Molecular signatures of ribosomal evolution

Elijah Roberts*, Anurag Sethi†, Jonathan Montoya†, Carl R. Woese‡§¶, and Zaida Luthey-Schulten*†§¶

*Center for Biophysics and Computational Biology, †Departments of Chemistry and ‡Microbiology, and §Institute for Genomic Biology, University of Illinois at Urbana–Champaign, Urbana, IL 61801

Contributed by Carl R. Woese, May 19, 2008 (sent for review May 6, 2008)

Ribosomal signatures, idiosyncrasies in the ribosomal RNA (rRNA) and/or proteins, are characteristic of the individual domains of life. As such, insight into the early evolution of the domains can be gained from a comparative analysis of their respective signatures in the translational apparatus. In this work, we identify signatures in both the sequence and structure of the rRNA and analyze their contributions to the universal phylogenetic tree using both sequence- and structure-based methods. Domain-specific ribosomal proteins can be considered signatures in their own right. Although it is commonly assumed that they developed after the universal ribosomal proteins, we present evidence that at least one may have been present before the divergence of the organismal lineages. We find correlations between the rRNA signatures and signatures in the ribosomal proteins showing that the rRNA signatures coevolved with both domain-specific and universal ribosomal proteins. Finally, we show that the genomic organization of the universal ribosomal components contains these signatures as well. From these studies, we propose the ribosomal signatures are remnants of an evolutionary-phase transition that occurred as the cell lineages began to coalesce and so should be reflected in corresponding signatures throughout the fabric of the cell and its genome.

three domains of life | genomic organization | environmental sequences

A huge and exponentially increasing dataset regarding the molecular makeup of cells has accumulated over the last several decades. Biologists today routinely ask questions of the data that are far more deeply probing than previously possible. What is not generally appreciated, however, is that large datasets of this type tend to bring into question the conceptual framework within which the questions themselves are posed. An especially informative example is our understanding of the cellular translation mechanism. In the past, the mechanism was conceptualized and probed in a reductionist "particle" framework, whereas understanding today comes increasingly from multimodal analyses. The questions and answers bespeak a highly integrated mechanism, whose essence would seem to lie in its delocalized collective properties.

This perceptual change not only obviously applies to translation but also embraces all biological organization, all things biological. Ultimate explanations in biology will come largely in terms of processes, a process perspective that unavoidably leads back to the dynamics of evolution, the process that gives rise to all of the subordinate biological processes constituting what we take to be biology today. The process of evolution is *a forteriori* nonuniform, and whereas its sporadic nature can be glimpsed throughout the fabric of the cell, perhaps its clearest markings are seen in the signatures of the translation apparatus, i.e., the ribosome and its translation factors.

Evidence today strongly suggests that a highly developed translation system was a necessary condition for the emergence of cells, as we know them (1). In the universal phylogenetic tree (UPT) format, this maturation of the translation system seems to be represented by the tree's basal branchings, where first the bacterial and then the archaeal and eukaryotic lineages appear individually to emerge. What lies beneath this "root" locus, the evolution leading up to it, cannot be captured in familiar tree representation. It would seem to be some distributed universal ancestral state from which the (three) primary organismal lineages materialized via one or a brief series of major evolutionary saltations in which the state of the evolving cellular organization and the accompanying evolutionary dynamic underwent dramatic change. The aboriginal evolutionary dynamic may have been "Lamarckian" in the sense that it seems likely to have involved massive pervasive horizontal transfer of genes (HGT), innovation sharing (2). The kind and frequency of the HGT envisioned would make evolution early on effectively communal. This communal evolutionary dynamic comes to an end relatively suddenly and transforms largely into the familiar genealogical dynamic when the evolving organisms in the community reach a stage of "critical complexity," wherein their organizations change significantly and rapidly, becoming more refined and individualized, more "self-composed." These we call Darwinian transitions (1). Certain signatures in the ribosome, i.e., idiosyncrasies in its RNA (rRNA) (3–6) and/or proteins (r-proteins) characteristic of the individual domains of life were locked in place at this time, becoming molecular fossils that are telling of the phase transitions.

The availability of genomic data and crystal structures for the bacterial small subunit (SSU) and the bacterial and archaeal large subunit (LSU) allows us now to extend the previous analyses of the ribosomal signatures both in depth, by including the r-protein(s), and in scope, by looking at signatures at the levels of structure and genomic organization. Using a variety of techniques, we herein investigate the evolution of the molecular signatures of translation. Understanding the characteristics of that process will help us gain insight into the early evolution of translation, and therefore, of early cellular life.

Results and Discussion

Evolution of rRNA Signatures. The 16S rRNA has become the molecular standard in studying evolutionary relationships between organisms (7). However, the 23S rRNA has followed a very similar (if not identical) evolutionary path, as shown by the congruence of its sequence phylogeny with the UPT [Fig. S1 in supporting information (SI) *Appendix*]. The 23S rRNA therefore provides additional complimentary data that can be tapped to study the evolution of the ribosome.

The 16S and 23S rRNAs each have a high degree of sequence identity, with 30–40% of the well aligned positions between bacteria and archaea being conserved. Yet despite this large degree of identity, there are significant phylogenetic signals in the pattern of change of the remaining nucleotides that can reveal the evolutionary history of the molecules. Among the strongest signals are the signatures, the regions that are constant and unique to, i.e., characteristic of, a particular domain of life. There appear to be two general kinds of signatures here.

Author contributions: E.R., A.S., J.M., C.R.W., and Z.L.-S. designed research; E.R., A.S., and J.M. performed research; and E.R., A.S., J.M., C.R.W., and Z.L.-S. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

¶To whom correspondence may be addressed. E-mail: carl@life.uiuc.edu or zan@uiuc.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0804861105/DCSupplemental.

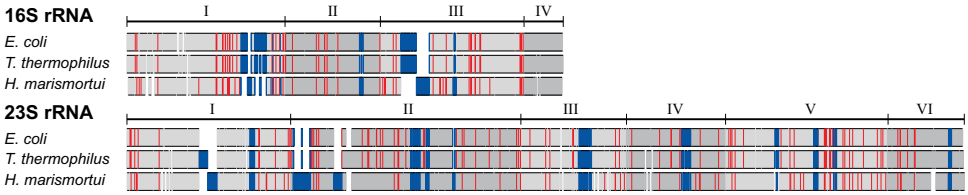

Fig. 1. Location of sequence signatures (red) and structural signatures (blue) in the aligned 16S and 23S rRNA from two bacteria and one archaea. Roman numerals show the domain numbering of the molecules. See Tables S2–S5 in *SI Appendix* for signature positions and domain boundaries.

Sequence signatures comprise positions in the primary structure whose compositions remain constant in one domain of life but occur rarely in the other domains. Structural signatures are regions in the secondary and/or tertiary structure that have a unique configuration in a given domain.

We identified 69 distinguishing sequence signatures between the bacterial and archaeal 16S rRNAs using the sequences of 2,735 organisms and 119 such between their 23S rRNAs based upon 441 sequences (Fig. 1 and Tables S2 and S3 in *SI Appendix*). The sequence signatures constitute ≈5% of the nucleotides in each molecule. Logically, because the compositions of these positions are conserved across the entirety of and are unique to one given domain, their idiosyncrasy must have evolved in the ancestral stem of that domain. Therefore, they should have a large impact on the measure of phylogenetic separation of the two domains; i.e., the distance between the roots of the bacterial and archaeal subbranches.

To estimate the contribution of the signatures to the overall phylogenetic signal, we performed a phylogenetic analysis of the 16S and 23S rRNAs both with and without the sequence signatures. Fig. 2 *A* and *B* show the effect of removing from the calculation the 5% of the sequence positions that constitute the characteristic sequence signatures: a 42% decrease in the separation between the bacterial and archaeal subbranches for the 16S rRNA and a 28% decrease for the 23S rRNA. The decrease in separation holds for a range of signature cutoffs from strict to lax conservation (95–80%). There is no appreciable change in the branchings or distances within the archaeal and bacterial subtrees; the sequence signatures carry only information distinguishing the two domains of life. With such a strong signal, it is not surprising that the three domains of life could be identified and distinguished in 1977 using only oligonucleotides created by T1 ribonuclease cleavage of 16S rRNA (8, 9).

In addition to the sequence signatures in the primary structure, there are regions of the rRNA that contain structural signatures in the secondary or tertiary structure. Such structural signatures can be of three types: (*i*) insertions or deletions (indels) that are characteristically present in one domain of life but absent in another, (*ii*) regions of the rRNA in which the secondary (and therefore tertiary) structure differs between two domains, or (*iii*) regions that are similar in secondary structure but differ in their tertiary conformation.

Using a combination of sequence- and structure-based techniques, we identified six structural signatures distinguishing the bacterial and archaeal 16S rRNAs and 14 distinguishing their 23S rRNAs (Tables S4 and S5 in *SI Appendix*). Because a crystal structure of the archaeal SSU has not been solved, structural signatures for the 16S rRNA are limited to types *i* and *ii* and are reliable only because of the high quality of the available 16S rRNA alignments (6). To exclude crystallization artifacts, type *iii* structural signatures in the 23S rRNA were included only if supported by specific interactions with r-proteins.

The contribution of the structural signatures to the separation between the bacteria and the archaea was calculated as above, by excluding regions containing the structural signature during a sequence phylogenetic analysis. The change in the separation between the bacterial and archaeal subbranches was significantly less than for the sequence signatures (8% for the 16S, 16% for the 23S).

Structural phylogenetic methods (10, 11), which include a framework for modeling indels, provide a (possibly more reliable) alternative for evaluating the phylogenetic contribution of the structural signatures. A structure-based phylogenetic tree (Fig. 2*C*) was generated by using the 23S rRNA structures of one

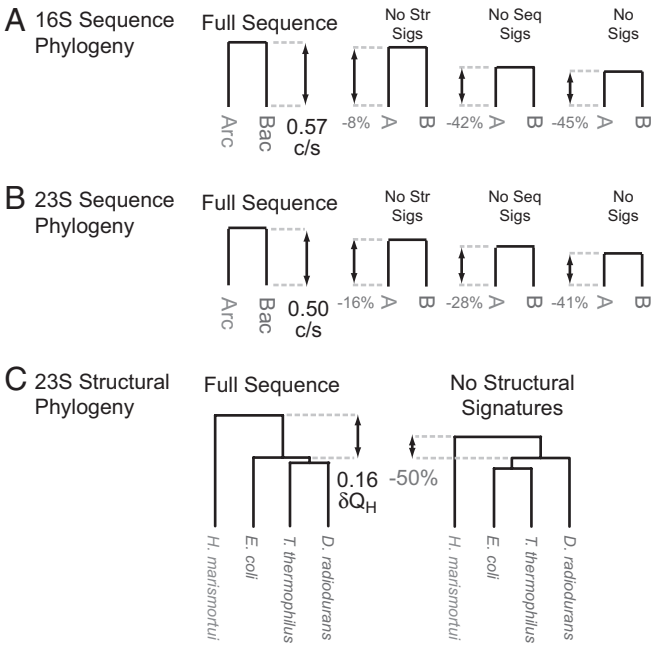

Fig. 2. Contribution of the rRNA sequence and structural signatures to phylogenetic division between the bacteria and archaea in the sequence phylogenies of (*A*) the 16S RNA and (*B*) the 23S RNA. Distances are in changes per site. (*C*) The contribution of the rRNA signatures to the structure phylogeny of the 23S in terms of change in *Q*H (a measure of structural similarity).

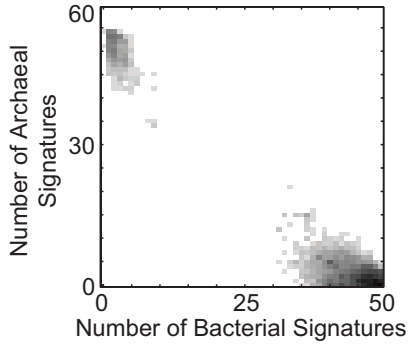

Fig. 3. Distribution of the presence of bacterial and archaeal signatures in 90,000 environmental 16S rRNA sequences.

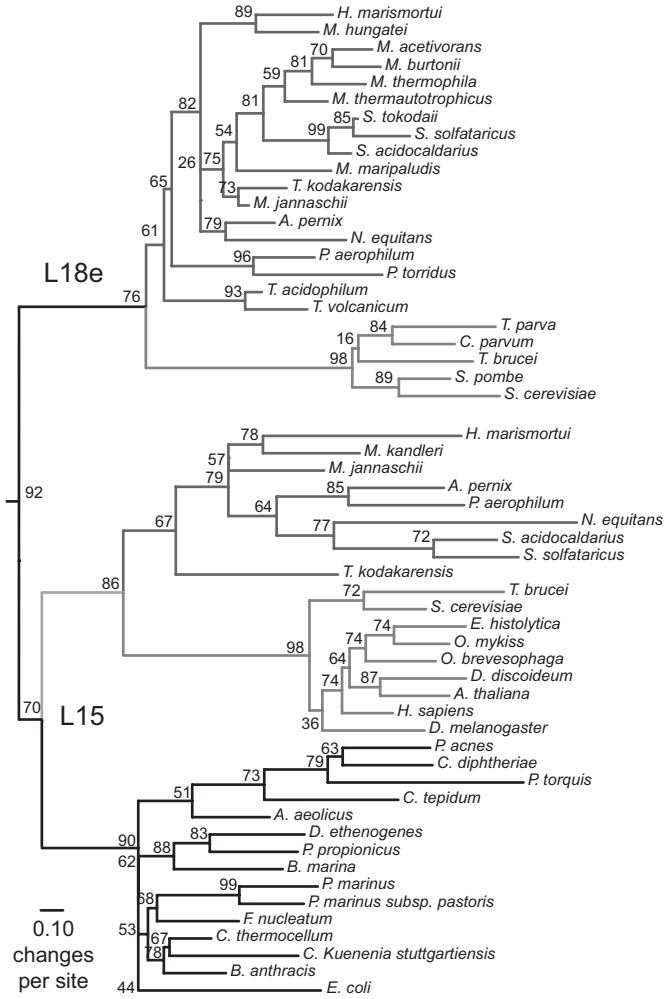

Fig. 4. Combined maximum-parsimony/maximum-likelihood phylogenetic tree of homologous ribosomal proteins L15 and L18e (rooted using L4 as an out-group). Branch values give the local bootstrap probabilities.

archaeal example and three bacterial: *Haloarcula marismortui* (12) and then *Deinococcus radiodurans* (13), *Escherichia coli* (14), and *Thermus thermophilus* (15). It reveals a deep separation between the archaeal and the bacterial 23S rRNA structures, similar to that seen in sequence-based phylogenetic trees. Removing the structural signatures from the structural phylogenetic analysis reduces the separation between the two domains by 50%. The sequence signatures make no contribution to the separation in the structural phylogeny, because the signature nucleotides (despite having different identities) occupy homologous positions in the overall structure. This structural phylogenetic analysis leads us to conclude that the structural signatures are as important as the sequence signatures in defining the differences between the domains of life.

One of the primary indications that the RNA signatures are, in fact, remnants of an evolutionary saltation is their discrete character. There is no signature continuum between the domains of life; organisms either have the bacterial, archaeal, or eukaryal character, with a sizeable two-domain signature that links the archaeal and eukaryal domains (7, 16). We have checked for the presence of the archaeal and bacterial 16S rRNA sequence signatures in >90,000 environmental sequences (see Fig. 3) from the Greengenes database (17). These sequences represent a much wider sampling from the organismal pool than the cultured sequences used initially to identify the signatures. Again, no exceptions are seen; no "gray area" exists between the archaeal and bacterial signatures: the ribosome is of either bacterial or archaeal nature.

Domain-Specific Ribosomal Proteins as Signatures. Comparative analysis of the available sequence and structure data allows us to infer whether a protein existed in the gene pool before the divergence of the primary organismal lineages. The universally distributed r-proteins exhibit what is called the canonical pattern as defined by Woese *et al*. (16), wherein the various taxa group into three distinct clusters (bacteria, archaea, eukarya), with the latter two showing the most structure and sequence similarity. Although the canonical pattern provides evidence that the universal r-proteins were present at the so-called base of the UPT, the situation is less clear with regard to the remaining ribosomal proteins.

It is well known that approximately half of all r-proteins are confined to a subset of the domains of life (ds-proteins; see Tables S1 and S6–S9 in *SI Appendix*). Practically all of the archaeal but none of the bacterial ds-proteins are present in eukarya, consistent with the notion that the bacterial lineage diverged from some ancestral "stem" before either the archaeal or eukaryal lineages. Because the presence of these ds-proteins within their specific domain(s) of life is conserved, with a few exceptions (18), their existence represents another of the signatures distinguishing the ribosome between the domains. The evolutionary history of the ds-proteins can therefore be informative as to the history of the signatures in general.

Many biologists assume that because they are not universal, ds-proteins are of relatively recent evolutionary origin. This need not be so. A phylogenetic analysis of all of the archaeal/eukaryal specific r-proteins shows a deep divergence between archaea and eukarya, as others have observed in specific cases (19). Such a divergence indicates that the ds-proteins developed well before the archaeal and eukaryal lineages diverged. However, a more detailed analysis of a protein's history is possible if it resulted from an earlier gene duplication event. In such a case, a combination of sequence and structural phylogenetic techniques can provide resolution of the phylogenetic relationship between the paralogs (20). Fortunately, there is at least one case of a ds-protein and a universal r-protein sharing common ancestry via gene duplication: L18e and L15. The question then arises as to whether ds-protein L18e is a recent innovation or, alternatively, present at the base of the UPT.

The globular domains of L15 and L18e are similar in both structure and sequence (Q_H = 0.6 and sequence identity of 20%), confirming that these proteins have a common evolutionary origin. Their tails, like many r-proteins, have no sequence or structural homology. Because of the low sequence identity, a structural alignment was used to guide a sequence alignment of L15 and L18e sequences. The phylogenetic tree shown in Fig. 4 is a map of the evolutionary history of L15 and L18e obtained from this alignment. As expected, the L15 sequences display the canonical phylogenetic pattern. The deep divide separating the bacterial and the archaeal/eukaryal versions of the molecule is clearly visible and, in turn, the eukarya are clearly distinguishable from the archaea. The point denoting the root of the L15 tree can be identified between the bacteria and the archaea. The portion of the phylogenetic tree showing the L18e sequences also exhibits a deep archaeal/eukaryal divide. Importantly, L18e appears to branch off before the root of the L15 tree, suggesting that the gene duplication event occurred before the three primary lineages diverged. Given the large evolutionary distance between these two proteins and the moderate length of the homologous region (≈80 residues), this tree must be treated with caution, but the support values give a reasonable probability that L18e is an ancient ribosomal protein, dating from before the divergence.

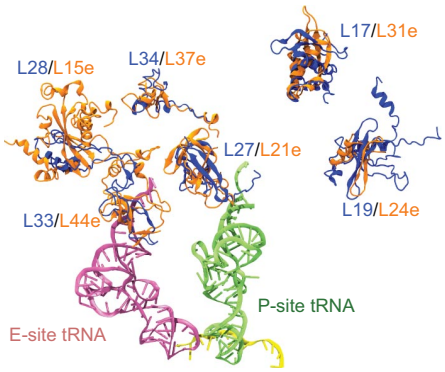

Fig. 5. Spatial analogs (nonhomologous domain-specific ribosomal proteins that occupy the same position in the ribosomal structure) in the large subunit from a 23S rRNA based structural alignment of *Thermus thermophilus* [blue; Protein Data Bank (PDB) ID code 2J01] and *Haloarcula marismortui* (green; PDB ID code 1S72). The P-site tRNA (green), E-site tRNA (red), and mRNA (yellow) are shown for orientation.

Further characteristic ribosomal signatures provided by the ds-proteins can be seen in their binding locations. A structural superposition of the LSUs from *T. thermophilus* and *H. marismortui* establishes six pairs of spatial analogues, ds-proteins that have no detectable structure or sequence homology but interact with the same region of the rRNA (12, 21) (Fig. 5 and Tables S8 and S9 in *SI Appendix*). Despite the lack of homology, spatial analogs often form similar contacts with the RNA. The overlap of the LSUs also reveals several cases in which ds-proteins superimpose on structural signatures in the 23S rRNA or on r-protein extensions in the other domain of life (also in Tables S8 and S9 in *SI Appendix*). Although the phylogenetic contribution of the ds-proteins to the separation between the domains cannot be strictly calculated, they absolutely give the ribosome a distinct structural character in each domain.

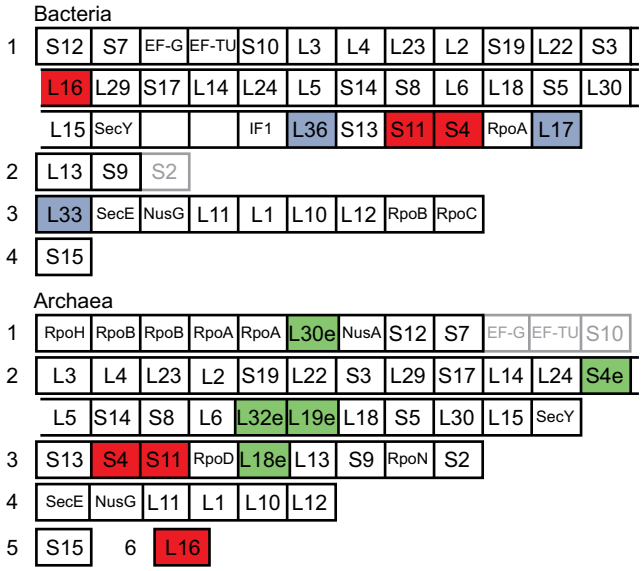

Fig. 6. Consensus diagram of the genomic clusters containing the genes of the universal r-proteins along with other translation and transcription genes in Bacteria and Archaea. Genes are labeled by their product with black indicating presence within the cluster in at least 50% of the genomes analyzed for a domain of life and gray at least 15%. Colors mark signature differences in the genomes between the two domains: universal r-proteins with differences in positioning (red), bacterial specific r-proteins (blue), and archaeal specific r-proteins (green).

Signatures in Genomic Organization. A well documented trait of the universal r-proteins is clustering of their genes in a genome. In many bacteria, all of the universal r-protein genes (except that of S15) are grouped into a few conserved genomic clusters along with the genes of other universally distributed proteins involved in the translation and transcription processes. Likewise, in many archaea the universal r-protein genes (except those of S15 and L16) are organized into similar groups (see Fig. 6). We have analyzed these genomic clusters in representative bacterial and archaeal genomes (listed in Table S10 in *SI Appendix*) looking for characteristic domain specific differences between them.

The relative ordering of universal r-protein genes within a gene cluster is very similar between the bacterial and archaeal genomes, showing just two signatures. First, the order of the genes of r-proteins S4 and S11 is reversed between the two domains (22), and second, the gene for r-protein L16, is missing from the corresponding gene cluster in archaea (23). Despite the conservation in relative ordering of the universal r-protein genes within a cluster, the clusters themselves are organized differently in the two domains of life. The boundaries dividing the clusters are located near different genes in each domain.

A majority of the genes of the ds-proteins are distributed either as isolated genes or in domain-specific clusters. Exceptions are the genes of r-proteins L36, L17, and L33 in bacteria and L30e, S4e, L32e, L19e, and L18e in archaea. Interestingly, these eight ds-protein genes are all located in the clusters containing the universal r-protein genes. The position of each ds-protein gene within a cluster is conserved within the domain of life, and its presence does not perturb the ordering of nearby universal r-protein genes. These ds-protein genes can be considered structural signatures of the bacterial and archaeal genomes. Two of the three bacterial-specific r-proteins whose genes are located in these clusters (L17 and L33) are known to have spatial analogs in the archaeal LSU, and L36 may have one as well (see below).

Correlation of Signatures in rRNA and Ribosomal Proteins. Conservation of the signatures within a given domain indicates they are functionally important, and it is also clear from their phylogenetic distribution that they must have evolved shortly after the divergence of the cellular lineages. Correlations between sequence, structure, and genomic signatures of the ribosome offer insight into their functional relationships and help answer questions about the ribosome's evolution.

Both evolutionary and dynamical correlations result from direct physical contact between signatures. Approximately half of the domain specific LSU r-proteins and nearly all of the 23S rRNA structural signatures interact with each other (Tables S8 and S9 in *SI Appendix*). In each interaction, a ds-protein and an rRNA structural signature create a domain-specific connection between distant regions of the 23S rRNA sequence. Expansion of the network of interactions within the ribosome in this manner is a well known theme in the evolution of the ribosome following the divergence of the lineages (12).

Some interactions between the ds-proteins and the rRNA structural signatures do not expand the interaction network but instead reconnect it in a different pattern. There are large differences in the tertiary structure of helices H15 and H58 of the 23S rRNA between the bacterial and archaeal crystal structures, with no significant differences in their primary or secondary structure. Both helices are held in different orientations by nearby ds-proteins. In bacteria, helix H15 interacts with ds-proteins L9 and L28, whereas in archaea, it contacts ds-proteins L7Ae and L15e. Similarly, helix H58 has no nearby ds-proteins in bacteria, but in archaea, it makes extensive contacts with ds-protein L37Ae. There are changes in the overall ribosomal interaction network as a result of the rearrangement of these two helices. Although it is possible the differences in the tertiary

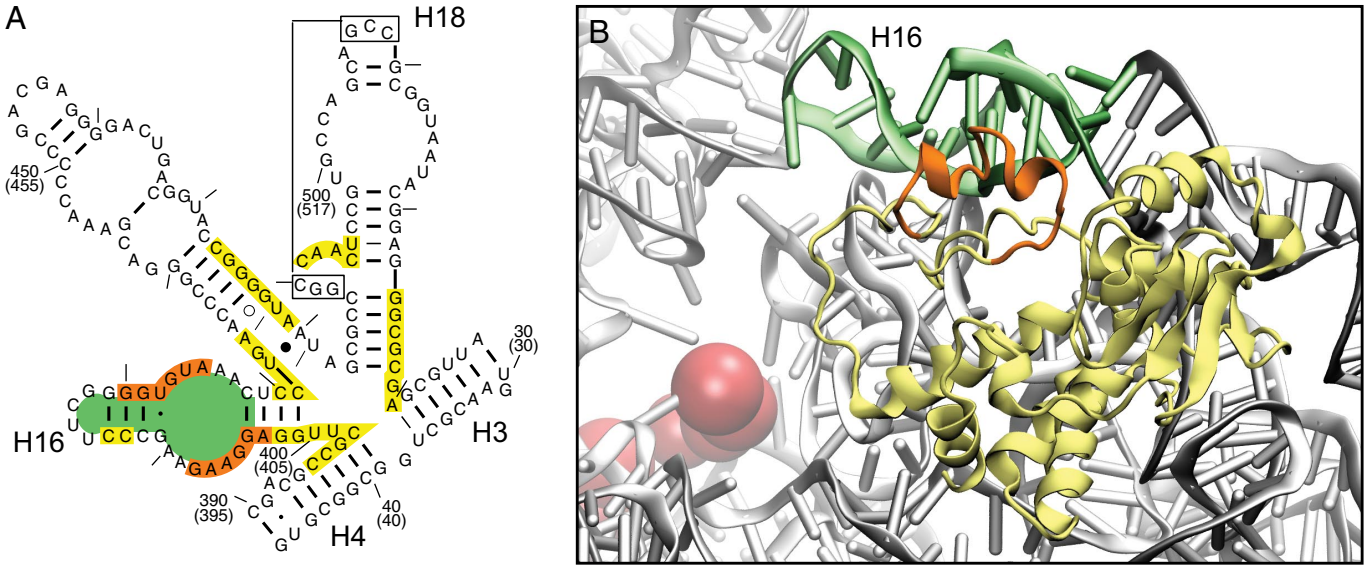

Fig. 7. Illustration of ribosomal protein S4 bound to the 16S rRNA in *T. thermophilus*. (*A*) A diagram of the secondary structure of the rRNA (6) and its interactions with the protein in the vicinity of the binding region and (*B*) a 3D view of the structure in the same region. Colors are consistent between two as follows: the rRNA structural signature (green), the protein structural signature (orange), and the remainder of S4 (yellow). Red spheres show the location of mRNA bound to the ribosome.

conformation of these two helices are crystallization artifacts, the interactions with ds-proteins make it likely that these are physically (perhaps physiologically) relevant differences between these domains of life.

In addition to ds-proteins *per se*, signatures distinguishing the domains are also evidenced by short indels in some universal r-proteins (24). We have looked for coevolution between these indels and rRNA structural signatures by analyzing the covariation between the r-proteins and the rRNA using mutual information. Our analysis identified a bacterial-specific insertion (≈12 aa in length) in the N-terminal domain of universal r-protein S4 that covaries with helix h16 of the bacterial 16S rRNA. Helix h16 was previously identified as one of the strongest signatures in the 16S rRNA that distinguished the archaea from the bacteria (5). The structural signatures in S4 and helix h16 make exclusive contact with each other (Fig. 7), and their coevolution is evidence that even within the universal r-proteins, characteristic signatures were evolving after the lineages diverged.

Ribosomal protein S4 is located near the decoding site of the ribosome and is a primary binding protein in the 30S subunit assembly map (25, 26). Therefore, one asks how the structural signatures in S4 and helix h16 of the 16S rRNA affect the translation process in the bacteria? One possibility is that, with S4's position near the decoding site, the two signatures affect the dynamics of the region through increased interactions between S4 and helix h16. In fact, mutations in S4 are known to affect the precision of translation (27–29), although none occurring within its structural signature have, to our knowledge, been studied. Another possibility, suggested by preliminary studies of the folding of S4 in the presence of 16S rRNA, is that the two structural signatures may make the initial contacts in the docking of S4 to the 16S rRNA during the assembly of the bacterial SSU (Z.L.-S., unpublished data).

Evolution of Ribosomal Proteins L36 and L40e. Another general pattern to emerge from studying the ribosomal signatures is the relationship between domain-specific r-proteins and the conserved core of the rRNA. There are a number of ds-proteins in the bacterial LSU that do not make contact with bacterial 23S rRNA structural signatures but interact only with the structurally invariant rRNA. In the archaeal LSU, these ds-proteins are consistently replaced by either archaeal rRNA structural signatures or archaeal-specific r-proteins (spatial analogs or r-protein extensions). The interaction network within the ribosome remains conserved between the two domains, even though the interactions are provided by different mechanisms. This observation leads us to predict an archaeal spatial analog to the bacterial-specific r-protein L36.

Protein L36 binds to helices from domains II, V, and VI in the bacterial 23S rRNA (these are H42, H89, H91, and H97) and, in some of the *E. coli* LSU crystal structures, it makes contacts with the L11-arm (H43 and H44), to which r-proteins L10 and L11 bind. Two bacterial 23S rRNA sequence signatures (base pairs G2526:U2537 and A1032:G1122) make base-specific contacts with r-protein L36, helping to establish its binding site in the bacteria. The analogous binding site is empty in the structure of the *H. marismortui* LSU, even though the structure of the rRNA in the region is highly conserved. L36 is known to be important for the structural stability of the bacterial LSU; deletion studies in *E. coli* have shown that its absence increases the accessibility of the region to reagents (dimethyl sulfate and hydroxyl radicals) and slows cell growth by 40−50% (30). The lack of any archaeal-specific 23S rRNA structural signatures near the binding site strongly suggests that the structural stabilization afforded to the bacterial LSU by L36 must be, in this case, provided by archaeal-specific r-protein interactions. Because there are no nearby archaeal ds-proteins in the *H. marismortui* LSU structure, the possibility of a yet-unresolved archaeal spatial analog to L36 has to be considered.

The crystal structure of the *H. marismortui* LSU is missing only two of the 14 ds-proteins whose genes are present in the species' genome: L40e and LX. These are the only candidates for an archaeal spatial analog of L36, and of these two, only L40e is present in all archaeal subbranches. Although L40e is slightly longer than L36 (48 vs. 38 aa on average), both proteins are highly basic and contain zinc finger motifs. A sequence comparison shows that the two proteins are not homologous. A solution structure of L40e has recently been determined by NMR spectroscopy (31), and the structures of L40e and L36 have a similar topology. From a superposition of L40e on L36 in

its binding site in the bacterial LSU, it is clear that L40e fits into the cavity created by the junction of the four rRNA helices (Fig. S2 in *SI Appendix*). Additionally, molecular dynamics simulations show that L40e is stable in this position in the archaeal LSU and provides interactions that could help to interconnect the 23S rRNA structure (data not shown).

Additional support for L36 having a spatial analog in the archaeal LSU comes from signatures in the genomic organization of the r-proteins. As discussed previously, only the three genes of bacterial-specific r-proteins L17, L33, and L36 are located in the conserved clusters of universal r-protein genes. Like L36, both L17 and L33 bind to conserved regions of the 23S rRNA with no nearby rRNA structural signatures. Both of these ds-proteins have known spatial analogs in the archaeal LSU (L31e and L44e, respectively). Assuming the shared organization of the genes of these three r-proteins correlates to other shared features, we would again anticipate r-protein L36 to have a spatial analog.

Although no single piece of the above evidence is by itself decisive, the consistency of the accumulated data within the signature framework implies that archaeal ds-protein L40e is the unresolved spatial analog to L36 in the archaeal LSU. Because the L11-arm appears to be open in the *H. marismortui* crystal structure, L40e may have been lost during the crystallization process. The presence of a ribosomal protein in this region of the archaeal LSU would have an impact on the dynamics of the ribosome during translation.

Final Remarks. The emergence of the primary organismal lineages was a profound event in the evolution of life. Through our analysis of ribosomal signatures, we have provided a glimpse into the evolutionary past, at the "base" of the UPT. This study has identified the ribosomal signatures and provided examples of how they are helpful in understanding the evolutionary dynamic by which the ribosome arose. These signatures give each phylogenetic domain a distinctive character and bespeak stages through which the evolution of the ribosome must have proceeded, both before the emergence of the individual lineages themselves (in the universal ancestral state) and subsequently, separately within each primary lineage.

Methods

Sequence alignments for the 16S and 23S rRNAs were obtained from the Comparative RNA Web Site (6) and environmental 16S rRNA sequence alignments from the Greengenes database (17). Genomic data were obtained from the Integrated Microbial Genomes system (32). All sequence and structural analyses, including identification of sequence and structural signatures, were performed by using MultiSeq (33) and VMD (34). Sequence phylogenetic trees were reconstructed by using a combination of maximum likelihood and Bayesian methods using PAUP (35), RAxML (36), and MrBayes (37). Structural phylogenetic trees were calculated by using the Q_H measure of structural similarity (11). The coevolution analysis of r-protein S4 and 16S rRNA was performed using mutual information. All-atom molecular dynamics simulations of r-protein L40e in the archaeal LSU were performed by using NAMD (38). Further details are provided in *SI Methods* in *SI Appendix*.

ACKNOWLEDGMENTS. This work was supported by the Department of Energy (Grant DE-FG02-05ER-64144) and the National Science Foundation (Grant MCB04-46227). We thank Dan Wright for assistance in collecting the data regarding the genomic organization of the ribosomal proteins.

1. Woese CR (2002) On the evolution of cells. *Proc Natl Acad Sci USA* 99:8742–8747.
2. Vetsigian K, Woese C, Goldenfeld N (2006) Collective evolution and the genetic code. *Proc Natl Acad Sci USA* 103:10696–10701.
3. Woese CR (1987) Bacterial evolution. *Microbiol Rev* 51:221–271.
4. Gutell RR, Woese CR (1990) Higher order structural elements in ribosomal RNAs: Pseudo-knots and the use of noncanonical pairs. *Proc Natl Acad Sci USA* 87:663–667.
5. Winker S, Woese CR (1991) A definition of the domains Archaea, Bacteria and Eucarya in terms of small subunit ribosomal RNA characteristics. *Syst Appl Microbiol* 14:305–310.
6. Cannone JJ, *et al.* (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* 3:2.
7. Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA* 87:4576–4579.
8. Fox G, Magrum L, Balch W, Wolfe R, Woese C (1977) Classification of methanogenic bacteria by 16S ribosomal RNA characterization. *Proc Natl Acad Sci USA* 74:4537–4541.
9. Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc Natl Acad Sci USA* 74:5088–5090.
10. O'Donoghue P, Luthey-Schulten Z (2003) On the evolution of structure in the aminocyl-tRNA synthetases. *Microbiol Mol Bio Rev* 67:550–573.
11. O'Donoghue P, Luthey-Schulten Z (2005) Evolutionary profiles derived from the QR factorization of multiple structural alignments gives an economy of information. *J Mol Biol* 346:875–894.
12. 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.
13. Schlünzen F, *et al.* (2005) The Binding Mode of the Trigger Factor on the Ribosome: Implications for Protein Folding and SRP Interaction. *Structure (London)* 13:1685–1694.
14. Berk V, Zhang W, Pai R, Doudna Cate J (2006) Structural basis for mRNA and tRNA positioning on the ribosome. *Proc Natl Acad Sci USA* 103:15830.
15. Selmer M, *et al.* (2006) Structure of the 70S ribosome complexed with mRNA and tRNA. *Science* 313:1935–1942.
16. Woese CR, Olsen GJ, Ibba M, Soll D (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol Mol Bio Rev* 64:202–236.
17. DeSantis T, *et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 72:5069–5072.
18. 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.
19. Yang D, Kusser I, Kopke AK, Koop BF, Matheson AT (1999) The structure and evolution of the ribosomal proteins encoded in the spc operon of the archaeon (Crenarchaeota) Sulfolobus acidocaldarius. *Mol Phylogenet Evol* 12:177–185.
20. O'Donoghue P, Sethi A, Woese CR, Luthey-Schulten ZA (2005) The evolutionary history of Cys-tRNA Cys formation. *Proc Natl Acad Sci USA* 102:19003–19008.
21. Mushegian A (2005) Protein content of minimal and ancestral ribosome. *RNA* 11:1400–1406.
22. Kunisawa T (2003) Gene arrangements and branching orders of gram-positive bacteria. *J Theor Biol* 222:495–503.
23. Hartmann E, Hartmann R (2003) The enigma of ribonuclease P evolution. *Trends Genet* 19:561–569.
24. Vishwanath P, Favaretto P, Hartman H, Mohr S, Smith T (2004) Ribosomal protein-sequence block structure suggests complex prokaryotic evolution with implications for the origin of eukaryotes. *Mol Phylogenet Evol* 33:615–625.
25. Mizushima S, Nomura M (1970) Assembly mapping of 30S ribosomal proteins from *E. coli*. *Nature* 226:1214.
26. Held WA, Ballou B, Mizushima S, Nomura M (1974) Assembly mapping of 30 S ribosomal proteins from *Escherichia coli*. Further studies. *J Biol Chem* 249:3103–3111.
27. Rosset R, Gorini L (1969) A ribosomal ambiguity mutation. *J Mol Biol* 39:95–112.
28. Allen P, Noller H (1989) Mutations in ribosomal proteins S4 and S12 influence the higher order structure of 16S ribosomal RNA. *J Mol Biol* 208:457–468.
29. Carter AP, *et al.* (2000) Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature* 407:340–348.
30. Maeder C, Draper D (2005) A small protein unique to bacteria organizes rRNA tertiary structure over an extensive region of the 50S ribosomal subunit. *J Mol Biol* 354:436–446.
31. Wu B, *et al.* (2008) Solution structure of ribosomal protein L40E, a unique C4 zinc finger-protein encoded by archaeon *Sulfolobus solfataricus*. *Protein Sci* 17:589–596.
32. Markowitz VM, *et al.* (2006) The integrated microbial genomes (IMG) system. *Nucleic Acids Res* 34:344–348.
33. Roberts E, Eargle J, Wright D, Luthey-Schulten Z (2006) MultiSeq: Unifying sequence and structure data for evolutionary analysis. *BMC Bioinformatics* 7:382.
34. Humphrey W, Dalke A, Schulten K (1996) VMD: Visual molecular dynamics. *J Mol Graphics* 14:33–38.
35. Swofford, D (2003) *PAUP* Phylogenetic Analysis Using Parsimony (* and Other Methods)* (Sinauer, Sunderland, MA), Vol. 4.
36. Stamatakis (2006) A RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688.
37. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.
38. Phillips J, *et al.* (2005) Scalable molecular dynamics with NAMD. *J Comp Chem* 26:1781–1802.