EXPERIMENTAL INVESTIGATION OF AN RNA SEQUENCE SPACE

YOUN-HYUNG LEE, LISA DSOUZA and GEORGE E. FOX*

Department of Biochemical and Biophysical Sciences, University of Houston, Houston, Texas 77204–5934, U.S.A.

(Received April, 1993)

Abstract. Modern rRNAs are the historic consequence of an ongoing evolutionary exploration of a sequence space. These extant sequences belong to a special subset of the sequence space that is comprised only of those primary sequences that can validly perform the biological function(s) required of the particular RNA. If it were possible to readily identify all such valid sequences, stochastic predictions could be made about the relative likelihood of various evolutionary pathways available to an RNA. Herein an experimental system which can assess whether a particular sequence is likely to have validity as a eubacterial 5S rRNA is described. A total of ten naturally occurring, and hence known to be valid, sequences and two point mutants of unknown validity were used to test the usefulness of the approach. Nine of the ten valid sequences tested positive whereas both mutants tested as clearly defective. The tenth valid sequence gave results that would be interpreted as reflecting a borderline status were the answer not known. These results demonstrate that it is possible to experimentally determine which sequences in local regions of the sequence space are potentially valid 5S rRNAs. This approach will allow direct study of the constraints governing RNA evolution and allow inquiry into how the last common ancestor of extant life apparently came to have very complex ribosomal RNAs that subsequently were very conserved.

Introduction

Ribosomal RNAs have been extensively used to determine evolutionary relationships between extant organisms and have succeeded in identifying three major lines of descent that developed from the common ancestor of all life (Woese, 1985). The history of the RNA itself and the closely tied origin of translation is far less clear. By comparing rRNAs from all three lines of descent one can deduce that even at the dawn of true life, the rRNAs were already likely to be quite large. It is nevertheless, not unreasonable to consider that in an earlier RNA world these rRNAs were represented as collections of independent smaller RNA fragments corresponding to functional regions, for example- the peptidyl transferase domain, that subsequently coalesced into larger rRNAs. In order to understand how this could have occurred, one needs to have a better understanding of the constraints that govern the evolution of small RNAs. It is the focus of the current effort to develop a combination theoretical and experimental framework to obtain the needed insights.

It is intuitively attractive to consider sequence changes that occur in any macromolecule over evolutionary time as being reflective of an exploration of a sequence space (Smith, 1970; Ninio, 1983; Eigen *et al.*, 1988; and others). Those sequences which are actually found in extant organisms belong to a special subset of that

* To whom all correspondence should be addressed.

Origins of Life and Evolution of the Biosphere **23**: 365–372, 1993.

sequence space which includes only those sequences that permit the macromolecule to function adequately. Typically such sequences have survived to the exclusion of others because they provide a three dimensional folding that allows proper positioning of important functional groups and, as required, interactions with other macromolecules. Such sets of allowed sequences can be considered to comprise the 'structure space' for a particular molecule. A fundamental relation is thereby established between three dimensional folding and evolution. It is then in principle possible to infer whether or not a particular primary sequence would be evolutionarily allowable by determining the extent to which it can satisfy known structural constraints such as secondary structure, etc.

The ability to reliably predict or determine which RNA sequences are allowable for a particular purpose would provide considerable insight to the evolutionary potential of the RNA. For example, by comparing the valid sequences with those actually found in nature, one could determine the extent to which the sequence space has been explored. Likewise, by examining the distribution (in terms of evolutionary events) of sequences one could determine how dependent the rates at which mutations can be accepted is, on a particular initial sequence. If all valid sequences for two historically related molecules with different functions were identifiable, it would be possible to recognize particular sequences that allow 'discovery' of the other function with the least possible number of evolutionary events. RNA structures are likely to be far less complex than those of proteins. Therefore, it might be possible to predict sets of valid sequences from structural constraints and even to explore them experimentally using *in vitro* methodology (Beaudry and Joyce, 1992; Burke and Berzal-Herranz, 1993) if proper selection criteria can be designed.

It is not necessarily obvious however, to what extent a given set of constraints must be satisfied. For example, within the confines of our current knowledge, converting a normal Watson-Crick base pair in a helical region to a G-U wobble pair may or may not result in a sequence that is valid. Indeed, we would expect that in some cases it would be allowable, whereas in others it would not. It thus would be exceedingly helpful to have an experimental system based on mutagenesis with which practical experience could be gained about the relationship between structural constraints and validity of sequences that satisfy them. Herein an experimental system which will allow us to explore the validity of sequences in a portion of the sequence space associated with 5S rRNA is described.

Mutagenesis is of course not new and a number of important studies with RNA molecules have already been made. For example, Yarus *et al.* (1986) made numerous mutations in the anticodon region of a tRNA and were able to determine their effect on aminoacylation, and translational efficiency, *in vivo*. Lumelsky and Altman (1988) have constructed mutants of the RNA component of RNAse P and examined their effect on catalytic activity *in vitro*. Studies on the large rRNAs have been numerous but have primarily focused on small numbers of mutants at positions that have been implicated with function in some way, e.g. protein binding sites,

sites of antibiotic interaction, etc. Although the available RNA data provide important insights to RNA evolution, they have not been collected with the goal of understanding the structure of the sequence space itself and therefore do not comprehensively address evolutionary issues. These earlier studies will however, be especially useful in determining how similar different RNA sequence spaces are.

5S rRNA was selected because it is large enough to serve as a very realistic model for rRNA evolution. It is an essential component of the translation machinery and interacts with several ribosomal proteins (Dohme and Nierhaus, 1976; Douthwaite *et al.*, 1982; Egebjerg *et al.*, 1989). A wide variety of functional roles have been proposed for 5S rRNA, including involvement with the ribosomal peptidyl transferase or GTPase activity, subunit association, translocation and tRNA binding to the ribosome via a highly conserved sequence segment. Convincing evidence has not been forthcoming for any of these hypothetical roles and one, tRNA binding, has been convincingly ruled out (Zagorska *et al.*, 1984). What is clear at present, is that 5S rRNA is essential for proper assembly of functional large subunits (Van Ryk *et al.*, 1992). Herein we establish the groundwork for an experimental exploration of the structure space of this RNA by establishing a means to explicitly determine the validity of individual sequences as 5S rRNAs.

Outline of Approach

The experimental system is focused on the examination of the sequence space in the vicinity of the 5S rRNA carried by the marine bacterium, *Vibrio proteolyticus*. An important advantage of choosing this 5S rRNA is that sequences from a large number of the other closely related species of the genus *Vibrio* are also known (MacDonell and Colwell, 1985; MacDonell *et al.*, 1986). This provides the opportunity to explore the validity of apparently equally parsimonious evolutionary pathways between known 5S rRNAs as well as to evaluate what portion of the sequences in the immediate vicinity of the *V. proteolyticus* sequence are valid 5S rRNAs. In order to experimentally determine the validity of any putative member of the 5S rRNA sequence set, one must first have assays that provide a criterion for making that decision. In order to establish the criterion, one needs to determine how a number of sequences that are known to belong to the valid sequence set behave in these assays. The best choice for these sequences are of course sequences that are known to exist in nature! Hence, as a first step we herein report the construction and functional evaluation of a number of sequences that occur in various extant *Vibrio* species.

The experimental system employed consists of a synthetic 5S rRNA embedded in a mini rRNA operon consisting of the two ribosomal RNA promoters, P1 and P2, a small fragment of 16S rRNA sequence and the two terminators, T1 ad T2, normally found in the *E. coli rrn*B operon. This mini-operon contains appropriate information for rRNA processing and when introduced into *E. coli* on a plasmid (Bronsius, 1984), the synthetic 5S rRNA is produced in large amounts. We wish

to be able to rapidly determine if any particular synthetic RNA is a valid 5S rRNA *in vivo* and hence a member of the 5S rRNA structure space. To this end two assays were developed. The first is a competitive fitness assay in which the rate of growth of the strain containing the test RNA is compared to a strain containing a standard RNA. The second is an *in vivo* ribosome analysis. Here the proportion of the test RNA in the pool of total 5S rRNA is determined, as well as the extent to which the RNA is incorporated into 50S subunits, 70S ribosomes and polysomes. A valid 5S rRNA would presumably be properly processed and accumulate to high levels in 50S subunits, 70S ribosomes and polysomes without obvious effect on organism fitness. If a 5S rRNA behaving in this manner were not functioning properly, then the numerous ribosomes carrying it would be detrimental to the organism with a resultant decline in fitness.

Materials and Methods

Materials and methods have been described in detail elsewhere (Hedenstierna *et al.*, 1993). In brief a *V. proteolyticus* 5S rRNA gene (MacDonell and Colwell, 1985) was constructed from 15 chemically synthesized deoxyoligonucleotides using a polymerase chain reaction approach described by Jayaraman *et al.*, (1989). The synthesized segment contained appropriate restriction sites such that it could be cloned into M13mp18 as a *Bam*HI-*Sal*I fragment. The construct was verified by sequencing and an expression vector, was made by replacing the 508 bp *Hin*dIII-*Hin*dIII fragment in pKK5-1 (Brosius,1984; Szeberenyi and Apirion, 1984) with our construct. Finally the normal *rrn*B terminator, T2 was added back to the construct. Additional wild type sequences and mutants, Table I, were made by successive rounds of site directed mutagenesis (Zoller and Smith, 1984; Vandeyar *et al.*, 1988) using commercially available kits (i.e., T7-Gen; USB).

Various 5S rRNA fractions were purified according to the following procedures. Total cellular RNA was prepared by low pH phenol extraction as described by Wallace (1987). Ribosomes, polysomes and 50S subunits were separated by sucrose gradient centrifugation following the protocol of Godson and Sinsheimer (1967). Gradient fractions containing ribosomes, polysomes or 50S subunits were then precipitated with ethanol, and phenol/chloroform extracted. RNA was purified by electrophoresis on a pH 3.5, 10% polyacrylamide gel by a modified version of the Jordan and Raymond (1969) procedure.

The proportion of *V. proteolyticus* 5S rRNA relative to total 5S rRNA in either whole cells, ribosomes, or 50S subunits was measured by hybridizing filterbound RNA to a probe complementary to *V. proteolyticus* 5S rRNA, stripping the filter, and hybridizing a second time with a probe complementary to *E. coli* 5S rRNA (Sambook *et al.*, 1989). Oligonucleotides HV2 (5′-GTCCAAATCGCTATGGTCGC-3′) and HE2 (5′-GACCACCGCGCTACTGCCGC-3′), complementary to bases 7 to 26 in *V. proteolyticus* and *E. coli* 5S rRNA respectively, were endlabelled with ^{32}P and used as probes. Known mixtures of *V. proteolyticus* 5S rRNA and *E. coli* 5S

rRNA were used to construct standard curves and blackening of autoradiogram bands was determined with a densitometer (Technology Resources Inc.).

The fitness of the strains carrying each of the various constructs was measured relative to a control strain carrying the *V. proteolyticus* construct by monitoring the ratio of cell numbers in actively growing mixed cultures (Hedenstierna *et al.*, 1993). The ratio of the two cell types was determined by spreading samples on indicator plates (McConkey agar), which discriminate between *lac*$^-$ *and lac*$^+$ strains by colony color (Chao and McBroom, 1985; Dykhuizen and Hartl 1983; Hartl *et al.*, 1983; Lenski, 1988a, b). The two test strains alone exhibit no significant difference in fitness. Two experiments were done in each case such that the construct was in one strain and the control in the other. The fitness difference, S, attributed to the construct as defined by Chao and McBroom (1985) was determined as the slope of a line fitted to data points plotted as t against $\ln[N_p(t)/N_r(t)]$.

Results

In the current studies an additional 9 wild type *Vibrio* 5S rRNAs were constructed

TABLE I

Properties of various 5S rRNAs in the *E. coli* host environment

5S rRNA[1]	Fitness Difference[2] (h^{-1})	% Incorporation			
		ΣRNA[3]	70S	50S	ΣPOLY[4]
V. proteolyticus	-0.028	26	36	29	34
A10C variant	0.047	5	tr[5]	tr	tr
A39C variant	-0.028	>50	tr	tr	tr
V. charcariae[7]	0.010	>50	>50	>50	>50
V. diazotrophicus[8]	-0.006	>50	>50	>50	>50
V. nereis[9]	N.E.[6]	43	35	26	27
V. alginolyticus[10]	0.020	>50	47	45	>50
V. gazogenes[11]	0.030	31	10	11	15
V. parahemolyticus[12]	N.E.	>50	>50	>50	>50
V. natriegenes[13]	0.012	>50	>50	>50	>50
V. patrii[14]	0.028	>50	>50	>50	>50
V. strain LT-1[15]	0.044	>50	>50	>50	>50

1. Footnotes 7-15 indicate the sequence differences of the various 5S rRNAs relative to *Vibrio proteolyticus.*
2. *V. proteolyticus* is compared to an *E. coli* strain carrying plasmid pKK5-1. All others are compared to an *E. coli* strain carrying the plasmid with the *V. proteolyticus* gene.
3. ΣRNA: total cellular RNA (= cytosolic RNA + ribosomal RNA).
4. ΣPOLY: polysomal pool (ie 2x, 3x, 4x).
5. tr: trace (< 5%).
6. N.E.: no effect.
7. U41A52; 8. A52C88; 9. C19::U34.1A52G64U107; 10. U19::U34.1A52A64
11. U19A64C65C88; 12. U17U19U41A52A64; 13. U19::C34.1A52A64
14. U19::C34.1U41A52A64; 15. U19::C34.1C35U41A52A64U114.

and characterized. The results pertaining to these alternative wild types as well as *V. proteolyticus* and two mutants are presented in Table I. It is found that in all cases the effect of the plasmid carrying the construct on growth rate is extremely small and in many cases slightly favorable. The wild type strains are all expressed to high levels and with the exception of *Vibrio gazogenes*, incorporated into 50S subunits, 70S ribosomes, and polysomes at similar levels.

Two mutants of *V. proteolyticus* of unknown validity were examined, Table I. In both cases the effect on fitness was minimal. The mutation of A10 to C10 results in the expression of an RNA that accumulates to only very low levels, apparently because it is unstable. The amount of A10C mutant is found in even lower levels in the ribosomes. An A39C mutant, Table I, results in an RNA which accumulates to very high levels but simply does not enter ribosomes to any significant extent. Hence one can conclude that despite their overall similarity to 5S rRNA, neither of these mutants is in fact a valid 5S rRNA in the internal environment of *E. coli*.

Discussion

The key result here is that ten 5S rRNA sequences of known evolutionary validity have been examined and all are found to be incorporated into *E. coli* ribosomes at high levels *in vivo* without significant effect on organism fitness. In contrast 5 two point mutations of the *V. proteolyticus* sequence behaved distinctly different. In one case the mutant RNA did not accumulate. In the other case the mutant 5S rRNA accumulated in the RNA pool, but did not reach the ribosome in significant amounts. These results are of critical significance to the long term effort because they imply that it will be possible in most cases to reach a definitive conclusion about whether or not a particular sequence variant does or does not belong to the 5S rRNA structure set.

A major concern was the possibility that valid sequences would exhibit a range of behavior such that it would be very difficult to decide if an unknown sequence were or were not valid. In fact, of the ten valid sequences examined to date with one exception, all have three key properties, high accumulation levels, levels of incorporation comparable to the level of accumulation and no major effect on fitness. When incorporation into polysomes is high, the lack of obvious effect on fitness as measured by competitive growth studies provides substantial evidence that the sequences being tested are functional.

Of the ten wild type sequences examined, two, *V. proteolyticus* and *V. gazogenes* accumulate to a lesser extent than the others. This suggests that minor differences in either stability or ease of processing may exist. Only in the case of *V. gazogenes* however, is there sufficient deviation in behavior to cast into doubt the validity of the sequence as a 5S rRNA. In this case, there is a clear decrease in the level of ribosome incorporation relative to total RNA accumulation and hence if we did not know its true status, this sequence would be classified as a borderline case since it does enter ribosomes to a considerable extent and is apparently not deleterious.

In principle it might have turned out that the co-evolution phenomenon would make an experimental study of the 5S rRNA structure space in isolation untenable. The presence of approximately 20 sequence changes between a typical *Vibrio* 5S rRNA and *E. coli* 5S rRNA taken in conjunction with evidence that even a single change can substantially change protein/5S rRNA binding constants (Goringer and Wagner, 1986) would appear to make it unlikely that any *Vibrio* 5S rRNA, let alone 10 out of 10, could function effectively in the internal environment of *E. coli*. In fact, earlier *in vitro* reconstitution studies had already shown that 5S rRNAs could be interchanged over greater phylogenetic distances (Wrede and Erdmann, 1973; Bellemare *et al.*, 1973; Hartmann *et al.*, 1988). Successful chimeras involved all prokaryotic 5S rRNAs tested including that of one archaebacterium. None of the chimeras involving eukaryotic 5S rRNAs were successful however. We speculate that it is the overall ability of the 5S rRNA/protein complex to enter 50S particles which is the paramount characteristic of a valid sequence rather than the strength of interaction with the individual proteins. Thus we suspect that eventual incorporation is largely responsive to changes in protein binding constants in an on/off way. Therefore even though individual sequence changes may significantly affect the binding constants of one or more of the proteins (Christianson *et al.*, 1985; Goringer and Wagner, 1986), the effect on final incorporation levels may be unimportant over a large range of affinities.

This is not to say that co-evolution of protein and RNA does not exist, as even here the *V. gazogenes* results imply that it does. What it does mean however, is that it will be manageable. Indeed, it may be possible to directly study such co-evolution by examining the effect of identical changes on different underlying sequences or by examining the behavior of an entire set of variants of one underlying sequence in an alternative host organism. In summary, the way to an experimental characterization of a portion of the 5S rRNA structure space is now clear. It will therefore be possible to determine the extent to which readily determinable constraints such as secondary structure, protein binding sites, etc. allow one to predict an RNA structure space.

Acknowledgements

This investigation was supported by grants from the National Aeronautics and Space Administration (NAGW-2198), the Whitehall Foundation and the University of Houston Energy Laboratory. We especially acknowledge Klas Hedenstierna and Yinqing Yang for helpful discussions.

References

Beaudry, A. A. and Joyce, G. F.: 1992, *Science* **257**, 635.
Bellemare, G., Vigne, R., and Jordan, B.: 1973, *Biochimie* **55**, 29.
Brosius, J.: 1984, *Gene* **27**, 161.
Burke, J. M. and Berzal-Herranz, A.: 1993, *FASEB J.* **7**, 106.

Chao, L. and McBroom, S. M.: 1985, *Molec. Biol. Evol.* **2**, 359.
Christiansen, J., Douthwaite, S. R., Christensen, A., and Garrett, R. A.: 1985, EMBO J. **4**, 1019.
Dohme, F. and Nierhaus, K. H.: 1976, *Proc. Natl. Acad. Sci. U.S.A.* **73**, 2221.
Douthwaite, S., Christensen, A., and Garrett, R. A.: 1982, *Biochemistry* **21**, 2313.
Dykhuizen, D. E. and Hartl, D. L.: 1983, *Microbiol. Rev.* **47**, 150.
Egebjerg, J., Christiansen, J., Brown, R. S., Larsen, N., and Garrett, R. A.: 1989, *J. Mol. Biol.* **206**, 651.
Eigen, M. and Winkler-Oswalitsch, R.: 1988, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 5913.
Godson, G. N. and Sinsheimer, R. L.: 1967, *Biochim, Biophys. Acta* **159**, 489.
Goringer, H. U. and Wagner, R.: 1986, *Z. Phys. Chem.* **367**, 769.
Hartl, D. L., Dykhuizen, D. E., Miller, R. D., Green, L., and de Framond, J.: 1983, *Cell* **35**, 503.
Hartmann, R. K., Vogel, D. W., Walker, R. T., and Erdmann, V.A.: 1988, *Nucl. Acids Res.* **16**, 3511.
Hedenstierna, K.O.F.., Lee, Y-H., Yang, Y., and Fox, G. E.: 1993, *System. Appl. Microbiol.* **16**, 280.
Jayaraman, K., Shah, J., and Fyles, J.: 1989, *Nucl. Acids Res.* **17**, 4403.
Jordan, E. M. and Raymond, S.: 1969, *Anal. Biochem.* **27**, 205.
Lenski, R. E.: 1988a, *Evolution* **42**, 425.
Lenski, R. E.: 1988b, *Evolution* **42**, 433.
Lumelsky, N. and Altman, S.: 1988, *J. Mol. Biol.* **202**, 443.
MacDonell, M. T. and Colwell, R. R.: 1985, *Syst. Appl. Microbiol.* **6**, 171.
MacDonell, M. T., Swartz, D. G., Ortiz-Conde, B. A., Last, G. A., and Colwell, R. R.: 1986, *Microbiol. Sci.* **3**, 171.
Ninio, J.: 1983, *Molecular Approaches to Evolution*, Princeton University Press, Princeton, N.J.
Sambrook, J., Fritsch, E. F., and Maniatis, T.: 1989 *Molecular Cloning: A Laboratory Manual*, 2nd ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor.
Smith, J. M.: 1970, *Nature* **225**, 563.
Szeberenyi, J. and Apirion, D.: 1984, *Biochim. Biophys, Acta* **783**, 15.
Vandeyar, M., Weiner, M., Hutton, C., and Batt, C.: 1988, *Gene* **65**, 129.
Van Ryk, D. I., Lee, Y., and Nazar, R. N.: 1992, *J. Biol. Chem.* **267**, 16177.
Wallace, D. M.: 1987, *Meth. Enzymol.* **152**, 33.
Woese, C. R.: 1987, *Microbiol. Rev.* **51**, 221.
Wrede, P. and Erdmann, V. A.: 1975, *FEBS Let.* **33**, 315.
Yarus, M., Cline, S. W., Wier, P., Breeden, L., and Thompson, R. C.: 1986, *J. Mol. Biol.* **192**, 235.
Zagorska, L., Can Duin, J., Noller, H. F., Pace, B., Johnson, K. D., and Pace, N. R.: *J. Biol. Chem.* **259**, 2798.
Zoller, M. J. and Smith, M.: 1984 *DNA* **3**, 479.