Volume 17 Number 6 1989 Nucleic Acids Research

Phylogenetic evidence for tertiary interactions in 16S-like ribosomal RNA

Tim Haselman, David G.Camp and George E.Fox*

Department of Biochemical and Biophysical Sciences, University of Houston, Houston, TX 77204-5500, USA

Received December 14, 1988; Accepted February 15, 1989

ABSTRACT

Major efforts are underway to elucidate the spatial distribution of ribosomal RNAs (rRNAs) in the ribosome. An especially informative approach is the identification of likely base−base tertiary interactions within the RNA by phylogenetic comparison. Herein evidence is presented for three heretofore unrecognized candidate tertiary interactions, G506/C525, C779/G803 and A994/U1380 (*1*) in 16S-like rRNAs. This brings to eight the number of such interactions that are strongly supported by phylogenetic evidence. The three newly identified interactions further define the folding within domains II and III of 16S-like rRNA. No interactions have yet been found that would serve to orient the domains relative to one another.

INTRODUCTION

Comparative sequence analysis has been used to successfully elucidate secondary structure in many RNA molecules including tRNA (*2*), rRNA (*3−6*), snRNA (*7−9*), and the RNA component of RNAse P (*10*). Although not as widely appreciated, the comparative approach has also been useful in recognizing tertiary interactions. The best known example is the Levitt pair (*11*) in tRNA which follows Watson−Crick pairing rules. Other studies have focused on tRNA (*12−13*), 16S rRNA (*14−15*) and 23S rRNA (*16−17*). The essence of a comparative search is to locate pairs of positions at which a high degree of coordinate base change or covariation (*11,18−19*) is found in an aligned set of sequences that exhibit variation over evolutionary time. Herein putative tertiary interactions in 16S-like rRNA that were recognized by a computer program sensitive to covariation (*13*) are reported and the status of interactions suggested earlier (*14*) are re-examined.

A previous search for stringently coordinated base changes (*14*) revealed a number of covariant positions. The most convincing of these were actually part of the secondary structure and thus served to emphasize the quality of the comparative data that underlies the standard structure. In addition ten promising candidate tertiary interactions were identified and a variety of less promising prospects tabulated. This earlier study had the single shortcoming that the sequences then available were limited in both number and phylogenetic diversity. Thus many positions showed correlations with several sites while it was obvious that positions that had not yet been found to vary might be concealing important interactions. The present effort takes advantage of a substantial expansion of the data base since late 1984.

RESULTS

The current status of interactions that follow Watson−Crick pairing rules is that five of the ten strongest candidates identified earlier, have clearly withstood the tests of additional

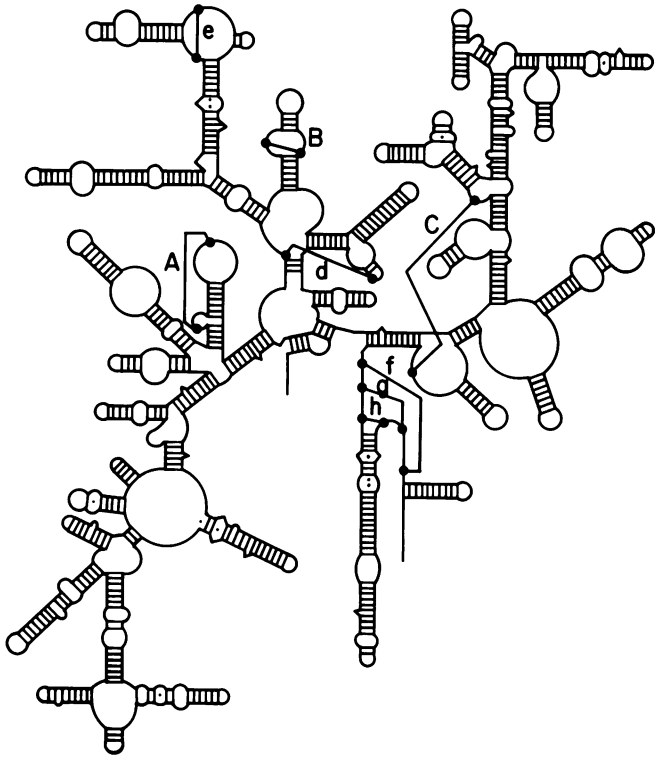

Figure 1. Location of the eight candidate tertiary interactions in 16S rRNA. The secondary structure of *E.coli* 16S rRNA (*14*) is shown in schematic format. The location of the tertiary interactions for which extensive phylogenetic evidence exists are indicated. For reference purposes each is labeled with a letter. The three new interactions documented here are assigned an upper case letter and the five previously recognized (*14*) interactions are assigned a lower case letter. The interactions are as follows: (A) G506/C525; (B) C779/G803; (C) A994/U1380; (d) G570/C866; (e) A673/U717; (f) C1399/G1504; (g) G1401/C1501 and (h) G1405/C1496.

data, two are now considered part of the secondary structure, one remains uncertain and two are probably incorrect. In addition three new strongly supported tertiary interactions are reported here. In total then at least eight strong candidate tertiary interactions that follow Watson–Crick pairing rules have been identified in the small subunit rRNAs. The locations of these eight are summarized in Figure 1.

The three new interactions that have been detected are G506/C525, C779/G803, and A994/U1380, Figure 1-interactions A, B, and C, respectively. The comparative data pertaining to these interactions is summarized in Tables 1, 2 and 3. The G506/C525 interaction was hidden in the earlier analysis (*14*) in a cluster of nine positions, all of which exhibited the same pattern of base variation. These candidate tertiary interactions have both been identified as a result of new data from extremely divergent RNAs. The need to have very diverse sequences in which the structure has been evolutionarily challenged (*ie.*, especially mitochondria) seems to be the rule. Interactions such as G570/C866, Figure 1-interaction d, that exhibit frequent variation seem to be the exception.

The comparative evidence for the G506/C525 interaction, Table 1, is very impressive

TABLE 1. Known sequence variation associated with base positions 506 and 525. Tables 2–5 are presented in the same format as described here. Bases at the key positions are in bold. Surrounding nucleotides are in lower case. They serve to define and illustrate the local alignment. The organism from which the particular sequence data shown was obtained is indicated. In addition the number of organisms in the same phylogenetic class with the same sequence version at the key positions is indicated in parenthesis. The number of sequences in particular phylogenetic classes varies for each position because the data base is comprised of partial sequences as well as complete sequences. General phylogenetic terminology is abbreviated such that mt=mitochondria, chl=chloroplast, eub=eubacteria, ani=animal and insect, arch=archaebacteria, pl=plant, pro=protist and fungal. Unless otherwise referenced, all sequence data shown in all tables is tabulated in the recent collection (*21*).

506	525	
caccg	**G** cuaa...gcag	**C** cgcgg	*Escherichia coli* (eub=54)
caucg	**G** cuaa...gcag	**C** cgcgg	*Zea mays chl.* (pl=5)
ggccg	**G** gcaa...gccg	**C** cgcgg	*Archaeoglobus fulgidus* (arch=13)
uggag	**G** gcaa...gcag	**C** cgcgg	*Homo sapiens* (ani=11)
uggag	**G** gcaa...gcag	**C** cgcgg	*Zea mays* (pl=7)
uggag	**G** gcaa...gcag	**C** cgcgg	*Tetrahymena thermophila* (pro=24)
gguug	**G** ucaa...gcca	**C** cgcgg	*Homo sapiens mt.* (ani=11)
uauug	**A** ccaa...gcag	**U** cgcgg	*Drosophila yakuba mt.* (ani)
ccccg	**G** cuaa...gcag	**C** cgcgg	*Zea mays mt.* (pl=5)
caccc	**C** ccaa...gaag	**G** gucgg	*Chlamydomonas reinhardtii mt.* (pl)
cucug	**G** cuaa...gcag	**C** cgcgg	*Tetrahymena pyriformis mt.* (pro=4)
ucuug	**A** ccaa...gcag	**U** cgcgg	*Aspergillus nidulans mt.* (pro)
uccug	**A** cuaa...gcag	**U** cgcgg	*Saccharomyces cerevisiae mt.* (pro)

with no wobble pairs (herein referring to G–U pairs only) combined with the existence of a strong sequence constraint. There is only one instance in which a pyrimidine has been found at position 506. The G506/C525 interaction places a strong conformational constraint on the structure of the 518–533 loop and the interaction is consistent with the inaccessibility of chemical probes specific for single-stranded RNA to positions 506 and 525 in 30S ribosomal subunits (*20*). In addition the residues involved in the newly proposed interactions at C779/G803 and A994/U1380 as well as the interactions at G570/C866 and A673/U717 are all protected from chemical modification by single-strand-specific probes in 30S subunits (*20*). The C779/G803 covariation has no exceptions, though a few wobble examples do exist, Table 2. This interaction would have rather local consequences in domain II. The A994/U1380 is likewise a perfect correlation with a wobble pair version occuring in *Drosophila yakuba* and *D. virilis* mitochondria (*21*). This interaction is extremely informative about the large scale folding of domain III of the RNA and appears to be consistent in this regard with a global three dimensional model of the small subunit that is based primarily on crosslinking data (*22*).

The ten interactions proposed earlier (*14*) will now be reviewed beginning with those in the three major structural domains. The evidence for G570/C866, Figure 1-interaction d, is overwhelming in all phylogenetic groups as has been documented previously (*14,15*) and continues to mount as new sequences are determined. Likewise, U921/A1396 is phylogenetically proven but is now regarded as part of the secondary structure due to the proximity of a conserved pair G922/C1395. The A673/U717 interaction, Figure 1-interaction e, is arguably part of either the secondary or tertiary structure. Regardless of which, it clearly is a 'local' interaction that serves to associate two areas of the structure that are obviously near each other from the existing secondary structure information. In

TABLE 2. Known sequence variation associated with base positions 779 and 803.

779		803		
ag	C aaaca...gua	G	ucc	*Escherichia coli* (eub=54)
ag	C aaaug...gua	G	ucc	*Zea mays chl.* (pl=5)
ag	C gaacc...gua	G	ucc	*Archaeoglobus fulgidus* (arch=13)
uu	C gaaga...gua	G	uuc	*Homo sapiens* (ani=8)
uu	U gaaga...gua	G	uuc	*Oryctolagus caniculus* (ani)
cu	C gaaga...cua	G	ucu	*Zea mays* (pl=7)
au	C aaaga...gua	G	ucu	*Tetrahymena thermophila* (pro=25)
ag	U gaaga...gua	A	ucu	*Plasmodium berghei* (pro)
cc	A aacug...cua	U	gcu	*Homo sapiens mt.* (ani=9)
ag	C gaaca...gua	G	ucc	*Zea mays mt.* (pl=6)
au	C gauga...gua	G	ucc	*Tetrahymena pyriformis mt.* (pro=4)
aa	U gaaaa...gua	G	ucu	*Paramecium primaurelia mt* (pro)

addition its variation pattern is more like that of secondary structure in that a wobble pair is sometimes found. Table 4 documents the comparative evidence for the A673/U717 interaction. It is established by evidence from the three major kingdoms, though in mitochondrial sequences this pair is not always supported by the comparative evidence.

Two interactions that were proposed earlier now seem at best highly suspect. The U118/A288 interaction is not supported by sequences from *Chloroflexus aurantiacus* and *Chlamydomonas reinhardtii (23)* mitochondria. In addition significant numbers of favorable cases have not turned up since the earlier study. The G9/C507 interaction is not supported by sequences from *Trypanosoma brucei*, *Crithidia fasiculata*, *Chlamdomonas reinhardtii* mt. and *Drosophila yukuba* mt. and furthermore does not seem consistent with other spatial data if the much more convincing G506/C525 interaction proposed here is correct. This is because the G506/C525 interaction would put G527 near C507. Nucleotide G527 is methylated and has been located by immune electron microscopy (*24*). Likewise, G9 is near the 5′ end which has also been mapped by immune electron microscopy (*25*). These studies indicate the 5′ end is not near G527. In addition a global model based primarily on crosslinking data (*22*) does not support a location for G9 that would place it in the vicinity of C507.

TABLE 3. Known sequence variation associated with base positions 994 and 1380.

	994		1380	
ccuggucuug	A ca...ugaauacg	U	ucccgggccuug	*Escherichia coli* (eub=54)
ccagggcuug	A ca...cgaauccg	U	ucccgggccuug	*Zea mays chl.* (pl=5)
ccgggga-g	A ca...ugaauacg	U	cccugcuccuug	*Archaeoglobus fulgidus* (arch=13)
cccggcccgg	A ca...ugauuaag	U	cccugcccuuug	*Homo sapiens* (ani=11)
ccagguccag	A ca...ugacuacg	U	cccugcccuuug	*Zea mays* (pl=7)
cgagcgcaag	A ca...ugauuaug	U	cccugccguuug	*Tetrahymena thermophila (pro=25)*
ccaccucuug	C uc...ugaacagg	G	cccugaagcgcg	*Homo sapiens mt.* (ani=9)
cuuaaauuug	U aa...ugauuuua	G	cucuaaaauaug	*Drosophila yakuba mt.* (ani)
ccagcccuug	A ca...ugaauaug	U	acccgggcccug	*Zea mays mt.* (pl=5)
ccacuuuuua	U ua...ugaa-ggu	A	accucuauugug	*Chlamydomonas reinhardtii mt.* (pl)
ccaacguuuu	A gg...ugaaauau	U	agucaaauuuug	*Tetrahymena pyriformis mt. (pro)*

TABLE 4. Known sequence variation associated with base positions 673 and 717.

673		717		
ggu	A gaauu...ggaa	U acc	*Escherichia coli* (eub=10)
ggu	G gaauu...ggaa	U acc	*Neisseria gonorrhoeae (27)* (eub=11)
agu	G gaauu...ggaa	C acc	*Bacillus subtilis* (eub=29)
gag	G gaauu...agaa	C acc	*Zea mays chl.* (pl=5)
ggg	G guauu...ggac	C acc	*Archaeoglobus fulgidus* (arch=13)
auu	C guauu...agac	G gac	*Homo sapiens* (ani=10)
auu	C guauu...agac	G aac	*Zea mays* (pl=7)
auu	A guauu...ggac	U aac	*Tetrahymena thermophila* (pro=18)
auc	G guauu...agac	U aac	*Saccharomyces cerevisiae* (pro)
auu	U guauu...agac	A aac	*Plasmodium berghei* (pro)
aaa	G auacg...agau	C cac	*Euglena gracilis* (pro).
gaa	C guacu...agac	G aac	*Trypanosoma brucei* (pro)
gcu	A aaacu...acaa	A aua	*Homo sapiens mt.* (ani=8)
agu	G gaauu...ggaa	C gcc	*Zea mays mt.* (pl=5)
ugg	G uaaua...ggaa	A gcc	*Chlamdomonas reinhardtii mt.* (pl)
gga	A gaagc...agac	U ggu	*Aspergillus nidulans mt.* (pro)

The remaining interactions that were proposed earlier are all at the interface of domain III and the 3′ subdomain. The status of the possible interactions in this region has been virtually unchanged by the expansion of the sequence data base. As many as five closely clustered candidate interactions were proposed originally (*14*). Of these, C1399/G1504, G1401/C1501, and G1405/C1496 continue to receive unequivocal, if somewhat limited, phylogenetic support, Table 5. Only one of these interactions, C1399/G1504 has been encountered in a wobble version. These three interactions form a nested set in the same vicinity and the main structural effect of this is to loosely extend helix 1420−1490, shown in Figure 1-interactions f, g, and h. As has been spelled out previously (*14*), an additional

TABLE 5. Known sequence variation associated with the putative base-pair interactions 1399/1504, 1401/1501 and 1405/1496.

1399	1401	1405	1496	1501	1504	
acaca	C c G	ccc G uca...aagu	C guaa	C	aa G g	*Escherichia coli* (eub=37)
acaca	C c G	ccc G uca...aagu	C guaa	C	aa G g	*Zea mays chl.* (pl=5)
acaca	C c G	ccc G uca...aagu	C guaa	C	aa G g	*Archaeoglobus fulgidus* (arch=13)
acaca	C c G	ccc G ucg...aagu	C guaa	C	aa G g	*Homo sapiens* (ani=12)
acaca	C c G	ccc G ucg...aagu	C guaa	C	aa G g	*Zea mays* (pl=7)
acaca	C c G	ccc G ucg...aagu	C guaa	C	aa G g	*Tetrahymena thermophila* (pro=36)
acaca	C c G	ccc G uca...aagu	C guaa	C	au G g	*Homo sapiens mt.* (ani=8)
acaca	U c G	ccc G ucg...aagu	C guaa	C	au A g	*Drosophila yakuba mt.* (ani)
acaua	U c G	ccc A ucg...aagu	U guaa	C	au A g	*Aedes albopictus mt. (28)* (ani)
acaca	C c G	ccc G uca...aagu	C guaa	C	aa G g	*Zea mays mt.* (pl=5)
acaca	U u G	ccc G uca...aagu	C guaa	C	a- G g	*Chlamydomonas reinhardtii mt.* (pl)
acuaa	C c A	ccc G uca...aagu	C gaaa	U	au G g	*Aspergillus nidulans mt.* (pro)
acuaa	U c A	cuc A uca...aagu	U gaaa	U	ac A g	*Saccharomyces cerevisiae mt.* (pro)
acaca	C u G	ccc A uca...aagu	U gaca	C	aa G g	*Paramecium primaurelia mt* (pro=2)

four base-pairs (1404/1497, 1407/1494, 1394/1506 and 1395/1505) could be added to the three interactions summarized in Table 5 to further stablize an extended anti-parallel interaction. All the positions comprising these additional four interactions remain universally conserved in the expanded sequence set and thus this earlier proposal (*14*) remains imminently reasonable. The other two interactions proposed earlier in the 1400 region remain problematical. The A1398/U1406 interaction remains largely untested as does C1399/G1405. The latter does, however, encounter negative evidence in both *Paramecium primaurelia* and *P.tetraurelia* mitochondrial RNAs. Thus the alternative interaction G1405/C1496 seems more likely. It is premature however to completely rule out a molecular switch involving G1405.

CONCLUSION

Recently determined sequences have allowed the detection of additional candidate tertiary interactions in 16S-like rRNAs and have clarified the status of interactions that were detected earlier. All of these interactions follow Watson–Crick pairing rules and hence would be expected to be either regular Watson–Crick pairs or reverse Watson–Crick pairs. It remains possible, indeed likely, that very conserved positions that have yet to exhibit significant sequence variation in the existing data set may conceal other interactions of this type. For example, G505/C526 is conserved in all published data except for wobble pairing in *Euglena gracilis* and *Chlamydia psittaci*, and a C/G pair in *Chlamydomonas reinhardtii* mitochondria. In addition, in less conserved regions of the molecule, the data set as presently constituted may be insufficient to find all significant interactions. This is because structural features in these regions may only be defined in phylogenetically coherent groups, *eg.*, the eubacteria or a subgroup such as the Gram positive bacteria. If this is the case, a large number of sequences from just the relevant phylogenetic subgroup must be available and analyzed. In a sense the A673/U717 interaction is of this type as it seems to be lost in mitochondrial sequences. If tRNA is a good indicator, however, the vast majority of base–base tertiary interactions will not follow the Watson–Crick pairing rules. An examination of tRNA sequences reveals that these interactions, when variable, typically have significantly less fidelity in their variation pattern and hence promise to be considerably more difficult to detect (*28*).

*To whom correspondence should be addressed

REFERENCES

1. All position numbers cited in the text refer to the *E. coli* 16S rRNA sequence.
2. Holley R.W. Apgar, J., Everett, G.A., Madison, J.T., Marquise, M. Merrill, S.H., Penswick, J.R., Zamir, A. (1965) *Science* **147** 1462–1465.
3. Fox, G.E. and Woese C.R. (1975) *Nature* **256** 505–507.
4. Nishikawa, K. and Takemura, S. (1974) *FEBS Lett.*. **40**, 106–109.
5. Noller,H.F. and Woese, C.R. (1981) *Science* **212**, 403–411.
6. Noller, H.F., Kop, J.A., Wheaton, V., Brosius, J., Gutell, R.R., Kopylov, A.M., Dohme, F., Herr, W., (1981) *Nucl.Acids Res.* **9**, 6167–6189.
7. Mount,S.M. and Steitz, J.A. (1981) *Nucl. Acids Res.* **9**, 6351.
8. Brow.D.A. and Guthrie, C. (1988) *Nature* **334**, 213–218.
9. Zucker-Aprison, E., Thomas, J.D., Blumenthal, T. (1988) *Nucl. Acids Res.* **16**, 7188.
10. James, B.D., Olsen, G.J.,Liu, J., Pace, N.R. (1988) *Cell* **52**, 19–26.
11. M. Levitt, (1969) *Nature* **224**, 759–763.
12. Olsen, G.J. (1984)*Comparative analysis of nucleotide sequence data*, Ph.D. Dissertation, University of Colorado Health Sciences Center, pp. 88–137.
13. Haselman, T. Chappelear, J.E. Fox, G.E. (1988)*Nucl. Acids Res.* **16**, 5673-5684.

14. Gutell, R.R., Weiser, B., Woese, C.R., Noller, H.F. (1985) *Prog. Nucleic Acid Res. Molec. Biol.* **32**, 155–216.
15. Gutell, R.R., Noller, H.F., Woese, C.R. (1986) *EMBO J.* **5**, 1111–1113.
16. Leffers, H., Kjems, J., Ostergaard, L., Larson, N., Garrett, R.A. (1987) *J. Mol. Biol.* **195**, 43–61.
17. Gutell, R.R. and Fox, G.E. (1988) *Nucl. Acids Res. Supp.* **16**, r175-r269.
18. Rich A. and RajBhandary, U.L. (1976) *Ann. Rev. Biochem.* **45**, 805–860.
19. Clark, B.G.C. in *Transfer RNA*, S.Altman, Ed. (MIT Press, Cambridge, 1978) pp.14–47.
20. Moazed, D., Stern, S., Noller, H.F. (1986)*J. Mol. Biol.* **187**, 399–416.
21. Unless specifically referenced all sequence data was taken from Dams E., Hendriks, L., Van de Peer, Y., Neefs, J., Smits, G., Vandenbempt, I., (1988) *Nucl. Acids Res. Supp.* **16**, r87-r173.
22. Brimacombe, R., Atmadja, J., Stiege, W., Schuler, D., (1988)*J. Mol. Biol.* **199**, 115–136.
23. Boer, P.H. and Gray, M.W. (1988) *Cell* **55**, 399–411 (1988).
24. Trempe, M.R. Ohgi, K. Glitz, D.G. (1982) *J. Biol. Chem.* **257**, 9822-9829.
25. Mochalova, L.V., Shatsky, I.N., Bogdanov, A.A., Vasiliev, V.D. (1982) *J. Mol. Biol.* **159**, 637–650.
26. Rossau, R., Heyndrickx, L., Heuverswyn, H.V. (1988) *Nucl. Acid Res.* **16**, 6227.
27. Dubin, D.T.and HsuChen, C.C. (1983) *Plasmid* **9**, 307–320.
28. Tim Haselman was supported by U.S. Army Fellowship Grant DAAL03–86–6-0031. Additional support was provided by NIH Grant GM-37655 and NASA Grant NSG-7440 to G.E.F.

This article, submitted on disc, has been automatically converted into this typeset format by the publisher.