Journal of Theoretical Biology 367 (2015) 130–158

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

The ribosome as a missing link in the evolution of life

Meredith Root-Bernstein [a,1], Robert Root-Bernstein [b,*]

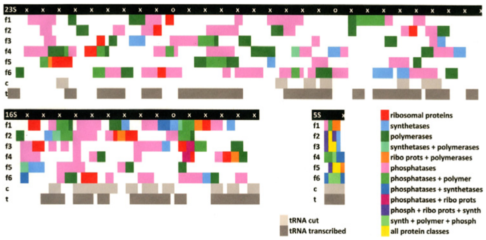
[a] School of Geography and the Environment, Oxford University, South Parks Road, Oxford, Oxfordshire OX1 3QY, United Kingdom
[b] Department of Physiology, Michigan State University, East Lansing, MI 48824, USA

HIGHLIGHTS

- Hypothesize that ribosome was self-replicating intermediate between compositional or RNA-world and cellular life.
- rRNA contains genetic information encoding self-replication machinery: all 20 tRNAs and active sites of key ribosomal proteins.
- Statistical analyses demonstrate rRNA-encodings are very unlikely to have occurred by chance.
- Contradicts view of rRNA as purely structural suggesting instead that rRNA, mRNA and tRNA had common ribosomal ancestor.
- Suggest that DNA and cells evolved to protect and optimize pre-existing ribosome functions.

GRAPHICAL ABSTRACT

Map Illustrating the Location of Transfer RNAs and Proteins in the Six Possible Reading Frames on the 23S, 16S and 5S Ribosomal RNAs of *E. coli* K12. Map suggests that rRNA once contained highly redundant and condensed genetic information encoding ribosome self-replication. "tRNA cut" means tRNAs can be excised from rRNAs; "tRNA transcribed" means tRNA production by transcribing the rRNA.



ARTICLE INFO

Article history:
Received 15 July 2014
Received in revised form
15 November 2014
Accepted 20 November 2014
Available online 10 December 2014

Keywords:
Ribosomal RNA
Transfer RNA
Messenger RNA
Protein synthesis
RNA world

ABSTRACT

Many steps in the evolution of cellular life are still mysterious. We suggest that the ribosome may represent one important missing link between compositional (or metabolism-first), RNA-world (or genes-first) and cellular (last universal common ancestor) approaches to the evolution of cells. We present evidence that the entire set of transfer RNAs for all twenty amino acids are encoded in both the 16S and 23S rRNAs of *Escherichia coli* K12; that nucleotide sequences that could encode key fragments of ribosomal proteins, polymerases, ligases, synthetases, and phosphatases are to be found in each of the six possible reading frames of the 16S and 23S rRNAs; and that every sequence of bases in rRNA has information encoding more than one of these functions in addition to acting as a structural component of the ribosome. Ribosomal RNA, in short, is not just a structural scaffold for proteins, but the vestigial remnant of a primordial genome that may have encoded a self-organizing, self-replicating, auto-catalytic intermediary between macromolecules and cellular life.

1. Introduction

A difficulty in accounting for the emergence of life is to explain how something as complex as a living cell could evolve. At present, several general approaches dominate evolutionary thinking. Working from simplicity to complexity, RNA-world, or "genetics-first" models (reviewed in Strobel, 2001; Neveu et al., 2013) and compositional replication, or "metabolism-first" models (reviewed in Hunding et al., 2006; Glansdorff et al., 2009; Schuster, 2010) together provide insights into early prebiotic evolution from simple molecules to the first polymers and polymer aggregates. Neither of these types of models fully explains the evolution of cells. RNA-world models cannot explain the evolution of metabolism and generally fail to take into account the

[*] Corresponding author.
E-mail address: rootbern@msu.edu (R. Root-Bernstein).
[1] Current address: Department of Bioscience, Aarhus University, Aarhus, Denmark.

http://dx.doi.org/10.1016/j.jtbi.2014.11.025

fact that amino acids (and therefore peptides and proteins) almost certainly were synthesized along with polynucleotides under prebiotic conditions, making it almost certain that these classes of molecules co-evolved (Caetano-Anolles and Seufferheld, 2013; Galadino et al., 2012). Compositional replication models can explain such co-evolution, but not how linear replication schemes became dominant (Schuster, 2010; Norris et al., 2012). Moreover, neither type of model accounts for how simple replicable molecules or aggregates of molecules evolved into complex cells with organized compartments and structures such as ribosomes, acidocalcisomes and functional membranes that incorporated specialized transporters and receptors. Models of the last universal common ancestor (LUCA) – the presumptive first cellular form of life (e.g., Koonin, 2003; Forterre et al., 2005; Mushegian, 2008; Douzounis et al., 2006) – attempt to resolve some of these problems by working from complexity toward simplicity. LUCA models provide insight (with much disagreement) into the minimum complexity required for cellularity but reveal little about the preceding evolutionary steps. The gap is enormous between the simplicity-toward-complexity models, which can suggest how simple replication of small sets of polymers may have emerged, and complexity-toward-simplicity models, which suggest a minimum of several hundred genes and their products networked within specialized metabolic compartments. What kind of evolvable entities might bridge this gap?

Evolvable entities existing between self-replicating polymers and fully functional cells would presumably have many, though not all, of the functions of a cell, yet be significantly simpler in composition and organization. These entities would be able to self-organize and replicate themselves; store information and replicate that information; translate the information into the components necessary to produce their functional structures; capture metabolic components and energy; and transform these into useful biochemical networks. Norris and his colleagues have called functional forms of organization midway between macromolecules and cells “hyperstructures” (Norris et al., 2007). Such hyperstructures had to be instantiated as evolvable entities, meaning that their components would be subject to variation, replication and natural selection. Most importantly, these evolvable hyperstructure entities should exist in a vestigial form in living systems today since evolution tends to be parsimonious, utilizing whatever modules have survived previous rounds of selection to evolve the next set. Indeed, it is this parsimony that produces a molecular paleontology permitting evolution to be studied.

We suggest that a ribosome-like entity was one of the key intermediaries between prebiotic and cellular evolution.

Ribosomes are prerequisites to all cellular life, ubiquitously conserved, with genetic roots that pre-date LUCA, and therefore entities that had to evolve prior to cellular life itself (Mushegian, 2008; Wang et al., 2009; Fox, 2010). While the ribosome may not be capable of the broad metabolic processes that characterize cellular life, the ribosome is a self-organizing complex composed of both polynucleotides and proteins that could link RNA-world to compositional replication concepts in the origins of life. Moreover, ribosomes carry out some of the most fundamental processes characteristic of living systems, including a coordinated series of chemical reactions capable of translating genetic information into functional proteins. What ribosomes are not thought to do is to carry genetic information, and in particular the genetic information required to encode their own structures and functions. But what if ribosomal RNA (rRNA), which is generally considered to be simply a structural component of ribosomes, actually represents a primitive genome encoding the genetic information needed to direct ribosomal replication, translation and self-organization?

It is important in evaluating the results reported below to keep in mind our hypothesis, which is that the ribosome evolved prior to cellular life and had the capability of genetically encoding its own transcription and translation apparatus. rRNA should therefore encode the tRNAs and proteins necessary to ribosomal function. (This statement must, of course, be moderated somewhat by the fact that ribosomes have existed within cells for billions of years so that any information they once contained will have become, by this point in time, somewhat degraded or vestigial in nature.) This hypothesis must be compared to the modern textbook view of ribosomal RNA, which is that it is purely structural in nature, encoding no genetic information. This textbook view might be thought of as the “null hypothesis”. An intermediate hypothesis might be that the amount of genetic information encoded in rRNA is purely random and therefore the number of tRNAs and ribosome-related proteins that rRNA encodes will be no more or less than any random assortment of any other set of randomly chosen RNAs. The tests reported below were chosen to differentiate between these three hypotheses.

2. Methods

We chose *Escherichia coli* K12 for our study of the possibility that rRNA encodes other functional molecules on the basis that that such a study should initially be performed on an organism such as a bacterium that is considered to be evolutionarily primitive. Moreover, the *E. coli* K12 genome and proteins have been very well characterized.

2.1. Sequence sources

The *E. coli* K12 rRNA sequences were obtained from the EcoliWiki (http://ecoliwiki.net/colipedia/index.php/16S_rRNA:Gene_Product%28s%29). The tRNA sequences were obtained as genes (i.e., DNA sequences) from Genomic tRNA Database at the University of California, Santa Cruz (http://gtrnadb.ucsc.edu/Esch_coli_K12/Esch_coli_K12-structs.html). Control mRNA sequences from the *E. coli* K12 genome were acquired from http://microbes.ucsc.edu/lists/eschColi_K12/refSeq-list.html. The control proteins used were: (1) the predicted fimbrial-like adhesin protein, b0135 (1239 bp); (2) a non-coding region of the genome, b0135 (769 bp); (3) broad specificity sugar efflux system protein, b0070 (1179 bp).

2.2. tRNA homology search

The possibility that *E. coli* K12 rRNA encodes tRNAs was explored using LALIGN (Huang and Miller, 1991) at www.expasy.ch. The alignment method was a “DNA” search for one alignment, “global without end gap penalty”. Parameters for opening and end gap penalties were left at their default values. For consistency's sake, the ribosomal sequence was put in the first box as a “Plain Text”. The tRNA sequence was entered in the second box, also as “Plain Text”. The tRNA sequences were then searched in two ways. The first method involved transcribing the DNA sequences (i.e., genes encoding tRNAs) into RNA sequences using a complementary strand program (http://clasher.myweb.uga.edu/testpages/seqconv.html or http://bioinfx.net/). This first method matches the tRNA produced by its gene to the existing rRNA. A match between the tRNA and rRNA would suggest that the rRNA itself is the direct source of the tRNAs, presumably by a function such as fragmentation. The second method used for matching involved substituting each thymidine (T) base in the tRNA gene into a uracil (U) base using the “Find and Replace” function of Microsoft Word. This form of matching tRNA to rRNA assumes that the tRNAs might be encoded in an RNA strand complementary to the rRNA. In other words, perhaps in a primitive pre-cellular system, the tRNAs were transcribed off of the rRNAs rather than existing as fragments within them.

We note that the choice of similarity search program is essential to the results reported here. For example, if one uses the GENBANK BLAST program at NCBI to search for tRNA–rRNA similarities, the search will yield “no results” even in the “somewhat similar (blastn)” mode. The problem is that this mode looks for identical sequences that have the number of identities that are preset in the algorithm parameters. The preset on 'somewhat similar' is 11, meaning that there must be at least 11 identical base matches in a row or nothing is output. Changing the preset to 7 yields "hits" when comparing *E. coli* tRNAs to *E. coli* rRNAs, but the output is only for that specific region of identity. The program does not show how much of the rest of the sequence matches nor does it provide information about gapped matches. In other words, the program does not, as LALIGN does, look for the best overall match across the entire input sequence. We therefore caution readers to pay close attention to the parameters built into search programs when attempting to perform the kinds of analysis reported here. Knowing the purpose, and how each program performs its search, is an essential element of such investigations.

2.3. tRNA secondary structure computation

The most probable secondary structures for the tRNA-like similarities observed in the tRNA Homology Search above were computed using the RNA Structure Web at the University of Rochester: http://rna.urmc.rochester.edu/RNAstructureWeb/Servers/Predict1/Predict1.html. Two features of the website were utilized. One was the calculation of each tRNA similarity individually; the other was the calculation of the most likely common structure shared by two tRNA. The latter was used specifically to determine whether an rRNA-encoded tRNA similarity retained sufficient identity to a modern *E. coli* K12 tRNA to fold into the same secondary pattern.

2.4. Ribosomal protein homology search

BLAST2.0 (Altschul et al., 1997) at www.expasy.ch was used to determine whether *E. coli* K12 rRNA encodes any proteins related to ribosomal function. The 5S, 16S and 23S rRNA sequences as well as the mRNAs of fimbrial protein, sugar efflux protein and a non-coding region (controls), were translated into all six possible reading frames, one through three, the standard directly-encoded ones ($5' \rightarrow 3'$) and four through six, the inverse complements ($3' \rightarrow 5'$). These translations were carried out using the Translate Tool at the Swiss Institute of Bioinformatics http://web.expasy.org/cgi-bin/translate/dna_aa. BLAST searches were carried through the Swiss Institute of Bioinformatics website (http://www.expasy.org) out against all six reading frames for each of the three rRNA sequences. Two types of BLAST searches were used. In the first, the translated rRNA sequences were entered in RAW format and compared with the ECOLI *Escherichia coli* K12 proteome from the "Select Microbial Proteome" list using a blastp program. Scoring sequences, best alignments to show, and *E* threshold were all set to "1000" and the "gapped alignment" was turned off. We emphasize that the use of the *E* threshold was not to produce statistical information – nor could it since every search was done using the same *E*-value cutoff and would therefore have essentially identical statistical probabilities – but rather to ensure that the data sets produced by the rRNA and control mRNA were of the same quality and value. Statistical analysis of the results was performed as an independent step described below after the data sets of rRNA and control mRNA had been evaluated for the presence of ribosome-associated proteins. The purpose of this method was to determine whether rRNA was more likely than random sets of mRNA to encode ribosome-like proteins. In the second type of BLAST search, the six rRNA proteins for each of the three mRNAs were again entered in RAW format into the blastp program, but "*Escherichia coli*" was chosen from the "Select a Database" section rather than the "Select a Microbial Proteome" section. This second search was broader than the first. The other parameters in this second search were the same as in the first search.

2.5. Statistics

While BLAST searches come with built-in statistics such as the *E* value, these built in statistics were ignored in the current study and used instead to generate data sets that were of equal probabilistic values from which to start the testing of the hypotheses presented in the paper. These hypotheses generally take the form of proposing that rRNA is much more likely than random sets of mRNA to encode ribosome-related functions such as tRNAs and ribosomal proteins. By using the same output values in all BLAST searches for both rRNAs and mRNAs, it was possible to assure that the datasets that were generated were comparable when performing additional statistical measures directly addressing the hypotheses.

Statistical comparisons between the tRNA encodings found in the rRNAs and mRNAs of the control sequences were made using a Kolmogorov–Smirnov test. In initial trials, the tRNA results for each rRNA or mRNA were sorted by the amino acid encoded or by total number of identities; no differences in results were found comparing these two sortings, so amino acid sorting was adopted.

To generate the statistical comparisons between the protein encodings of the rRNAs and the mRNAs of the control sequences, only the similarities from the blastp search described above that had an *E* value of less than 100 were utilized. Using this cutoff limited the amount of data that needed to be compared to manageable proportions, since the number of total protein similarities for all six reading frames combined from any given RNA sequence was under 160 at $E=100$, whereas using $E=1000$ yielded between 50 and 300 per reading frame. The Fisher Exact Test was obtained from http://graphpad.com/quickcalcs/contingency2/ The Kolmogorov–Smirnov (K–S) test was run in R version 3.1.0 using the ks.test command. The Bonferroni correction was applied to all tests in order to control the family wise error rate, due to multiple testing of each data set (Gould and Gould, 2002).

To generate the statistical comparisons between active sites within rRNA-encoded proteins and mRNA-control-encoded proteins, all of the ribosomal-function-related proteins generated by the method described above were manually entered one at a time into the UniProt database system via www.expasy.org. The number of protein sequences for which there was no structural information was accumulated as was the number for which there was structural information. Those for which structural information was available were categorized as either falling into a region of the protein that did not have a known function or, if it overlapped or encompassed a functional region, was listed as having a known function. The percentage of rRNA-encoded protein similarities having a known function was calculated by dividing the number of proteins with known function by the total number of proteins with known and unknown functions (but ignoring the proteins for which no functionality was available). The percentage of active proteins from rRNA-encoded proteins was then compared by chi-squared analysis to the percentages of rRNA-encoded active proteins from mRNA-encoded proteins. Since each percentage was used in multiple tests, a Bonferroni correction was applied to all tests (Gould and Gould, 2002).

3. Results

Overall, our results clearly favor the hypothesis that the ribosome may have been a primordial self-replicating entity over the alternative hypotheses that rRNA contains no genetic information or that it contains random genetic information such as might be found by chance in any string of RNA.

3.1. tRNAs encoded in rRNAs

In order to determine whether rRNA encodes any tRNA-like sequences, the LALIGN [16] similarity program on the Swiss Institute of Bioinformatics (www.expasy.ch) was used to compare *E. coli* K12 rRNA 5S, 16S and 23S sequences with *E. coli* K12 tRNA sequences (see Section 2 for sources of sequences) in a pairwise fashion using the DNA function. The LALIGN results were performed in two ways. One was a global search without end-gap penalties and the other was a local search with end-gap penalties at the default settings. All reported similarities between the rRNA and tRNAs were at least 50% identical over the entire tRNA sequence, and as much as 70% identical. In most cases, the reported similarities include an identity at the anticodon site for the particular amino acid tRNA in question. These results were compared with the similarities derived from an identical search of an *E. coli* K12 fimbrial protein RNA, sugar efflux protein RNA and a non-coding region of the genome for tRNA similarities. The control mRNAs had significantly fewer tRNA similarities of any given degree of identity than did the rRNAs. The results of the global search are shown in Table 1. In the global comparisons, the 16S and 23S sequences do not differ significantly from each other in the degree to which tRNA sequences are found in their sequences ($p=0.818$, Kolmogorov–Smirnov test), nor do the control sequences differ significantly from one another (see Table 1). Both the 16S and 23S sequences, however, differ significantly from the control sequences (all combinations $p<0.017$, Kolmogorov–Smirnov test with Bonferroni correction for 3 comparisons ($\alpha=0.017$)). Very similar differences were found in the local similarity search (Table 2).

Encoding of tRNAs in rRNAs was found to occur in two ways. All genetically encoded tRNAs for all twenty standard amino acids are encoded indirectly in both the 16S and 23S rRNAs. Fig. 1 shows the actual results of the homology searches for the indirect encodings found on the 16S rRNA, while Fig. 2 provides a graphical summary of the same results. Fig. 3 shows the graphical summary for the indirect encodings of tRNA on the 23S rRNA. These indirectly encoded tRNAs can be produced by replicating the appropriate rRNA sequence to produce a complementary RNA that would function as a tRNA. Alternatively, if the entire rRNA could itself be replicated into a complementary rRNA, that complementary rRNA could be cut or edited into fragments to produce appropriate tRNA sequences.

The entire set of genetically encoded tRNAs are also encoded directly in the 16S rRNAs so that it would be possible to generate the tRNA sequences by cutting or editing the rRNA itself into appropriate fragments (graphically represented in Fig. 4). The 23S rRNA, however, directly encodes only six tRNAs (graphically represented in Fig. 5). The 5S rRNA contained one tRNA-like sequence similar to the alanine and arginine tRNA (not shown). The fact that the search program did not identify tRNA-like sequences for the vast majority of the amino acids by direct homology in either the 23S or 5S rRNAs provides a good negative control helping to confirm that the positive results described above are unlikely to be due to chance.

It is striking that the entire set of tRNAs appear to be encoded in a redundant fashion within the 16S and 23S rRNAs. Not only does the entire set of tRNAs appear both directly and by replication in the 16S rRNA, but they are repeated in the 23S rRNA. In some cases, the redundancy is such that tRNAs encoding more than one amino acid overlap within the same sequence. In other cases, the best match for several different tRNAs localizes to an identical sequence, suggesting that primitive tRNAs may have been less specific than those that evolved more recently—a result that is not unexpected in a primitive system.

3.2. rRNA-encoded tRNAs fold properly

While it is not possible to say with certainty that any of the overlapping tRNAs observed here were functional in a pre-cellular world without performing appropriate experimental tests (see Discussion), it is notable that the vestigial sequences retained in the ribosomal RNA do retain, at least theoretically, the ability to fold into tRNA-like structures. In order to demonstrate the possibility that the rRNA-encoded tRNA-like sequences might have had actual tRNA functions, the homologous sequences were input into the RNA Structure Web at the University of Rochester and the most

Table 1
Statistical comparison of probabilities that differences in tRNA sequence appearance in rRNAs and control RNAs are due to chance. Each tRNA encoded in *E. coli* K12 was compared with each rRNA (16S and 23S) of *E. coli* K12 and with the *E. coli* K12 fimbrial protein, sugar efflux protein, and non-coding region mRNAs to determine whether that tRNA sequence appeared in the rRNA or mRNA. Only those sequences having at least 50% identity over the entire tRNA sequence were considered to be a "match". tRNA–RNA sequence similarities were determined using an LALIGN *global* DNA search that looks for the best overall match for the entire search (tRNA) sequence. The number of "matches" was compared. The top row of statistics for each match are the *p* values from the Kolmogorov–Smirnov test. Bonferroni correction for 3 comparisons for each data set means that significance at the $p=0.05$ level is accepted at $p=0.017$ (i.e., $\alpha=0.017$). *p* Values that remain significant are in bold. *D* gives the effect size. 16S is the 16S rRNA; 23S is the 23S rRNA; FIMBRIAL is the predicted fimbrial-like adhesin protein, b0135; SUG EFFL is the broad specificity sugar efflux system protein, b00702; and NON-CODE is a non-coding region of the genome, b0135. The results clearly demonstrate that rRNA encodes tRNAs at a significantly higher rate than a random assortment of mRNAs, and certainly higher than would be predicted from the "null hypothesis".

tRNA GLOBAL COMPARISONS	**23S**	**FIMBRIAL**	**SUG EFFL**	**NON-CODE**
	AVG 53.9	AVG 35.0	AVG 36.1	AVG 24.8
16S	$p=1.0$	$p=\mathbf{0.0015}$	$p=\mathbf{0.0047}$	$p<\mathbf{0.0001}$
AVG 55.9	$D=0.1$	$D=0.6$	$D=0.55$	$D=0.75$
23S		$p=\mathbf{0.0015}$	$p=\mathbf{0.0047}$	$p=<\mathbf{0.0001}$
AVG 53.9		$D=0.6$	$D=0.55$	$D=0.7$
FIMBRIAL			$p=0.978$	$p=0.3291$
AVG 35.0			$D=0.15$	$D=0.3$
SUG EFFL				$p=0.3291$
AVG 36.1				$D=0.3$

Table 2
Statistical comparison of probabilities that differences in tRNA sequence appearance in rRNAs and control RNAs are due to chance. tRNA-RNA sequence similarities were determined using an LALIGN *local* DNA search, which looks for the best match between any two regions of the sequences being compared. As in Table 1, only those sequences having at least 50% identity over the entire tRNA sequence were considered to be a "match". The number of "matches" was compared for the 16S and 23S rRNAs and the fimbrial, sugar efflux and non-coding mRNA controls. Bonferroni correction for 3 comparisons for each data set means that significance at the $p=0.05$ level is accepted at $p=0.017$ (i.e., $\alpha=0.017$). *p* Values that remain significant are in bold. *D* gives the effect size. See Table 1 for key to RNA identities. As in Table 1, the results clearly demonstrate that rRNA encodes tRNAs at a significantly higher rate than a random assortment of mRNAs, and certainly higher than would be predicted from the "null hypothesis".

tRNA LOCAL COMPARISONS	**23S**	**FIMBRIAL**	**SUG EFFL**	**NON-CODE**
	AVG 58.0	AVG 40.5	AVG 42.1	AVG 44.4
16S	$p=0.8186$	$p=\mathbf{0.0047}$	$p=\mathbf{0.0047}$	$p=\mathbf{0.0135}$
AVG 54.8	$D=0.2$	$D=0.55$	$D=0.55$	$D=0.5$
23S		$p=\mathbf{0.0135}$	$p=\mathbf{0.0135}$	$p=\mathbf{0.0047}$
AVG 58.0		$D=0.5$	$D=0.5$	$D=0.55$
FIMBRIAL			$p=0.5596$	$p=0.1725$
AVG 40.5			$D=0.25$	$D=0.35$
SUG EFFL				$p=0.978$
AVG 42.1				$D=0.15$

likely secondary structures computed. A selection of these computed secondary structures is shown in Figs. 6–8. Fig. 6A and B display the lowest energy secondary conformations of replicated 16S and 23S Asp tRNA homologues. Fig. 7A and B display the lowest energy homologue of the 16S Asn tRNA compared with the lowest energy secondary conformation of the normal *E. coli* K12 Asn tRNA. Notably, all four structures share common features such as a loop formed near residue 20, another loop near residue 40 and

```
E_ E|  AAAUUGAAGAGUUUGAUCAUGGCUCAGAUUGAACGCUGGCGGCAGGCCUAACACAUGCAA

E_     ------------------------------------------------------------
               70        80        90       100       110       120
E_ E|  GUCGAACGGUAACAGGAAGCAGCUUGCUGCUUCGCUGACGAGUGGCGGACGGGUGAGUAA

E_     ------------------------------------------------------------
              130       140       150       160       170       180
E_ E|  UGUCUGGGAAGCUGCCUGAUGGAGGGGGAUAACUACUGGAAACGGUAGCUAAUACCGCAU

E_     ------------------------------------------------------------
              190       200       210       220       230       240
E_ E  AAUGUCGCAAGACCAAAGAGGGGGACCUUCGGGCCUCUUGCCAUCGGAUGUGCCCAGAUG
                                    |||| |  | | ||  |  | || | | |
Trp    ---------------------------AGGGGCGUAGUUCAAUUGGUAGAGCACCGGUC
                                   TRP     10        20        30
              250       260         270       280       290      300
E_ E  GGAUUAGCUUGUUGGUGGGGUAACGG--CUCACCAAGGCGACGAUCCCUAGCUGGUCUGAGA
           | |  | || ||||  ||     ||| || |  |    |||| |||||   |||
Trp   UCCAAAACCGGGUGUUGGGAGUUCGAGUCUCUCCGCCCCUG-CAUCCGUAGCUCAGCUGGAU
    THR & GLN 40        50        60        70 ARG & MET10        20
           70
               310        320       330       340       350
E_ E|  GGAUGAC-CAGCCAC-ACUGGAACUGAGACACGGUCCAGACUCCUACGGGAGGCAGCAGU
        ||  || | || || |   || | |    | | ||  || |||| | ||| ||| ||
E_     AGAGUACUCGGCUACGAACCGAGCGGUCGGAGGUUC--GAAUCCUCCCGGAUGCACCA--
              30        40        50          60        70
              370       380       390       400       410       420
E_ E  GGAAUAUUGCACAAUGGGCGCAAGCCUGAUGCAGCCAUGCCGCGUGUAUGAAGAAGGCCU
                                                       |  ||  | |
Pro    -----------------------------------------------CGGUGAUUGGCG
                                                       PRO     10
              430       440       450       460       470       480
E_ E  UCGGGUUGUAAAGUACUUUCAGCGGGGAGGAAGGGAGUAAAGUUAAUACCUUUGCUCAUU
       |   ||  | ||    ||||  | | |||    | |   | |    ||
Pro    CAGCCUGGUAGCGCACUUCGUUCGGGACGAAGGGGUCGGAGGUUCGAAUCCUCUAUCACC
             20        30        40        50        60        70
              430       440       450       460       470       480
E_ E|  UCGGGUUGUAAAGUACUUUCAGCGGGGAGGAAGGGAGUAAAGUUAAUACCUUUGCUCAUU
             |     ||| || |||  || || || || | |  |||||| | | || ||| |
E_     ------UCCUCUGUAGUU-CAGUCGGUAG-AACGGCGGACUGUUAAUCCGUAUG-UCACU
           ASN & TYR 10         20         30        40         50
                490        500       510       520       530
E_ E|  GA--CGUUACCCG-CAGAAGAAGCACCGGCUAACUCCGUGCCAGCAGCCGCGGUAAUACG
       |   ||   || | |||| ||  ||
E_     GGUUCGAGUCCAGUCAGAGGAGCCA-----------------------------------
              60        70
              550       560         570       580       590
E_ E|  GGUGCAAGCGUUAAUCGGAAUUACUGGGC--GUAAAGCGCACGCA-GGCGGUUUGUUAAGU
                       |  ||  ||  ||  | | |||||  ||| |||  |     | ||
E_     -------------GGGGCUAUAGCUCAGCUGGGAGAGCGCUUGCAUGGC-AUGCAAGAGGU
                    ALA     10        20        30        40

      600       610         620       630       640       650
E_ E|  CAGAUGUGAAAUCCCCGG--GCUCAACCUGGGAACUGCAUCUGAUACUGGCAAGCUUGAG
       |||  ||   |||||     |||| |||
E_     CAGCGGUUCGAUCCCGCUUAGCUCCACCA-------------------------------
         50        60        70
```

Fig. 1. Mapping of *E. coli* K12 tRNA homologies onto the *E. coli* K12 16S rRNA assuming that the tRNAs are transcribed from the rRNA. Top row (E1) is the 16S rRNA (1542 base pairs in length); bottom row shows where the best homology for each tRNA maps; solid bars between the lines indicate identities between the base pairs at that position. Underlined base pairs indicate where the anticodon is in each tRNA and its homology. Note that many of the encodings of tRNAs overlap.

```
        660       670       680       690       700       710       720
E_ E|  UCUCGUAGAGGGGGGUAGAAUUCCAGGUGUAGCGGUGAAAUGCGUAGAGAUCUGGAGGAAUACC

E_     ----------------------------------------------------------------
              730       740       750       760       770
E_ E|  GGUGGCGAAGGCGGCCCCCUGGACGAAGACUGACGCUCAGGUGCGAAAGCG--UGGGGAG
                                  | ||   |||||| || || ||||  ||    |
E_     -------------------------GGGGCUAUAGCUCAGCUGGGAGAGCGCCUGCUUUG
                                ALA     10        20        30
      780                790       800        810       820
E_ E|  CAAACAGGA---------UUAGAUACCCUGGUAG-UCCACGCCGUAAACGAUGUCGACUU
       ||  |||||         || ||| |||   ||| |||||
E_     CACGCAGGAGGUCUGCGGUUCGAU-CCCGCAUAGCUCCACCA------------------
          40        50         60        70
              790       800       810       820       830       840
E_ E  AACAGGAUUAGAUACCCUGGUAGUCCACGCCGUAAACGAUGUCGACUUGGAGGUUGUGCC
        | ||| ||       ||||||  |||| | |      || | |  | | ||| | |
Gly   -GCGGGAAUAGCUCAGUUGGUAGAGCACGACCU------UGCCAAGGUCGGGGUCGCGAG
      GLY       10        20        30              40        50
              850       860       870       880       890       900
E_ E  CUUGAGGCGUGGCUUCCGGAGCUAACGCGUUAAGUCGACCGCCUGGGGAGUACGGCCGCA
       | ||| | | | |||| |  | |
Gly   UUCGAGUC-UCGUUUCCCGCUCCA------------------------------------
            60         70
              850       860       870       880         890
E_ E  CUUGAGGCGUGGCUUCCGGAGCUAACGCGUUAAGUCGACCGCCUGGGGA--GUACGGCCG
                 || | |  |||| |   | | ||   ||| ||| |  |  ||  ||
Val    ----------GCGUUCAUAGCUCAGUUGGUUAGAGCACCACCUUGACAUGGUGGGGGUC
                  VAL     10        20        30        40
      900       910       920       930       940       950
E_ E  CAAGGUUAAAACUCAAAUGAAUUGACGGGGGCCCGCACAAGCGGUGGAGCAUGUGGU--UUA
         ||||  |   ||| |||      ||||   ||| ||||||||   |||  | ||  | |
Val   GUUGGUUCGAGUCCAAUUGAACG---GGGGUAUCGC-CAAGCGGUAAGGCACCGGAUUCUGA    50
       50        60        70  0        10 ILE, GLN, PHE, LYS    30
     960       970       980       990      1000      1010
E_ E|  UUCGAUGCAACGCGAAGAACCUUACCUGGUCUUGACAUCCACGGAAGUUUUCAGAGAUGA
       |||   |||   ||| |  |    ||| ||     || |||
E_     UUCCG-GCAUUCCGAGGUUCGAAUCCUCGUACCC-CAGCCA-------------------
         40         50        60         70
             1030      1040      1050      1060        1070
E_ E  AAUGUGCCUUCGGGAACCGUGAGACAGGUGCUGCAUGGCUG-UCGUCAG-CUCGUGUUGU
                       || | |   || |  |  |||  | ||||| | ||| |
Met    ----------------CGCGGGGU-GGAGCAGCCUGGUAGCUCGUCGGGCUCAUAACCC
                        MET      10        20        30        40
     1080       1090       1100      1110      1120      1130
E_ E  GAA-AUGUUGGGUU-AAGUCCCGCAACGAGCGCAACCCUUAUCCUUUGUUGCCAGCGGUC
      ||| ||  | |||| || ||| ||  |   |||||||
Met   GAAGAUCGUCGGUUCAAAUCCGGCCCC---CGCAACCA----------------------
             50        60           70



             1150      1160      1170      1180      1190      1200
E_ E  CGGGAACUCAAAGGAGACUGCCAGUGAUAAACUGGAGGAAGGUGGGGAUGACGUCAAGUC
                                                                ||
Glu    ---------------------------------------------------------GUC

             1210      1220      1230      1240      1250      1260
E_ E  AUCAUGGCCCUUACGACCAGGGCUACACACGUGCUACAAUGGCGCAUACAAAGAGAAGCG
        | | | |   | | |||||   |||| ||  ||   | ||||   |||  | |   ||
```

Fig. 1. (*continued*)

a third loop located near residue 60. Thus, some of the canonical aspects of the cloverleaf pattern associated with tRNA structures are found in these ribosomally-derived sequences as well.

One unexpected (for us), but widely repeated, result of our tRNA structural investigation is shown in Fig. 8A and B, where the secondary structure of the normal *E. coli* K12 Ala tRNA is compared with the favored secondary structure of the transcribed 23S Ala tRNA homologue. The lowest energy conformation calculated by the program was not the typical cloverleaf pattern, but an even more energetically favored conformation (energy for Ala tRNA −29.0 versus −27.4 for the more typical cloverleaf pattern). This alternative conformation still has loops near residues 20, 40 and 60, but with an overall arrangement that is more linear. This linear organization was repeated seen in other tRNAs and homologues such as those for the 16S and 23S Gly tRNA homologues (data not shown).

Such linear arrangements of tRNA have been observed experimentally. Some mitochondrial tRNA also lack one of the three "leaves" of the typical cloverleaf pattern (see, e.g., Belostotsky et al., 2011; Pereira and Baker, 2004; Ohtsuki and Watanabe, 2007; Watanabe et al., 2014). This alternative structure may therefore

```
Glu   CCCUUCGUCUAGAGGCCCAGG---ACAC-CGCCCUUUCACGGCGGUAACA--GGGGUUCG
       GLU  10        20            30        40          50
             1270      1280      1290      1300      1310      1320
E_ E  ACCUCGCGAGAGCAAGCGGACCUCAUAAAGUGCGUCGUAGUCCGGAUUGGAGUCUGCAAC
      |   | | || | | ||
Glu   AAUCCCCUAGGGGACGCCA-----------------------------------------
        60        70
             1210      1220      1230      1240      1250      1260
E_ E|  AUCAUGGCCCUUACGACCAGGGCUACACACGUGCUACAAUGGCGCAUACAAAGAGAAGCG
                                               ||||| |  | |   |||||
E_     ----------------------------------------GGCGCGUUAACA---AAGCG
                                             CYS & SER 10
             1270      1280      1290      1300      1310      1320
E_ E|  ACCUCGCGAGAGCAAGCGGACCUCAUAAAGUGCGUCGUAGUCCGGAUUGGAGUCUGCAAC
              |  |   ||||||   ||   | | |||| |||||||| || || || | |||
E_     -------GUUAUGUAGCGGAUUGCA---AAUCCGUC-UAGUCCGG-UUCGACUCCGGAAC
               20        30            40         50         60
             1330        1340      1350      1360      1370
E_ E  UCGACUCCAUGAAGU--CGGAAUCGCUAGUAAUCGUGGAUCAGAAUGCCACGGUG--AAU
                 ||||  || ||||| |||     || |||   ||  | || ||   |||
Leu   -------GCCGAAGUGGCGAAAUCGGUAGACGCAGUUGAUUCAAAAUCAACCGUAGAAAU
        LEU & GLY     10        20        30        40        50
       1380      1390      1400       1410      1420      1430     1440
E_ E  ACGUUCCCGGGCCUUGUACACACCGCCCGUCA-CACCAUGGGAGUGGGUUGCAAAAGAAGUAGGU
      |||| || ||  |  ||     ||| || ||  |||||                         ||
Leu   ACGUGCC-GGUUCGAGU-----CCGGCCUUCGGCACCA-------------------------GU
            60              70        80
             1450         1460      1470      1480      1490
E_ E|  AGCUUAACCU---UCGGGAGGGCGCUUACCACUUUGUGAUUCAUGACUGGGGUGAAGUCG
        |||  | ||   | || || || ||       |||||||||  |  ||  |||   |||
E_     GGCUAUAGCUCAGUUGGUAGAGCCCUGG----AUUGUGAUUCCAGU-UGUCGUGGGUUCG
       HIS   10        20        30            40         50
               60        70
               1450      1460      1470      1480      1490      1500
E_ E  AGCUUAACCUUCGGGAGGGCGCUUACCACUUUGUGAUUCAUGACUGGGGUGAAGUCGUAA
                                                   || | || | || |
Ser    --------------------------------------------GGAGAGAUGCCGGAG
                                                 SER & VAL  10
                  1510      1520      1530      1540
E_ E  C-----AAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCCUUA-------------
      |     || | | | ||   | || | |  || ||  || |||  ||
Ser   CGGCUGAACGGACCGGUCUCGAAAACCGGAGUAGGGGCAACUC--UACCGGGGGUUCAAA
          20        30        40        50          60        70
             1510      1520         1530      1540
E_ E|  CAAGGUAACCGUAGGGGAACCUG---CGGUUGGAUCACCUCCUUA---------------
                 | || || |  |    ||||| ||  |||| | |
E_     ----------GGAGCGGUAGUUCAGUCGGUUAGAAUACCUGCCUG
                 ASP     10        20        30
```

Fig. 1. (*continued*)

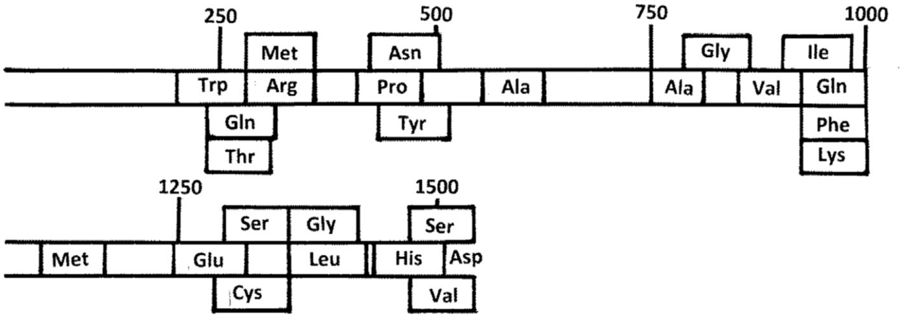


Fig. 2. Mapping of transcribed tRNA-like regions derived from Table 1 onto the 16S rRNA of *E. coli* K12. The central double lines represent the 16S rRNA. The numbers above the double line are base pair markers. As noted in Fig. 1, many of the tRNA encodings overlap.

represent the vestiges of a more primitive tRNA structure that evolved primordially and functioned as a simple translation molecule (Seligmann and Krishnan, 2006; Seligmann et al., 2006; Seligmann, 2008, 2010a; Seligmann and Labra, 2014). In addition this result may have a link with evidence that tRNA lateral arms also have anticodon functions (Seligmann, 2013a,b, 2014), which is consistent with experimental observations on "armless" tRNA (Juehling et al., 2012; Wende et al., 2014).

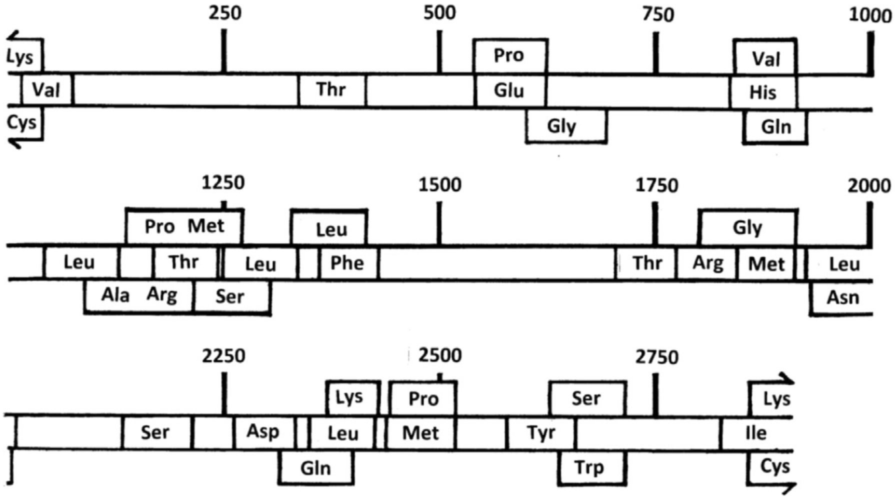


Fig. 3. Mapping of transcribed tRNA-like regions onto the 23S rRNA of *E. coli* K12. This map was derived from an initial plot similar to Fig. 1. In the map, the central double lines represent the 23S rRNA. The numbers above the double line are base pair markers. As noted in Fig. 2, many of the tRNA encodings overlap.

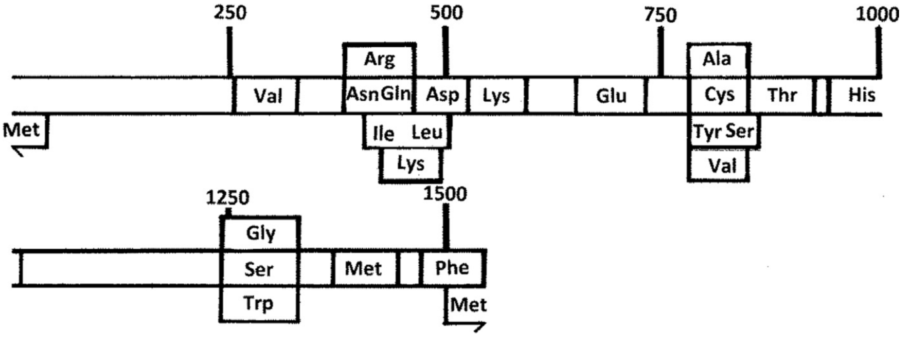


Fig. 4. Mapping of direct homologies between tRNAs onto the 16S rRNA of *E. coli* K12. This map was derived from an initial plot similar to Fig. 1, but in this case, the assumption is that fragmentation or editing of the rRNA could directly yield tRNA-like sequences. In the map, the central double lines represent the 23S rRNA. The numbers above the double line are base pair markers. As previously noted, many of the tRNA encodings overlap.

3.3. rRNAs encode ribosome-related protein modules

In order to determine whether *E. coli* K12 rRNA also encodes proteins, a BLAST2.0 search was performed using the Swiss Institute of Bioinformatics ExPASy website (www.expasy.ch). The 5S, 16S and 23S rRNA and the three randomly selected control sequences (fimbrial protein RNA, sugar efflux protein RNA and non-coding region) of *E. coli* K12 were translated into each of their six possible reading frames (one through, three 5′→3′, and four through six, inverse complements [3′→5′]). Each resulting protein sequence was entered in the BLAST program and compared with the *E. coli* K12 proteome from the “Select Microbial Proteome” section. Those sequences having an *E* value of less than 1000 were output. The use of the *E* value of 1000 was not intended to imply any statistical evaluation of the results, but was used simply as a standard benchmark that would produce data from each search that was equivalent to every other search in the quality and number of similarities generated. The object of this part of the study was to determine whether rRNA is more likely to encode ribosome-related proteins than are mRNAs from any other part of the *E. coli* genome. Therefore, the resulting similarities were filtered for their direct relevance to ribosome function focusing on tRNA synthetases and transferases that would permit amino acid loading of tRNAs; RNA and DNA polymerases, ligases, recombinant hot spot proteins, etc. that might foster transcription of polynucleotide sequences; proteins specifically identifiable as ribosomal proteins; and phosphatases and related phosphate binding and transporting proteins that would be needed to synthesize RNA and DNA as well as provide the energy for transcription and translation. For the 5S, 16S and 23S rRNAs, these four sets of ribosomal function-related proteins represent up to 50% of the total matches generated from the BLAST search whereas an identical BLAST search yielded only about 8% ribosome-related proteins when RNA sequences for the fimbrial protein, sugar efflux protein and non-coding region of *E. coli* K12 were used. These results suggest that one in four or five of the protein similarities found from the rRNA search may have occurred by chance. A Fischer’s Exact Test with Bonferroni correction revealed very significant differences between the rRNAs and the fimbrial protein, sugar efflux protein and non-coding region RNAs, with all $p < 0.017$ (Table 3), but no significant differences were found in the number of ribosome-related proteins encoded by the 16S and 23S rRNAs, nor any significant differences between the number of such protein encoded by the various control RNAs.

Key similarity results for the ribosomal proteins encoded in rRNAs are summarized in Figs. 9–12. These Figures reinforce the statistical results showing that rRNA sequences encode an unusually large proportion of ribosome-related proteins. While many additional matches were also found to non-ribosomal proteins, these have been omitted in the Figures in order to keep the present paper to a reasonable length. These additional protein matches suggest the possibility that fragments of ribosomally-encoded proteins were incorporated into many other proteins involved in functions such as replication, sugar metabolism, etc., during cellular evolution, a possibility that we will explore in a later paper.

Fig. 9 summarizes protein sequences from tRNA synthetases and transferases that are encoded in rRNA. Synthetases and transferases are

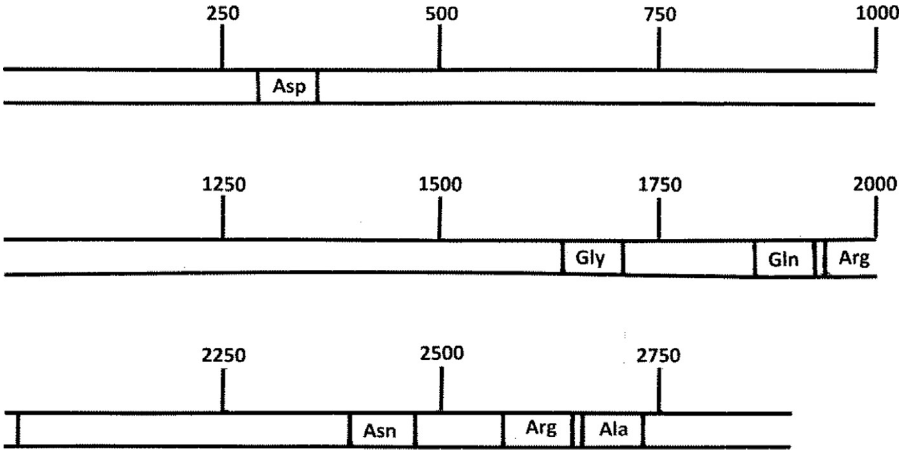


Fig. 5. Mapping of direct homologies between tRNAs onto the 23S rRNA of *E. coli* K12. This map was derived from an initial plot similar to Fig. 1. In the map, the central double lines represent the 23S rRNA. The numbers above the double line are base pair markers. Note the paucity of homologies compared with Figs. 1–4, suggesting that the tRNA-like sequences found in the previous Figures are not due to chance.

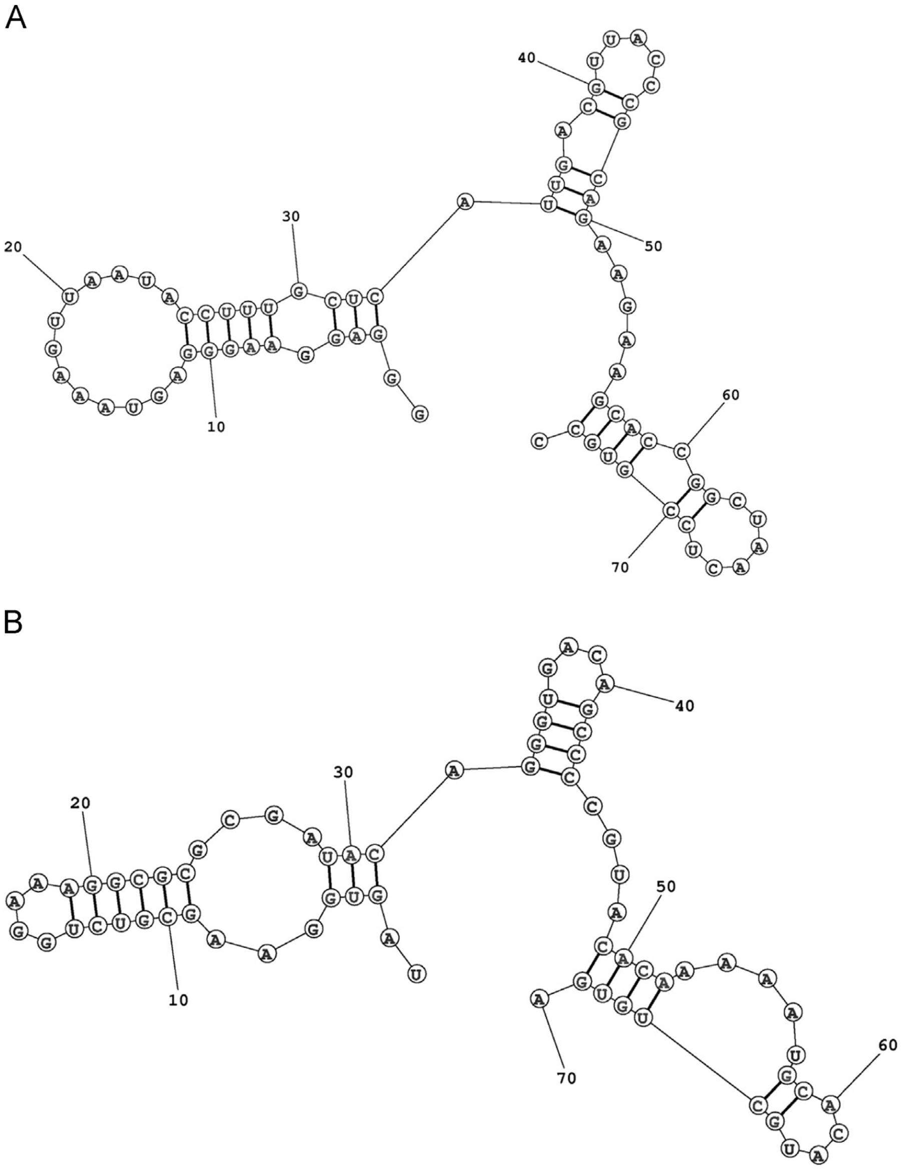


Fig. 6. (A) 16S rRNA Asp homologue (energy −17.2). (B) 23S Asp tRNA homologue (energy −20.0).

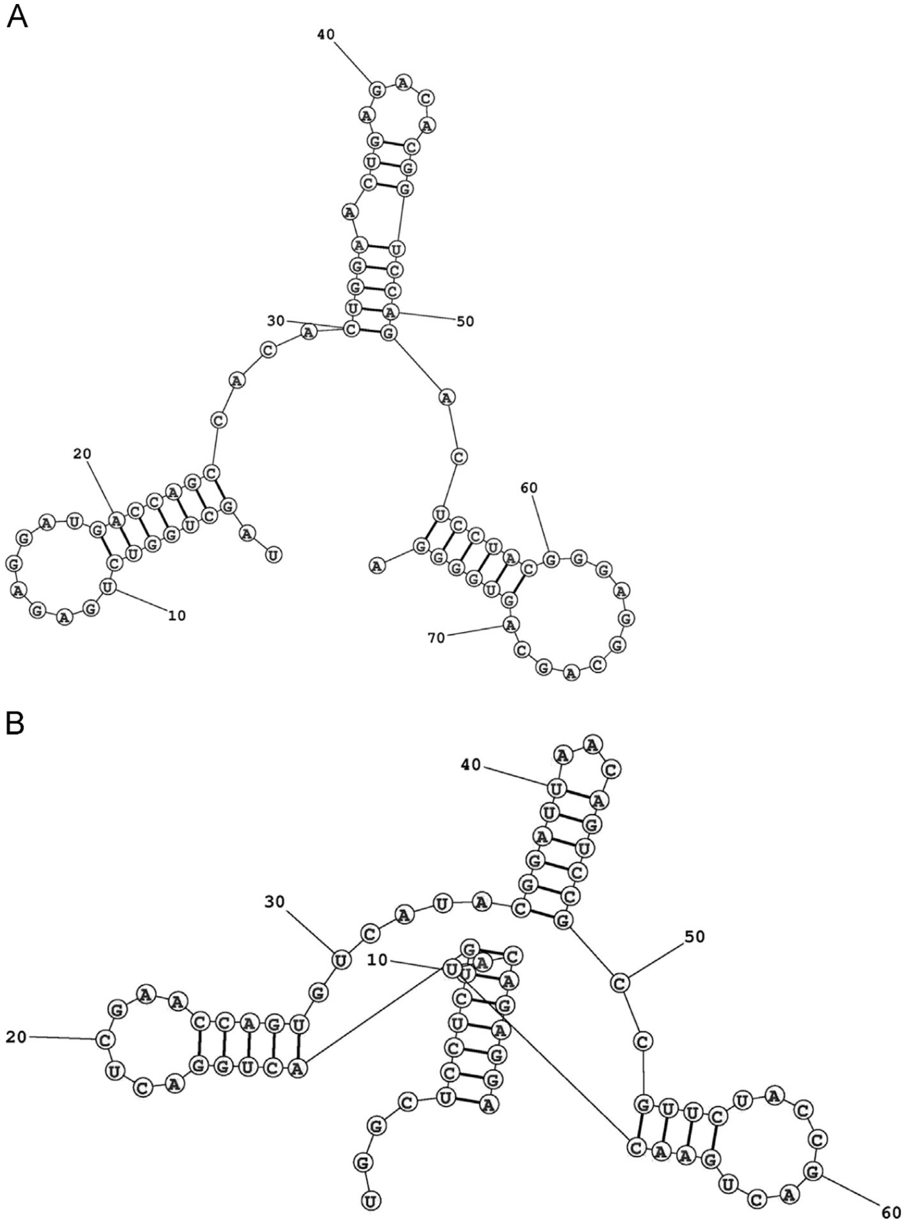


Fig. 7. (A) 16S rRNA Asn homologue (energy −27.2). (B) *E. coli* k12 tRNA Asn (energy −27.4).

enzymes that catalyze the covalent attachment of specific amino acids to their corresponding tRNAs. These sequences are the high-similarity regions and are often fragments of longer sequences of lesser similarity. Sequences matching modules within proteins required to attach most of the standard 20 amino acids to their tRNAs are present in Fig. 9.

Fig. 10 summarizes protein sequences from RNA- and DNA-polymerases that are encoded in rRNA. Polymerases are enzymes that catalyze replication and transcription of polynucleotides. Since RNA- and DNA-polymerases evolved from common ancestral enzymes containing a single highly conserved peptide at their catalytic core that seems to be able to polymerize both types of polynucleotides (Steitz, 1998; Cramer, 2002; Iyer et al., 2004), care must be employed in accepting the modern identification of these proteins as being RNA- or DNA-specific. What can be said is that rRNA encodes possible polyribonucleotide polymerase fragments that could potentially have participated in the replication of RNA, the replication of DNA, the reverse transcription of RNA into DNA or the transcription of DNA into RNA. It is also notable that the rRNAs encode several highly conserved recombinant hot spot protein (rhs) modules. Rhs proteins regulate many functions including transcription, RNA processing, nucleotide biosynthesis and metabolism, and tRNA expression (Aggarwal and Lee, 2011) as well as polynucleotide recombination.

Fig. 11 summarizes protein sequences encoded in rRNA that are similar to ribosomal proteins. Ribosomal proteins have many functions including structural ones; creating binding sites for mRNAs, tRNAs and peptide chains; providing orientation for these molecules relative to each other; acting as catalytic sites for ribosomal reactions; and mechanical functions such as moving the growing peptide chain past the mRNA encoding its synthesis. Fragments of many of the key ribosomal proteins from the 50S and 30S ribosome subunits are present in this list and include various synthases, transferases, and ligases.

Fig. 12 summarizes protein sequences that involve phosphorylation reactions including the synthesis of RNA nucleosides and DNA nucleotides, phosphate uptake and transport, and phosphatases. Such proteins are essential for energy storage and transduction and their presence in the rRNA provides evidence that primitive ribosome-like entities may have encoded the basic elements of the energy metabolism system required to drive ribosomal

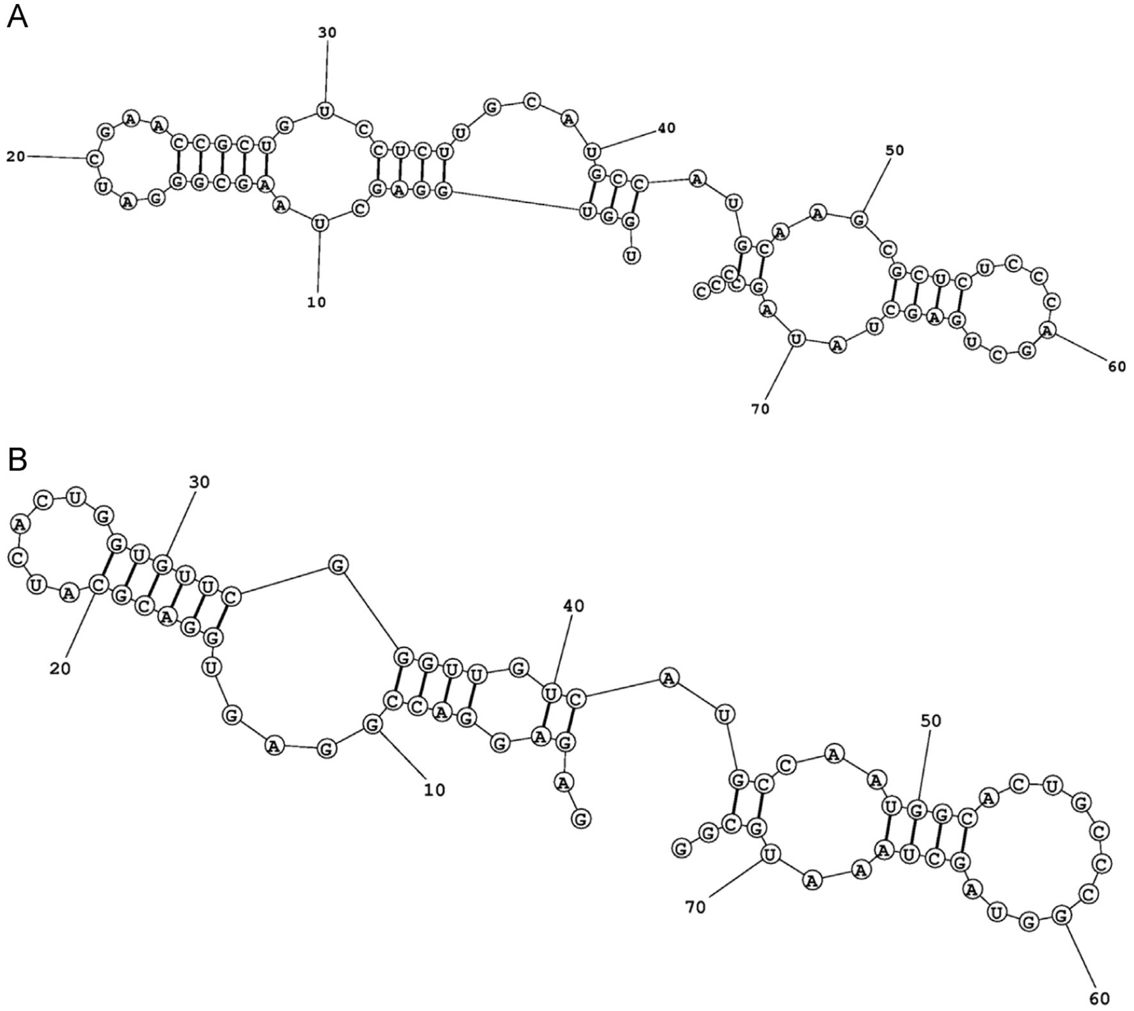


Fig. 8. (A) *E. coli* K12 tRNA Ala (energy −29.0). (B) 23S rRNA homologue to tRNA Ala (energy −29.0).

Table 3
Statistical comparison of probabilities that differences in the appearance of ribosome-related proteins in rRNAs and control RNAs is due to chance. Each RNA was translated into all six possible reading frames and the resulting "proteins" compared with *E. coli* K12 genome by means of BLAST 2.0. A Fischer's exact test was employed; there is no test-specific effect size statistic so none is reported. Bonferroni correction for 3 comparisons for each data set means that significance at the $p=0.05$ level is accepted at $p=0.017$ (i.e., $\alpha=0.017$). p Values that remain significant are in bold. See Table 1 for key to RNA identities. As in Tables 1 and 2, the results clearly demonstrate that rRNA encodes tRNAs at a significantly higher rate than a random assortment of mRNAs, and certainly higher than would be predicted from the "null hypothesis".

PROTEIN COMPARISONS	23S	FIMBRIAL	SUG EFFL	NON-CODE
	40/120	7/103	10/112	12/120
16S 39/153	$p=0.1793$	$p=$**0.0001**	$p=$**0.0007**	$p=$**0.0010**
23S 40/120		$p<$**0.0001**	$p<$**0.0001**	$p<$**0.0001**
FIMBRIAL 7/103			$p=0.6202$	$p=0.4745$
SUG EFFL 10/112				$p=0.8258$

functions. Among these proteins are epimerases and other enzyme-related sequences involved in the synthesis of ribonucleotides and deoxyribonucleotides. Note that the sequences presented in Fig. 12 represent a selection of the results that have been winnowed down for space reasons.

3.4. *Most rRNA-encoded protein Modules represent active sites*

An obvious question is whether the ribosome-related protein sequences listed in Figs. 9–12 are functional. This question can only be answered definitively through experiment, but it is also possible to evaluate whether each similarity is located in a region that is known to be functionally active in the modern protein with reference to the annotations associated with the UniProt protein database (www.expasy.org). Every sequence listed in Figs. 9–12 as well as every control protein similarity that was generated in determining the statistics reported in Table 3 was evaluated for whether it overlapped or included a known active site. Approximately a third of the proteins analyzed lacked sufficient information to make such a determination and these data were discarded. Data were then gathered and analyzed by chi-squared analysis (Table 4) for all the proteins for which active sites were known or for which mutagenesis studies could identify key regions of protein activity. These data are summarized in Table 4, which clearly demonstrates that the number of active site-related similarities is much higher (55%) in rRNA-encoded proteins fragments than in fragments of proteins encoded by the control proteins (average 19% active sites). The differences in frequency of active-site related rRNA proteins compared with the control proteins is very highly statistically significant by chi-squared analysis after Bonferonni correction, but no significant differences were found between the different control proteins. The key active site data for the rRNA-encoded sequences follows.

Fig. 9 shows the various amino acid synthetases and transferases encoded by rRNA. Of 32 sequences, no information concerning active and binding sites was available for 11 sequences. Thirteen of the sequences fell into regions of the modern proteins for which no specific function has been identified, but where information was available, all of the sequences were parts of beta strand regions. The remaining eight sequences matched recognized functional regions of modern proteins. NP_414736.1| prolyl-tRNA synthetase: 531–545 is in the anticodon binding domain of the enzyme (uniprot/P16659); NP_418047.1| selenocysteinyl-tRNA-specific translation factor 168–175 is within the tr-typeG domain (uniprot/P14081); NP_418063.1| tRNA Leu mC34, mU34 2′-*O*-methyltransferase 80–116 is within the binding site of for tRNA (uniprot/P0AGJ7). NP_416154.1| tyrosyl-tRNA synthetase 236–246 cross-links to tRNA and binds ATP (uniprot/P0AGJ9). NP_416969.1| elongator methionine tRNA (ac4C34) acetyl-transferase 273–319 contains ATP binding site (uniprot/P76562). NP_417635.1| tRNApseudouridine (55) synthase: 205–220 is probably part of the substrate binding site that includes residue 202 (uniprot/P60340). NP_415722.1| peptidyl-tRNA hydrolase 125–137 mutation of 134 results in functional failure (uniprot/P0A7D1). NP_416154.1| tyrosyl-tRNA synthetase: 224–248 is the ATP binding site and part of the tRNA cross-linking site (uniprot/P0AGJ9). NP_416380.1| aspartyl-tRNA synthetase 440–450 is the aspartate binding site (uniprot/P21889).

```
23S rRNA Frame 2:                             883 SLTLGSRNVEDDDVD 897
                                                  ++ LG+RN+++DD++
NP_414736.1| prolyl-tRNAsynthetase:           531 TIVLGDRNLDNDDIE 545

23S rRNA Frame 3:              249 FSPKAIVAPREFISGGRALFRQGGH--PDLPT 278
                                   +SP  ++ P+  +SG+  ++R+  H   DLP+
                               166 YSPTELIEPKSVVSGATPVMRDSEHFFFDLPS 197
NP_416617.1| methionyl-tRNA synthetase:

23S rRNA Frame 4:                             581 LPRRPTRPVSYYAFFK 596
                                                  L RRPTRP+++ A  K
NP_418016.1| glycine tRNA synthetase:         550 LARRPTRPADFDARMK 565

23S rRNA Frame 4:                             661 ETVLYPRRIHEALPKLSG 678
                                                  E VL P  +HE + +LSG
                                              141 EAVLKPEIVHERMQQLSG 158
NP_416321.1|tRNA(ANN) t(6)A37 threonylcarbamoyladenosine modification protein

23S rRNA Frame 4:                             725 LYPATRPVK 733
                                                  L+PA RPVK
NP_417366.1| lysine tRNA synthetase:          497 LFPAMRPVK 505

23S rRNA Frame 5:                             539 QHASQHTF 546
                                                  +HASQH+F
                                              168 EHASQHSF 175
NP_418047.1| selenocysteinyl-tRNA-specifictranslation factor:

23S rRNA Frame 6:      753 TKGTQSHAACSHCLYVHGFRFFFTP-----LAGVLFAFPS 787
                           TKGT +H+A S   Y  G    F P      A +L A P+
                       80  TKGTPAHSAVS---YQDGDYLMFGPETRGLPASILDALPA 116
NP_418063.1| tRNA Leu mC34,mU34 2'-O-methyltransferase

23S rRNA  Frame 6:                               797 GVFSLGGWSPHIQTGYH 813
                                                     GV + + W P I+TG+H
NP_417528.1| fused tRNA nucleotidyl transferase: 215 GVPAPAKWHPEIDTGIH 231

16S rRNA Frame 1: 308 FNSMQREEP-----YLVLTSTEVFRDENVPSGTVR-QVLHGCRQLVLNVG 351
                      F+ +Q +E      YL L S     DE+  +  VR Q+L G RQ+++N G
                  157 FHGLQDQEARYRQRYLDLISN----DESRNTFKVRSQILSGIRQFMVNRG 202
NP_417366.1| lysine tRNA synthetase

16S rRNA Frame 1:                                  398 QWRIQREATSREQA 411
                                                       +W  QREA + EQA
NP_417225.1| tRNA(Glu) pseudouridine(13) synthase: 264 EWGTQREALAFEQA 277

16S rRNA Frame 1:                             237 GPLDEDRSGAKAWGANRI 254
                                                  G L ++RSGA+ +G +R+
                                              182 GDLRQNRSGAEHFGLQRL 199
NP_417271.1| tRNA(Ile1,Asp) pseudouridine(65)synthase:

16S rRNA Frame 1:                             453 WEWVAKEVGSLT 464
                                                  WEW A/  G++T
NP_418679.1| valyl-tRNAsynthetase:            115 WEWKAESGGTIT 126

16S rRNA Frame 2:              72  DVPRWDLVGGVTAHQGDDPLVED-DQ-PHWNDTVQT 105
                                   ++P+W +   +TA+   D L++D D+  HW DTV+T
                               187 EIPQWFI--KITAYA--DELLNDLDKLDHWPDTVKT 218
```

Fig. 9. tRNA synthetases and transferases encoded in *E. coli* K.12 rRNA. Sequences are listed using the single letter amino acid abbreviations. The middle rows list the amino acids shared by the ribosomally-encoded protein sequence and that of the modern *E. coli* K12 protein. The "+" sign in the middle rows indicates substitution of a similar amino acid. The reading frames for the rRNA sequences are: frame 1, forward (5′–3′) starting at the first base pair; frame 2, forward starting at the second base pair; frame 3, forward starting at the third base pair; frame 4, inverse complement (3′–5′) starting at the first base pair; frame 5, inverse complement starting at the second base pair; frame 6, inverse complement starting at the third base pair.

```
NP_415175.1| leucyl-tRNAsynthetase

16S rRNA Frame 3:                            323 FGNREТGAAWL 333
                                                 FG/ E GA+WL
NP_416154.1| tyrosyl-tRNA synthetase:        236 FGKTEGGAVWL 246

16S rRNA Frame 5:
23  HKVVSALPKVKLPTSFATHSHGVTGGVYKARERIHRGILIHD-YRF-RLHGVELQTPIR 79
    H++VS +P+  L T+   +            E   RG+L++   RF +LH +ELQ+PIR
273 HQLVSRFPRTLLTTTVQGY------------EGTGRGFLLKFCARFPHLHRFELQQPIR 319
NP_416969.1| elongator methionine tRNA (ac4C34) acetyltransferase

16S rRNA Frame 6:                            23  WAPSRRLSYLLLL 35
                                                 WAP+R  +YL +L
                                             401 WAPARSQAYLGVL 413
NP_418197.1| 5-methylaminomethyl-2-thiouridinemodification at tRNA U34

16S rRNA Frame 6:                            404 VSRYPTNKLIPSGHIR 419
                                                 VS+YP ++++   H+R
NP_417635.1| tRNApseudouridine(55) synthase: 205 VSKYPVERMVTLEHLR 220

5S rRNA Frame 2:                             24  GVSPCESRELP 34
                                                 G+SP E+RE+P
NP_418473.2| tRNA-dihydrouridine synthase A:  179 GLSPKENREIP 189

5S rRNA Frame 2:                               5  RGGPTPHAEL 14
                                                  R  PT HAE+
NP_417054.2| tRNA-specific adenosine deaminase: 51 RHDPTAHAEI 60

5S rRNA Frame 2:              7   GPTPHAELRSETPRRW--CGVSPCESREL 33
                                  GP   AEL   T R W   G+S   + EL
                              137 GPDIDAELIMLTARWWRALGISEHVTLEL 165
NP_417009.1| histidyl tRNA synthetase

5S rRNA Frame 3:              5  SAVVPPDPMPNSEVKR-RSADGSVGSPHAR 33
                                  SA  P DP    ++ +  + DG+ G+P +R
                              10 SATTPVDPRVAEKMMQFMTMDGTFGNPASR 39
YP_026169.1| cysteine desulfurase; tRNA sulfurtransferase

5S rRNA Frame 3:                               12  PMPNSEVKR 20
                                                   P+P +EVKR
NP_417726.1| tRNA-dihydrouridine synthase B:   249 PLPLAEVKR 257

5S rRNA Frame 4:                      12 TPHYHRRYGVSLLSSAWGQVG 32
                                         TP +  R+  +LLSS    VG
                                      12 TPDFAARHLDALLSSGHNVVG 32
NP_417746.1|L-methionyl-tRNA(fMet)N-formyltransferase

5S rRNA Frame 4:                            14  HYHRRYGVSLLSSAWGQVGPPRYGR 38
                                                H+  R G   L  +W ++G P + R
NP_414730.1| tRNA(Ile)-lysidine synthetase: 367 HIVGRNGGRKLKKIWQELGVPPWLR 391

5S rRNA Frame 4:                              16 HRRYGVSLLSSAW 28
                                                 H ++G+S  ++AW
NP_414730.1| tRNA(Ile)-lysidine synthetase:   48 HVHHGLSANADAW 60

5S rRNA Frame 4:                             10  GETPHYHR-RYGV 21
                                                 G++P++HR R+G+
NP_415722.1| peptidyl-tRNA hydrolase:        125 GNNPNFHRLRIGI 137
```

Fig. 9. (*continued*)

Fig. 10 lists the various RNA- and DNA-related enzymes found within the rRNA sequences. Of the 25 sequences listed there, no information about either the structure or function of their modern equivalents was found for four sequences. Two sequences, NP_415702.1| DNA polymerase V, subunit C 349–372 and NP_418415.1| RNA polymerase, β prime subunit, 241–251, are probably in an inactive, stuctural regions of their proteins according to data available on the SwissProt database. Three of the remaining 19 sequences are in regions that have essential enzymatic functions. NP_418415.1| RNA polymerase, β prime subunit 546–567 is in the Rpb2 domain 3 region of the polymerase which is also known as the fork domain and is proximal to the catalytic domain (uniprot/P0A8T7). NP_416906.1| DNA ligase 204–230 contains both helix and beta strand and mutation of residue 208 eliminates 99% of enzyme activity suggesting that this is a critical region of the active site (uniprot/P15042). NP_418300.1| fused DNA polymerase I 5′–3′ polymerase 8–35 is within the 5′–3′ exonuclease region (uniprot/P00852). And finally, the remaining 16 sequences are highly conserved recombinant hot spot (rhs) element core protein fragments, all of which are within recognized protein domain repeats, suggesting that they are very likely to be important functional elements of these proteins (uniprot entries P16916; P16917; P16918).

```
5S rRNA Frame 5:                              17  GATAFHFVR 25
                                                  GA+ +HFVR
NP_418016.1| glycine tRNA synthetase:         151 GASDVHFVR 159

5S rRNA Frame 5:                              20  AFHFVRHGVRWDHR 33
                                                  AFHFV  +  WD R
                                              195 AFHFVIPADEWDER 208
NP_417286.1| 23S rRNA C2498 2'-O-ribose methyltransferase:

5S rRNA Frame 6:                      2   WQFPTLAWGDPTLPSALRRFTSEFGMGSG 30
                                          W+F TL+W + T   ALR+F \\  M SG
NP_418679.1| valyl-tRNA synthetase:   482 WTFSTLGWPENT--DALRQFHPTSVMVSG 508

5S rRNA Frame 6:                      13  TLPSALRRFTSEFGMGSGGTTALRP 37
                                          T+P + +   ++FG   GG   L P
NP_416154.1| tyrosyl-tRNA synthetase: 224 TVPLITKADGTKFGKTEGGAVWLDP 248

5S rRNA Frame 6:                              28  GSGGTTALRPP 38
                                                  G+GG TA++ P
NP_416380.1| aspartyl-tRNA synthetase:        440 GEGGLTAMHHP 450

5S rRNA Frame 6:                      10 GDPTLPSALRRFTSEFGMGSGGTTALRP 37
                                         G   +PS ++ + +E G+      +LRP
                                      44 GKKLMPSPVKVLAEEKGLPVFQPVSLRP 71
NP_417746.1|L-methionyl-tRNA(fMet)N-formyltransferase
```

Fig. 9. (*continued*)

Fig. 10 also lists the various protein- and peptide-related enzymes found within the rRNA sequences. Of 49 such enzymes, no information concerning active regions was available for 17. Of the remaining 32 sequences, 16 overlapped established functional regions, while 16 fell into regions for which no function is yet known: NP_416115.1 putative peptidase: 220–248 straddles the active site (uniprot/P76176); NP_416494.1|murein L,D-transpeptidase:252–282 straddles the active site (uniprot/P39176); NP_416280.1| protease IV: 413–432 straddles the active site (uniprot/P08395); NP_416989.1|metalloprotease:440–449 comprises most of the TPR 4 domain repeat (uniprot/P66948); NP_414691.1|transpeptidase 493–520 overlaps the active site (uniprot/P02919): NP_416223.1| putative peptidase: 104–115 straddles the active site (uniprot/P23898); NP_416115.1| putative peptidase: 42–76 overlaps the active site (uniprot/P76176); NP_417695.1| ClpXP protease: 69–86 binds SspB and ssrA; ssrA is a degradation tag (AANDENYALAA) added trans-translationally to proteins that are stalled on the ribosome, freeing the ribosome and targeting stalled peptides for degradation (uniprot/P0AFZ3); NP_416005.1| D-ala-D-aladipeptidase: 168–192 forms the catalytic site (161–165) (uniprot/P77790); NP_415360.1| D-alanyl-D-alanine carboxypeptidase: 7–30 forms the signal peptide of the enzymem (uniprot/P08506); NP_418725.4| Zn-dep. exopeptidase domain: 320–341 is part of the Zn-binding catalytic domain 2 (uniprot/P39366); NP_417384.1| proline aminopeptidase P II: 385–406 contains two mettal binding sites (uniprot/P15034); NP_415722.1| peptidyl-tRNA hydrolase: 125–137 overlaps the active site (uniprot/P0A7D1); NP_414691.1|transpeptidase: 233–247 straddles the active site (uniprot/P02919); NP_416989.1|metalloprotease: 447–457 comprises the TPR 4 domain repeat (uniprot/P66948).

Fig. 11 shows the similarities between rRNA-encoded proteins and ribosomal proteins. Of 25 sequences listed in the Figure, no information about active sites was available for seven. Of the remaining eighteen, half overlapped known active sites while half did not. The nine that overlapped ribosomal protein active regions were the following: NP_415785.1| 23S rRNA pseudouridine synthase: 9–23, which represents the S4 RNA binding region (uniprot/P37765); NP_417954.1| 16S rRNA m(2)G1516methyltransferase 172–236 straddles the 16S rRNA binding site for methylation (uniprot/P68567); NP_417099.4 16S rRNA processing protein: 101–112 overlaps the binding site for 30S and S19 (uniprot/P0A7 × 6); NP_415373.1 ribosomal protein S6 modification protein: 203–216 comprises the nucleotide (ATP)binding site (uniprot/P0C0U4); NP_418410.1| 50S ribosomal subunit protein L11: 49–84 contains 3 methylated lysines suggesting active region (uniprot/P0A7J7); NP_417747.1| 16S rRNA m(5)C967 methyltransferase: 307–312 is the binding site for 16S rRNA (uniprot/P36929); NP_415785.1| 23S rRNA pseudouridine(2605) synthase: 85–106 forms active site (uniprot/P37765); YP_026225.1| fused ribosome-associated ATPase: 261–281 forms part of the ABC transporter 2 region (uniprot/P37624).

Fig. 12 shows the similarities between rRNA-encoded proteins and phosphotases or related enzymes. Of the 54 sequences listed, there is no information available on functional regions for 24. Seventeen of the similarities do not match any region with a function that has so far been identified by experiment. Thirteen of the similarities do match regions of enzymes with known functions: The similarity with NP_417388.1| D-3-phosphoglycerate dehydrogenase 246–271 includes NAD binding site and part of the enzyme active site (uniprot/P0A9T0); NP_417665.1| 3-deoxy-D-manno-octulosonate 8-phosphate phosphatase 70–94 contains two elements of the substrate binding region (uniprot/P0ABZ4); NP_417633.4| polynucleotide phosphorylase/polyadenylase 323–345 interacts with RNAase E (uniprot/P05055); NP_416170.1| putative ATP-dependent helicase 271–322 helicase C terminal (uniprot/P30015); NP_418040.1| L-ribulose-5-phosphate 4-epimerase 156–186 contains part of zinc binding site of enzyme (uniprot/P37680); NP_414917.2| bacterial alkaline phosphatase 174–209 contains magnesium binding site (uniprot/P00634); NP_414641.1| nucleoside triphosphate pyrophosphohydrolase 33–61 contains all four elements of magnesium binding site (uniprot/P08337); NP_417242.1| 3′-phosphoadenosine 5′-phosphosulfate reductase 168–219, mutagenesis of 209 reduces enzyme Vmax suggesting this region contains active site (uniprot/P17854); NP_415279.1| galactose-1-phosphate uridylyltransferase 295–307 contains the iron-binding site of the enzyme (uniprot/P09148); NP_416995.1| phosphoribosylglycinamideformyltransferase 170–181 contains the 5′-phosphoribosylglycinamide binding site (uniprot/P08179); NP_418391.1| phosphoenolpyruvate carboxylase: 587–601—mutation 587 results in complete loss of enzyme activity arguing this region is in the active site (uniprot/P00864); NP_415500.1| phosphoanhydridephosphorylase: 38–48 contains substrate binding site and nucleophile active site (uniprot/P07102); NP_416006.4|c-di-GMP phosphodiesterase: 424–437 is located in the GGDEF domain (uniprot/P76129).

Overall, approximately one third of all of the rRNA-encoded proteins have similarities to identified active regions of the proteins they mimic; that proportion increases to over one half (55%) if only proteins with information regarding active regions are included in the calculation; and this calculation assumes that information regarding active regions of the proteins in question are known in complete detail, which is rarely the case in reality. Thus, there is a reasonable probability that any protein fragment encoded by an rRNA in any reading frame would have had some ribosome-related functionality in a pre-cellular world. Moreover, rRNA sequences incorporate active regions at a much higher frequency than do any of the control mRNAs studied. Each of the control mRNA (for fimbrial protein, sugar efflux protein and the non-coding region) were analyzed as was done for the rRNA sequences just described. Overall, only one in ten sequences could be identified with an active region of the protein it mimicked as compared with one third of the rRNA-encoded proteins. Among the proteins for which functional information was available (about 2/3 of the sequences), active regions were identifiable in an average of only 18% of the homologous regions of these proteins as compared with 55% for the rRNA sequences. Thus, rRNA is far more likely to encode active protein segments than a random selection of mRNA sequences. As Table 4 demonstrates, the difference in the probability that an rRNA sequence will encode active-site regions of proteins is statistically significantly much higher than that a random mRNA sequence will do so ($p < 0.001$ after Bonferonni correction) whereas none of the control mRNA sequences encode such active regions at rates significantly different from one another or from their average.

Polymerases, Ligases and Peptidases Encoded in rRNA of *E. coli* K.12

RNA Polymerases:

```
23S rRNA Frame 4:                             186 LMPLHPPDVRP 196
                                                  ++P+ PPD+RP
NP_418415.1| RNA polymerase, β prime subunit: 241 VLPVLPPDLRP 251

5S rRNA Frame 2:                        13  ELRSETPRRWCGVSPCESRELP 34
                                            E+R+  P ++/ V P E+ E P
NP_418414.1| RNA polymerase, β subunit: 546 EVRDVHPTHYGRVCPIETPEGP 567
```

DNA Polymerases

```
23S rRNA Frame 1:              251 AGSPRKLFRRLVNSSPGVEHCFGKGVI 277
                                   AGS R+L  R+    P    C+G GV+
NP_416906.1| DNA ligase:       204 AGSLRQLDPRITAKRPLTFFCYGVGVL 230

23S rRNA Frame 1:
718 SWSDIRRLVQWHKPALR--ARREQVR---KQVI-----VIRWFMEGPSLNGKVLRGQADTAQEFI 772
    +WS+++RL++    ALR   R EQ+R   K++I     +  W  E    +  +    ADT++F+
150 AWSNLKRLLKQRNAALRQVTRYEQLRPWDKELIPLAEQISTWRAE---YSAGIAADMADTCKQFL 211
NP_418155.1| gap repair protein

23S rRNA Frame2:                          880 DEFSLTLGSRNVEDDDVDRPGVAQ 903
                                              D FS  +++ N+ DD++ RPG  Q
NP_415702.1| DNA polymerase V, subunit C: 349 DFFSQGVAQLNLFDDNAPRPGSEQ 372

23S rRNA Frame 3:                             396 IKRVKSPLAGRPRVPV 411
                                                  +++V/ PL+GRP +P+
NP_418156.1| DNA polymerase III, β subunit:    14 LQQVSGPLGGRPTLPI 29

23S rRNA Frame 5:
383 HLHISVPSPEV-------TAPFCLV------PSPEFS--QAPWYSLPDHLCR-FGVRFDV 426
    H ++ +P+PEV        AP  ++      P P+ +  QAP   LP+   + ++ R +
360 HPRMPLPEPEVPRQSFAPVAPTAVMTPTQVPPQPQSAPQQAPTVPLPETTSQVLAARQQ- 418

427 TCLEAFPGSRAFVASAPCLVITPQPFSGFAWKTSLHATGTTVARPTPS 474
      L+ + G+ \   S P  +/  +P++  A + /+T    ARP PS
419 --LQRVQGATKAKKSEPAAATRARPVNNAALERLASVTDRVQARPVPS 464
NP_415003.1| DNA polymerase III/DNA elongationfactor III, tau and gamma subunits

23S rRNA Frame 6:                     635 PVITFSGIRSLHRVGKSGPPCRNSALPP 662
                                          P+I ++G / L+R+  + PP  NSA  P
                                      8   PLILVDGSSYLYRAYHAFPPLTNSAGEP 35
NP_418300.1| fused DNA polymerase I 5'-_3'polymerase

16S rRNA Frame 3:                             461 NLREGAYHFVIHDWGEVV 478
                                                  ++ +G YHF ++D GE+V
NP_414726.1| DNA polymerase III α subunit:    815 DINSGLYHFHVNDDGEIV 832

16S rRNA Frame 5:                             18  TPVMNHKVVSALPK 31
                                                  TPVM++K V A PK
NP_415003.1| DNA polymerase III:              482 TPVMQQKEVVATPK 495
```

Fig. 10. Polymerases, ligases and peptidases encoded in rRNA of *E. coli* K12. Sequences are listed using the single letter amino acid abbreviations. The middle rows list the amino acids shared by the ribosomally-encoded protein sequence and that of the modern *E. coli* K12 protein. The "+" sign in the middle rows indicates substitution of a similar amino acid. The reading frames for the rRNA sequences are: frame 1, forward (5′–3′) starting at the first base pair; frame 2, forward starting at the second base pair; frame 3, forward starting at the third base pair; frame 4, inverse complement (3′–5′) starting at the first base pair; frame 5, inverse complement starting at the second base pair; frame 6, inverse complement starting at the third base pair.

```
5S rRNA Frame 1:                              7   RWSHLTPCRTQKNAVA 22
                                                  RW   TP   QK++VA
NP_415003.1| DNA polymerase III:              477 RWKATTPVMQQKEVVA 492

5S rRNA Frame 1:                                      30  PMREGTA 36
                                                          P REGTA
NP_414685.4| poly(A) polymerase:                      459 PRREGTA 465

5S rRNA Frame 5:                              5   PYSRMGRPHTTIGATAFH 22
                                                  P  RMG   T + A AFH
NP_415003.1| DNA polymerase III:              343 PDRRMGVEMTLLRALAFH 360

5S rRNA Frame 5:                                      32  HRATAA 37
                                                          HRATAA
NP_418300.1| fused DNA polymerase I 5'-3'polymerase: 734 HRATAA 739
```

Recombinant Hot Spot Proteins (RHS elements)

```
23S rRNA Frame 1:                             121 RGILSEYGGTILQG 134
                                                  +G +++YGG+I+QG
NP_418050.1| rhsA element core protein:       9   QGDMTQYGGSIVQG 22
YP_026224.1|rhsB element core protein
NP_415229.1| rhsC element core protein
NP_415030.1| rhsD element core protein

23S rRNA Frame 2:                             690  PFNVCSNVDPSGLRTV 705
                                                   P N  SN+DP GL T+
NP_415229.1| rhsC element core protein:       1235 PLNPISNIDPLGLETL 1250

23S rRNA Frame 3:                     556 VYKHSTVQTRKWTYTVRLPG 575
                                          VY +S \Q R +TY ++ PG
NP_415030.1| rhsD element protein:    329 VYDRSNTQVRAFTYDAQHPG 348

23S rRNA Frame 3:                             549 RYQLAATVYKHSTVQTRKWTY 569
                                                  R +LAA VY +S \Q R +TY
NP_415229.1| rhsC element core protein:       319 RGELAA-VYDRSNTQVRSFTY 338

23S rRNA Frame 4:                             328 AAVYRGFDQELRLRPHQ 344
                                                  AA+++++ELRL PH+
NP_418050.1| rhsA element core protein:       156 AALWQALPEELRLSPHR 172
YP_026224.1| rhsB element core protein
NP_415229.1| rhsC element core protein

16S rRNA Frame 1:      97  GPATLELRHGPDSYGRQQWGILHNGRKPDAAMPRVR 132
                           GP  LELR+  D  GR +WG L +   PD+ + R R
                       479 GPDGLELRREYD--GR-EWGRLIQETAPDGDITRYR 509
NP_418050.1| rhsA element core protein
YP_026224.1| rhsB element core protein
NP_415229.1| rhsC element core protein

16S rRNA Frame 2:                             9    NAGGRPNTCKSNGN 22
                                                   N  G PN C+ +GN
YP_026224.1| rhsB element core protein:       1397 NRKGLPNVCRVHGN 1410

16S rRNA Frame 2:                             386  LHTCYNGAYKEKRPRES 402
                                                   +  C  G +KE RPR S
NP_415229.1| rhsC element core protein:       1332 VENCLKGKFKEVRPRYS 1348
```

Fig. 10. (*continued*)

3.5. *rRNAs contain massive amounts of genetic information in overlapping encodings*

Fig. 13 presents a map illustrating the locations of all of the tRNA and the selected protein sequences described above that are encoded in the 5S, 16S and 23S rRNAs. In most cases, the blank spaces in the protein translation frames actually encode proteins as well, but not those selected here for their relevance to ribosomal function. The map clearly demonstrates that all three rRNAs encode massive amounts of ribosomal function-related information. As noted above, the 16S and 23S rRNAs can be transcribed into a complete set of tRNAs and the 16S rRNA also contains many of the tRNAs encoded in such a way that they would be yielded by the fragmentation of the rRNA itself. In addition, all six possible reading frames of each of the rRNAs are utilized to encode ribosome-related proteins. In some cases, especially in the 5S rRNA, but to a lesser extent in the 16S and23S rRNAs as well, the proteins are encoded in an overlapping fashion within each reading frame so that translating a sequence at one amino acid yields a protein similar to a phosphatase, while beginning the translation several amino acids later yields a protein that is similar to a peptide ligase or protease, etc.

A second type of protein BLAST search was performed selecting "*Escherichia coli*" or "all proteomes" from the "Database" section of the program rather than the "Select Microbial Proteome" section

```
5S rRNA Frame 3:                              18  VKRRSADGSV 27
                                                  VK+  ADGSV
NP_418050.1| rhsA element core protein:       411 VKKEHADGSV 420
YP_026224.1|rhsB element core protein
NP_415229.1| rhsC element core protein
NP_415030.1| rhsD element core protein

5S rRNA Frame 4:                              10  GETPHYHRRYGVSL 23
                                                  G+ \  HR  G+SL
NP_415030.1| rhsD element core protein:       562 GQMTAVHREEGISL 575
```

Peptidases and Proteinases

```
23S rRNA Frame 1:                     701 VVLGRSPPKERRSTKVGSW 719
                                          + LG +   +RR+TKVG W
NP_417628.2| putative protease:       32  IYLGEAVCSKRRATKVGDW 50

23S rRNA Frame 1:      705  RSPPKERRSTKVGSWSDIRRLVQWHKPALRARREQVRKQVIV 746
                            RS   ++RS +  + + +  L+Q+  PA    +EQ+ KQ+ V
NP_416011.1| peptidase: 865 RSLDIQQRSVQQLANTIVNSLIQYDDPAAWTEQEQLLKQMTV 906

23S rRNA Frame 1:
656 IAHTLSLDVDRWEAKCGRQSAWSRPNTTLCLMFRPVIRVADSVWWVV---LGRSP 707
    + +  SL +  W+AKCG  S            FR VIR      WV+    G SP
149 VHYNFSLPMAFWQAKCGDISGADAKEKISAGYFR-VIRNYYRFGWVIPYLFGASP 202
NP_417173.1| glutamate-cysteine ligase

23S rRNA Frame 2:              143 DSEPVPGKGEKNPGEGSEKEP 163
                                   D EP+P KG+   G+G  ++P
NP_414691.1|transpeptidase:    42  DEEPMPRKGK---GKGKGRKP 59

23S rRNA Frame 2:                     539 LREKARYVGE 548
                                          +RE ARY GE
NP_415475.1| putative peptidase:      272 IREAARYTGE 281

23S rRNA Frame 2:                     545 YVGEATCSWSNQSKI 559
                                          Y+GEA CS    +K+
NP_417628.2| putative protease:       33  YLGEAVCSKRRATKV 47

23S rRNA Frame 3:      469  SGKSRLRRDDEALRCSNKCPASRKSLASGNIKSYPK 504
                            +GK+RL++ DE L+       +RK+L +   +S+PK
NP_415657.1| peptidase: 99  TGKNRLKKSDELLK------WARKNLQTTGCESWPK 128

23S rRNA Frame 3:                      410 PVQRSGQGESTPKARPKGVVDGKQ 433
                                           PV  S QGE+ P+A+   ++ G Q
NP_415445.1| mureinL,D-transpeptidase: 41  PVAVSEQGEALPQAQATAIMAGIQ 64

23S rRNA Frame 3:                                     208 PKPGDLAMGRL 218
                                                          PKPGD+A  R+
NP_417649.1| D-alanyl-D-alanine carboxypeptidase:     182 PKPGDMAFIRV 192

23S rRNA Frame 3:                768 PGITGYRPRVHIDGGVWHL-DVGSSHPGAE 796
                                     PG +G  / +H D G W+L  V SS P+A+
NP_416115.1| putative peptidase: 220 PGDSGSPLMLHTDDG-WQLIGVQSSAPAAK 248

23S rRNA Frame 4:                      74 HSGPLVLGAAPLSSPAPTADRD 95
                                          H G  +LGA  +  PAP+A+R+
NP_415842.2| murein peptide amidase A: 19 HYGRSLLGAPLIWFPAPAASRE 40
```

Fig. 10. (*continued*)

(see Section 2). This second type of search examined possible homologies with all *E. coli* subspecies. This search generally confirmed the first type of BLAST search and demonstrated that the same types of homologies that we found for *E. coli* K12 occur throughout *E. coli* subspecies and in many other bacterial species as well (data not shown). Given the length of the current paper, a fuller account of these data demonstrating the generalizability of our results will be presented elsewhere. The critical point here is simply that the results we are reporting are not unique to the species of bacterium that we have analyzed but occur in a much wider range of microbes as well.

4. Discussion

The interpretation of our results must be made as a function of the three hypotheses we set out to test against each other: (1) that the ribosome evolved prior to cellular life and had the capability of genetically encoding its own transcription and translation apparatus. rRNA should therefore encode (at least in vestigial manner) the tRNAs and proteins necessary to ribosomal function; (2) that ribosomal RNA is purely structural in nature, encoding no genetic information related to tRNAs or ribosomal proteins (the “null hypothesis”); and (3) that the amount of genetic information

```
23S rRNA Frame 4:                     84  PLSSPAPTADRDRTVSRRSKPSSRTTLNGEQ 114
                                          PLS  \   + DR V     PS R  +NG++
NP_416494.1|murein L,D-transpeptidase:252 PLSRNRAEYESDRKVPLPVTPSLRAFINGQE 282

23S rRNA Frame 4:                     853 LWAAPRSLAATGGISV 868
                                          +WA P+ +   GGISV
NP_414626.1| transpeptidase:          91  IWADPKEVHDAGGISV 106

23S rRNA Frame 4:                     495 PVSHRLRLSASPGSTHPAPINVGQEPLVFRRAG 527
                                          PV H +RL  SP           +E  V RRAG
NP_417384.1|prolineaminopeptidase:    168 PVVHEMRLFKSP-----------EEIAVLRRAG 189

23S rRNA Frame 4:                     84  PLSSPAPTADRDRTVSRRSKPSSRTTLNGEQ 114
                                          PLS   \  + DR V     PS R  +NG++
NP_416494.1| L,D-transpeptidase:      252 PLSRNRAEYESDRKVPLPVTPSLRAFINGQE 282

23S rRNA Frame 4:                     585 PTRPVSYYAFFKWLLLSQ 602
                                          P RP    AF++WLL +Q
NP_417695.1| ClpXP protease:          8   PRRPYLLRAFYEWLLDNQ 25

23S rRNA Frame 5:                             390 SPEVTAPFCLVPSPEFSQ 407
                                                  +P+VTA F   PSP F+Q
NP_415845.2| mureintripeptide transporter:    338 TPDVTAGFTPEPSP-FEQ 354

23S rRNA Frame 5:                             682 WPFTPSHKSSANFSTLVGSV 701
                                                  W  TP++   AN STL GS+
NP_416280.1| protease IV:                     413 WISTPANYIVANPSTLTGSI 432

23S rRNA Frame 5:                             575 AGRLDQAITL 584
                                                  AGRLDQAI+L
NP_416989.1|metalloprotease:                  440 AGRLDQAISL 449

23S rRNA Frame 5:                             63 QCHWHDNPNTSDASTPVLSYEQPPSVLQRP 92
                                                 Q +WH     SDAS  +L+ ++P  + \RP
NP_414915.1| D-alanine-D-alanine ligase:      46 QGQWH----VSDASNYLLNADDPAHIALRP 71

23S rRNA Frame 6:                             173 AVTLAVKLA-YAIALTSCPTRISQPSCSS 200
                                                  A+ LAV  A  A+A T+  T  SQP  +S
NP_416638.4|D-alanyl-D-alanine peptidase:     12  ALMLAVPFAPQAVAKTAAATTASQPEIAS 40

23S rRNA Frame 6:                     19  RPINVVVFNVPSGPLKGQGELISGQ 43
                                          RP+  +VFNVP+     + EL SG+
NP_418681.1|cysteinylglycinase: 1     52  RPLRKMVFNVPT-----RRELTSGE 171

23S rRNA Frame 6:                     901 EIAGYNGSYHLTDAYRRLARPSSPLTA 928
                                          + AGYN +     +   LA+P++ LTA
                                      493 QFAGYNRAMQARRSIGSLAKPATYLTA 520
NP_414691.1|transpeptidase

23S rRNA Frame 6:                     234 YFKVGSMQTGVH 245
                                          +FK GS Q G+H
NP_416223.1| putative peptidase:      104 FFKTGSGQNGLH 115

16S rRNA Frame 1:      367 SGRELKGDCQTGGRWGRQVIMALTTR----ATHVLQWRIQREATSREQA 411
                           S R +  D + GG W  Q+I+ LT+     +TH L  + + EA + E+A
                       407 SERRVAVDIELGG-WQEQLILTLTSEEGVSITHTLDGQFD-EANNAEKA 453
NP_415952.2| putative peptidase
```

Fig. 10. (*continued*)

encoded in rRNA is purely random and therefore the number of tRNAs and ribosome-related proteins that rRNA encodes will be no more or less than any random assortment of any other set of randomly chosen RNAs.

Our results clearly favor the hypothesis that the ribosome could have been a self-organizing, self-replicating pre-cellular entity. To summarize, our study demonstrates that the rRNA of *E. coli* K12 is not merely a structural component of ribosomes but also encodes, at least in a vestigial manner, essential elements of many key components of the transcription and translation mechanisms of modern cells. Sequences homologous to all of the tRNAs required to translate mRNA into proteins are present in the 16S and 23S rRNAs. These tRNAs are encoded in two different ways. They are encoded directly within the rRNAs so that fragmentation of the rRNAs can result in the tRNA sequences. The tRNAs are also encoded as complementary sequences, so that they can be produced by either replication of short sequences of the rRNA or by fragmentation of the entire sequence of a replicated rRNA. Fragments of many of the synthetases required to load the tRNAs with their appropriate amino acids are encoded in the rRNA sequences. Fragments of many proteins making up the structure of ribosomes are also encoded in the rRNAs. And fragments of synthases and polymerases required to reverse transcribe rRNA into DNA and then transcribe the DNA back into rRNA or, perhaps, to directly replicate rRNA into complementary RNA, are also encoded in the rRNA. Finally, fragments of many proteins necessary to transduce energy from ATP or NADH into synthesizing proteins and RNAs are also present in rRNA. The statistical analyses indicate that ribosome-related information is not carried by a random selection of RNA sequences other than rRNAs.

4.1. tRNA–rRNA similarities as clues to ribosome evolution

Looking specifically at the tRNA–rRNA similarities, one might question whether such similarities occur simply because tRNA and rRNA are both high structure polyribonucleotides made up of

```
16S rRNA Frame 1:                             212 LVEGGRIPGVAVKCV 226
                                                  L+E  +IPG+AV  +
NP_418574.1|D-alanine carboxypeptidase:       35  LIEQQKIPGMAVAVI 49

16S rRNA Frame 2:                     94  DDQPHWNDTVQTP---TG---GSSGEYCTMGASLMQP 124
                                          DD+   NDT Q+P    G    +SG  CT  A+L+ P
NP_416115.1| putative peptidase:      42  DDRVPVNDTTQSPWDAVGQLETASGNLCT--ATLIAP 76

16S rRNA Frame 2:                     213 QVR-NARSGGIPVAKAAP 229
                                          +VR NAR GGIP   ++P
NP_417695.1| ClpXP protease:          69  EVRFNARFGGIPRQVSVP 86

16S rRNA Frame 2:             343 VPQRAQ--PLSFVASGPAGNSKETASDKLEEGG 373
                                  +PQR++  PL+F/\   A  +     D+L+++G
                              235 LPQRSEPLPLAFAVQDGASYAGAILKDELKQAG 267
NP_417649.1| D-alanyl-D-alanine carboxypeptidase

16S rRNA Frame 3:                     385 DQGYTRATMAHTK---------RSDLARAS 405
                                          D  Y +   AH +         R DLARAS
NP_415952.2| putative peptidase:      247 DMSYVKNITAHYRQMLDAIIEERGDLARAS 276

16S rRNA Frame 4:                     452 LHQAASQTLLTRPPLVSEAASCFLLPFDLHV 482
                                          L QAAS  LL      ++  SCF+ P   HV
NP_416005.1| D-ala-D-ala dipeptidase: 168 LPQAASYPLL------ADQFSCFISPGTQHV 192

16S rRNA Frame 4:                     188 LHRIKPHAP 196
                                          LH+I+PH P
NP_415952.2| putative peptidase:      381 LHKIRPHHP 389

16S rRNA Frame 5: 2                  238 LRRGLPG---YLILFAPHAFAPER 258
                                         L RGL +   +L+LFAP AFA E+
                                      7  LLRGLAAGSAFLFLFAPTAFAAEQ 30
NP_415360.1| D-alanyl-D-alanine carboxypeptidase

16S rRNA Frame 5:                       338 SPVLLLRVTSMSKGINFTPFLPA 360
                                            +P   +R  S+ +GI   P +PA
NP_416193.1| murein L,D-transpeptidase: 148 TPTAGIRQRSLERGIKLPPVVPA 170

16S rRNA Frame 6:                     414 PSGHIRWQEA 423
                                          P G++RWQ A
NP_417627.1| putative peptidase:      71  PDGYARWQRA 80

16S rRNA Frame 6:                             118 SLLVPGRTAGNKGGLRSLRDLT 139
                                                  SL +P R   + + + SLRDLT
NP_418725.4| Zn-dep. exopeptidase domain:     320 SLSIPCRYTHSPAEVASLRDLT 341

5S rRNA Frame 2:                              17  ETPRRW 22
                                                  ETPR+W
NP_416494.1| mureinL,D-transpeptidase:        130 ETPRNW 135

5S rRNA Frame 3:                      2  GGRSAVVPPDP 12
                                         G R+AVV  DP
NP_415168.1| transpeptidase:         268 GSRAAVVVTDP 278

5S rRNA Frame 3:                      1   PGGRSAVVPPDPMPNSEVKRR 21
                                          PG   A V PD +  +  KR+
NP_415952.2| putative peptidase:      632 PGSVVASVSPDELLKTLPKRK 652
```

Fig. 10. (*continued*)

multiple stem-loop structures. This is, of course, a possibility, but there are four arguments against this factor explaining the results we have reported here. First, while one might expect tRNA to mimic by chance one or two regions within a rRNA, there is no reason to expect to find all twenty tRNA encoded by chance in separate places within the rRNA. Second, stem-loop structures can be formed by any appropriate sequence of bases so that an infinite set of possible sequences exists. The hypothesis we are testing here is whether specific tRNA *sequences* occur in the rRNA, which they do at higher frequency than expected by chance. Third, one would not, *a priori*, expect any rRNA fragments to fold into tRNA-like structures, as we have reported here, since these rRNA sequences have presumably been selected for ribosomal functions, not for tRNA-like functions. Yet many of the tRNA-like rRNA sequences do fold into tRNA-like structures with loops at the appropriate places and in the proper order. Once again, since there are an infinite number of permutations of stem-loop structures that RNA sequences could theoretically take on, it is a priori unexpected to find them folding into cloverleaf-like patterns typical of tRNA. Fourth, we have internal controls that suggest that the incorporation of tRNA into rRNA is not simply due to rRNA and tRNA both sharing the ability to self-order: the normal reading frame of the 23S rRNA contains only a handful of tRNA-like regions in distinct contrast to the transcribed 23S and 16S rRNA reading frames. Simply having a high degree of stem-loop structures cannot therefore account for the appearance of all 20 tRNA in multiple copies and in multiple reading frames in the 16S rRNA or for the appearance of the 20 tRNA in the transcribed reading from of the 23S rRNA. And finally, even if one were to deny all of these arguments and assert that our findings could be due to chance, chance does not obviate the observation that tRNA appear to be encoded in rRNA, that rRNA may have been the evolutionary source of tRNA or that tRNA may have, conversely, given to rRNA. Evolution works by chance. The issue is not whether the appearance of tRNA in rRNA is by chance, but whether there was selection for such chance events that has caused these homologies to be retained through evolution and we claim there was because of additional

```
5S rRNA Frame 3:                              2   GGRSAVVPPDPMPNSEVKRRSA 23
                                                  GG    +P D  P+ V   SA
NP_415445.1| murein L,D-transpeptidase:       267 GGPKITLPGDDTPTDAVVSPSA 288

5S rRNA Frame 4:                              2   PGSSLLSHGETPHYHRRYGVSL 23
                                                  PG  +   +E P  +R  G+ +
NP_417384.1| proline aminopeptidase P II:     385 PGLYIAPDAEVPEQYRGIGIRI 406

5S rRNA Frame 4:                                      4   SSLLSHGETPHYH 16
                                                          S  + HGE /+YH
NP_416514.4| D-alanyl-D-alanine carboxypeptidase:     188 SRAIIHGEPEFYH 200

5S rRNA Frame 4:                              10  GETPHYHR-RYGV 21
                                                  G  P++HR R G+
NP_415722.1| peptidyl-tRNA hydrolase:         125 GNNPNFHRLRIGI 137

5S rRNA Frame 4:                              11  ETPHYHRRYGVSLLS 25
                                                  E  H++   G+SL S
NP_414691.1|transpeptidase:                   233 EDRHFYEHDGISLYS 247

5S rRNA Frame 4:                              14 HYHRRYGVS 22
                                                 HY R+YG++
NP_414568.1| signal peptidase II:             49 HYARNYGAA 57

5S rRNA Frame 4:                              21  VSLLSSAWGQV 31
                                                  +SLLSSA  QV
NP_416989.1|metalloprotease:                  447 ISLLSSASSQV 457

5S rRNA Frame 5:                      12  PHTTIGATAFHFVRHGVRWDHR 33
                                          PH   G  +  F   GVR ++R
NP_418656.1| putative peptidase:      295 PHLLKGLASTPFDSEGVRTERR 316

5S rRNA Frame 5:                      11  RPHTTIGATAFHFVRHGVRWD 31
                                          R H        +F+ HG R D
NP_415403.1|serine protease:          125 RKHEVSRLDVVNFISHGTRKD 145

5S rRNA Frame 6:                              11  DPTLPSALRRFTSEFGMGSGG 31
                                                  D  L  A++RF +  G+G+ G
NP_415445.1| murein L,D-transpeptidase:       318 DNELVEAVKRFQAWQGLGADG 338
```

Fig. 10. (*continued*)

data we offer concerning the unusually high degree of active site protein modules associated with ribosome function that are also encoded in the rRNA.

The observation that tRNAs are encoded in ribosomal RNA in an overlapping manner may seem quite odd to many readers, but overlapping tRNA encoding is very common in mitochondria from an extremely wide range of species (Seligmann, 2010a,b, 2011a,b, 2013b, 2014;), including all metazoans (Mörl and Marchfelder, 2001; Reichert and Mörl, 2000). Some of these overlapping tRNA encodings are, like the ribosomal ones observed here, found on the antisense strand of the gene (Seligmann, 2006; Faure et al., 2011). Many of these overlaps are known to be functional: Functional mitochondrial tRNA gene overlaps exist in human mitochondria (Reichert et al., 1998) and in all metazoans (e.g., Mörl and Marchfelder, 2001; Reichert and Mörl, 2000; Hatzoglou et al., 1995), but apparently not in organisms such as yeast or *Escherichia coli* (Schuster et al., 2005). Schuster et al. (2005), however, experimentally introduced such overlaps into yeast demonstrating that they can still process these overlaps functionally. Schuster et al. (2005) interpret their results to suggest that yeast is "on its way to evolving tRNA editing"; we suggest instead that yeast retains the vestigial mechanisms for such editing and that overlapping tRNA were once universal, originating in the ribosome itself.

The observation that tRNA are encoded sequentially and sometimes in an overlapping manner is also very interesting in light of the research of de Farias (2013), de Farias et al. (2014) who have shown that the protein translation cores (PTC) embedded in the 23S rRNA subunit of *Thermus thermophilus* ribosome is very similar to a concatenation of sequential tRNAs. "In this study the information content between the concatamers of ancestral tRNAs and the catalytic regions of PTC of various organisms were also compared, and a positive correlation among all molecules was observed, demonstrating that, despite the long evolutionary time, this molecule has vestiges of its early origin." Bloch et al. (1984, 1989) have similarly proposed that rRNAs may have evolved from concatenations of primitive tRNA-like modules, which is certainly consistent with the observations we have made here. Thus, the ribosome may contain within itself a "molecular paleontology" revealing the key step by which it evolved and the components involved in that evolution (Root-Bernstein, 2012).

4.2. Overlapping protein encodings

The observation that rRNAs encode overlapping genes in multiple frames is also to be expected since overlapping genes have been demonstrated to exist in functional forms in almost every organism from viruses and bacteria to vertebrates (e.g., Firth, 2014; Fonseca et al., 2014; Fukuda et al., 2003; Huvet and Stumpf, 2014; Makalowska et al., 2005; Mir and Schober, 2014; Pallejà et al., 2008; Seligmann, 2012a,b,c, 2013a). Such gene overlaps minimize the amount of genetic material required to encode the maximum number of proteins. Such overlaps also suggest that apparently unrelated proteins may share inobvious common genetic information due to frameshifting. What is surprising about the multiple overlaps in gene encodings, and the use of all six possible frames of translation, is that such overlapping encodings would have required that multiple selection criteria be at work simultaneously in the evolution of rRNA for a significant amount of time to yield such an information-dense rRNA genome. Such selection could only evolve through a very rigorous interaction between genes and gene products such as might be expected of a self-organizing RNA-protein complex such as a ribosome.

In short, two conclusions are inescapable. First, the ribosome-related information encoded in rRNA is extremely dense—so dense as to make it extremely likely that extensive selection over a very long geological period of time must have been at work to incorporate its many facets. rRNA appears to have been used evolutionarily as structural components of the ribosomes themselves; as tRNAs to translate the sequences; as mRNAs, using all possible reading frames, to encode key ribosomal proteins; and it is also highly redundant, encoding some functional elements such as tRNAs, polymerases and ligases in multiple ways. A second conclusion is that rRNA specifically encodes molecules associated with functions that could potentially have permitted a primitive ribosome to reproduce itself. The fact that all of this information resides in the ribosomes of present-day *E. coli* (and other bacterial) species must be considered in light of several billion years of evolution that have occurred since ribosomes were incorporated into cells. The remaining homologies are almost certainly vestigial and represent fragments remaining after gene loss or transfer to the cellular genome. Thus, the primordial ribosome may have been more complex or complete than that represented by our search strategy.

Another likely implication of our results is that RNA co-evolved with proteins to yield a self-organizing, self-replicating entity. Given the high information density found in the ribosome, selection is likely to have been for peptides that could bind to the RNA sequences

```
23S rRNA Frame 1:                             328 QTARMLAKQP 337
                                                  +TARM+++QP
NP 417697 30S ribosomal subunit protein S9:    42 ETARMVVRQP 51

23S rRNA Frame 2:                               301 TAGANVRREEGNNPDRQLRS 320
                                                    T+ A+VR+E+G    R+LR+
NP_416690.1| 50S ribosomal subunit protein L25: 3   TINAEVRKEQGKGASRRLRA 22

23S rRNA Frame 2:                              824 VRSLSAVGAGELRGAAPSTRG 844
                                                   +R L  +GA+  RG+ P+ RG
NP_417776.1| 50S ribosomal subunit protein L2: 202 LRVLGKAGAARWRGVRPTVRG 222

23S rRNA Frame 3:                             185 VARLTEGSRRETESL 199
                                                  +AR   GSRRE ES+
NP_415785.1| 23S rRNA pseudouridine synthase: 9   LARAGHGSRREIESI 23

23S rRNA Frame 4:                     155 SVERWPFHSEPPDHYDLLSHLLAPSRSQ 182
                                          +V+RW F ++   H + LSH +  S  Q
                                      121 TVKRWNFRTQDATHGNSLSHRVPGSIGQ 148
NP_417779.1| 50S ribosomal subunit protein L3:

23S rRNA Frame 5:
78  PVLSYEQPPSVLQRPRQI-----GTELSHD-VLNPARVPLMANSITLGTYFSPRMADIEVPNTAV 136
    P++ ++Q  +++++  ++     G +L+ D +L+PAR+/     +     ++P +A++  PN/ V
172 PMFPHKQKSALVKKEMRVFQSLVGPDLDADGLLEPARLLATKRVVVKRPDYAPPLANVATPNAVV 236
NP_417954.1| 16S rRNA m(2)G1516methyltransferase

23S rRNA Frame 6:                     22  NVVVFNVPSGPLKGQGELIS 41
                                          +VVV N+ +GPL+  + LIS
                                      225 DVVVANILAGPLRELAPLIS 244
NP_417725.1| methyltransferase for 50S ribosomal subunitprotein L11

23S rRNA Frame 6:                     307 RQGISLPDRYSYGRR 321
                                          RQG+SL / +SY RR
                                      535 RQGFSLRRLFSYSRR 549
YP_026225.1| fused ribosome-associated ATPase:ATP-binding protein

16S rRNA Frame 1:                             43  LMEGDNYWKRLI 54
                                                  L EGD YWK L+
NP_417099.4 16S rRNA processing protein:      101 LEEGDYYWKDLM 112

16S rRNA Frame 1:                     389 LTTRATHVLQWRIQREATSREQADL 413
                                          L+  A  +++  +QR +T +EQAD+
                                      60  LSKEAQKLMKMPFQRAITKKEQADM 84
NP_418059.1| conserved protein, ribosome-associated

16S rRNA Frame 2: 221 GIPVAKAAPWTKTDAQVRKRGEQTGLDTLVV--HAVNDVDLEVVPLR 265
                      G+PV++ \ W K+D + R++GE      L+V  H +++ DL +  LR
                  207 GAPVGELLAWVKEDEN-RRKGEM----VLIVEGHKAQEEDLPADALR 248
NP_417615.1| 16S rRNA C1402 2'-O-ribose methyltransferase

16S rRNA Frame 3:                     346 RNERNPYPLLPAVRPGTQR-RLPVINW 371
                                          R  + P P  P++RP T+R R  + NW
                                      20  RGRKLPVPDSPGLRPTTDRVRETLFNW 46
NP_417922.1| 16S rRNA m(2)G966 methyltransferase
```

Fig. 11. Ribosomal Protein subunits encoded in *E. coli* K.12 rRNA. Sequences are listed using the single letter amino acid abbreviations. The middle rows list the amino acids shared by the ribosomally-encoded protein sequence and that of the modern *E. coli* K12 protein. The "+" sign in the middle rows indicates substitution of a similar amino acid. The reading frames for the rRNA sequences are: frame 1, forward (5′–3′) starting at the first base pair; frame 2, forward starting at the second base pair; frame 3, forward starting at the third base pair; frame 4, inverse complement (3′–5′) starting at the first base pair; frame 5, inverse complement starting at the second base pair; frame 6, inverse complement starting at the third base pair.

```
16S rRNA Frame 3:                                       57  RKTKEGDLRASCHR 70
                                                            R+ KEGD+R++ HR
NP_415373.1 ribosomal protein S6 modification protein: 203 RRAKEGDFRSNLHR 216

16S rRNA Frame 4:              440 IMRYLPFPVVIPLHQAASQTLLTRPP----LVSEAA 471
                                   I + LP PVVI ++/  S T +T+ P    L+ +AA
                               49  IEKGLPIPVVITVYADRSFTFVTKTPPAAVLLKKAA 84
NP_418410.1| 50S ribosomal subunit protein L11

16S rRNA Frame 4:                              356 PRKYFTTRRPSSYTR 370
                                                   PR+ F+ RR  SY+R
YP_026225.1| fused ribosome-associated ATPase: 534 PRQGFSLRRLFSYSR 548

16S rRNA Frame 6:                             153 PVSRFPKAHSHLKLPWMSRP 172
                                                  P+SR P \++H+ + \M +P
NP_417085.1| 23S rRNA pseudouridine synthase: 188 PISRHPTKRTHMAVHPMGKP 207

5S rRNA Frame 1 :                             17  QKNAVAPMVVWGLPMRE 33
                                                  ++N+ +P V  GLPM E
NP 414624.1 16S rRNA methyltransferase:       254 RENSRGPQVPAGLPMTE 270

5S rRNA Frame 2:                                      19  PRRWCG 24
                                                          P +WCG
NP_417747.1| 16S rRNA m(5)C967 methyltransferase:     307 PSQWCG 312

5S rRNA Frame 2:                      4   PRGGPTPHAELRSETPRRWCGV 25
                                          P G PT + /L\     RW +V
                                      85  PEGRPTVFDRLPKLRGARWIAV 106
NP_415785.1| 23S rRNA pseudouridine(2605) synthase

5S rRNA Frame 3:                      1   PGGRSAVVPPDPMPNSEVKRRSADGSVG 28
                                          PGGR +++    + +  VKR   + S G
                                      232 PGGRLSIISFHSLEDRIVKRFMRENSRG 259
NP_414624.1| 16S rRNA m(4)C1402 methyltransferase

5S rRNA Frame 3:                               5   SAVV_PPDPMPNSEVKRRSAD 24
                                                   +AVV+PP\\  N+E+   + D
YP_026225.1| fused ribosome-associated ATPase: 261 QAVVIPPYQPENAEIAIEARD 281

5S rRNA Frame 5:                                   27  GVRWDHRATAA 37
                                                       G RW H A  A
NP_416349.2| 16S rRNA m(5)C1407 methyltransferase: 394 GYRWQHEAVIA 404

5S rRNA Frame 5:                                      29  RWDHRATAAR 38
                                                         +W HR+T+ R
NP_416688.1| 16S rRNA pseudouridine(516) synthase:   116 QWSHRITSPR 125

5S rRNA Frame 5:                                      13  HTTIGATAFHFVRHG 27
                                                          HTT +A  +HF //G
                                                      248 HTTTAARLYHFPHGG 262
NP_418585.4| ribosome small subunit-dependent GTPase A
```

Fig. 11. (*continued*)

encoding them (i.e., for molecular complementarity) and the resulting RNA-peptide interactions would additionally have been selected for their functions (ability to form platforms that bound other RNA sequences; promoted peptide formation; had RNA or DNA polymerase or ligase activity; stabilized RNA and/or peptides against degradation; etc.) (Hunding et al., 2006; Root-Bernstein and Dillon, 1997). Prebiotic tRNAs, for example, may not have been just tRNAs, but also mRNAs that encoded crucial peptide sequences with various enzyme or structural functions that were enhanced by binding to their own, or other tRNA sequences. Specialization of RNA into ribosomal, messenger and transfer types likely came later in evolution.

4.3. *Redundancy of encodings Ensures stability*

Special note should be made of several aspects of the ways in which the rRNAs encode redundant information. The same protein segment is sometimes encoded in more than one rRNA. For example, the same sequence of DNA polymerase III is mimicked in 16S rRNA frame 5 and in the 5S rRNA frame 1 (Fig. 10). Similarly, a shared region of D-tagatose 1,6-bisphosphatealdolase is encoded in both the 23S (frame 2) and 16S (frame 2) rRNAs, and a shared tagatose 6-phosphate aldolase region is encoded in both the 23S (frame 3) and 5S (frame 1) rRNAs (Fig. 12). Thus, one aspect of the selection process that yielded the rRNAs very likely involved selection for redundancy in the encodings. This redundancy is also evident in the repetition of the tRNA encodings in both the 16S and 23S rRNAs and in the encoding of many of the tRNAs both by transcription and by fragmentation of the rRNAs. Redundancy of information is also evident in the use of repetitive modules. For example, the recombinant hot spot core element proteins (rhsA, rhsB, rhsC and rhsD) share several common sequences not only among themselves but with several rRNA segments (see section on rhs in Fig. 10). We speculate that these shared and repetitively encoded elements represent key active modules from which larger proteins were subsequently assembled and thus provide clues as to evolution of macromolecular activity. Experiments might show, for example, that these modules contain a low level of whatever activity is currently embodied in the larger proteins into which they have been incorporated.

It is also notable that the encoded proteins are not present randomly in the rRNAs. Proteins with direct ribosomal functions make up about 55% of the similarities yielded by our BLAST search, strongly suggesting that rRNAs evolved to encode the information necessary to carry out their own functions. Other evidence of non-randomness can be found within the sets of proteins that are encoded as well. For example, modules of all the enzymes required to catalyze the reaction ATP+D-tagatose 6-phosphate ⇌ADP+D-tagatose 1,6-bisphosphate are present among the phosphatases listed in Fig. 12. Similarly, protein segments involved in the ligation of glutamate to cysteine and of cysteine to glycine are both present among the peptide ligases (Fig. 9) permitting the synthesis of glutathione, a key peptide involved in ascorbic acid recycling and antioxidant functions that is enzymatically synthesized rather than translated from a gene. These observations suggest that the evolution of information encoded in the rRNAs was directed by selection for integrated functionality. A fuller mapping of the metabolic relationships of the proteins identified within the Figures here, as well as fuller investigation of the protein similarities not described in this paper, may reveal interesting clues about rRNA-encoding of other metabolic pathways.

4.4. Tests of the hypothesis

Several testable predictions follow from the implications just stated. If the ribosome predates the origins of cellular life, then tRNA encoded in rRNA will be found in all forms of microbes and all

```
23S rRNA Frame 1:              39  GETQCVSTHYHLNPVNEANRGNNIVPRGKEINRDSPSSGER 79
                                   G T+ + T Y L  +    RG  I+  GKEIN+   S+GER
                               301 GRTELAETLYGLRTL----RGGRIMLNGKEINK--LSTGER 335
NP_416030.1| autoinducer 2 import ATP-binding protein

23S rRNA Frame 1:              71  RDSPSSGERTGSSPEPESVCVLVEASG 97
                                   R  PSSG   G+S    \+++L EA+G
NP_417393.1| membrane ATPase: 120 RPVPSSGHLGGASQRARELMLLCEAAG 146

23S rRNA Frame 1:               803 RAGFRTSDSSVPICRGR 819
                                   + G++  DS +PI RG+
NP_418190.1| ATP synthase:     147 QTGYKAVDSMIPIGRGQ 163

23S rRNA Frame 2:              559 IPAGCNCLLKTQHCANTKVDVYGVTPA 585
                                   IPA C+ L  ++H A   +DV+   PA
                               246 IPALCDAL-ASKHLAGAAIDVFPTEPA 271
NP_417388.1| D-3-phosphoglycerate dehydrogenase

23S rRNA Frame 2:                      410 EKPARRKTKGSCPTLIGAG 428
                                           +KP +R+ K + P ++GAG
NP_415929.1| putative phosphatase:     202 QKPDQRRIKIALPYVVGAG 220

23S rRNA Frame 2:                                      869 SAESIARNLPRDEFSLTLGS 888
                                                           +AES+A +  R++ S  +G+
NP_416598.1| D-tagatose 1,6-bisphosphate aldolase:     151 AAESVATDCQREQLSYVIGT 170

23S rRNA Frame 3:              332 LEAAIIRKRNSSLVESACAEDVTGLNH 358
                                   +E AII  R + LVE  CA    G+ H
                               70  IEVAIITGRKAKLVEDRCA--TLGITH 94
NP_417665.1| 3-deoxy-D-manno-octulosonate 8-phosphate phosphatase

23S rRNA Frame 3:                      385 EVSEVRMLTVTIKRVKSPL---AGRPRVPVQRS 414
                                           E SE+  L  \+KR+  PL   +GRP   + R+
D-arabinose 5-phosphate isomerase:     108 ESSEITALIPVLKRLHVPLICITGRPESSMARA 140

23S rRNA Frame 3:                              870 ETCPEMSSPP-FKGPEGTLKTTTL 892
                                                   E C EM SP    G \GT K   L
NP_417606.1| tagatose 6-phosphate aldolase:    36  EVCSEMRSPVILAGTPGTFKHIAL 59

23S rRNA Frame 4:              432 MLRGFSWKQGICCFSTVVPRHHASALIFR 460
                                   M+RG   + G      V+PR H SAL/ R
                               323 MIRGLDVRTG------VLPRTHGSALFTR 345
NP_417633.4| polynucleotide phosphorylase/polyadenylase

23S rRNA Frame 4:              873 SGYLDVSVPPVRL---INLWIQLMIVCRNTLGF 902
                                   SGY+DVS+ P  L   +N+   LM  C    GF
                               227 SGYIDVSIVPEELGFAVNVGELLMTECEMVNGF 259
NP_418523.1| ribophosphonate triphosphate synthase

23S rRNA Frame 4:                      595 FKWLLLSQHPGCLGLPTSFPTPLWD 619
                                           F+ LL S +PG   +P + PT  WD
                                       179 FRPLLESGNPGTAQIPVTLPT--WD 201
NP_416759.1| undecaprenyl phosphate-alpha-L-ara4FNdeformylase

23S rRNA Frame 5:                      789 LVHYRSVRSIPWR 801
                                           ++H RS++S+PWR
                                       77  VLHERSLQSLPWR 89
NP_417388.1| D-3-phosphoglycerate dehydrogenase
```

Fig. 12. Phosphatases and related proteins. Only a selection of phosphatases has been selected from the results of the BLAST search. Approximately three times as many satisfied the search criteria (see Section 2). Sequences are listed using the single letter amino acid abbreviations. The middle rows list the amino acids shared by the ribosomally-encoded protein sequence and that of the modern *E. coli* K12 protein. The "+" sign in the middle rows indicates substitution of a similar amino acid. The reading frames for the rRNA sequences are: frame 1, forward (5′–3′) starting at the first base pair; frame 2, forward starting at the second base pair; frame 3, forward starting at the third base pair; frame 4, inverse complement (3′–5′) starting at the first base pair; frame 5, inverse complement starting at the second base pair; frame 6, inverse complement starting at the third base pair.

```
23S rRNA Frame 5:             57  PHLATGQCHWHDNPNTSDASTPVLSYEQPPS 87
                                  P L  G+  W +    S  S  + S+EQP S
                              385 PWLNNGELDWREGAEKSLDSNVIASFEQPFS 415
NP_416365.1| 6-phosphogluconate dehydratase

23S rRNA Frame 5:             619 CFPLHDGRHPPCVSRDNILRYSQFASGWV 647
                                  C PL + R   C+  + +LR++    GW+
                              286 CQPLLNARSQQCIGVEILLRWNNPRQGWI 314
NP_416329.4|cyclic-di-GMP phosphodiesterase

23S rRNA Frame 6:             322 LPGLRSRASLALTPSINLPAPGRRHTVYVH 351
                                  LPG +  A + L   I+LPAP R   VY H
                              386 LPGSQEPAEVTLRKVISLPAPLRGSAVYRH 415
NP_416744.1| sn-glycerol-3-phosphate dehydrogenase

23S rRNA Frame 6:
198 CSSVTLEETAPV-----KLPTRHCPQPGLRVNVRTSNIKG------WYFKVGSMQTGV 244
    CSS+      PV     +LP    PQP LR+ +  + IKG      W    GS   GV
109 CSSIFGYRNVPVVDILAELPA---PQPLLRLTIDRALIKGSPVLIQWTPAAGSSNAGV 163
NP_416329.4| putative cyclic-di-GMP phosphodiesterase

23S rRNA Frame 6:                            433 HLLLQHRSAS 442
                                                 HLLLQH SAS
NP_418480.1| thiamin phosphate synthase:     39  HLLLQHTSAS 48

16S rRNA Frame 1:
142 AGRKGVKLIP-----LLIDVTRRRSTGLRASSRGNTE--GASVNRNYWASARR 187
    AGR+G  + P     +L +V R RST +  +SRG  E   A +N  Y A  +R
271 AGREG-SIWPYIETGILDEVLRHRSTIVFTNSRGLAEKLTARLNELYAARLQR 322
NP_416170.1| putative ATP-dependent helicase

16S rRNA Frame 1:      88  RRRSLAGLRGPATLELRHGPDSYGRQQWGILHN 120
                           R RS A +  PA L   HGP ++G+     +HN
                       156 RGRSPAQI--PAVLVHSHGPFAWGKNAADAVHN 186
NP_418040.1| L-ribulose-5-phosphate 4-epimerase

16S rRNA Frame 1:      118 LHNGRKPDAAMPRVRRPSGCKVLSAGRKGVKLIPLLI 154
                           +H   KPD A+ /         + AGR+GV  +P L+
                       508 IHRSSKPDLAIEVA--------MEAGRRGVDSVPTLL 536
NP_415397.1| nucleoside triphosphate hydrolase domain

16S rRNA Frame 2:      345 QRAQPLSFVAS-------GPAGNSKETASDKLEEGG 373
                           Q A P + VA        GP+  S++   + LE+GG
                       174 QDATPAALVAHVTSRKCYGPSATSEKCPGNALEKGG 209
NP_414917.2| bacterial alkaline phosphatase

16S rRNA Frame 2:             150 LTLPAEEAPANSVPAAAVIRRVQALIGIT 178
                                  L  P+ +      P  AV+R +Q  +GIT
                              33  LEFPGGKIEMGETPEQAVVRELQEEVGIT 61
NP_414641.1| nucleoside triphosphate pyrophosphohydrolase

16S rRNA Frame 2:                                    235 AQVRKRGEQTGLDTLVV 251
                                                         AQ R+  E TG+D+L V
NP_416599.2| D-tagatose 1,6-bisphosphate aldolase:   158 AQAREFAEATGIDSLAV 174

16S rRNA Frame 3:
       353 PLLPAVRPGTQRRLPVINW--RKVGMTSSHHG----P-YDQGYTRATMAHTKR 398
           P+L A++ G  + LP+I+W  R +      HG    P +D+GY      HT R
       168 PVL-AIQRGVFKVLPIIDWDNRTIYQYLQKHGLKYHPLWDEGYLSVGDTHTTR 219
NP_417242.1| 3'-phosphoadenosine 5'-phosphosulfate reductase
```

Fig. 12. (*continued*)

microbes will share similar tRNAs encoded in similar regions of the rRNA. More specifically, if one were to examine a bacterium other than *E. coli* K12, e.g., *Bacillus subtilis*, then we would predict that *B. subtilis* rRNA–tRNA regions would be very similar to those shared by the *E. coli* rRNA–tRNA regions. Given an early role for a ribosome-based genome, we further expect to find common *E. coli* rRNA-*B. subtilis* tRNA regions corresponding to *B. subtilis* rRNA-*E. coli* tRNA regions as well.

Similarly, if the ribosome predates the origins of cellular life, then the protein modules encoded in the *E. coli* K12 rRNA should also appear in the rRNAs of other microbes and be encoded in similar regions of their rRNAs. Thus, to be more specific, the list of ribosome-related proteins generated here for *E. coli* should be mirrored closely in the *B. subtilis* rRNA and should be encoded in the same order and in the same rRNA subunits. These predictions are, of course, subject to the rRNAs of the various microbial organisms being very highly conserved. Although this is often the case, it is not universally so and the exceptions may prove to be enlightening tests of our theory.

An additional test of the hypothesis concerns the possible functionality of the protein modules encoded by rRNA. Since many of these modules overlap or include known active sites of modern proteins, it is possible that these modules will themselves exhibit biological activity. Such activity should be relatively easily established by synthesizing the modules in question and testing them for the activity found in their modern counterparts. While a positive outcome to such experiments would greatly strengthen the hypothesis

```
16S rRNA Frame 3:      293 AQAVEHVVFDATRRTLPGLDIHGSFQRECA 322
                           AQ \+H   D   R+LP L      Q+ CA
                       124 AQKAQHSALDDIPRSLPALMRAQKIQKRCA 153
NP_417261.1| nucleoside triphosphate pyrophosphohydrolase

16S rRNA Frame 3:      22 VTGSSLLLRRVADGVMSGKLPDGGG 46
                          V ++ LL++  AD   SG+LP G G
                       68 VRAARLLMKTAAD---SGRLPTGSG 89
NP_415726.1| 4-diphosphocytidyl-2-C-methylerythritol kinase

16S rRNA Frame 4:      23  QSGKRPPEGATYFFCNPLPWCDGRCVQGPGTYSPWHSDP 61
                           QSGK       + + + +PW      Q PGT++PW   P
                       88  QSGKGKSRKYLHTYDEAVPWN-----QVPGTFTPWQPLP 121
NP_417814.1| phosphoribulokinase

16S rRNA Frame 4:                     132 AGPLATKDKGCA---RCG 146
                                          A PLAT +KG A   RCG
                                      89  AAPLATVEKGHAMAQRCG 106
NP_418038.1| 3-keto-L-gulonate 6-phosphate decarboxylase:

16S rRNA Frame 4:              29  PEGATYFFCNPLPWCDGRCVQGPGT 53
                                   P G T ++  PP   G+C  GPGT
                               191 PRGNTLYWIGP-P--GGKCDAGPGT 212
NP_417224.1| broad specificity 5'(3')-nucleotidase and polyphosphatase

16S rRNA Frame 5:      101 CSTCVALVVRAMMTRHPH---LPPVYHWQSPLSSRPDRWQQRIRVALVA 146
                           C+  VAL   AMM   P    LPP+  ++    +  PD + ++    L A
                       196 CAILVALFAFAMMRDTPQSCGLPPIEEYK---NDYPDDYNEKAEQELTA 241
NP_416743.1| sn-glycerol-3-phosphate transporter

16S rRNA Frame 5:              229 RAQPPSRHRLRRGLP 243
                                   R   P RHR+RR  P
                               81  RESLPGRHRVRRYRP 95
NP_415397.1|nucleoside triphosphate hydrolase domain

16S rRNA Frame 5:                             138 QRIRVALVAGLNP 150
                                                  +RI+V L AGL+P
NP_417306.1| PEP-protein phosphotransferase:  433 ERIKVMLNAGLSP 445

16S rRNA Frame 6:                     197 LTLRPYSP 204
                                          LTLRPYSP
                                      54  LTLRPYSP 61
NP_414594.1| 4-hydroxy-L-threonine phosphate dehydrogenase:


16S rRNA Frame 6:      368 AHCAIFPTAASRRSLDRVSVPVWLVILSDQLGI 400
                           AH  I P  A+ +   R ++  WL    DQLGI
                       225 AHQQISPDLANSQ---RAALAAWLEEYPDQLGI 254
NP_415451.1| nicotinatephosphoribosyltransferase

16S rRNA Frame 6:              269 LHLEFYPPLRDSS 281
                                   LH  FYPPL  S+
                               295 LHAHFYPPLLRSA 307
NP_415279.1| galactose-1-phosphate uridylyltransferase:

5S rRNA Frame 1:       1   CLAAVARWSHLTPCRTQKNAVAPMVVWGLPMREGTAR 37
                           CL+A   W  +     Q+      V + LP+R+G  R
                       161 CLSAQIIWQAMGHKLYQRLQSWYRVCFALPIRKGWVR 197
NP_415362.1| undecaprenyl pyrophosphate phosphatase
```

Fig. 12. (*continued*)

proposed here, a negative result may simply indicate that these modules are vestigial remnants of a more complex ribosome complex that has off-loaded many of its functions to other organismal genes.

Similarly, it is possible that some of the tRNAs encoded in rRNAs retain functionality. Again, synthesis of these tRNA would permit their activity to be tested in a modern system to determine whether they can be primed with appropriate amino acids and recognize appropriate codons. And again, while a positive outcome would add significantly to the validity of the hypothesis, these tRNA may be vestigial and yield negative results. In this regard, however, it would be interesting to determine how few base substitutions would render such vestigial tRNA active.

Another prediction that follows from our data is that many genes in microbes, besides those encoding the ribosome itself, should have a ribosomal origin. This prediction follows directly from the fact that rRNA appears to encode large numbers of proteins with ribosome-related functions such as the synthetases, ligases, proteases, and phosphatases. These classes of proteins are so essential to cellular life that one would expect that if cells evolved to incorporate pre-existing ribosomes, then rRNA would be the basis for the class of genes encoding synthestases, ligases, protease, phosphatases, etc. for the cell as a whole. Thus, an examination of microbial genomes for rRNA-like regions should reveal significant proportions of these genomes to have originated as rRNA sequences.

4.5. New questions raised by the hypothesis

Many questions are raised by this study and new possibilities realized. Billions of years of evolution have occurred since

```
5S rRNA Frame 1:                     16 TQKNAVAPMVVWGLPMREGTARH 38
                                        T  N  AP+++ G P   GT  H
                                     37 TAANLHAPVIIAGTP---GTFTH 56
NP_416599.2| D-tagatose 1,6-bisphosphate aldolase

5S rRNA Frame 1:                     15  RTQKNAVAPMVV 26
                                         +TQ++A+ P+V+
                                     170 QTQEHAIYPLVI 181
NP_416995.1| phosphoribosylglycinamideformyltransferase

5S rRNA Frame 2:                     18 TPRRWCGVS 26
                                        +PR WCGV+
                                     90 SPRHWCGVA 98
YP_588462.1| undecaprenyl phosphate-alpha-L-ara4N exporter:

5S rRNA Frame 2:                                     12  AELRSETPRRW 22
                                                         A++R++ PRRW
NP_418087.1| kinase that phosphorylates heptoses:    191 AQIRAKVPRRW 201

5S rRNA Frame 2:                                     5   RGGPTPHAELRSETP 19
                                                         RGG   HA L S+ P
NP_418391.1| phosphoenolpyruvate carboxylase:        587 RGGAPAHAALLSQPP 601

5S rRNA Frame 3:                             2  GGRSAVVPPDP 12
                                                G+  A+VP DP
NP_418275.1| uridinephosphorylase:           18 GATLAIVPGDP 28

5S rRNA Frame 3:                             3   GRSAVVPPDPMPNSEVKRR 21
                                                 GR+ ++P   +    V+R+
NP_416815.1| amidophosphoribosyltransferase: 332 GRTFIMPGQQLRRKSVRRK 350

5S rRNA Frame 3:                                     15  NSEVKRRSADG 25
                                                         N+E +RR  DG
NP_418618.1| L-xylulose 5-phosphate 3-epimerase:     127 NNETRRRFRDG 137

5S rRNA Frame 4:                             3  GSSLLSHGETPHYHRRYGVSL 23
                                                GSS  S++  +/+  YG++L
NP_416815.1| amidophosphoribosyltransferase: 79 GSSSASEAQPFYVNSPYGITL 99

5S rRNA Frame 4:                             8   SHGETPHYHRR 18
                                                 SHG++ + HRR
                                             266 SHGNCQKQHRR 276
NP_416219.1|3-deoxy-D-arabino-heptulosonate-7-phosphate synthase

5S rRNA Frame 4:                                     10  GETPHYHRRYGV 21
                                                         GE PH+ R YG+
NP_418523.1| ribophosphonate triphosphate synthase:  264 GEPPHFTRGYGL 275

5SrRNA Frame 5:                              11  RPHTTIGATAFH 22
                                                 RPH +IG   F+
NP_417986.4| cyclic-di-GMP phosphodiesterase: 357 RPHCSIGVAMFY 368

5S rRNA Frame 5:                             30  WDHRATAAR 38
                                                 WDH A  AR
                                             108 WDHHAWQAR 116
NP_416759.1| undecaprenyl phosphate-alpha-L-ara4FN deformylase

5S rRNA Frame 5:                             25 RHGVRWDHRAT 35
                                                RHGVR   +AT
NP_415500.1| phosphoanhydridephosphorylase:  38 RHGVRAPTKAT 48
```

Fig. 12. (*continued*)

ribosomes were incorporated into all living cells. Ribosomal rRNA and the proteins making up the functional structure of ribosomes are now encoded in a separate DNA-based genome. The rRNA sequence may therefore be the vestige of an RNA-protein-based world that has been incorporated into a much more diverse and complex system. How much of that primitive world remains within the rRNA sequence is open to investigation. For example, fragments of most of the synthetases required to charge tRNAs with their appropriate amino acids are encoded in the rRNA sequences, but whether these fragments are the key, functional peptide sequences from which more specific and efficient modern protein synthetases evolved, or whether these are fragments of larger rRNA "genes" that have been shifted over to the DNA genome and now exist only as truncated vestiges, are possibilities that each need to be investigated further. We assume that present-day ribosomes have been stripped of some of their genes and proteins as a result of symbiosis with cells and the incorporation of ribosomal genes into the cellular genome.

The observation that rRNAs encode tRNAs and modules essential to ribosome structure and functions may also force us to reconsider how translation machinery and its associated processes evolved. Most theories of ribosome evolution seem to focus on the

```
5S rRNA Frame 6:       4  FP-TLAWGDPTLPSALRRFTSEFGMGSGGTTALRPPG 39
                          FP T  WG  T  + +     E G G   T+ L+P G
                       4  FPETFLWGGATAANQVEGAWQEDGKGI-STSDLQPHG 39
NP_418177.1| cryptic phospho-beta-glucosidase B

5S rRNA Frame 6:                              8   AWGDPTLPSALRRF 21
                                                  AW+D  L   + RF
NP_416006.4|c-di-GMP phosphodiesterase:       424 AWADQALLEVVNRF 437

5S rRNA Frame 6:               15 PSALRRFTSEFGMGSGGTTALRPPG 39
                                  P  LR+F    G+   G   + P G
                               48 PFELRQFALSHGVAMDGLQVIDPHG 72
NP_416953.1| phosphate acetyltransferase
```

Fig. 12. (*continued*)

Table 4
Chi squared statistical analysis of probabilities that differences between the observed appearance of active site homologies in rRNA-encoded and mRNA-encoded protein controls. The possible protein-encodings of rRNAs listed in Figs. 9–12 were evaluated for whether they overlap identified active regions of the proteins they mimic as listed in the UniProt protein database. The protein homologies found for the control sequences used to calculate Table 2 (fimbrial protein mRNA, sugar efflux protein mRNA and a non-coding mRNA) were evaluated in the same way. Bonferroni correction for 3 comparisons for each data set means that significance at the $p=0.05$ level is accepted at $p=0.017$ (i.e., $\alpha=0.017$). p Values that remain significant are in bold. As in Tables 1–3, the results clearly demonstrate that rRNA encodes tRNAs at a significantly higher rate than a random assortment of mRNAs, and certainly higher than would be predicted from the "null hypothesis".

ACTIVE SITES	FIMBRIAL PROTEIN	SUGAR EFFLUX	NON-CODING
	22% of 112	21% of 100	12% of 110
rRNA PROTEINS 55% of 115	**$Chi^2=51.65$** **$p<0.0001$**	**$Chi^2=46.71$** **$p<0.0001$**	**$Chi^2=74.71$** **$p<0.0001$**
FIMBRIAL PROTEIN 22% of 112		$Chi^2=0.060$ $p=0.86$	$Chi^2=5.83$ $p=0.016$
SUGAR EFFLUX 21% of 100			$Chi^2=4.88$ $p=0.027$

origins of the protein translation center (PTC), since logically it would seem that ribosomal structures necessitating large proteins as a component could not evolve prior to the PTC itself (e.g., Fox, 2010; Tamura, 2011; Hsiao et al., 2013; Mushegian, 2005). If rRNAs themselves encode tRNAs and key protein modules involved in forming the translation machinery, then the evolution of protein translation becomes a boot-strapping problem in which sequences of RNA and protein were mutually selected for encoding integrated transcription and translation functions along with the property of being able to self-aggregate into semi-stable translation platforms. In a prebiotic world, selection would have been for RNAs that could function simultaneously as mRNAs encoding functional protein modules, as primitive rRNAs capable of stabilizing these functional protein or peptide complexes, as tRNAs that could translate RNA sequences into peptides, and as "genes" that could replicate themselves. The evolution of specialized transfer RNA, ribosomal RNA and messenger RNA functions would have evolved only after, and perhaps as a result of, incorporation of the proto-ribosome into cells. The PTC may not be the origin of translation but the result of its evolution.

Our work does not explain the evolution of DNA, but it provides significant hints about how and why DNA storage of information may have evolved. Our guess is that DNA was a natural by-product of RNA replication since many of the RNA-encoded protein sequences (e.g., the ligases and polymerases) we have identified have cognate DNA-related functions. The synthesis of DNA as a by-product of RNA replication would have resulted in an unintended but evolutionarily valuable effect: in "hard" times, during which RNA-based structures became unstable (perhaps due to heat, or changes in salinity or dessication) DNA would have served as a more stable template for ribosomal survival. Indeed, this conjecture gains some credibility from the discovery of transpovirons composed of short segments of DNA that "infect" viruses (Desnues et al., 2012; Yutin et al., 2013). In other words, we suggest that genes evolved in response to protein translation, and to increase its survivability. Genes, then, may be the products of "selfish ribosomes" rather than their origin.

There are obvious limitations to this study. We have examined in detail only a single bacterial rRNA. As with any single species study, there is the possibility that our results are aberrant. The methods and materials used in this study are, however, readily available and easy to apply to the genomes of other bacteria, archaea and protista. Our own brief peek into several of these genomes suggests that our results will be replicated in studies of other evolutionarily primitive rRNAs and therefore that the general principles revealed here will be widely applicable.

The evolution of independently replicating ribosomes assumes that the necessary precursor molecules (sugars, bases, nucleosides, nucleotides, amino acids, etc.) were either readily available in the environment through inorganic, prebiotic reactions or that these are being provided by the simultaneous evolution of other hyperstructures capable of catalyzing these prebiotic reactions (Hunding et al., 2006; Norris et al., 2007, 2012; Root-Bernstein and Dillon, 1997). Acidocalcisomes, for example, could have co-existed separately from ribosomes but in the same environment, providing polyphosphate, polybutyric acid, calcium ions, and other cofactors required for ribosome function. Independent acidocalcisomes may also have buffered the local environment within which the ribosomes evolved. Co-localizing ribosomes and acidocalcisomes within a common membrane would have had obvious evolutionary advantages in facilitating homeostasis and shared metabolic functions. These shared functions would have needed to be incorporated as an integrated set into the first cells and the nature of this integration is not apparent from the present analysis.

We believe that our results provide tantalizing insights into evolution processes that bridge the RNA-world and compositional approaches to the origins of life with LUCA approaches to provide an intermediary state of organization that integrates self-replication with protein translation. A self-replicating ribosomal entity would provide a logical intermediary between self-replicating RNAs or compositionally-organized aggregates of molecules and highly organized, cell-encapsulated genomes. "Selfish" ribosomes, in short, provide one potential intermediary in the process of evolution from the first macromolecules to hyperstructures and finally cells.

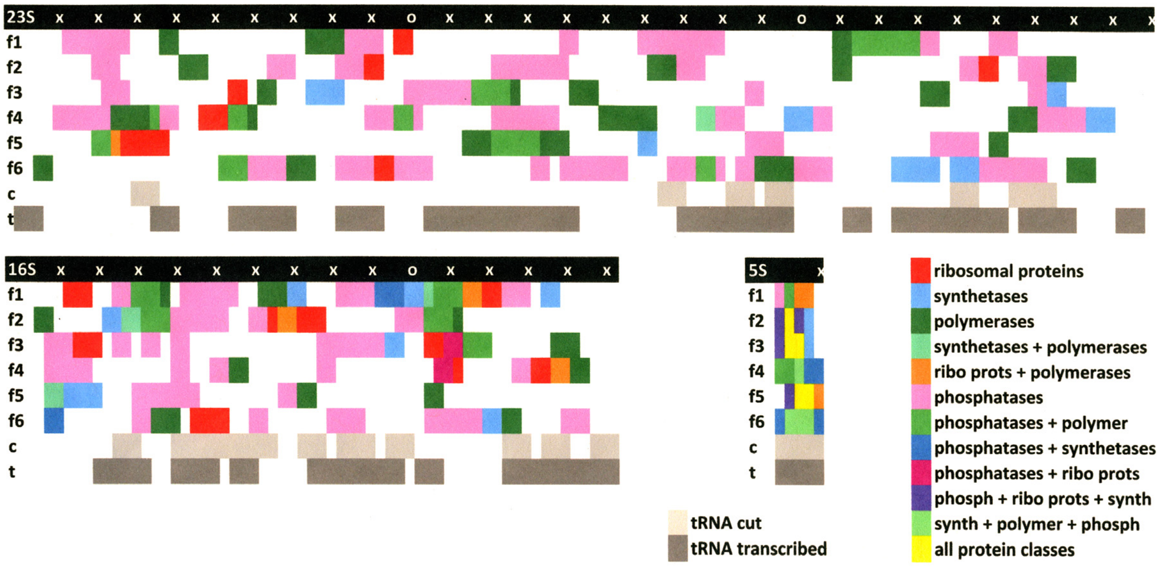


Fig. 13. Map illustrating the location of tRNAs and proteins in various reading frames. This figure summarizes and integrates all of the data from the previous nine Figures. "tRNA cut" refers to direct homologies between tRNAs and rRNAs (implying that tRNAs could be generated by cutting or editing the rRNA itself (Figs. 3 and 5). "tRNA transcribed" refers to the production of tRNA-like sequences from rRNAs by transcribing the rRNA (Figs. 1, 2 and 4). "Synthetases" or "Synth" refers to the sequences in Fig. 9. "Polymerases" or "Poly" refers to the sequences in Fig. 10. "Ribosomal proteins" or "Ribo Prot" refers to the sequences in Fig. 11. "Phosphatases" or "Phosph" refers to the sequences in Fig. 12. Note that all of the 5S, 16S and 23S rRNA sequences, in one or more reading frames, encodes either one or more proteins associated with ribosomal function and/or one or more tRNA sequences. Note also that there is high redundancy in the encoding of the classes of proteins and of the tRNAs.

References

Aggarwal, K., Lee, K.H., 2011. Overexpression of cloned *RhsA* sequences perturbs the cellular translational machinery in *Escherichia coli*. J. Bacteriol. 193 (18), 4869–4880.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Belostotsky, R., Ben-Shalom, E., Rinat, C., Becker-Cohen, R., Feinstein, S., Zeligson, S., Segel, S., Elpelg, O., Nassar, S., Frishberg, Y., 2011. Mutations in the mitochondrial seryl-tRNA synthetase cause hyperuricemia, pulmonary hypertension, renal failure in infancy and alkalosis, HUPRA syndrome. AJHG 88 (2), 193–200. http://dx.doi.org/10.1016/j.ajhg.2010.12.010.

Bloch, D., McArthur, B., Widdowson, R., Spector, D., Guimaraes, R.C., Smith, J., 1984. tRNA–rRNA sequence homologies: a model for the origin of a common ancestral molecule, and prospects for its reconstruction. Orig. Life 14, 571–578. http://dx.doi.org/10.1007/BF00933706.

Bloch, D.P., McArthur, B., Guimarães, R.C., Smith, J., Staves, M.P., 1989. tRNA–rRNA sequence matches from inter-and intraspecies comparisons suggest common origins for the two RNAs. Braz. J. Med. Biol. Res. 22, 931–944.

Caetano-Anolles, G., Seufferheld, M.J., 2013. The coevolutionary roots of biochemistry and cellular organization challenge the RNA world paradigm. J. Mol. Microbiol. Biotechnol. 23 (1-2), 152–177.

Cramer, P., 2002. Multisubunit RNA polymerases. Curr. Opin. Struct. Biol. 12, 89–97.

de Farias, S.T., 2013. Suggested phylogeny of tRNAs based on the construction of ancestral sequences. J. Theor. Biol. 335, 245–248. http://dx.doi.org/10.1016/j.jtbi 2013.06.033.

de Farias, S.T., do Rêgo, T.G., José, M.V., 2014. Evolution of transfer RNA and the origin of the translation system. Front. Genet. 28 (5), 303. http://dx.doi.org/10.3389/fgene.2014.00303 (Aug).

Desnues, C., La Scola, B., Yutin, N., Fournous, G., Robert, C., Azza, S., Jardot, P., Monteil, S., Campocasso, A., Koonin, E.V., Raoult, D., 2012. Provirophages and transpovirons as the diverse mobilome of giant viruses. Proc. Natl. Acad. Sci. U.S.A. 109 (44), 18078–18083. http://dx.doi.org/10.1073/pnas.1208835109 (Oct 30).

Douzounis, C.A., Junin, V., Darzentas, N., Goldovsky, L., 2006. A minimal estimate for the gene content of the last universal common ancestor—exobiology from a terrestrial perspective. Res. Microbiol. 157 (1), 57–68.

Faure, E., Delaye, L., Tribolo, S., Levasseur, A., Seligmann, H., Barthélémy, R.-M., 2011. Probable presence of an ubiquitous cryptic mitochondrial gene on the antisense strand of the cytochrome oxidase I gene. Biol. Direct 6, 56.

Firth, A.E., 2014. Mapping overlapping functional elements embedded within the protein-coding regions of RNA viruses. Nucleic Acids Res. (Oct 17. pii: gku981).

Fonseca, M.M., Harris, D.J., Posada, D., 2014. Origin and length distribution of unidirectional prokaryotic overlapping genes. G3 (Bethesda) 4 (1), 19–27. http://dx.doi.org/10.1534/g3.113.005652 (Jan 10).

Forterre, P., Gribaldo, S., Brochier, C., 2005. LUCA: the last universal common ancestor. Med. Sci. (Paris) 21 (910), 860–865.

Fox, G.E., 2010. Origin and evolution of the ribosome. Cold Spring Harbor Perspect. Biol. , http://dx.doi.org/10.1101/cshperspect.a003483.

Fukuda, Y., Nakayama, Y., Tomita, M., 2003. On dynamics of overlapping genes in bacterial genomes. Gene 323, 181–187 (Dec 24).

Galadino, R., Botta, G., Pino, S., Costanzo, G., DiMauro, E., 2012. Genetics first or metabolism first? The formamide clue. Chem. Soc. Rev. 41 (16), 5526–5565.

Glansdorff, N., Xu, Y., Labedan, B., 2009. The origin of life and the last universal common ancestor: do we need a change of perspective? Res. Microbiol. 160 (7), 522–528.

Gould, J.L., Gould, G.F., 2002. Biostats Basics: A Student Handbook. Freeman, New York, NY.

Hatzoglou, E., Rodakis, G.C., Lecanidou, R., 1995. Complete sequence and gene organization of the mitochondrial genome of the land snail *Albinaria coerulea*. Genetics 140 (4), 1353–1366 (Aug).

Hsiao, C., Lenz, T.K., Peters, J.K., Fang, P.-Y., Schneider, D.M., Anderson, E.J., Preeprem, T., Bowman, J.C., O'Neill, E.B., Lie, L., Athavele, S.S., Gossett, J.J., Trippe, C., Murray, J., Petrov, A.S., Wartell, R.M., Harvey, S.C., Hud, N.V., Williams, L.D., 2013. Molecular paleontology: a biochemical model of the ancestral ribosome. Nucleic Acid Res. 41 (5), 3373–3385.

Huang, X., Miller, W., 1991. A time-efficient linear-space local similarity algorithm. Adv. Appl. Math. 12, 337–357.

Hunding, A., Kepes, F., Lancet, D., Minsky, A., Norris, V., Raine, D., Sriram, K., Root-Bernstein, R., 2006. Compositional complementarity and prebiotic ecology in the origin of life. Bioessays 28 (4), 399–412.

Huvet, M., Stumpf, M.P., 2014. Overlapping genes: a window on gene evolvability. BMC Genomics 15, 721. http://dx.doi.org/10.1186/1471-2164-15-721 (Aug 27).

Iyer, L.M., Koonin, E.V., Aravind, L., 2004. Evolution of bacterial RNA polymerase: implications for large-scale bacterial phylogeny, domain accretion, and horizontal gene transfer. Gene 335, 73–88.

Juehling, F., Puetz, J., Florentz, C., Stadler, P.F., 2012. Armless mitochondrial tRNAs in enoplea (nematoda). RNA Biol. 9 (9), 1161–1166. http://dx.doi.org/10.4161/rna.21630.

Koonin, E.V., 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat. Rev. Micriobiol. 1 (2), 127–136.

Makalowska, I., Lin, C.F., Makalowski, W., 2005. Overlapping genes in vertebrate genomes. Comput. Biol. Chem. 29 (1), 1–12 (Feb).

Mir, K., Schober, S., 2014. Selection pressure in alternative reading frames. PLoS One 9 (10), e108768. http://dx.doi.org/10.1371/journal.pone.0108768 (Oct 1).

Mörl, M., Marchfelder, A., 2001. The final cut. The importance of tRNA 3′-processing. EMBO Rep. 2 (1), 17–20 (Review).

Mushegian, A., 2005. Protein content of minimal and ancestral ribosome. RNA 11, 1400–1406.

Mushegian, A., 2008. Gene content of LUCA, the last universal common ancestor. Front. Biosci. 13, 4657–4666.

Neveu, M., Kim, H.J., Benner, S.A., 2013. The "strong" RNA world hypothesis: fifty years old. Astrobiology 13 (4), 391–403.

Norris, V., den Blaauwen, T., Doi, R.H., Harshey, R.M., Jannier, L., Jimenez-Sanchez, A., Jin, D.J., Levin, P.A., Mileykovskaya, E., Minsky, A., Misevic, G., Ripoll, C., Saier Jr., M., Skarstad, K., Thellier, M., 2007. Toward a hyperstructure taxonomy. Annu. Rev. Microbiol. 61, 309–329.

Norris, V., Loutelier-Bourhis, C., Thierry, A., 2012. How did metabolism and genetic replication get married? Orig. Life Evol. Biosph. 42 (5), 487–495.

Ohtsuki, T., Watanabe, Y., 2007. T-armless tRNAs and elongated elongation factor Tu. IUBMB Life 59 (2), 68–75 (Feb).

Pallejà, A., Harrington, E.D., Bork, P., 2008. Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? BMC Genomics 9, 335. http://dx.doi.org/10.1186/1471-2164-9-335 (Jul 15).

Pereira, S.L., Baker, A.J., 2004. Low number of mitochondrial pseudogenes in the chicken (*Gallus gallus*) nuclear genome: implications for molecular inference of population history and phylogenetics. BMC Evol. Biol. 4, 17. http://dx.doi.org/10.1186/1471-2148-4-17.

Reichert, A.S., Mörl, M., 2000. Repair of tRNAs in metazoan mitochondria. Nucleic Acids Res. 28 (10), 2043–2048 (May 15).

Reichert, A., Rothbauer, U., Mörl, M., 1998. Processing and editing of overlapping tRNAs in human mitochondria. J. Biol. Chem. 273 (48), 31977–31984 (Nov 27).

Root-Bernstein, R.S., Dillon, P.F., 1997. Molecular complementarity 1: The molecular complementarity theory of the origin and evolution of life. J. Theor. Biol. 188, 447–479.

Root-Bernstein, R., 2012. A modular hierarchy-based theory of the chemical origins of life based on molecular complementarity. Acc. Chem. Res. 45 (12), 2169–-2177. http://dx.doi.org/10.1021/ar200209k (2012)⟨http://pubs.acs.org/toc/achre4/45/12⟩.

Schuster, J., Betat, H., Mörl, M., 2005. Is yeast on its way to evolving tRNA editing? EMBO Rep. 6 (4), 367–372 (Apr).

Schuster, P., 2010. Origins of life. Concepts, data, and debates. Complexity 15 (3), 7–10. http://dx.doi.org/10.1002/cplx.20302.

Seligmann, H., Krishnan, N.M., Rao, B.J., 2006. Mitochondrial tRNA sequences as unusual replication origins: pathogenic implications for Homo sapiens. J. Theor. Biol. 243 (3), 375–385 (Epub 2006 Jul 1).

Seligmann, H., 2006. Error propagation across levels of organization: from chemical stability of ribosomal RNA to developmental stability. J. Theor. Biol. 242 (1), 69–80 (Epub 2006 Apr 3).

Seligmann, H., Krishnan, N.M., 2006. Mitochondrial replication origin stability and propensity of adjacent tRNA genes to form putative replication origins increase developmental stability in lizards. J. Exp. Zool. B Mol. Dev. Evol. 306 (5), 433–449 (Sep 15).

Seligmann, H., 2008. Hybridization between mitochondrial heavy strand tDNA and expressed light strand tRNA modulates the function of heavy strand tDNA as light strand replication origin. J. Mol. Biol. 379 (1), 188–199. http://dx.doi.org/10.1016/j.jmb.2008.03.066 (May 23).

Seligmann, H., 2010a. Avoidance of antisense antiterminator tRNA anticodons in vertebrate mitochondria. Biosystems 101, 42–50.

Seligmann, H., 2010b. Undetected antisense tRNAs in mitochondria? Biol. Direct 5, 39.

Seligmann, H., 2011a. Two genetic codes, one genome: frameshifted primate mitochondrial genes code for additional proteins in presence of antisense antitermination tRNAs. Biosystems 105, 271–285.

Seligmann, H., 2011b. Pathogenic mutations in antisense mitochondrial tRNAs. J. Theor. Biol. 269, 287–296.

Seligmann, H., 2012a. An overlapping genetic code for frameshifted overlapping genes in Drosophila mitochondria: antisense antitermination tRNAs UAR insert serine. J. Theor. Biol. 298, 51–76.

Seligmann, H., 2012b. Overlapping genetic codes for overlapping frameshifted genes in Testudines, and *Lepidochelys olivacea* as a special case. Comput. Biol. Chem. 41, 18–34.

Seligmann, H., 2012c. Coding constraints modulate chemically spontaneous mutational replication gradients in mitochondrial genomes. Curr. Genomics 13, 37–54.

Seligmann H. 2013a. Putative protein-encoding genes within mitochondrial rDNA and the D-Loop region. In: Lin, Z., Liu, W. (Eds.), Ribosomes: Molecular Structure, Role in Biological Functions and Implications for Genetic Diseases. Chapter 4, pp. 67–86.

Seligmann, H., 2013b. Pocketknife tRNA hypothesis: anticodons in mammal mitochondrial tRNA side-arm loops translate proteins? BioSystems 113, 165–176.

Seligmann, H., 2014. Putative anticodons in mitochondrial tRNA sidearm loops: pocketknife tRNAs? J. Theor. Biol. 340, 155–163.

Seligmann, H., Labra, A., 2014. The relation between hairpin formation by mitochondrial WANCY tRNAs and the occurrence of the light strand replication origin in Lepidosauria. Gene 542 (2), 48–57. http://dx.doi.org/10.1016/j.gene.2014.02.021 (Jun 1).

Steitz, T.A., 1998. A mechanism for all polymerases. Nature 391, 231–232.

Strobel, S.A., 2001. Repopulating the RNA world. Nature 411, 1003–1006.

Tamura, K., 2011. Ribosome evolution: emergence of peptide synthesis machinery. J. Biosci. 36, 921–928.

Wang, J., Dasgupta, I., Fox, G.E., 2009. Many nonuniversal archaeal ribosomal proteins are found in conserved gene clusters. Archaea 2 (4), 241–251.

Watanabe, Y., Suematsu, T., Ohtsuki, T., 2014. Losing the stem-loop structure from metazoan mitochondrial tRNAs and co-evolution of interacting factors. Front. Genet. 5, 109. http://dx.doi.org/10.3389/fgene.2014.00109 (May 1).

Wende, S., Platzer, E.G., Jühling, F., Pütz, J., Florentz, C., Stadler, P.F., Mörl, M., 2014. Biological evidence for the world's smallest tRNAs. Biochimie 100, 151–158. http://dx.doi.org/10.1016/j.biochi.2013.07.034.

Yutin, N., Raoult, D., Koonin, E.V., 2013. Virophages, polintons, and transpovirons: a complex evolutionary network of diverse selfish genetic elements with different reproduction strategies. Virol. J. 10, 158. http://dx.doi.org/10.1186/1743-422X-10-158 (May 23).