Published in final edited form as:

Biopolymers. 2018 August ; 109(8): e23086. doi:10.1002/bip.23086.

The burst-phase folding intermediate of ribonuclease H changes conformation over evolutionary history

Shion An Lim^{a,b} and Susan Marqusee^{a,b}

^aDepartment of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA, United States

^bInstitute for Quantitative Biosciences (QB3), University of California, Berkeley, Berkeley, CA, United States

Abstract

The amino acid sequence encodes the energy landscape of a protein. Therefore, we expect evolutionary mutations to change features of the protein energy landscape, including the conformations adopted by a polypeptide as it folds to its native state. Ribonucleases H (RNase H) from *E. coli* and *T. thermophilus* both fold via a partially folded intermediate in which the core region of the protein (helices A-D and strands 4-5) is structured. Strand 1, however, uniquely contributes to the *T. thermophilus* RNase H folding intermediate ($I_{\text{core}+1}$), but not the *E. coli* RNase H intermediate (I_{core}) (Rosen & Marqusee, *PLoS One* 2015). We explore the origin of this difference by characterizing the folding intermediate of seven ancestral RNases H spanning the evolutionary history of these two homologs. Using fragment models with or without strand 1 and FRET probes to characterize the folding intermediate of each ancestor, we find a distinct evolutionary trend across the family—the involvement of strand 1 in the folding intermediate is an ancestral feature that is maintained in the thermophilic lineage and is gradually lost in the mesophilic lineage. Evolutionary sequence changes indeed modulate the conformations present on the folding landscape and altered the folding trajectory of RNase H.

Graphical abstract

Keywords

protein folding; energy landscape; protein evolution; ancestral sequence reconstruction; folding intermediates

Introduction

A major effort in protein biophysics is to understand how the amino acid sequence encodes a protein's energy landscape. This energy landscape represents all of the conformations that a given polypeptide chain can adopt, their relative stabilities, and the dynamics of inter-conversion between these states.¹ Importantly, this landscape also encodes the protein's folding pathway—the process by which an unfolded polypeptide navigates its energy landscape to the native state, which is usually its thermodynamically most stable conformation.² A deeper understanding of how the amino acid sequence encodes a protein's energy landscape and the evolutionary pressures that act on this landscape will allow us to more accurately predict the function of uncharacterized proteins, provide tools to engineer protein properties of interest, and better understand when and how protein folding goes wrong.

This multi-dimensional energy landscape ultimately defines a protein's biological activity and organismal fitness, and, as such, is subject to selection over the course of molecular evolution.^{3,4} Mutations will alter specific features of the landscape in response to changing evolutionary pressures, while maintaining the overall fitness of the protein.^{5,6} While the native structure of a protein is fairly robust to small sequence changes (typically two homologs with approximately >30% sequence identity will adopt the same fold), other regions of the landscape, such as partially folded states, may be more labile to mutation.^{7,8} These high-energy states are known to contribute to a protein's function, folding, and fitness.⁹ To date, however, we lack a clear understanding of how evolutionary processes modulate a

protein's energy landscape and thus generate the diversity of properties observed across the proteome.

One approach to understanding the subtleties encoded in the sequence of a protein has been to characterize the biophysical properties of homologous proteins. These studies have revealed how proteins with nearly identical folds can have different stabilities, folding pathways, and conformational states.^{7,10–12} One such well-studied example is *E. coli* RNase H (ecRNH*) and *T. thermophilus* RNase H (ttRNH*). (The asterisk (*) refers the cysteine-free variant of the protein, which has been used in the majority of biophysical studies of RNases H.¹³) While these two proteins catalyze the same reaction and have nearly identical folds, ttRNH*, which is from a thermophilic organism, is more stable across a wide range of temperatures than ecRNH*, which is from a mesophilic organism.¹⁴ Both ecRNH* and ttRNH* are known to fold through a three-state mechanism with a rapidly forming partially folded intermediate that forms within the dead time of the stopped-flow (milliseconds), followed by a slower rate-limiting step that occurs on the order of seconds ($U \rightleftharpoons I \rightleftharpoons N$).^{15,16} NMR, hydrogen exchange, and mutational studies have revealed that the α -helical core region of the protein involving helices A-D and strands 4 and 5 (residues 43 to 122, ecRNH* numbering) is structured in the folding intermediate, while the rest of the protein, the periphery, remains unstructured until the rate-limiting step (Figure 1A).^{16–20} A fragment of RNase H, with only residues 43 to 122 of the full-length sequence, referred to as I_{core} , was created as a structural model of the folding intermediate. This fragment folds autonomously and can be used as an equilibrium model of the transient folding intermediate.^{18,20,21}

Curiously, recent studies have suggested that the first β -strand (Strand 1) can dock on to this I_{core} region in ttRNH* but not in ecRNH*, and that this first β -strand is structured in the folding intermediate of ttRNH*, but not of ecRNH*.^{21,22} A mimic of this intermediate ($I_{\text{core}+1}$), created using an non-natural linker between strand 1 and helix A, was found to fold autonomously and had a distinctly higher stability than I_{core} for ttRNH*.²¹ This implies that strand 1 contributes to the structure and stability of $I_{\text{core}+1}$, and therefore I_{core} and $I_{\text{core}+1}$ are separate minima on the energy landscape of ttRNH*. Kinetic experiments using FRET confirmed that $I_{\text{core}+1}$ is the likely structure of the ttRNH* folding intermediate (Figure 1B). In contrast, for ecRNH*, addition of strand 1 to create $I_{\text{core}+1}$ has an indistinguishable stability from I_{core} , indicating that strand 1 does not contribute any stability to the core and is likely unstructured in ecRNH*.²¹ This, along with additional experiments, suggest that the folding intermediate of ecRNH* is comprised of I_{core} and does not involve strand 1 (Figure 1B).^{16–18,20,21}

Here, we have explored the origin of this structural difference by harnessing information from evolutionary history. Previously, we employed ancestral sequence reconstruction (ASR) to reconstruct the ancestral proteins along the evolutionary lineages of these two RNase H homologs (Figure 1B).^{23,24} ASR is a computational technique that involves creating a phylogenetic tree of a protein family and applying an evolutionary model to infer the most likely protein sequence at various “nodes” along the phylogenetic tree.^{25–28} These resurrected proteins can then be generated and characterized in the laboratory. ASR has been used for a variety of proteins; it has yielded insights into the properties of ancient proteins, uncovered evolutionary trends in biophysical properties, and provided a tool to identify

sequence determinants of the energy landscape.^{29–37} For the RNase H family, we reconstructed the last common ancestor of ttRNH* and ecRNH* (Anc1*), two ancestors along the thermophilic lineage (Anc2* and Anc3*), and four ancestors along the mesophilic lineage (AncA*, AncB*, AncC*, AncD*) (Figure 1B).^{23,24} Previous thermodynamic and kinetic analyses of these proteins revealed distinct evolutionary trends in the kinetics and thermodynamic parameters of the landscape, and confirmed that all of these ancestral proteins fold via a folding intermediate, a remarkable conservation of the three-state folding pathway over ~3 billion years of evolution.^{23,24} While these studies reveal a conservation of the folding mechanism, they do not divulge structural features of these intermediates, so whether the folding pathway of RNase H was *structurally* conserved was not addressed.

Here, we generate and characterize fragment models of the folding intermediate (I_{core} and $I_{\text{core}+1}$) for each ancestral protein along the lineages of ttRNH* and ecRNH* to determine 1) what conformations are present on the RNase H energy landscape over evolution, and 2) which partially folded structure (I_{core} or $I_{\text{core}+1}$) each ancestral protein populates *en route* to its native state. We find a distinct evolutionary trend in the structure of the folding intermediate across the two RNase H lineages—the involvement of strand 1 in the folding intermediate is an ancestral feature of RNase H that is maintained in the thermophilic lineage, and is gradually lost in the mesophilic lineage. Our study reveals evidence that sequence changes over evolution indeed alter the conformations along the folding pathway of a protein, and provide a system to address how selective pressures might modulate the different features of the protein energy landscape.

Materials & Methods

Constructs of RNase H variants

I_{core} fragments were subcloned from full-length constructs of the corresponding ancestral protein.²⁴ Insertion of the strand 1 sequence to create $I_{\text{core}+1}$ from I_{core} was done via round-the-horn PCR. Single-site mutations were generated by site-directed mutagenesis and were sequence verified by Sanger sequencing.

Protein expression and purification

The I_{core} and $I_{\text{core}+1}$ fragments for the different ancestral proteins were recombinantly over-expressed in *E. coli* and purified as described previously.^{18,20,21} The purity and mass of the purified proteins were confirmed by SDS-PAGE and mass spectrometry. Anc1*-K4C-W22Y was expressed in *E. coli* and purified using established full-length RNase H purification protocols.¹⁶

CD Equilibrium Experiments

CD spectra and urea denaturation melts were measured using an Aviv 410 CD spectrophotometer. For spectra, the CD signal from 200 nm to 300 nm was collected in a 0.1-cm path length cuvette at 25°C. Samples contained 0.4 mg/mL of protein in 20 mM sodium acetate 50 mM potassium chloride pH 5.5 (RNase H Buffer). All CD spectra were blanked with RNase H buffer in the absence of protein. For urea-denaturation studies, samples containing 40 µg/mL of protein and varying [urea] in RNase H buffer were

equilibrated overnight. The CD signal at 222 nm was measured in a 1-cm path length cuvette at 25°C with stirring. Melts were obtained in triplicate, and the data were fit to a two-state model with a linear free-energy extrapolation using Igor Pro.³⁸

TNB Labeling

TNB labeling of the cysteine residue in Anc1*-K4C-W22Y was conducted as described previously.²¹ Mass spectrometry confirmed ~100% labeling efficiency.

Fluorescence Spectra

Fluorescence spectra of TNB-labeled and unlabeled Anc1*-K4C-W22Y were obtained on a FluoroMax-3 fluorimeter. Intrinsic tryptophan fluorescence at 25°C was measured in a 1-cm path length cuvette with stirring using an excitation wavelength of 295 nm and collecting emission from 300 nm to 450 nm in 1 nm intervals. Protein concentration was 0.04 mg/mL and all spectra were buffer corrected.

Kinetic Experiments

Kinetic experiments monitored by fluorescence were performed on a Biologic SFM-400 stopped-flow instrument. Unfolded protein at 6-8 mg/mL in high [urea] in RNase H buffer was diluted 10-fold into final refolding conditions (RNase H buffer, varying concentrations of [urea]) at 25°C. Fluorescence signal was monitored by excitation at 295 nm, and the emission was collected using a 375/10 nm band-pass filter. The dead time of the stopped-flow was 7.3 milliseconds.

Kinetic experiments monitored by CD were performed on an Aviv 202 stopped-flow spectrophotometer or by manual mixing on an Aviv 410 spectrophotometer. For stopped-flow refolding experiments, unfolded protein at 6-8 mg/mL in high [urea] in RNase H buffer was diluted 11-fold into final refolding conditions (RNase H buffer, varying concentrations of [urea]) and CD signal at 222 nm was monitored at 25°C in a 0.1-cm path length cuvette. The dead time for the stopped-flow CD was 18 milliseconds. For manual mixing refolding or unfolding experiments, either unfolded or folded protein, respectively, at 1.5 mg/mL was manually diluted 30-fold into RNase H buffer with varying concentrations of [urea] and CD signal at 222 nm was monitored at 25°C in a 1-cm path length cuvette, with stirring. The dead time for manual mixing experiments was ~10 seconds.

Results

I_{core} and I_{core+1} fragments of ancestral RNases H

We generated fragment mimics, I_{core} and I_{core+1}, for each of the RNase H ancestral proteins (Figure 1B). For each ancestral protein, the sequence defining I_{core} and I_{core+1} was determined from the multiple sequence alignment of the RNase H family and corresponds to the same regions as the extant proteins, ecRNH* and ttRNH*. I_{core} fragment spans residues 43 and 122 (ecRNH* numbering), and the I_{core+1} fragment has strand 1 (residues 1 to 20) as an N-terminal addition before helix A, the first helix of the core.^{18,22} The sequence spanning the I_{core+1} fragment is non-contiguous, so an unnatural junction between strand 1 and helix A is created for this construct (the last residue of strand 1 (residue 20) is followed by the

first residue of helix A (residue 43)).²² All fragment mimics expressed and purified to homogeneity using previously established methods for expression and purification.^{18,20,21}

All of the fragments generated (I_{core} and $I_{\text{core}+1}$ for each ancestral protein) form autonomous folding units as determined by circular dichroism spectroscopy (CD). All spectra show minima at 208 nm and 222 nm, consistent with a largely α -helical structure (see Figure 2A for an example).

The last common ancestor, Anc1*, populates two different partially folded states on the energy landscape

To determine whether Anc1*, the last common ancestor between ecRNH* and ttRNH*, populates two different partially folded structures on its energy landscape, we monitored the thermodynamic stabilities of the fragments using CD. Equilibrium urea-denaturation melts, which resulted in cooperative sigmoidal curves, fit well to a two-state model, yielding stabilities (G_{unf}) and associated m -values (Figure 2B).³⁸ As evident from the denaturation profiles, Anc1* I_{core} and $I_{\text{core}+1}$ have different stabilities and m -values, indicating that they are distinct autonomously folded structures that are representative of different minima on the energy landscape, similar to results from the fragments of ttRNH* (Figure 2B, Table 1).²¹

To assess which fragment model, I_{core} or $I_{\text{core}+1}$, is more representative of the kinetic folding intermediate of Anc1*, we compared the stability of the fragments to the calculated stability of the folding intermediate determined from kinetic refolding experiments of the full-length protein.³⁹ We used the data from a global three-state analysis of the kinetic chevron plot, which yielded the kinetic and thermodynamic parameters of the three-state folding landscape of each ancestral protein ($U \rightleftharpoons I \rightleftharpoons N$).²⁴ We then compared the C_m , the [urea] at which half the protein is unfolded, between the fragments and the burst-phase intermediate from the kinetic analysis. Note that we used the C_m rather than G_{unf} because Anc1* I_{core} does not have a well-defined folded baseline and C_m , which corresponds to the inflection point of a sigmoidal curve, is more robust to errors in the fit than G_{unf} . The C_m of the folding intermediate of Anc1*, 3.71 ± 0.35 M, matches closely to the C_m of $I_{\text{core}+1}$ (3.56 ± 0.09 M), but not of I_{core} (2.09 ± 0.17 M) (Figure 1B, Table 1). Thus, the properties of the $I_{\text{core}+1}$ fragment appear to be more consistent with the kinetic folding intermediate populated during refolding for Anc1*.

$I_{\text{core}+1}$ as the folding intermediate is maintained in the thermophilic lineage

Characterization of the I_{core} and $I_{\text{core}+1}$ fragments of ancestors along the thermophilic lineage revealed that, similar to Anc1* and ttRNH*, I_{core} and $I_{\text{core}+1}$ are two different minima on the energy landscape for the thermophilic ancestors. The CD spectra of the fragments of Anc2* and Anc3* show they are all well-folded, largely α -helical structures (Figure 3A, 3B). Equilibrium urea-denaturation curves indicate that I_{core} and $I_{\text{core}+1}$ have different stabilities and m -values, indicating they are distinct autonomously folded structures representing different minima on the energy landscape for both Anc2* and Anc3* (Figure 3A, 3B, Table 1). Additionally, we compared the stabilities of I_{core} and $I_{\text{core}+1}$ to the stability of the kinetic folding intermediate from refolding experiments (as described previously for Anc1*). The C_m of the Anc2* folding intermediate (3.26 ± 0.60 M) matches closely to the

C_m of the I_{core+1} fragment (3.45 ± 0.04 M) and not to the I_{core} fragment (1.67 ± 0.20 M). The C_m of the Anc3* folding intermediate (4.31 ± 0.58 M) matches closely to the C_m of the I_{core+1} fragment (4.50 ± 0.07 M) and not to the I_{core} fragment (3.23 ± 0.11 M). Thus, I_{core+1} appears to be a better model of the kinetic folding intermediate populated during refolding for all proteins along the thermophilic lineage.

There is a gradual shift in the intermediate structure from I_{core+1} to I_{core} along the mesophilic lineage

We then characterized fragment models of ancestors along the mesophilic lineage. Although ecRNH* folds via I_{core} , our studies on Anc1* suggest that the folding intermediate of its oldest ancestor, involves strand 1.^{16–18,21} Characterizing the properties of I_{core} and I_{core+1} fragments of ancestors along the mesophilic lineage might uncover an evolutionary pattern of interest.

The I_{core} and I_{core+1} fragments of the ancestors along the mesophilic lineage reveal that the two distinct partially folded states on the energy landscape gradually converge. The CD spectra of all I_{core} and I_{core+1} fragments of the mesophilic ancestors, AncA*, AncB*, AncC*, AncD* show they are well-folded, largely α -helical structures (Figure 4A–D). Equilibrium urea-denaturation curves show that for AncA*, I_{core} and I_{core+1} have different stabilities, corresponding to distinct minima on the energy landscape (Figure 4A, Table 1). In fact, the I_{core} fragment of AncA* was quite destabilized with no folded baseline, and was thus not fit to a two-state model. The C_m of the folding intermediate from kinetic experiments (1.75 ± 0.34 M) overlaps with the C_m of AncA* I_{core+1} (1.57 ± 0.05 M) but not to I_{core} ($C_m \sim 0$ M), suggesting that strand 1 is involved in the folding intermediate of AncA*. For the next ancestor along the mesophilic lineage, AncB*, I_{core} and I_{core+1} have different stabilities, showing that these two structures remain distinct minima on the energy landscape (Figure 4B, Table 1). The C_m of the AncB* folding intermediate from kinetics (3.95 ± 0.93 M) is more similar to the C_m of I_{core+1} (4.44 ± 0.06 M) than of I_{core} (3.19 ± 0.16 M). However, since the errors overlap, the identity of the folding intermediate cannot be determined unambiguously. A similar outcome is found for AncC*, but the stabilities of I_{core} and I_{core+1} are even more similar to each other, and the identity of the folding intermediate again cannot be determined unambiguously (Figure 4C, Table 1). The stabilities of I_{core} and I_{core+1} for the final ancestor along the mesophilic lineage, AncD*, are nearly identical, indicating they are no longer distinct minima on the energy landscape (Figure 4D, Table 1). The addition of strand 1 to the helical core of AncD* does not appear to affect the stability nor the m -value, indicating that AncD* does not populate the I_{core+1} intermediate, similar to ecRNH*. The C_m of the AncD* folding intermediate from kinetic experiments overlaps with the C_m of the AncD* fragments, indicating that I_{core} is the likely intermediate of AncD*.

Strand 1 docks before the rate-limiting step during Anc1* folding

To further confirm that the folding intermediate of the last common ancestor, Anc1*, involves strand 1 in addition the α -helical core, we employed FRET labeling to monitor strand 1 docking to the core during refolding of the full-length protein.²¹ If strand 1 is involved in the structure of the folding intermediate, which forms within milliseconds of

refolding, then strand 1 should come in contact with the core on this milliseconds time scale, much faster than the subsequent rate-limiting step of global folding when the periphery folds (~seconds to minutes).

To assess this, residue 4 on strand 1 of Anc1* (ecRNH* numbering) was mutated to cysteine (K4C) and labeled with thionitrobenzoate (TNB) (Figure 1A). The TNB is used to quench the intrinsic fluorescence of the cluster of tryptophan residues located in the α -helical core region of the protein in a distance-dependent manner. A tryptophan on strand 2 of Anc1* (W22Y) was mutated to tyrosine to remove any contribution from the periphery of the protein. This is the same approach that was used previously on tRNH* to confirm that strand 1 is involved in its folding intermediate.²¹

TNB-labeled Anc1* was characterized by CD to confirm that the mutations and the labeling did not significantly perturb the protein. The CD spectrum of TNB-labeled Anc1* indicates that the overall fold of the protein is unchanged from wild-type Anc1* (Figure 5A). Equilibrium urea-denaturation melts yield a cooperative transition that fit well to a two-state model, with $G_{\text{unf}} = 6.80 \pm 0.6 \text{ kcal mol}^{-1}$ and m -value of $1.82 \pm 0.16 \text{ kcal mol}^{-1} \text{ M}^{-1}$, a $3.4 \text{ kcal mol}^{-1}$ destabilization from wild-type Anc1* (Figure 5B). Since there is no change in CD spectra or the m -value and cooperativity of the protein, TNB-labeled Anc1*-K4C-W22Y appears to be a destabilized, but properly folded variant of Anc1*.

We confirmed that TNB on strand 1 quenches the intrinsic fluorescence of the tryptophan residues in the core region of the protein in the folded state but not when the protein is unfolded, and thus can be used as a probe for refolding kinetics (Figure 5C). We chose to monitor kinetics at 375 nm, since there is a large difference in the fluorescence signal between the folded and unfolded state for the labeled protein, but there is little difference in signal between folded and unfolded states for the unlabeled protein. Thus, changes in signal at 375 nm during refolding should largely report on the FRET-based quenching of tryptophan fluorescence by TNB, and correspond to the process of strand 1 docking to the core.

Refolding kinetic traces of TNB-labeled Anc1* at different [urea] were monitored by both fluorescence (monitors strand 1 docking to core) and CD (monitors global folding). Fluorescence refolding traces show a large decrease in signal that occurs within the dead-time of the stopped-flow fluorimeter (~7.3 milliseconds) (Figure 6A). This decrease in fluorescence is not observed with unlabeled protein, indicating that this fast phase is unique to quenching by TNB (Figure 6B). Although we cannot capture the transient kinetics of TNB quenching, we can use the dead time of the stopped-flow (7.3 milliseconds) to determine that the rate of TNB quenching is quite fast ($k_{\text{quench}} \gg 136 \text{ sec}^{-1}$) (Figure 6D).

Refolding monitored by CD using manual mixing and stopped-flow shows that global folding is significantly slower than TNB quenching (Figure 6C, 6D). Similar to wild-type Anc1*, upon refolding of TNB-labeled Anc1*, there is a burst phase in the CD signal within the dead time of the stopped-flow instrument (18 milliseconds) that corresponds to the formation of the folding intermediate. This is followed by an observable phase corresponding to global folding to the native state over the rate-limiting step. This takes

place in seconds to minutes for TNB-labeled Anc1*, on the same time scale as wild-type Anc1* (Figure 6C, 6D). Thus, consistent with the fragment data, TNB kinetic experiments show that strand 1 contacts the core well before the rate-limiting step and is likely involved in the structure of the Anc1* folding intermediate.

Discussion

In this study, we turn to ancestral proteins to investigate the evolutionary significance of a structural difference in the energy landscapes of two homologs, *E. coli* RNase H and *T. thermophilus* RNase H. By generating fragment mimics of the core folding intermediate with or without strand 1 for each ancestral protein along the lineages of ecRNH* or ttRNH*, we determined whether I_{core} and $I_{\text{core}+1}$ form two distinct partially folded states on the energy landscape of each ancestor. Additionally, by comparing the stabilities of the fragments to the calculated stability of the folding intermediate in the full-length protein, we determined which structure, I_{core} or $I_{\text{core}+1}$, is representative of the partially folded state populated before the rate-limiting step of folding. Our results indicate a clear change in the conformations populated on the RNase H energy landscape over ~3 billion years of evolution, and provides an evolutionary record of how sequence changes alter a protein's folding pathway.

The conformations on the energy landscape change over RNase H evolution

Previous studies revealed a clear difference between the energy landscapes of two homologous proteins, ecRNH* and ttRNH*. ²¹ ttRNH* populates two distinct partially folded states on the energy landscape (I_{core} and $I_{\text{core}+1}$), whereas ecRNH* is able to populate only one of these states (I_{core}). We find that, similar to ttRNH*, I_{core} and $I_{\text{core}+1}$ are distinct minima on the energy landscape of Anc1*, the last common ancestor of these two homologs. This ancestral feature remains preserved in the thermophilic lineage, as Anc2* and Anc3* exhibit similar behavior. Along the mesophilic lineage, we find a gradual convergence in the stability of the I_{core} and $I_{\text{core}+1}$ states, such that by AncD*, the most recent ancestor of ecRNH*, there is no difference in energetics between the presence or absence of strand 1. This indicates that mutations along the mesophilic lineage resulted in the inability of strand 1 to dock to the core region of the protein without the rest of the periphery.

There are subtle changes in the shape of the CD spectra and m -values of the I_{core} and $I_{\text{core}+1}$ fragments across the different ancestral proteins. Although we do not attempt to elucidate the molecular details of these differences in spectral shape and solvent-accessible surface area, they may be indicative of subtle differences in the conformations of these partially folded states across the RNase H family that future experiments using, for example, NMR or hydrogen exchange, may uncover.

The structure of the folding intermediate changes over RNase H evolution

The folding trajectories of the ancestral RNase H proteins were previously characterized by kinetic studies, which yielded a detailed description of the energetics, but not the structural details of the folding pathway. ²⁴ By comparing the stability of the folding intermediate to the fragment models, we were able to infer which structure each ancestral protein folds

through. Anc1*, the last common ancestor, the ancestors and homologs along the thermophilic lineage (Anc2*, Anc3*, and tRNH*), and the first ancestor along the mesophilic lineage (AncA*), all fold via a folding intermediate that involves the core and strand 1. The identity of the folding intermediate for AncB* and AncC* cannot be unambiguously determined because the stabilities of I_{core} and $I_{\text{core}+1}$ are too similar. However, strand 1 does not appear to be involved in the folding intermediate of AncD* and ecRNH*. There appears to be a clear shift in the folding intermediate structure along the mesophilic lineage.

I_{core} and $I_{\text{core}+1}$ are truncations of the respective full-length protein, and although they contain the predicted structured regions of the protein in the folding intermediate, they are fragment models that involve deletions and non-natural junctions. It is possible that these autonomously folded fragments at equilibrium do not fully capture the conformation of the kinetic folding intermediate. There may be some role of the apparently disordered periphery, and there may be non-native or partially formed interactions in the folding intermediate during refolding. The NMR structure of $I_{\text{core}+1}$ of tRNH*, shows however, a native-like conformation of strand 1.²² Additionally, corroborating data obtained by NMR, mutations, hydrogen exchange, and FRET-based TNB quenching experiments suggest that these fragments are consistently good models of the intermediate for RNases H, and are useful tools to study the properties of transiently formed species that typically cannot be captured in a kinetic experiment.^{16–18,20,21}

The RNase H ancestors and homologs are between 52 and 93% identical in sequence to each other, with ecRNH* and tRNH* being the most dissimilar.²³ The core region is more conserved than the periphery across the proteins.²⁴ Our thermodynamic studies and a previously solved NMR structure of tRNH* $I_{\text{core}+1}$ suggest that the folding intermediate likely adopts a structure similar to the fully folded state.^{16–18,22} Thus, we can expect that in the folding intermediate, strand 1 makes contacts with residues in helix A and strand 5, and mutations at that interface are likely to contribute to the loss of strand 1 docking in the mesophilic lineage. However, the gradual convergence of the two partially folded states along the mesophilic lineage suggests that there is not a single mutation, but rather, a series of mutations that are responsible for the undocking of strand 1. This, along with epistatic effects, are likely to complicate the identification of sequence determinants for the early docking of strand 1.⁴⁰ Preliminary analyses of sequence alignments have not identified obvious sequence determinants— future mutational studies will be employed to uncover the mechanistic details underlying this structural shift in the folding intermediate.

Addressing uncertainty in ancestral sequence reconstruction

Since ancestral sequence reconstruction produces inferred maximum likelihood sequences based on a phylogenetic relationship of extant homologs, it is important to consider how uncertainty in the reconstruction may affect the measured biophysical properties.^{27,28,41} This has been addressed extensively in previous works on the ancestral proteins of the ribonuclease H family by generating and characterizing alternative sequences, and it was found that the thermodynamic and kinetic properties were robust to uncertainty in the reconstruction.^{23,24} Although we have not characterized the fragment models of the

alternative sequences in this study, we can infer from sequence identities and posterior probabilities whether our measured properties of the fragments would be robust to the reconstruction.

First, the mean posterior probability in the region spanning core and strand 1 for each ancestral RNase H (92.0-98.1%) indicates that the reconstruction is well supported in this region of the protein with little ambiguity in the inferred amino acid. Second, the core and strand 1 region of the protein are highly similar across the alternate sequences of Anc1*. In the region of the protein spanning the core and strand 1, the mean pairwise sequence identity between Anc1* and the alternative sequences of Anc1* is 96.9% (range 92.8-99.0%). Most of the alternative sequences of Anc1* are invariant in the interface between the core and strand 1. Of those that do have changes, there are only two potential conservative changes (L7I and R57K, *E. coli* RNase H numbering). Given this high sequence conservation, we believe that the conclusions from our study, drawn from the fragment conformations observed across multiple ancestors in the two lineages, are likely robust to reconstruction uncertainty.

Evolutionary significance of the folding intermediate structure

What might be the evolutionary significance of early docking of strand 1 onto the core, and what impact might it have on the protein's biological function? The early docking of strand 1 to the core region of the protein poses an entropic challenge, as strand 1 is located at the N-termini of the protein and the sequence of the core does not begin until further along the polypeptide. There would be a conformational penalty for this strand to dock onto the core region that is non-contiguous in sequence space. Thus, we believe that the maintenance of this entropically unfavorable interaction along the thermophilic lineage for ~3 billion years is significant, especially since the interaction can be lost without compromising folding, as shown in the mesophilic lineage. Whether a more thermophilic protein requires a more structured intermediate is not established, but the unique environmental conditions that thermophilic proteins experience may play a role in the nature of the structures populated along the folding pathway of RNase H.^{42,43} Additionally, analytical ultracentrifugation experiments have shown that I_{core} , but not $I_{\text{core}+1}$, homodimerizes in a head-to-head manner with a K_d of ~150 μM for ttRNH*.^{20,21} Perhaps, docking of strand 1 onto the core masks the dimerization interface and prevents any unwanted protein-protein interactions, particularly since partially folded states on the energy landscape of proteins have been attributed to be gateways for misfolding and aggregation.^{44,45}

Alternatively, it is entirely possible that strand 1 docking is not directly under selection, but instead, coupled to another unknown property under evolutionary pressure. All of the RNase H proteins studied fold efficiently with high fidelity, so the involvement of strand 1 in the folding intermediate may be a neutral trait. Although the true evolutionary implication of strand 1 during RNase H evolution may be lost in history, our study nonetheless illuminates the evolutionary process during which features on the RNase H energy landscape changed to yield extant homologs that fold via different conformations.

Conclusion

In this study, we characterize the conformations present on the energy landscape of ancestral proteins of the RNase H family and identify the putative intermediate structure that each ancestor folds through to reach its native state. The involvement of the first β -strand in the folding intermediate is conserved in a lineage leading to a thermophilic homolog, whereas this feature is gradually lost in a separate lineage to a mesophilic RNase H. Our study shows how partially folded states can appear or disappear over evolution without altering the native fold and activity the protein, and that a protein's folding pathway is not necessarily structurally conserved over evolution. Future efforts to understand how all of these different conformational states are encoded and altered by the sequence will reveal the mechanisms underlying these trends, and inform our efforts to engineer and design specific features on the protein energy landscape.

Acknowledgments

We thank the Marqusee Lab for helpful comments and discussion. This work was funded by NIH grant GM050945 (to S.M.) and the National Science Foundation Graduate Research Fellowship (to S.A.L.).

References

1. Dill KA, MacCallum JL. *Science*. 2012; 338(6110):1042–1046. [PubMed: 23180855]
2. Anfinsen CB. *Science* (80-). 1973; 181(4096):223–230.
3. Godoy-Ruiz R, Ariza F, Rodriguez-Larrea D, Perez-Jimenez R, Ibarra-Molero B, Sanchez-Ruiz JM. *J Mol Biol*. 2006; 362(5):966–978. [PubMed: 16935299]
4. DePristo MA, Weinreich DM, Hartl DL. *Nat Rev Genet*. 2005; 6(9):678–687. [PubMed: 16074985]
5. Louis JM, Roche J. *J Mol Biol*. 2016:1–13.
6. Gong LI, Suchard MA, Bloom JD. *Elife*. 2013; 2:e00631. [PubMed: 23682315]
7. Ferguson N, Capaldi AP, James R, Kleanthous C, Radford SE. *J Mol Biol*. 1999; 286(5):1597–1608. [PubMed: 10064717]
8. Pearson WR. *Curr Protoc Bioinforma*. 2013; (SUPPL 42):1–8.
9. Bershtein S, Mu W, Serohijos AWR, Zhou J, Shakhnovich EI. *Mol Cell*. 2013; 49(1):133–144. [PubMed: 23219534]
10. Kwa LG, Wensley BG, Alexander CG, Browning SJ, Lichman BR, Clarke J. *J Mol Biol*. 2014; 426(7):1600–1610. [PubMed: 24373753]
11. Nickson AA, Clarke J. *Methods*. 2010; 52(1):38–50. [PubMed: 20570731]
12. Nickson AA, Wensley BG, Clarke J. *Curr Opin Struct Biol*. 2013; 23(1):66–74. [PubMed: 23265640]
13. Dabora JM, Marqusee S. *Protein Sci*. 1994; 3(9):1401–1408. [PubMed: 7833802]
14. Hollien J, Marqusee S. *Biochemistry*. 1999; 38(12):3831–3836. [PubMed: 10090773]
15. Hollien J, Marqusee S. *J Mol Biol*. 2002; 316(2):327–340. [PubMed: 11851342]
16. Raschke TM, Kho J, Marqusee S. *Nat Struct Biol*. 1999; 6(9):825–831. [PubMed: 10467093]
17. Parker MJ, Marqusee S. *J Mol Biol*. 2001; 305(3):593–602. [PubMed: 11152615]
18. Chamberlain AK, Fischer KF, Reardon D, Handel TM, Marqusee AS. *Protein Sci*. 1999; 8(11):2251–2257. [PubMed: 10595528]
19. Hu W, Walters BT, Kan Z, Mayne L, Rosen LE, Marqusee S, Englander SW. *Proc Natl Acad Sci USA*. 2013; 110(19):7684–7689. [PubMed: 23603271]
20. Rosen LE, Connell KB, Marqusee S. *Proc Natl Acad Sci U S A*. 2014; 111(41):14746–14751. [PubMed: 25258414]
21. Rosen LE, Marqusee S. *PLoS One*. 2015; 10(3):e0119640. [PubMed: 25803034]

22. Zhou Z, Feng H, Ghirlando R, Bai Y. *J Mol Biol.* 2008; 384(2):531–539. [PubMed: 18848567]
23. Hart KM, Harms MJ, Schmidt BH, Elya C, Thornton JW, Marqusee S. *PLoS Biol.* 2014; 12(11):e1001994. [PubMed: 25386647]
24. Lim SA, Hart KM, Harms MJ, Marqusee S. *Proc Natl Acad Sci U S A.* 2016; 113(46):13045–13050. [PubMed: 27799545]
25. Zuckerkandl E, Pauling L. *J Theor Biol.* 1965; 8(2):357–366. [PubMed: 5876245]
26. Harms MJ, Thornton JW. *Curr Opin Struct Biol.* 2010; 20(3):360–366. [PubMed: 20413295]
27. Harms MJ, Thornton JW. *Nat Rev Genet.* 2013; 14(8):559–571. [PubMed: 23864121]
28. Wheeler LC, Lim SA, Marqusee S, Harms MJ. *Curr Opin Struct Biol.* 2016; 38:37–43. [PubMed: 27288744]
29. Bridgham JT, Ortlund EA, Thornton JW. *Nature.* 2009; 461(7263):515–519. [PubMed: 19779450]
30. Perez-Jimenez R, Inglés-Prieto A, Zhao Z-M, Sanchez-Romero I, Alegre-Cebollada J, Kosuri P, Garcia-Manyes S, Kappock TJ, Tanokura M, Holmgren A, Sanchez-Ruiz JM, Gaucher EA, Fernandez JM. *Nat Struct Mol Biol.* 2011; 18(5):592–596. [PubMed: 21460845]
31. Risso VA, Gavira JA, Mejia-Carmona DF, Gaucher EA, Sanchez-Ruiz JM. *J Am Chem Soc.* 2013; 135(8):2899–2902. [PubMed: 23394108]
32. Akanuma S, Nakajima Y, Yokobori S, Kimura M, Nemoto N, Mase T, Miyazono K, Tanokura M, Yamagishi A. *Proc Natl Acad Sci U S A.* 2013; 110(27):11067–11072. [PubMed: 23776221]
33. Gaucher EA, Govindarajan S, Ganesh OK. *Nature.* 2008; 451(7179):704–707. [PubMed: 18256669]
34. Hobbs JK, Shepherd C, Saul DJ, Demetras NJ, Haaning S, Monk CR, Daniel RM, Arcus VL. *Mol Biol Evol.* 2012; 29(2):825–835. [PubMed: 21998276]
35. Clifton BE, Jackson CJ. *Cell Chem Biol.* 2016; 23(2):1–10. [PubMed: 26933730]
36. Howard CJ, Hanson-Smith V, Kennedy KJ, Miller CJ, Lou HJ, Johnson AD, Turk BE, Holt LJ. *Elife.* 2014; 3:e04126.
37. Smock RG, Yadid I, Dym O, Clarke J, Tawfik DS. *Cell.* 2016; 164:1–11.
38. Street TO, Courtemanche N, Barrick D. *Methods Cell Biol.* 2008; 84(7):295–325. [PubMed: 17964936]
39. Chamberlain AK, Marqusee S. *Advances in protein chemistry.* Vol. 53. Elsevier; 2000. 283–328.
40. Sailer ZR, Harms MJ. *PLoS Comput Biol.* 2017; 13(5):1–16.
41. Eick GN, Bridgham JT, Anderson DP, Harms MJ, Thornton JW. 2016; 34(2):247–261.
42. Ratcliff K, Marqusee S. *Biochemistry.* 2010; 49(25):5167–5175. [PubMed: 20491485]
43. Robic S, Guzman-Casado M, Sanchez-Ruiz JM, Marqusee S. *Proc Natl Acad Sci U S A.* 2003; 100(20):11345–11349. [PubMed: 14504401]
44. Dobson CM. *Nature.* 2003; 426(18):884–890. [PubMed: 14685248]
45. Jahn TR, Radford SE. *Arch Biochem Biophys.* 2008; 469(1):100–117. [PubMed: 17588526]

Figure 1. RNase H structure and evolutionary tree

A) *T. thermophilus* RNase H structure (PDB: 1RIL). The I_{core} region of the protein is depicted in green, and strand 1, additionally involved in $I_{\text{core}+1}$, is depicted in black. Numbers represent β -strands (strands 1-5) and letters represent α -helices (helix A-E). Tryptophan residues are highlighted in blue, and residue 4, the TNB labeling site, is highlighted in red.

B) Phylogenetic tree of the RNase H family illustrating the reconstructed ancestors, adapted from Hart et al. *PLoS Biology* 2014) and Lim et al. *PNAS* 2016.^{23,24} Branches are labeled with their branch lengths and circles represent the nodes of the reconstructed RNase H ancestors. Anc1* is the last common ancestor between ecRNH* and ttRNH*. Two ancestors (Anc2*, Anc3*) exist along the thermophilic lineage leading to ttRNH*. Four ancestors (AncA*, AncB*, AncC*, AncD*) exist along the mesophilic lineage leading to ecRNH*. The structure of the folding intermediate for the two extant homologs, ecRNH* and ttRNH*, are shown on the right.

Figure 2. Anc1* folds via the I_{core+1} intermediate

A) CD spectra of Anc1* I_{core} fragment (black circles) and I_{core+1} fragment (open circles). B) Representative urea denaturation melts of Anc1* I_{core} fragment (black circles) and I_{core+1} fragment (open circles) monitored by CD at 222 nm. The red shaded region represents the C_m range of the kinetic folding intermediate for each ancestor, as determined previously from a chevron analysis of the full-length protein.²⁴

Figure 3. The $I_{\text{core}+1}$ intermediate is maintained in the thermophilic lineage

CD spectra (left) and representative urea denaturation melts (right) of I_{core} fragment (black circles) and $I_{\text{core}+1}$ fragment (open circles) of A) Anc2* and B) Anc3*. The red shaded region on right panels represents the C_m range of the kinetic folding intermediate, as determined previously from a chevron analysis of the full-length protein.²⁴

Figure 4. The $I_{\text{core}+1}$ intermediate is lost in the mesophilic lineage

CD spectra (left) and representative urea denaturation melts (right) of I_{core} fragment (black circles) and $I_{\text{core}+1}$ fragment (open circles) of A) AncA*, B) AncB*, C) AncC*, D) AncD*. The red shaded region on right panels represents the C_m range of the kinetic folding intermediate for each ancestor, as determined previously from a chevron analysis of the full-length protein.²⁴

Figure 5. Anc1*-K4C-W22Y labeled with TNB can be used to monitor strand 1 docking

A) CD spectra of Anc1*-K4C-W22Y labeled with TNB (black circles) and Anc1* wild-type (open circles). B) Representative urea denaturation melts of Anc1*-K4C-W22Y labeled with TNB (black circles) and Anc1* wild-type (open circles) monitored by CD at 222 nm. C) Fluorescence emission spectra of Anc1*-K4C-W22Y folded (0M [urea], black circles) and unfolded (7M [urea], open circles) labeled with TNB (left) and unlabeled (right). The red solid line marks emission at 375 nm.

Figure 6. Strand 1 docking occurs much faster than global folding for Anc1*-K4C-W22Y labeled with TNB

Fluorescence stopped-flow refolding kinetics of Anc1*-K4C-W22Y A) labeled with TNB or B) unlabeled. Representative traces at 1 M [urea] (top) and 1.5 M [urea] (bottom) are shown. The red circle represents the expected signal for an unfolded protein. C) Representative refolding kinetics of Anc1*-K4C-W22Y labeled with TNB monitored by stopped-flow CD. Final refolding conditions are 1 M [urea]. The red circle represents the expected CD signal for an unfolded protein and the red line represents a fit of the observed kinetics to a single exponential. D) Chevron plot of Anc1*-K4C-W22Y labeled with TNB. The black circles represent global folding kinetics monitored by CD of Anc1*-K4C-W22Y labeled with TNB. The chevron for wild-type Anc1* is shown by a gray line for comparison.²⁴ Red circles

represent the lower limit of the rate constant for TNB quenching, estimated from the dead time of the stopped-flow fluorimeter.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Thermodynamic parameters of ancestral RNase H fragment models

	Anc3*	Anc2*	Anc1*	AncA*	AncB*	AncC*	AncD*
I_{core} fragment							
G _{unf} (kcal mol ⁻¹)	2.81 ± 0.08	1.27 ± 0.12	1.71 ± 0.11	not fitted [‡]	3.41 ± 0.11	2.58 ± 0.10	2.79 ± 0.10
m _{unf} (kcal mol ⁻¹ M ⁻¹)	0.87 ± 0.05	0.76 ± 0.06	0.82 ± 0.04	not fitted [‡]	1.07 ± 0.04	1.26 ± 0.03	1.18 ± 0.03
C _m (M)	3.23 ± 0.11	1.67 ± 0.20	2.09 ± 0.17	not fitted [‡]	3.19 ± 0.16	2.05 ± 0.09	2.36 ± 0.10
I_{core+1} fragment							
G _{unf} (kcal mol ⁻¹)	5.17 ± 0.07	4.21 ± 0.09	4.14 ± 0.20	2.42 ± 0.07	5.56 ± 0.08	3.64 ± 0.08	2.78 ± 0.06
m _{unf} (kcal mol ⁻¹ M ⁻¹)	1.15 ± 0.01	1.22 ± 0.04	1.15 ± 0.07	1.54 ± 0.03	1.25 ± 0.03	1.35 ± 0.03	1.13 ± 0.02
C _m (M)	4.50 ± 0.07	3.45 ± 0.04	3.56 ± 0.09	1.57 ± 0.05	4.44 ± 0.06	2.69 ± 0.04	2.46 ± 0.03
Kinetic Intermediate							
C _m (M) ^{‡‡}	4.31 ± 0.58	3.26 ± 0.6	3.71 ± 0.35	1.75 ± 0.34	3.95 ± 0.93	2.91 ± 0.77	2.58 ± 0.56

Errors are in S.D.

[‡]Not sufficient folded baseline to fit to a two-state model.

^{‡‡}Calculated from data published in Lim et al. (2016).²⁴