doi:10.1016/j.jmb.2008.09.044 J. Mol. Biol. (2008) 384, 531–539

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

The High-Resolution NMR Structure of the Early Folding Intermediate of the *Thermus thermophilus* Ribonuclease H

Zheng Zhou[1]†, Hanqiao Feng[1]†, Rodolfo Ghirlando[2] and Yawen Bai[1]*

[1]*The Laboratory of Biochemistry and Molecular Biology, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA*

[2]*The Laboratory of Molecular Biology, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892, USA*

Received 15 July 2008; received in revised form 16 September 2008; accepted 17 September 2008
Available online 26 September 2008

Edited by P. Wright

Elucidation of the high-resolution structures of folding intermediates is a necessary but difficult step toward the ultimate understanding of the mechanism of protein folding. Here, using hydrogen-exchange-directed protein engineering, we populated the folding intermediate of the *Thermus thermophilus* ribonuclease H, which forms before the rate-limiting transition state, by removing the unfolded regions of the intermediate, including an α-helix and two β-strands (51 folded residues). Using multidimensional NMR, we solved the structure of this intermediate mimic to an atomic resolution (backbone rmsd, 0.51 Å). It has a native-like backbone topology and shows some local deviations from the native structure, revealing that the structure of the folded region of an early folding intermediate can be as well defined as the native structure. The topological parameters calculated from the structures of the intermediate mimic and the native state predict that the intermediate should fold on a millisecond time scale or less and form much faster than the native state. Other factors that may lead to the slow folding of the native state and the accumulation of the intermediate before the rate-limiting transition state are also discussed.

Published by Elsevier Ltd.

Keywords: protein folding intermediate; protein folding; NMR structure; hydrogen exchange; protein engineering

Introduction

A major question in protein folding has been how protein molecules find their native states in a vast possible conformation space on a biologically meaningful time scale.[1] One hypothesis suggests that they solve the conformation search problem by folding through partially unfolded intermediates.[2] Indeed, a large number of proteins have been reported to have partially unfolded intermediates on their folding pathways.[3–6] However, the key features of the folding intermediates and their exact roles in protein folding are still not fully understood. For example, does the folded region of a folding intermediate have a specific structure or comprise an ensemble of very different structures?[7–9] Is the folding intermediate native-like or non-native-like? Why do some proteins populate early folding intermediates before the rate-limiting transition states, whereas the intermediates of others exist after the rate-limiting transition states?[10] These issues are difficult to resolve without high-resolution structural information on the folding intermediates.

The main obstacle to obtaining high-resolution structural information of folding intermediates is that protein folding is a fast process and folding intermediates can only populate transiently during folding. The available methods for protein structure determination, such as X-ray crystallography and multidimensional NMR, are not directly applicable. In addition, partially unfolded intermediates do not populate significantly under equilibrium native conditions because the native state has the lowest free energy. Moreover, partially unfolded intermediates at high protein concentration (approximately millimolar), which are required for their structure

*Corresponding author. E-mail address: yawen@helix.nih.gov.

† Z.Z. and H.F. contributed equally to this work.

Abbreviations used: Rd-apocyt b_{562}, redesigned apocyt b_{562}; En-HD, engrailed homeodomain; RNase H, ribonuclease H.

0022-2836/$ - see front matter. Published by Elsevier Ltd.

determination by the current solution NMR methods, tend to aggregate due to the exposure of hydrophobic side chains.

Despite such difficulties, some success toward determining the structures of folding intermediates has been achieved by the use of peptide models to represent the folding intermediates associated with the formation of disulfide bonds[11] and by populating partially unfolded states at low pH and equilibrium conditions.[12] In these cases, secondary structure for the intermediates have been obtained; information for their tertiary structures, however, is still lacking. More recently, the high-resolution structures of the folding intermediates of a redesigned apocyt b_{562} (Rd-apocyt b_{562})[13,14] and T_4 lysozyme,[15] which exist after the rate-limiting steps, and the early folding intermediate of engrailed homeodomain (En-HD),[16] which exists before the rate-limiting step, have been determined. In these three cases, the structures of the folding intermediates were first characterized with the use of hydrogen-exchange and mutation studies at the level of residues and, subsequently, protein engineering was used to populate the intermediates by deleting or mutating the residues in the unfolded regions of the intermediates.[17–20] A common feature of the structures of these intermediates is that they have native-like backbone topology with local non-native side-chain interactions.

Although the high-resolution structure of the early folding intermediate of En-HD has provided significant insights into the early folding events,[16] En-HD is small (61 amino acids) and its native state can form on submillisecond time scale as the intermediate does, making it difficult to reveal the cause for the population of the intermediate. By contrast, many larger proteins (>100 amino acids) have submillisecond early folding intermediates and fold to the native state on the time scale of seconds. Therefore, a detailed study of the structural features and roles of the intermediates of these proteins may help to reveal the general principles that govern protein folding. One of the well-studied proteins with early folding intermediates is ribonuclease H (RNase H). For the past decade, Marqusee and coworkers have characterized the folding behavior of several homologues of RNase H, including those from *Escherichia coli*,[21] *Thermus thermophilus*,[22] and HIV-1.[23] Several experimental results indicate that these proteins fold similarly with on-pathway early intermediates, although the folded regions in the intermediates of these proteins are not identical: (i) the formation of a significant amount of secondary structure was detected in the dead time of the stopped-flow circular dichroism experiments;[22,24] (ii) the structures of the intermediates were characterized in the kinetic hydrogen/deuterium-exchange pulse-labeling experiments;[22,24] (iii) two partially unfolded intermediates were detected in the native-state hydrogen-exchange experiments and one of them was similar to that identified in the pulse-labeling experiment for both *E. coli* and *T. thermophilus* RNase H;[21,25] (iv) the kinetic intermediate identified in the hydrogen/deuterium-exchange method was confirmed by a protein engineering study for *E. coli* RNase H;[26] (v) a discrete on-pathway intermediate with properties similar to that of the kinetic intermediate was identified in a single-molecule experiment also for *E. coli* RNase H.[27]

Previously, Chamberlain *et al.* used hydrogen-exchange-directed protein engineering to populate the folding intermediate of *E. coli* RNase H by deleting the regions that are believed to be unfolded in the intermediate.[28] This intermediate mimic forms a dimer ($K_d = 0.5$ μM) at high protein concentrations. Here, we used a similar protein engineering approach to populate the folding intermediate of the cysteine-free *T. thermophilus* RNase H (RNase H*), which was used in the earlier folding studies.[22] This intermediate involves more folded regions than the intermediate of *E. coli* RNase H. We found that the intermediate mimic is highly soluble and exists as a monomer at ~1 mM concentration, which allows us to use multidimensional NMR to solve its structure to a high resolution and to address the issues concerning the early intermediates.

Results

Identification of the unfolded region of the early folding intermediate

The native structure of *T. thermophilus* RNase H* is composed of five α-helices (A to E) and five β-strands (I to V) (Fig. 1a). In the earlier pulse-labeling experiment, seven amide protons in strand II are fully labeled in the early intermediate, indicating that strand II is unfolded in the intermediate.[22] There are two amide protons in strand III that can be used as probes in the pulse-labeling experiment; both are labeled. No amide protons in the E-helix can be used as probes to monitor its structure in the intermediate because they exchange too fast in the native state. Nevertheless, it is reasonable to conclude that strand III and the E-helix are unfolded in the intermediates, since strand III is stabilized by forming hydrogen bonds with strand II and the E-helix mainly packs against strand II. By contrast, several amide protons in helices A and D and strand IV are fully protected in the intermediate, indicating they are folded in the intermediate. Strand I has one amide proton as probe, which is partially protected. Strand V also has one amide proton as probe, which is fully protected. These results indicate that strands I and V cannot be fully unfolded in the intermediate. The amide protons in helices B and C also exchange too fast in the native state to provide probes for characterizing their structures in the intermediate. However, an earlier protein engineering study on the folding of *E. coli* RNase H indicates that they are likely folded in the intermediate.[26]

The above conclusions are further supported by the native-state hydrogen-exchange experiment, in which two partially unfolded intermediates were identified.[25] One intermediate has the E-helix un-