Frozen in Time: The History of Proteins

Nicholas A. Kovacs,[†,1] Anton S. Petrov,[†,1] Kathryn A. Lanier,[1] and Loren Dean Williams*[,1]

[1]School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, GA

[†]These authors contributed equally to this work.

*Corresponding author: E-mail: loren.williams@chemistry.gatech.edu.

Associate editor: Nicole Perna

Abstract

The ribosome is imprinted with a detailed molecular chronology of the origins and early evolution of proteins. Here we show that when arranged by evolutionary phase of ribosomal evolution, ribosomal protein (rProtein) segments reveal an atomic level history of protein folding. The data support a model in which aboriginal oligomers evolved into globular proteins in a hierarchical step-wise process. Complexity of assembly and folding of polypeptide increased incrementally in concert with expansion of rRNA. (i) Short random coil proto-peptides bound to rRNA, and (ii) lengthened over time and coalesced into β–β secondary elements. These secondary elements (iii) accreted and collapsed, primarily into β-domains. Domains (iv) accumulated and gained complex super-secondary structures composed of mixtures of α-helices and β-strands. Early protein evolution was guided and accelerated by interactions with rRNA. rRNA and proto-peptide provided mutual protection from chemical degradation and disassembly. rRNA stabilized polypeptide assemblies, which evolved in a stepwise process into globular domains, bypassing the immense space of random unproductive sequences. Coded proteins originated as oligomers and polymers created by the ribosome, on the ribosome and for the ribosome. Synthesis of increasingly longer products was iteratively coupled with lengthening and maturation of the ribosomal exit tunnel. Protein catalysis appears to be a late byproduct of selection for sophisticated and finely controlled assembly.

Key words: **protein evolution, ribosomal protein, origin of life, ribosomal origins and evolution, origins of protein folding, β-harpin.**

Introduction

Translation of mRNA to protein underpins the macromolecular partnership that has dominated the biological earth for nearly 4 billion years, and provides a blueprint of the common origin and interrelatedness of all living systems (Woese and Fox 1977). Structures of ribosomes in three dimensions, from across the tree of life (Ramakrishnan 2011), reveal macromolecules from deep biological history and provide a guidebook to the pre-biological evolution of biopolymers. Ribosomal RNAs (rRNAs) and ribosomal proteins (rProteins) are molecular fossils from before the last universal common ancestor of life (Lecompte et al. 2002; Söding and Lupas 2003; Fox and Naik 2004).

In the previously described Accretion Model of ribosomal evolution, rRNA recursively accreted and froze, increasing in mass over time (Bokov and Steinberg 2009; Hsiao et al. 2009; Petrov et al. 2014). The ribosome sequentially acquired capabilities for RNA folding, noncoded condensation of amino acids to form peptides, subunit association, correlated subunit evolution and decoding, and energy transduction (Petrov et al. 2015). rRNA growth is partitioned into six phases in prokaryotes (fig. 1A) with two additional eukaryotic phases. The first phase contributes the most ancient rRNA while the final phase contains the most recent rRNA. A consistent theme of Phases 1–6 of ribosomal evolution is extension and elaboration of the exit tunnel.

Here we incorporate rProteins into the Accretion Model of ribosomal evolution by establishing temporal correlations between acquisition of rRNA elements and acquisition of rProtein segments. These correlations assume that the age of a given segment of rProtein is the same as that of the rRNA with which it interacts. The results provide a test of the Accretion Model.

This extension of the Accretion Model allows us to construct a molecular level “movie” of protein evolution. The rRNA to rProtein temporal mapping provides frames of a movie suggestive of incremental and hierarchical evolution of proteins. The movie shows step-wise formation of protein domains. (i) Initially, short random coil (RC) peptides bound to rRNA. (ii) The peptides lengthened and coalesced into secondary elements, with β–β structures more frequent than α-helices. Polypeptide secondary elements (iii) accreted and collapsed into domains composed primarily of β-strands. Protein domains (iv) accumulated and gained increasingly complex super-secondary structures composed of mixtures of α-helices and β-strands. Protein evolution was continuously guided and accelerated by interactions with rRNA (Söding and Lupas 2003). Throughout this process polypeptide was immersed in rRNA. RNA folding evolved (Hsiao et al. 2009) in parallel with protein folding.

Results

We integrated rProteins from the Large Ribosomal Subunit (LSU) into the Accretion Model (fig. 1B). The phases of the Accretion Model are a series of course grained states incorporating the highly detailed temporal information provided

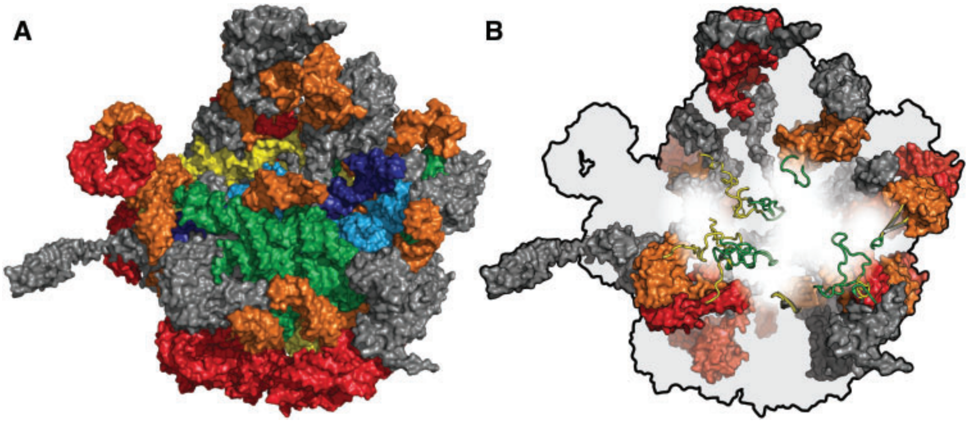


Fig. 1. (*A*) The rRNA of the large subunit of the *T. thermophilus* ribosome colored by relative age. Phase 1, the most ancient phase, is dark blue. Phase 2 is light blue. Phase 3 is green. Phase 4 is yellow. Phase 5 is orange. Phase 6, the most recent prokaryotic phase, is red. rProteins are grey. (*B*) The orientation is maintained but rRNA is colored in light grey, universal rProteins are colored by evolutionary phase, and bacterial rProteins are colored dark grey. Phases 3 (green) and 4 (yellow) are shown in cartoon representation. Phases 5 (orange) and 6 (red) are shown in surface representation. From PDB entry 1VY4.

by insertion fingerprints, A-minor interactions and other elements within the ribosome (Petrov et al. 2014, 2015). rProteins were computationally segmented (cleaved) and the segments were partitioned into phases (fig. 2, table 1, supplementary figs. S1 and S2, Supplementary Material online) corresponding to those of the rRNA in the Accretion Model. The phase of each segment is determined by the phase of the rRNA with which it interacts. Segments were terminated where polypeptide backbone passes through rRNA phase boundaries. Figure 2 illustrates the segmenting of rProteins uL22 and uL13, and their assignment into phases. The correspondence of phases and a geological timeline is thus far indeterminate. The absolute age of any phase is unknown except that the beginning of Phase 6 corresponds roughly with the last universal common ancestor, around 3.8 billion years ago (Fox 2016) and Phase 1 was at or near the origin of life. Alternative phase models, with increased or decreased granularities, do not change the general trends observed here.

Here we focus on universal rProteins, to maximize the universality of the results. The patterns observed here appear to be general and robust in that the trends and the structural characteristics for each phase are the same for universal and for all LSU rProteins of ribosomes of two bacteria (*Thermus thermophilus* or *Escherichia coli*) or an archaean (*Pyrococcus furiosus*) (figs. 3 and 4, and supplementary figs. S3–S6, Supplementary Material online). All 15 universal rProteins in the LSU of the *T. thermophilus* ribosome were segmented and partitioned into phases of rRNA evolution (table 1, supplementary figs. S1 and S2, Supplementary Material online). The *T. thermophilus* ribosome was used here because it is the most accurately determined bacterial ribosome, with the best resolution.

Hallmarks of Protein Evolution

rProtein segments demonstrate increasing extent and complexity of folding with increasing phase. Changes in rProtein folding with phase highlight the time directionality of protein evolution; and reveal evolutionary transitions from unstructured to simple to complex structure. All universal rProteins in the LSU contain globular domains (table 1) (Klein et al. 2004). Some contain multiple domains (uL1, uL2, uL3, uL6, uL10, uL11, uL14, and uL16) and others contain single domains (uL4, uL5, uL13, uL15, uL18, uL22, uL23, uL24, uL29, and uL30). Eight universal LSU rProteins contain nonglobular extensions of idiosyncratic conformation (table 1), which appear to be frozen random coil that penetrates the ribosomal core. Extensions of rProteins uL2, uL3, uL4, uL13, uL14, uL15, and uL16 interact with Phase 3 rRNA. Extensions of rProteins uL2, uL3, uL4, uL13, uL15, and uL22 interact with Phase 4 rRNA. Extensions of uL13 and uL22 exhibit secondary structure (anti-parallel β-strands or β-hairpins). Globular rProtein domains that interact with Phase 5 rRNA are β-barrel domains (uL2, uL3, and uL14) or $\alpha - \beta$ domains (uL1, uL6, uL10, uL11, uL16, uL22, and uL30). These particular $\alpha - \beta$ domain rProteins mimic β-barrel topology except that one or more β-strands of a β-barrel have been converted to α-helices (Grishin 2001). rProtein domains that interact with Phase 6 rRNA (uL2, uL3, uL4, uL13, uL14, uL15, and uL16) display more complex topologies that are mixtures of $\alpha + \beta$ elements or are purely α-helical. Many rProteins (uL1, uL5, uL6, uL10, uL11, uL18, uL23, uL24, uL29, and uL30) are localized exclusively within Phases 5 and/or 6, including multidomain proteins (uL1, uL6, uL10, and uL11).

Reaction Coordinate for Protein Evolution

The results support previous proposals that universal rProteins provide records of ancient processes (Söding and Lupas 2003; Hartman and Smith 2014), contain intermediates representing a molecular level reaction coordinate for the evolution of protein folding and provide support for the Accretion Model of rRNA evolution. Our partitioning of universal rProteins into segments is shown in table 1. The information contained in the rProtein segments in figure 1 was "read" by various analytical methods. For each segment, the

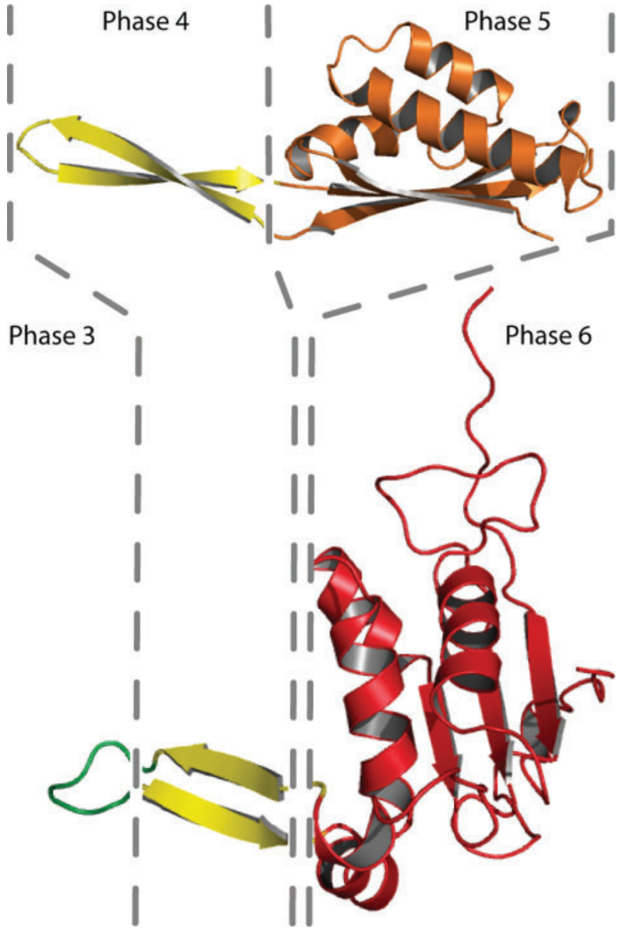


Fig. 2. The history of protein folding illustrated by LSU rProteins uL22 (top) and uL13 (bottom). rProtein segments are colored by their phase, in accordance with rRNA and rProtein phases in figure 1. Segment boundaries are indicated by dashed lines. rProtein uL22 has segments in Phases 4 and 5. uL13 has segments in Phases 3, 4, and 6. The Phase 3 segment of uL13 is random coil. Phase 4 segments of uL22 and uL13 contain isolated $\beta-\beta$ structures. The Phase 5 and 6 segments of uL22 and uL13 contain globular domains with extensive intramolecular hydrogen bonds and reduced solvent accessible surface area. These domains contain hydrophobic cores and hydrophilic surfaces. rProtein segments in lower numbered phases are more ancient than those in higher numbered phases. Structures are extracted from the *T. thermophilus* ribosome.

frequency of intramolecular hydrogen bonding of backbone atoms (IMHB^{BA}) indicates extent of formation of α-helices and β-strands (Sticke et al. 1992), while the solvent accessible surface area (SSA, calculated in absence of rRNA) shows collapse from extended to globular structures (Livingstone et al. 1991). IMHB^{BA} of rProtein segments was characterized with the program STRIDE (Frishman and Argos 1995). The SSA of rProtein segments was characterized by the program Naccess (Hubbard and Thornton 1993). rProtein segments in Phase 3 exclusively form extended and irregular structures consistent with frozen RC (fig. 1*B* and table 1). These protein segments are constrained by surrounding rRNA and interact extensively with it. These RC segments have low IMHB^{BA} and high SSA (fig. 3 and supplementary figs. S3 and S5, Supplementary Material online). rProtein segments in Phase 4 are seen to form both secondary structures and RC. The most frequent secondary structures in Phase 4 are antiparallel β–β structures, composed of intramolecular β-hairpins or β–β dimers between amino acids that are remote in primary structure but belong to a common peptide chain (fig. 4 and supplementary figs. S4 and S6, Supplementary Material online). Our results are in agreement with Hartman and Smith (2014), who previously noted the importance of β-hairpins in early protein evolution.

The frequency of secondary structure of polypeptides increases from Phase 3 through Phase 5 (fig. 4 and supplementary figs. S4 and S6, Supplementary Material online). Some of the rProtein in Phase 5 has collapsed into globular domains causing the average SSA per amino acid to decrease in Phase 5 compared with Phase 4 (fig. 3 and supplementary figs. S3 and S5, Supplementary Material online). These domains, which are composed primarily of anti-parallel β-sheets, have hydrophobic cores and hydrophilic surfaces (Wang and Hecht 2002). The β-barrel domains that are common in Phase 5 give the appearance of arising from collapse of the isolated β–β structures observed in Phase 4. In Phase 6, complex domains of β-sheets and α-helices form assemblies linked by quaternary interactions. The fraction of polypeptide in α-helices increases from Phase 4 to Phase 5 to Phase 6 (fig. 4 and supplementary figs. S4 and S6, Supplementary Material online).

Folding Fitness Landscape

Protein folding is commonly represented as a funnel with depth related to the stability of the native folded state and cross-sectional area related to conformational entropy (Bryngelson et al. 1995; Dill and Chan 1997). The surface outside the funnel is high and flat to indicate heterogeneity in conformations of the random coil state. As a protein folds, the number of configurational substrates decreases. In contrast, evolutionary processes are commonly represented by fitness landscapes in which the surface represents genotype and the height of a peak is related to replicative success (Wright 1932).

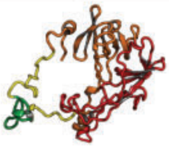
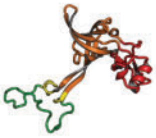
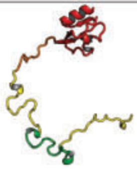
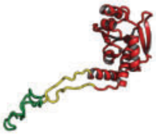
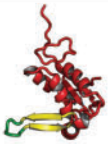
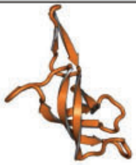
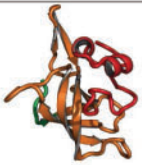
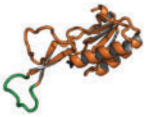
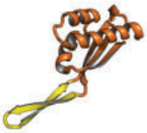
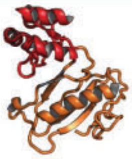
The protein folding funnel can be inverted and integrated into a fitness landscape to form a “folding fitness landscape” with a peak where modern proteins fold to form mature globular domains (fig. 5). The surface of the landscape is an abstraction that might represent a combination of chemical composition of the backbone, polypeptide sequences, along with configuration entropy of the polymer backbone. This landscape describes incremental change from heterogeneous, unstructured oligomers, to secondary elements, to globular domains. An increase in fitness and a decrease in chemical/configurational entropy are associated with ascent of the folding fitness peak.

Discussion

Ancestral Folding

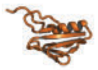
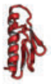
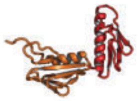
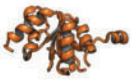
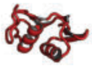
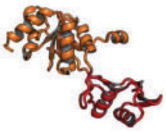
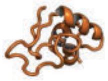
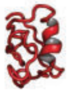
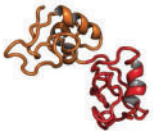
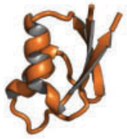
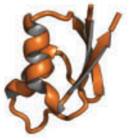
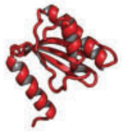
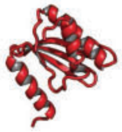
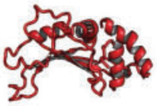
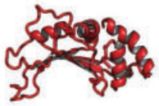
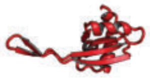
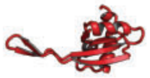
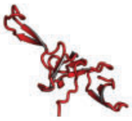
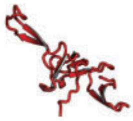
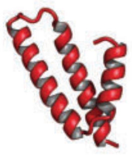
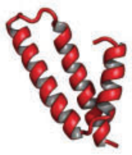
Extant proteins are composed of domains, which fold autonomously and cooperatively (Orengo et al. 1994; Porter and Rose 2012) and are evolutionarily persistent

Table 1. Segmentation of Universal rProteins from *Thermus thermophilus* Reveals the Evolution of Protein Folding.

Protein	Phase 3	Phase 4	Phase 5	Phase 6	Complete
uL2					
uL3					
uL15					
uL4					
uL13					
uL14					
uL16					
uL22					
uL1[a]					

(continued)

Table 1. Continued

uL6					
uL10[a]					
uL11[b]					
uL30					
uL18					
uL5					
uL23					
uL24					
uL29					

Note.—Each rProtein segment is colored by phase following the coloring scheme in figure 1. Empty cells indicate that no rProtein segment is contained in that phase for that protein. rProteins are from the *T. thermophilus* crystal structure.

[a] These rProteins are from the *P. furiosus* ribosomal structure because it is absent from the *T. thermophilus* structure.

[b] This rProtein is from *E. coli* ribosomal structure because it is absent from the *T. thermophilus* structure.

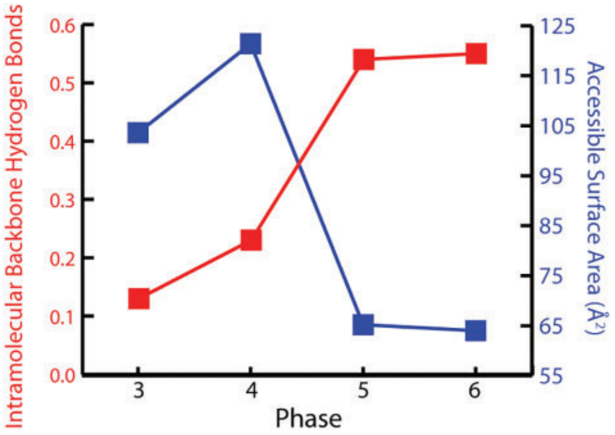


FIG. 3. Structural attributes of rProtein segments in Phases 3–6 of ribosomal evolution. The number of IMHB^{BA} per amino acid is plotted in red and the SSA per amino acid is plotted in blue. IMHB^{BA} increases from Phase 3 through 6, indicating increase in secondary structure. SSA, calculated when rRNA is computationally omitted, decreases from Phase 3 through 6, indicating collapse to globular domains. rProteins are from the *T. thermophilus* ribosome.

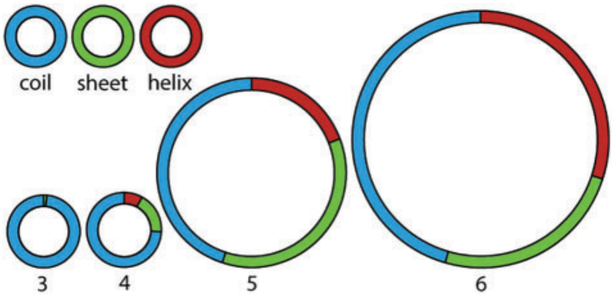


FIG. 4. Protein structural elements (coil, β-sheet and α-helix) decomposed in Phases 3–6 of ribosomal evolution. Protein segments transition from random coil to secondary structure from Phase 3 to Phase 4. Secondary structure converts from predominantly β-strand in Phase 4 and 5 to mixed β-strand and α-helix in Phase 6. The area of each pie chart is proportional to the number of amino acids within that phase. Structures are from the *T. thermophilus* ribosome.

(Lupas and Koretke 2008). The ribosome contains a molecular level history of the incorporation of protein into biological systems (Vishwanath et al. 2004). A chronology for the evolution of protein and protein domains is revealed by the conformations and interactions of rProtein segments within the ribosome. When relative ages of rRNA are mapped onto rProtein segments, the genesis and evolution of protein folding is recapitulated.

In this chronology, polypeptide was immersed in RNA throughout the early evolution of protein and protein domains. RNA chaperoned the ancestral conversion of RC peptide oligomers into full length globular domains. The protein–RNA partnership was initiated by co-assembly of oligopeptides and rRNA. Noncoded peptides synthesized by ancient rRNA would have conferred advantage by protecting rRNA against thermal unfolding and chemical degradation. The rRNA facilitated conversion of RC oligopeptides to secondary structures, which accreted on the rRNA. A subset of these collapsed into primitive globular domains. Domains matured and diversified, gaining complexity of super secondary motifs. The stabilization of folded rRNA by rProteins (Woodson 2011) is consistent with the model proposed here.

Our input and output is three-dimensional structure, in part because structure is more conserved over evolution than sequence (Illergard et al. 2009). The most ancient events in biology are best recorded in structure, not sequence. The results support the suggestion by Lupas (Lupas and Koretke 2008; Alva et al. 2015) that the broad diversity of protein domains in nature descended from a limited number of ribosomal prototypes.

Emergent Environment

Folding from short random coil peptides to functional domains was an emergent phenomena, depending on interactions with RNA (Söding and Lupas 2003). Evidently, nucleic acids have retained a generalized ability to chaperone protein folding (Docter et al. 2016). Conversely, it appears that complex folding of RNA was emergent on interactions with polypeptide. Polypeptide induced changes in RNA folding and interactions with other assembly cofactors such as magnesium (Hsiao et al. 2009). The co-evolution of RNA and protein was accomplished in the context of the ribosome, which was therefore the cradle of early evolution. Initial protein domains were created by the ribosome, on the ribosome and for the ribosome.

It appears that evolution of protein and RNA folding were coupled to each other in processes mediated by direct protein-RNA interactions. The data here support the model of Fox (Fox and Naik 2004) that the products of the very early ribosome were noncoded random sequences of peptides without propensity to fold. The products of ribosomes at intermediate stages of evolution had intermediate propensities to fold (to isolated secondary elements), and would have been characterized by rudimentary specificity in sequence such as binary hydrophobic/hydrophilic bias. The products of the extant ribosome contain specific amino acid sequences that fold efficiently to complex globular domains. Therefore the incremental increase in polypeptide length and assembly competence co-evolved with the genetic code.

Directed Search

The properties of biological proteins are not the same as those of random sequences of amino acids. Proteins fold to domains, characterized by unique three-dimensional structures. Random sequence polypeptides do not fold (Taylor et al. 2001; Dobson 2004). The frequency of protein sequences in random space that are competent to fold to domains is very nearly zero. The production of folded protein domains is impossible by random searching of full length *sequence* space, in analogy to the Levinthal paradox (Levinthal 1969), addressing the impossibility of protein folding by random searching of *conformational* space.

The results here suggest a mechanism of discovery of the rare protein sequences that are competent to fold to discrete globular domains. It appears that domain evolution was

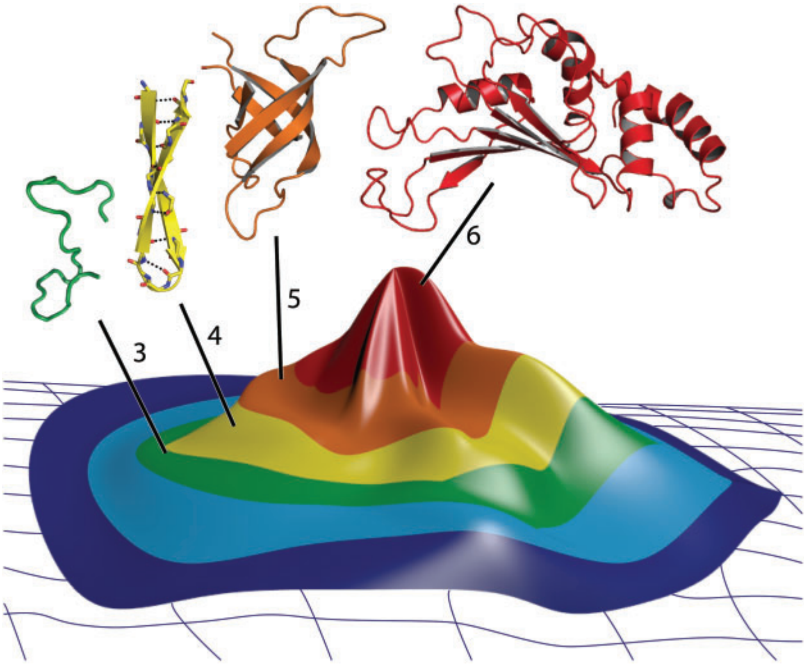


FIG. 5. A folding fitness peak describing relationships of protein folding, fitness, and ribosomal development. The color of the surface indicates the phase of ribosomal evolution. Fitness is maximized where proteins fold to complex three-dimensional structures. Representative rProtein segments, also colored by phase are shown above the funnel. The rProtein segments shown here were extracted from appropriate phase of the *T. thermophilus* ribosome.

hierarchical and was directed and accelerated by interactions with rRNA as illustrated by the folding fitness peak in figure 5. The ribosome gained advantage by discovery of short RC peptides with affinity for rRNA. Formation of secondary structures from short random oligopeptides has reasonable probability. The ribosome gained further advantage by the discovery of oligopeptides that formed secondary structures in association with rRNA. Formation of globular structures from the collapse of preformed secondary structures also has reasonable probability. The ribosome gained additional advantage by production of true protein domains subsequent to discovery of secondary structure. We presume this process to be a continuum; ribosomal products of successively greater size and ability to fold conferred ever-increasing advantage to the ribosome. Therefore, biology discovered folding competent proteins by an incremental directed pathway, bypassing the immense unproductive space of random sequences.

Folding and Fitness

We have combined the concepts of the protein folding funnel (Bryngelson et al. 1995; Dill and Chan 1997) and the fitness landscape (Wright 1932) to create a “folding fitness landscape” (fig. 5). The surface of this landscape is represented by performance, possibly defined by replicative success, which is at a maximum where proteins fold. This surface has a peak where proteins fold to mature globular domains because folded proteins are a dominant contributor to performance. A system performs better, and is more successfully replicated, when the proto-ribosome produces folded protein. The configurational entropy of a polypeptide chain is reduced as a discrete folded state is approached. Therefore decreasing polypeptide entropy correlates with increasing performance of the system. This model extends the concept of fitness beyond the Darwinian threshold. The basal regions in the folding fitness landscape predate polymers, protein-based polymerases and genetics. These regions of the surface describe a pre-biological world of chemical evolution—about which we know little, and where performance may not be measured by classically understood replicative success.

Bootstrapping and the Exit Tunnel

The data are consistent with a bootstrapping process in which synthesis of increasingly longer oligomers by the ribosome was coupled with lengthening and maturation of the exit tunnel. The ancestral PTC produced heterogeneous noncoded oligomers (Fox et al. 2012; Petrov et al. 2014). A subpopulation of these oligopeptides bound to the rRNA and conferred advantage. Increases in the length of the oligopeptides facilitated extension and rigidification of the exit tunnel (Petrov et al. 2015). Continued exit tunnel development facilitated production of longer oligopeptides. Longer products stabilized assemblies with more extended tunnels. This coupling helped drive early ribosomal evolution.

Driven by Assembly

It is widely believed that the general catalytic superiority of proteins over RNA drove the merger of polypeptide and RNA (Gilbert 1986; Cech 2009). However, this model appears to require foresight, which is not available to evolutionary processes. Indeed, our results here fail to support an early role for catalysis in protein evolution. Instead, it appears that the primary driving force for early protein evolution was co-assembly of RC peptides with RNA, then self-assembly to form secondary structures, followed by formation of primitive domains. Catalytic properties of protein appear to be a late byproduct of a process that selected for ever more sophisticated assembly and co-assembly.

Chicken and Egg

The extreme improbability that two fundamentally different biopolymers such as RNA and protein would emerge simultaneously and independently has been noted many times (Bernhardt 2012; Neveu et al. 2013). One came first, it seems, either RNA or protein. The RNA World, with a single polymer, provides one resolution to this chicken-or-egg predicament. Our results indicate that the evolution of RNA and protein are not independent, it appears that they co-evolved. The near-simultaneous emergence of RNA and protein in an emergent environment in which each polymer chaperoned the evolution of the other may provide an alternative to the RNA World, a simpler and more direct evolutionary pathway to current biology, and a resolution of the chicken and egg dilemma. Linked evolution of RNA and protein backbone and sidechain elements eliminates the chicken and egg dilemma.

Materials and Methods

Phase Assignments of rProtein Segments

Phase assignments of rProtein segments were performed by the following rules. (i) Protein segments interacting extensively with or bridging rRNA of two phases were assigned to the later phase. (ii) A segment localized at the ribosome surface and interacting primarily with rRNA from a single phase, was assigned to that phase. For example, the globular domain of uL16 interacts with H38 and H42 (both emerge in phase 5) and was assigned to Phase 5. The globular Domain of uL23 binds to H53 and H9 (both in Phase 6) and was assigned to Phase 6. (iii) Globular domains were not segmented. A globular domain was assigned to a single segment and to a single phase and (iv) two globular domains of a single protein are not restricted to the same phase.

In a few cases, adjacent amino acids interact with rRNA from multiple phases. In these cases, the preponderance of the interactions was used to establish the phase assignment. The amino acids contacting Phase 1 and Phase 2 rRNA are rare and are isolated along the peptide chain from one another. These amino acids are found in rProteins uL2, uL3, uL4, uL6, uL13, uL14, uL15, uL16, bL27, bL28, bL32, and bL36. These segments were assigned to Phase 3 because the preponderance of the interactions are with that phase. The allocation of all early phase segments to Phase 3 does not impact the general trends or overall results described here. LSU rProteins from three ribosomes (*T. Thermophilus*, PDB 1VY4, 2.6 Å resolution, X-ray; *E. coli*, PDB 4V9D, 3.0 Å, X-ray; *P. furiosus*, PDB 4V6U, 6.6 Å, Cryo-EM) were analyzed independently.

rRNA Secondary Structures

Secondary structures of LSU rRNAs are taken from our public gallery (http://apollo.chemistry.gatech.edu/RibosomeGallery, last accessed 24 February 2017). Data were mapped onto rRNA secondary structures with the program RiboVision (supplementary fig. S1, Supplementary Material online) (Bernier et al. 2014).

rProtein Segments from Bacterial and Archaeal LSU Particles

Segmentation of universal rProteins was performed for ribosomes of *T. thermophilus* (Polikanov et al. 2014) (PDB 1VY4, bacterium, 2.6 Å resolution, X-ray), *E. coli* (Dunkle et al. 2011) (PDB 4V9D, bacterium, 3.0 Å, X-ray), and *P. furiosus* (Armache et al. 2013) (PDB 4V6U, archaea, 6.6 Å, Cryo-EM). The statistics of segment conformations are essentially the same for each of these ribosomes (supplementary figs. S3 and S4, Supplementary Material online). Global superimpositions of the LSU particles were performed using the CEAlign functionality of PyMOL (Schrodinger 2016).

Assigning rProtein Segments to Phases

LSU rProteins were incorporated into the previously described Accretion Model of rRNA evolution (Petrov et al. 2014, 2015). rProteins from three different ribosomal structures (*T. thermophilus, E. coli*, and *P. furiosus)* were analyzed independently. rProteins were computationally segmented based on the phase of the surrounding rRNA, as described in the supplementary text, Supplementary Material online.

Protein Secondary Structures

The secondary structures of rProtein segments were characterized by the frequency of intramolecular hydrogen bonding of backbone atoms (IMHB^{BA}) and by solvent accessible surface area (SSA) with the programs STRIDE (Frishman and Argos 1995), Naccess (Hubbard and Thornton 1993), and by visual inspection. IMHB^{BA}s primarily indicate extent of backbone–backbone interactions in α-helix or β-sheet (Sticke et al. 1992), while the SSA shows extent of collapse from extended to globular structures (Livingstone et al. 1991). DSSP secondary structural elements "Turn", "Bend", and "-" were grouped as Random Coil (RC); "Extended Strand" and "Isolated Bridge" were grouped as "Sheet"; and "α-Helix", "3_{10}-Helix", and "π-Helix" were grouped as "Helix".

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Author Contributions

L.D.W. and A.S.P. conceived the study; N.A.K. and A.S.P. defined the rProtein segments; N.A.K. and K.A.L. defined rRNA phases in TT and PF; N.A.K. analyzed the data and generated all figures and tables; and L.D.W., A.S.P., N.A.K., and K.A.L. wrote the manuscript.

Acknowledgment

This work was supported by NASA Grant NNX16AJ29G.

References

Alva V, Söding J, Lupas AN. 2015. A vocabulary of ancient peptides at the origin of folded proteins. *eLife* 4:e09410.

Armache J-P, Anger AM, Márquez V, Franckenberg S, Fröhlich T, Villa E, Berninghausen O, Thomm M, Arnold GJ, Beckmann R, Wilson DN. 2013. Promiscuous behaviour of archaeal ribosomal proteins: implications for eukaryotic ribosome evolution. *Nucleic Acids Res.* 41:1284–1293.

Bernhardt HS. 2012. The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others). *Biol Direct.* 7:23.

Bernier C, Petrov AS, Waterbury C, Jett J, Li F, Freil LE, Xiong B, Wang L, Le A, Milhouse BL, et al. 2014. Ribovision: visualization and analysis of ribosomes. *Faraday Discuss.* 169:195–207.

Bokov K, Steinberg SV. 2009. A hierarchical model for evolution of 23S ribosomal RNA. *Nature* 457:977–980.

Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. 1995. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins Struct Funct Bioinformatics* 21:167–195.

Cech TR. 2009. Crawling out of the RNA world. *Cell* 136:599–602.

Dill KA, Chan HS. 1997. From Levinthal to pathways to funnels. *Nat Struct Mol Biol.* 4:10–19.

Dobson CM. 2004. Principles of protein folding, misfolding and aggregation. *Semin Cell Dev Biol.* 15:3–16.

Docter BE, Horowitz S, Gray MJ, Jakob U, Bardwell JC. 2016. Do nucleic acids moonlight as molecular chaperones? *Nucleic Acids Res.* 44:4835–4845.

Dunkle JA, Wang LY, Feldman MB, Pulk A, Chen VB, Kapral GJ, Noeske J, Richardson JS, Blanchard SC, Cate JHD. 2011. Structures of the bacterial ribosome in classical and hybrid states of tRNA binding. *Science* 332:981–984.

Fox GE. 2016. Origins and early evolution of the ribosome. In: Hernández G, Jagus R, editors. Evolution of the protein synthesis machinery and its regulation. Springer. p. 31–60.

Fox GE, Naik AK. 2004. The evolutionary history of the translation machinery. In: de Pouplana LR, editor. The genetic code and the origin of life. Kluwer Academic/Plenum Publishers, New York, p. 92–105.

Fox GE, Tran Q, Yonath A. 2012. An exit cavity was crucial to the polymerase activity of the early ribosome. *Astrobiology* 12:57–60.

Frishman D, Argos P. 1995. Knowledge-based protein secondary structure assignment. *Proteins Struct Funct Bioinformatics* 23:566–579.

Gilbert W. 1986. Origin of life: the RNA world. *Nature* 319:618.

Grishin NV. 2001. Fold change in evolution of protein structures. *J Struct Biol.* 134:167–185.

Hartman H, Smith T. 2014. The evolution of the ribosome and the genetic code. *Life* 4:227–249.

Hsiao C, Mohan S, Kalahar BK, Williams LD. 2009. Peeling the onion: ribosomes are ancient molecular fossils. *Mol Biol Evol.* 26:2415–2425.

Hubbard SJ, Thornton JM. 1993. Naccess. Computer Program, Department of Biochemistry and Molecular Biology, University College, London.

Illergard K, Ardell DH, Elofsson A. 2009. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins* 77:499–508.

Klein DJ, Moore PB, Steitz TA. 2004. The roles of ribosomal proteins in the structure, assembly, and evolution of the large ribosomal subunit. *J Mol Biol.* 340:141–177.

Lecompte O, Ripp R, Thierry JC, Moras D, Poch O. 2002. Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Res.* 30:5382–5390.

Levinthal C. 1969. How to fold graciously. In: DeBrunner JTP, Munck E, editors. Mossbauer spectroscopy in biological systems. Allerton House, Monticello (IL): University of Illinois Press, p. 22–24.

Livingstone JR, Spolar RS, Record Jr MT. 1991. Contribution to the thermodynamics of protein folding from the reduction in water-accessible nonpolar surface area. *Biochemistry* 30:4237–4244.

Lupas A, Koretke K. 2008. Evolution of protein folds. In: Schwede T, Peitsch MC, editors. Computational structural biology: methods and applications. Singapore: World Scientific. p. 131–151.

Neveu M, Kim H-J, Benner SA. 2013. The "Strong" RNA world hypothesis: fifty years old. *Astrobiology* 13:391–403.

Orengo CA, Jones DT, Thornton JM. 1994. Protein superfamilies and domain superfolds. *Nature* 372:631–634.

Petrov AS, Bernier CR, Hsiao C, Norris AM, Kovacs NA, Waterbury CC, Stepanov VG, Harvey SC, Fox GE, Wartell RM, et al. 2014. Evolution of the ribosome at atomic resolution. *Proc Natl Acad Sci U S A.* 111:10251–10256.

Petrov AS, Gulen B, Norris AM, Kovacs NA, Bernier CR, Lanier KA, Fox GE, Harvey SC, Wartell RM, Hud NV, Williams LD. 2015. History of the ribosome and the origin of translation. *Proc Natl Acad Sci U S A.* 112:15396–15401.

Polikanov YS, Steitz TA, Innis CA. 2014. A proton wire to couple aminoacyl-tRNA accommodation and peptide-bond formation on the ribosome. *Nat Struct Mol Biol.* 21:787–793.

Porter LL, Rose GD. 2012. A thermodynamic definition of protein domains. *Proc Natl Acad Sci U S A.* 109:9420–9425.

Ramakrishnan V. 2011. The eukaryotic ribosome. *Science* 331:681–682.

Schrödinger, LLC. 2016. The Pymol Molecular Graphics System, Version 1.8. New York.

Söding J, Lupas AN. 2003. More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays* 25:837–846.

Sticke DF, Presta LG, Dill KA, Rose GD. 1992. Hydrogen bonding in globular proteins. *J Mol Biol.* 226:1143–1159.

Taylor SV, Walter KU, Kast P, Hilvert D. 2001. Searching sequence space for protein catalysts. *Proc Natl Acad Sci U S A.* 98:10596–10601.

Vishwanath P, Favaretto P, Hartman H, Mohr SC, Smith TF. 2004. Ribosomal protein-sequence block structure suggests complex prokaryotic evolution with implications for the origin of eukaryotes. *Mol Phylogen Evol.* 33:615–625.

Wang W, Hecht MH. 2002. Rationally designed mutations convert de novo amyloid-like fibrils into monomeric β-sheet proteins. *Proc Natl Acad Sci U S A.* 99:2760–2765.

Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A.* 74:5088–5090.

Woodson SA. 2011. RNA folding pathways and the self-assembly of ribosomes. *Acc Chem Res.* 44:1312–1319.

Wright S. 1932. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In: Jones DF, editor. Proceedings of the Sixth International Congress of Genetics. Austin (TX): Genetics Society of America.