolyl bonds whose rate of isomerization from the trans in the unfolded state would otherwise limit the rate of the major phase of refolding. Crucially, CI2 has no disulfide crosslinks, so that it may be refolded from its maximally unfolded forms. Further, CI2 is a rare example of a protein whose folding and unfolding behavior, under both equilibrium and nonequilibrium conditions, has been shown to follow the two-state model (32): no intermediates accumulate and there is only one kinetically significant transition state (or set of transition states). This protein thus provides us with one of the simplest systems possible with which to investigate protein folding. The important consequence of the two-state behavior is that the energetics and structure of the transition state can be probed in both the direction of folding and the direction of unfolding by use of the protein engineering method (32–35).

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: CI2, chymotrypsin inhibitor 2.

MATERIALS AND METHODS

Sixty novel mutants were produced, expressed, purified, and analyzed kinetically and thermodynamically as described (32–35). The kinetics of unfolding were measured by mixing native protein at 25°C in 50 mM 2-(*N*-morpholino)ethanesulfonic acid/HCl buffer (pH 6.2) with various concentrated solutions of guanidinium chloride in the same buffer in a stopped-flow fluorimeter. Refolding was monitored under the same conditions by mixing acid-denatured protein with a more alkaline buffer to give the same final pH. The fast refolding phase, representing some 70–80% of the amplitude of the fluorescence signal, which corresponds to the refolding of the protein with all its peptidyl-prolyl residues in the trans conformation, was used for analysis.

RESULTS

We have prepared the ground for this study by presenting detailed analyses of kinetics and thermodynamics of folding and unfolding of wild-type CI2 and hydrophobic core mutants (32–35). Those mutants were shown to fold and unfold by a two-state mechanism (simple V-shaped dependence of logarithms of rate constants versus guanidinium chloride concentration, ratio of refolding to unfolding rate constants gives the correct equilibrium constant for protein folding, etc.). The new mutants described here also fully fit two-state behavior. Importantly, there is indeed good general agreement between Φ values from refolding experiments (Φ_F) and from unfolding experiments (Φ_U). [For a two-state mechanism, $\Phi_F = (1 - \Phi_U)$.] In addition, Φ values do not change much when different mutations are constructed at the same site. Thus, the overall formation of structure in the transition state parallels the formation of interactions between individual side chains.

Structure of the Transition State. Almost all the mutations analyzed have fractional Φ values, and there are no Φ values of 1, indicating that no region of the protein can be detected with its full native-like structure in the transition state. At the majority of sites, Φ values are <0.5, indicating that the transition state is only weakly structured. Fig. 1 shows a schematic representation of the structure of CI2 in which the degree of structure formation in the transition state for folding is represented by color—the two extreme colors are defined to be red, in regions where the structure is completely disordered in the transition state, and blue, in regions where the structure is native-like in the transition state. Fractional values of Φ are indicated by a mixture of the two colors. The colors in Fig. 1 are seen to be basically a mixture of blue and red throughout.

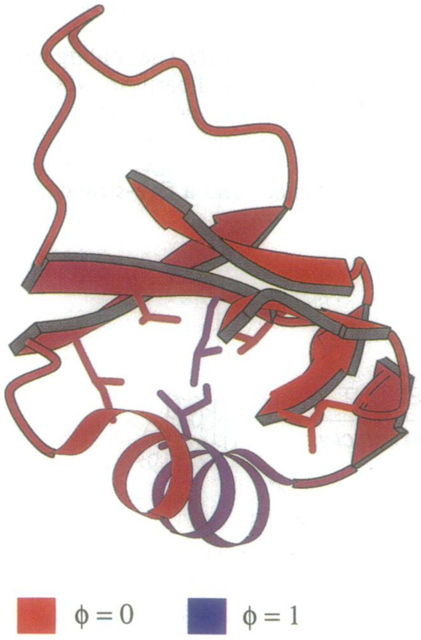

FIG. 1. Transition state for the folding of CI2, color coded according to the degree of structure formation. Bright red is fully unfolded, deep blue is fully folded, and partial formation is indicated by a mixture of the two colors. In practice, there is little bright red and no deep blue.

We now describe individual structural elements. The secondary structure of CI2, defined from an NMR analysis of the solution structure (30, 31), is as follows (numbering is according to the long form): residues 22–24, β-strand 1; 24–27, type III reverse turn; 27–30, type II reverse turn; 30–32, β-strand 2; 31–43, α-helix; 44–47, type I reverse turn; 47–53, β-strand 3; 54–61, reactive-site loop (extended structure); 64–70, β-strand 4; 71–73, turn; 74–77, β-strand 5; and 79–83, β-strand 6. An analysis of the secondary structure by ϕ and φ angles (36) finds only three β-strands, corresponding in our notation to residues 46–53, β-strand 3; residues 65–71, β-strand 4; and residues 79–82, β-strand 5. The α-helix runs from 32 to 42 according to ϕ and φ angles, but we use the definition of caps by Richardson and Richardson (37) to include Ser-31 as the N cap of the helix and Lys-43 as its C cap.

α-Helix (residues 31–43). Eight sites were probed in the α-helix of CI2. A variation of Φ values is observed as the probe traverses along the helix, averaging about 0.4. The C cap of the helix (residue 43) is almost completely disordered in the transition state, with Φ values close to zero. From residues 40 to 31, however, there is a weakened structure with Φ values between 0.3 and 0.7.

β-Sheet. There are seven sites probed in the β-sheet. There is a gradation of Φ values. The antiparallel β-structure formed between β-strand 1 (mutated at position 22) and β-strand 6 (mutated at positions 77, 79, and 82) appears to be almost completely unstructured in the transition state, whereas the parallel β-structure (strand 3 mutated at position 49, to remove interactions with strand 4) is partially formed in the transition state. This is perhaps not surprising. The parallel strand pair, with seven internal hydrogen bonds, forms the longest stretch of uninterrupted structure within the β-sheet and is likely to have greater intrinsic stability than the shorter, antiparallel strands. The interactions between β-strands 2, 3, 4, and 5 have Φ values between 0.2 and 0.4.

Hydrophobic cores. Seven sites of the major hydrophobic core have been probed previously (34, 35). The behavior of these mutants can be divided into two classes. The first includes L27A (Leu-27 → Ala), V66A, and I76A, which exhibit Φ values close to zero, indicating that the interactions made at these sites are not formed in the transition state; these mutations are all located on the edge of the hydrophobic core, either at the beginning or at the end of a β-strand or in turns. The second class of mutants includes I39V, I48V, L68A, V70A, and I48A/I76V. These exhibit fractional Φ values, between 0.30 and 0.65, indicating partial formation of interactions at these sites, and are all located in the centre of the core. Therefore, the core could form from the innermost positions outwards.

Residues Leu-51, Val-57, and Phe-69 form a hydrophobic pocket near one end of the reactive loop, on the other side of the β-sheet relative to the hydrophobic core. We call this cluster the minicore. The minicore appears to be less well formed in the transition state than the major hydrophobic core. Φ values for the three probes at positions 51, 57, and 69 cluster around 0.3. This is not significantly higher than the Φ values of the surrounding residues, suggesting that this is not a local nucleus for folding.

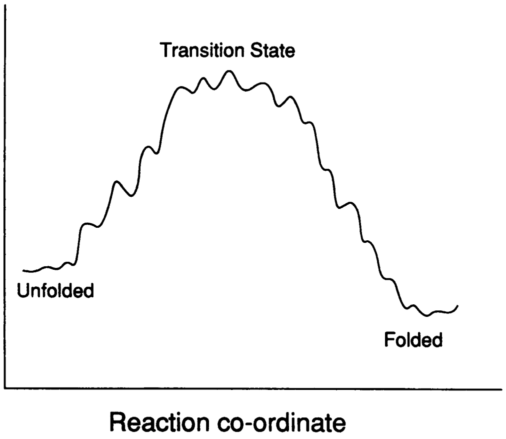

FIG. 2. Reaction-coordinate diagram for the folding and unfolding of CI2. The Gibbs energy (G) of reaching the transition state contains the contributions of all the conformational equilibria that precede it in either direction of the reaction.

Turns and reactive-site loop. All mutations in these regions were of polar or charged residues to nonpolar or uncharged residues. The solvation energies of the wild-type and mutant side chain can be significantly different in these cases and so fractional Φ values are difficult to interpret (20). However, by the use of double mutant cycles it is possible to calculate $\Delta\Delta G_{\text{int}(\ddagger-\text{U})}$, the interaction energy between the two residues in the transition state for folding, and compare this to $\Delta\Delta G_{\text{int}(\text{U}-\text{F})}$, the interaction energy in the folded state. Thus, a Φ value can be calculated on the basis of the ratio $\Delta\Delta G_{\text{int}(\ddagger-\text{U})}/\Delta\Delta G_{\text{int}(\text{U}-\text{F})}$. The theory, assumptions, and limitations of this approach have been discussed extensively (20). Application of the procedure gives Φ values of about 0.3 for these regions.

DISCUSSION

The transition state for the folding of CI2 was shown previously to be quite compact from its thermodynamic properties and from the average degree of exposure of residues to solvent, which indicates that the transition state for folding is approximately two-thirds folded (32). The transition state for the reaction must be a highly crenulated surface (see ref. 17), consisting of many maxima and minima (Fig. 2). The energy of reaching the transition state contains the energetics of all the preequilibria and hence all the bonds that are made or broken. We find from the Φ-value analysis that no region of CI2 is fully formed in the transition state and that interactions in the transition state are weakened by >50% ($\Phi_F < 0.5$) relative to the native protein at the majority of sites probed, averaging 66% weaker. The transition state for folding is thus a generally expanded form of the folded structure with no fully formed elements of structure. The main region of structure that has higher Φ values is the α-helix, especially toward the N terminus, but this region is significantly weakened in energy. A computational study of the unfolding of CI2, performed without prior knowledge of the experimental results of this study, has produced results of remarkable similarity (36).

Interpretation of Fractional Φ Values. There is an ambiguity in that fractional Φ values may arise from either partial structure formation or a mixture of folded and unfolded states (20). For example, there is a Φ_F value of 0.4 for the mutation of Ser-31 (which is the N cap of the helix) to Gly-31. Either the transition state is basically a single species in which the N terminus of the α-helix is weakened by 60% or there are parallel pathways of folding and unfolding in which, for example, 40% of the species have the N terminus fully intact in the transition state and 60% fully unfolded. In the accompanying paper (38), we show that there is not a mixture of fully intact and fully unfolded structures and there must be partial structure formation. The transition state is a collection of similar structures that fluctuate around the lowest-energy transition state.

Mechanistic Implications of Fractional Φ Values. A Φ value of 0.3, which is typical for many of the mutations, indicates that the structure is present but significantly weakened. Thus, there is a weakened helix present in the transition state, as well as a weakened β-sheet consisting of β-strands 2–5, weakened minicore, and weakened binding loop. The most important conclusion from these data is that there are no fully developed elements of secondary structure in the transition state for the folding of CI2. By inference, therefore, folding is not initiated according to a framework model for folding in which the elements of secondary structure are fully formed. The data instead appear to fit a model in which weakly formed elements of secondary and tertiary structure have coalesced, perhaps consistent with the rearrangement of a "globally collapsed" structure.

Are the Conclusions General? The folding pathway of CI2 appears to contrast with that of barnase, which has also been comprehensively analyzed by protein engineering and other methods (24). Barnase is a monomeric, single-chain ribonuclease of 110 amino acids. It is larger than CI2 and has more structure (three α-helices and a five-stranded β-sheet, with

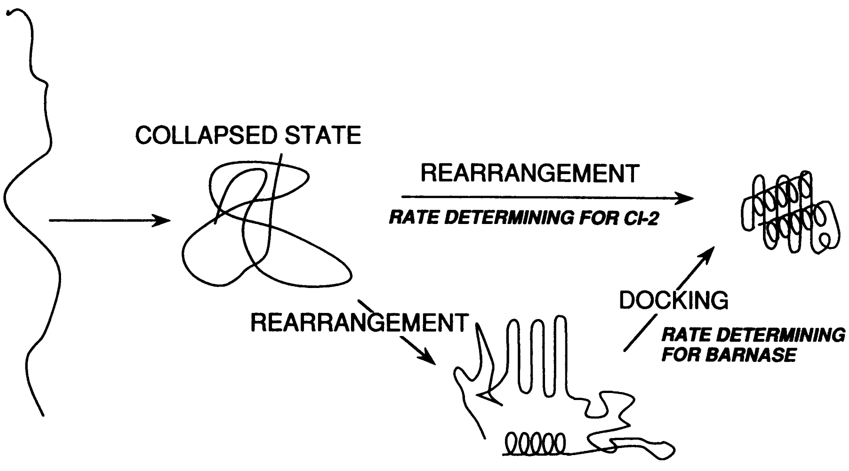

FIG. 3. A unified scheme for the folding of barnase and CI2. The fully unfolded chain collapses. The formation of the single domain CI2 has as its rate-determining step the rearrangement of this collapsed state. The multimodule barnase has as its rate-determining step the consolidation and rearrangement of the hydrophobic core during the docking of the modules.

three distinct hydrophobic cores). The main distinctions between the folding pathways of the two proteins are as follows. (*i*) The refolding of barnase is more complex than that of CI2 in that the former has a kinetically important intermediate on the pathway (18). (*ii*) There are very many examples of mutants of barnase with Φ values for the intermediate and the major transition state of close to 0 or 1.0, corresponding to the situations in which the interactions being probed by the mutation are completely absent or fully formed, respectively. (*iii*) The rate-determining step in the folding of the protein is the rearrangement of the hydrophobic core in a complex in which the major α-helix has associated with the β-sheet, and both are largely formed. Fragments of barnase that contain separately the major helix and sheet have a small component of each in the correct structure and rapidly associate to form active enzyme (39). Barnase can thus fold in parts that associate. The simplest interpretation is that barnase folds via a framework model with the aforementioned helix and the sheet being the initiation sites. But this may not be so. The pathway of folding of barnase may be perturbed by radical mutations in its major helix that cause the helix to form late in the pathway, and so the helix is not an obligatory initiation site (J. M. Matthews and A.R.F., unpublished work).

The major difference between barnase and CI2 is that the larger protein, barnase, appears to be constructed of smaller "modules" (40). CI2, on the other hand, appears to be closer to a single module. Barnase is perhaps representative of larger proteins that contain separate modules. These can fold separately and then associate. CI2 may be representative of the folding of an individual module and so illustrates an earlier stage of folding. A unified scheme is presented in Fig. 3.

This work was stimulated by discussions at the North Atlantic Treaty Organization Advanced Study Institute on Protein Structure, Function and Design, held at Spetsai, Greece, in September 1993. D.E.O. was supported by a fellowship from the Danish Natural Science Research Council.

1. Levinthal, C. (1968) *J. Chim. Phys.* **85,** 44–45.
2. Ptitsyn, O. B. (1973) *Dokl. Acad. Nauk.* **210,** 1213–1215.
3. Kim, P. S. & Baldwin, R. L. (1990) *Annu. Rev. Biochem.* **59,** 631–660.
4. Richards, F. M. (1992) in *Protein Folding*, ed. Creighton, T. E. (Freeman, New York), pp. 1–58.
5. Ptitsyn, O. B. (1987) *J. Protein Chem.* **6,** 273–293.
6. Ptitsyn, O. B. (1992) in *Protein Folding*, ed. Creighton, T. E. (Freeman, New York), p. 243–300.
7. Ptitsyn, O. B. (1994) *Protein Eng.* **7,** 593–596.
8. Karplus, M. & Weaver, D. L. (1994) *Protein Sci.* **3,** 650–668.
9. Dill, K. A., Fiebig, K. M. & Chan, H. S. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 1942–1946.
10. Ptitsyn, O. B. (1991) *FEBS Lett.* **285,** 176–181.
11. Karplus, M. & Weaver, D. C. (1976) *Nature (London)* **260,** 404–406.
12. Kim, P. S. & Baldwin, R. L. (1982) *Annu. Rev. Biochem.* **51,** 459–489.
13. Wetlaufer, D. B. (1973) *Proc. Natl. Acad. Sci. USA* **70,** 697–701.
14. Moult, J. & Unger, R. (1991) *Biochemistry* **30,** 3816–3824.
15. Harrison, S. C. & Durbin, R. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 4028–4030.
16. Sali, A., Shakhnovich, E. & Karplus, M. (1994) *J. Mol. Biol.* **235,** 1614–1636.
17. Matouschek, A., Kellis, J. T., Jr., Serrano, L. & Fersht, A. R. (1989) *Nature (London)* **342,** 122–126.
18. Matouschek, A., Kellis, J. T., Jr., Serrano, L., Bycroft, M. & Fersht, A. R. (1990) *Nature (London)* **346,** 440–445.
19. Matouschek, A., Serrano, L. & Fersht, A. R. (1992) *J. Mol. Biol.* **224,** 819–835.
20. Fersht, A. R., Matouschek, A. & Serrano, L. (1992) *J. Mol. Biol.* **224,** 771–782.
21. Serrano, L., Kellis, J. T., Cann, P., Matouschek, A. & Fersht, A. R. (1992) *J. Mol. Biol.* **224,** 783–804.
22. Serrano, L., Matouschek, A. & Fersht, A. R. (1992) *J. Mol. Biol.* **224,** 805–818.
23. Serrano, L., Matouschek, A. & Fersht, A. R. (1992) *J. Mol. Biol.* **224,** 847–859.
24. Fersht, A. R. (1993) *FEBS Lett.* **325,** 5–16.
25. Caflisch, A. & Karplus, M. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 1746–1750.
26. McPhalen, C. A. & James, M. N. G. (1987) *Biochemistry* **26,** 261–269.
27. Harpaz, Y., ElMasry, N. F., Fersht, A. R. & Henrick, K. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 311–315.
28. Clore, G. M., Gronenborn, A. M., Kjær, M. & Poulsen, F. M. (1987) *Protein Eng.* **1,** 305–311.
29. Clore, G. M., Gronenborn, A. M., James, M. N. G., Kjær, M., McPhalen, C. A. & Poulsen, F. M. (1987) *Protein Eng.* **1,** 313–318.
30. Kjær, M. & Poulsen, F. M. (1987) *Carlsberg Res. Commun.* **52,** 355–362.
31. Ludvigsen, S., Shen, H., Kjær, M., Madsen, J. C. & Poulsen, F. M. (1991) *J. Mol. Biol.* **222,** 621–635.
32. Jackson, S. E. & Fersht, A. R. (1991) *Biochemistry* **30,** 10428–10435.
33. Jackson, S. E. & Fersht, A. R. (1991) *Biochemistry* **30,** 10436–10443.
34. Jackson, S. E., Moracci, M., ElMasry, N., Johnson, C. M. & Fersht, A. R. (1993) *Biochemistry* **32,** 11259–11269.
35. Jackson, S. E., ElMasry, N. & Fersht, A. R. (1993) *Biochemistry* **32,** 11270–11278.
36. Li, A. & Daggett, V. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 10430–10434.
37. Richardson, J. S. & Richardson, D. C. (1988) *Science* **240,** 1648–1652.
38. Fersht, A. R., Itzhaki, L. S., ElMasry, N., Matthews, J. M. & Otzen, D. E. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 10426–10429.
39. Kippen, A., Sancho, J. & Fersht, A. R. (1994) *Biochemistry* **33,** 3778–3786.
40. Yanagawa, H., Yoshida, K., Torigoe, C., Park, J. S., Sato, K., Shirai, T. & Go, M. (1993) *J. Biol. Chem.* **268,** 5861–5865.