RNA

Identity and geometry of a base triple in 16S rRNA determined by comparative sequence analysis and molecular modeling

P. Babin, M. Dolan, P. Wollenzien and R. R. Gutell

RNA 1999 5: 1430-1439

References

Article cited in:

<http://www.rnajournal.org/cgi/content/abstract/5/11/1430#otherarticles>

Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#)

Notes

To subscribe to *RNA* go to:
<http://www.rnajournal.org/subscriptions/>

Identity and geometry of a base triple in 16S rRNA determined by comparative sequence analysis and molecular modeling

PATRICIA BABIN,¹ MICHAEL DOLAN,¹ PAUL WOLLENZIEN,¹ and ROBIN R. GUTELL²

¹Department of Biochemistry, North Carolina State University, Raleigh, North Carolina 27695-7622, USA

²Institute for Cellular and Molecular Biology and School of Biological Sciences, University of Texas, Austin, Texas 78712-1095, USA

ABSTRACT

Comparative sequence analysis complements experimental methods for the determination of RNA three-dimensional structure. This approach is based on the concept that different sequences within the same gene family form similar higher-order structures. The large number of rRNA sequences with sufficient variation, along with improved covariation algorithms, are providing us with the opportunity to identify new base triples in 16S rRNA. The three-dimensional conformations for one of our strongest candidates involving U121 (C124:G237) and/or U121 (U125:A236) (*Escherichia coli* sequence and numbering) are analyzed here with different molecular modeling tools. Molecular modeling shows that U121 interacts with C124 in the U121 (C124:G237) base triple. This arrangement maintains isomorphous structures for the three most frequent sequence motifs (approximately 93% of known bacterial and archaeal sequences), is consistent with chemical reactivity of U121 in *E. coli* ribosomes, and is geometrically favorable. Further, the restricted set of observed canonical (GU, AU, GC) base-pair types at positions 124:237 and 125:236 is consistent with the fact that the canonical base-pair sets (for both base pairs) that are not observed in nature prevent the formation of the 121(124:237) base triple. The analysis described here serves as a general scheme for the prediction of specific secondary and tertiary structure base pairing where there is a network of correlated base changes.

Keywords: 16S rRNA; base triple; comparative sequence analysis; molecular modeling

INTRODUCTION

Current approaches to determine the three-dimensional structure for large RNA molecules and ribonucleoproteins include electron microscopy (see Penczek et al., 1994) and low-angle scattering experiments (Svergun et al., 1997) to directly investigate global structures. Structure determination of large RNA molecules additionally is being guided by the approaches taken to determine the high-resolution structures of smaller RNA molecules. This list includes most notably high-resolution crystal structures of several tRNAs (Kim et al., 1973; see Basavappa & Sigler, 1991), crystal structures of the P4–P6 domain (Cate et al., 1996) and catalytic domain (Golden et al., 1998) of the *Tetrahymena* Group I intron ribozyme, the crystal structure of the hammerhead ribozyme (Pley et al., 1994), the crystal

structure of the loop E region of the 5S rRNA (Correll et al., 1997), the crystal structure of the Hepatitis Delta Virus ribozyme (Ferré-D'Amaré et al., 1998), and NMR-derived structures for a number of synthetic RNA aptamers (see Uhlenbeck et al., 1997).

Crystals of ribosomes and ribosomal subunits with good diffraction properties have been available for a number of years (van Bohlen et al., 1991), but it has been difficult, because of the size and asymmetry of the ribosome and the need to obtain isomorphous heavy atom derivatives, to determine the diffraction phasing necessary for electron-density maps and three-dimensional structural determination. The phasing problem has been solved recently for the *Haloarcula marismortui* 50S ribosomal subunit (Ban et al., 1998) leading to a 9-Å electron density map of the structure. This should provide critical information for the organization of the whole subunit including the active site and its supporting structures when the fold of the rRNA in that structure is determined. So far, it has been necessary to infer details of rRNA structure from constraints determined by different experiments and to incorporate this

Reprint requests to: Paul Wollenzien, Department of Biochemistry, Box 7622, North Carolina State University, Raleigh, North Carolina 27695-7622, USA; e-mail: wollenz@bchserver.bch.ncsu.edu.

information into comprehensive models of the structure. Chemical probing experiments to determine RNA reactivity patterns, the determination of structures of isolated rRNA regions by NMR and the determination of three-dimensional arrangement by RNA cross-linking are among the types of experiments that have been employed in this approach (see Green & Noller, 1997). It is anticipated that RNA structure motifs recognized in smaller RNAs can be applied to the larger rRNA structure based on chemical reactivity and patterns of sequence variance (Brion & Westhof, 1997; Leontis & Westhof, 1998).

Base triples, a prominent tertiary structure motif, are important components in the tertiary structures of tRNAs and the group I introns. Several years ago, Gautheret et al. (1995) presented improvements in the prediction of base triples with comparative sequence analysis. This algorithm identifies the best covariations between all unpaired nucleotides and known base pairs. We have used this method to predict several base triples in 16S and 23S rRNA (Gutell, 1996; R.R. Gutell, unpubl. data). Here we present one of the best base triple candidates in 16S rRNA involving nt 121 and bp 124:237 and/or 125:236 (*Escherichia coli* numbering). Chemical probing data and molecular modeling are used to help ascertain which of the proposed base-triple interactions occurs in the RNA, and the conformation of these interactions.

RESULTS

Identification of base-triple candidates by comparative sequence analysis

The 122–128/233–239 16S rRNA helix first was proposed with comparative analysis based on the small number of sequences available in 1980 (Woese et al., 1980). (We will refer to this interaction as helix 122 here). With a significant increase in the number of 16S rRNA sequences and the development of more powerful covariation algorithms (Gutell et al., 1992; Maidak et al., 1999; R.R. Gutell, S. Subrashchandran, M. Schnare, Y. Du, N. Lin, L. Madabusi, K. Muller, N. Pande, N. Yu, Z. Shang, S. Date, D. Konings, V. Schweiker, B. Weiser, & J. Cannone, in prep.), all of the positions within the 122–128/233–239 helix have their strongest statistically significant covariation with their previously predicted base-pair partner, except for the 125:236 base pair, where position 125 is nearly always a U in approximately 5,000 prokaryotic sequences, and position 236 is G in 75% and A in 25% of the sequences, (forming UG and UA base pairs).

With a smaller set of rRNA sequences, our earlier analysis revealed several strong base-triple candidates in 16S and 23S rRNA (Gutell, 1996), including 1072(1092:1099) in 23S rRNA, and 121(124:237)/(125:236) in 16S rRNA. Recently the putative 23S rRNA

base triple [1072(1092:1099)] has been substantiated with experimental methods (Conn et al., 1998, 1999). We have repeated this base-triple analysis utilizing the same methods, but applied them to a larger prokaryotic 16S rRNA alignment (Maidak et al., 1999). Several strong candidates have been identified, including 595(596:644) and our previous 121(124:237)/(125:236) base triple. There is a mutual best covariation between the unpaired position 595 and the base pair 596:644 with all three algorithms used to evaluate the significance of sequence covariations—chi square, pseudo-phylogenetic event counting (ec), and covary methods (see Materials and Methods). The phylogenetic distribution reveals that alternate versions of the triplets have evolved independently several times, thus increasing the likelihood that this comparatively inferred base triple is real. Nuclear magnetic resonance analysis of this region of the 16S rRNA (Kalurachchi & Nikonowicz, 1998) revealed a base triple at these positions. The experimental support for the 16S [595(596:644)] and 23S [1072(1092:1099)] rRNA interactions lends credibility to the comparative methods employed here.

Our base-triple covariation analysis was performed on the most recent (July 1998, version 7.0) release of the Ribosomal Database Project (RDP) prokaryotic alignment of 16S rRNA (6205 sequences, including complete and partial sequences) (<http://www.cme.msu.edu/RDP/download.html>; Maidak et al., 1999). The structure in the 120–130/230–240 region of the 16S rRNA is conserved in all prokaryotes and chloroplasts. This same region is slightly different in the Eucarya nuclear and mitochondrial 16S-like rRNAs. Because of this difference in structural homology, only the prokaryotic and chloroplast 16S rRNA sequence sets are analyzed here.

The covariation signal between position 121 and the base pairs 124:237 and 125:236 is very strong. The coordinated variation involves five positions. The chi square and ec methods identify mutual best covariations between position 121 and the base pair 125:236, whereas the newer covary method identifies a mutual best covariation between 121 and 124:237. Approximately 5,000 sequences in the prokaryotic RDP alignment had nucleotide information at positions 121 and 124:237 and 125:236. The dominant triplet sequences at 121(124:237) are C(GC) [74%], U(AU) [10%], and U(CG) [7%]. At 121(125:236) the dominant sequences are C(UG) [73%] and U(UA) [21%] (Table 1A,B).

The nucleotide at position 121 determines the base pairs and their arrangement at 124:237 and 125:236. Note that position 121 is a pyrimidine in 98% of the prokaryotic 16S rRNA sequences, with C occurring in 76% of these sequences, U in 22%, and G and A each with 1% (Table 1A). When position 121 is a C, then 124:237 is a GC pair in 98% of the sequences, and 125:236 is a UG in 97% of the sequences (Table 1B). When position 121 is a U, then 124:237 is an AU (45%), CG (30%), UA (15%), or GC (9%), and 125:236 is a UA

TABLE 1. Nucleotide frequencies at positions 121(124:237) (A) and 121(125:236) (B).

A ^a	121	A	C	G	U	
		— ^b	—	—	10	AU
		—	—	—	7	CG
		—	74	—	2	GC
		—	—	—	3	UA
						124:237
B ^c	121	A	C	G	U	
		—	2	—	21	UA
		—	73	—	—	UG
						125:236

^aEntries are percentages of 5,056 sequences with nucleotide information at positions 121(124:237).

^bAll percentages less than 1.5 are shown with a dash.

^cEntries are percentages of 4,939 sequences with nucleotide information at positions 121(125:236).

(94%), or UG (5%). Taken all together, four sequence motifs occur for these sequences (Table 2): when position 121 is a C, the 124:237 and 125:236 base pairs are predominantly GC/UG (76%, motif A), and when 121 is a U, then the 124:237 and 125:236 base pairs are predominantly AU/UA (10%, motif B), followed by CG/UA (7%, motif C) or UA/UA (3%, motif D).

Comparative analysis should reveal, in addition to the nucleotide frequencies for the positions of interest, an approximate number of events or times that the nucleotides have changed coordinately during the evolution of these rRNAs (Gutell et al. 1986). These “phylogenetic events” are a gauge for the authenticity of every base-pair and base-triple interaction predicted with comparative methods. Our confidence for each proposed interaction is proportional to the number of times a mutual change (covariation) occurs in the phylogenetic tree.

The four most frequent sequence combinations at positions 121(124:237)/(125:236) were mapped onto the RDP's prokaryotic phylogenetic tree (Maidak et al. 1999), providing us with the number of occurrences for

TABLE 2. Distribution (in percentage of 5,000 prokaryotic sequences analyzed) of nucleotide identity at positions 121 (124:237)/(125:236).

Sequence positions			Motif Name
121	(124:237)/(125:236)		
C	GC/UG	76%	A
U	AU/UA	10%	B
U	CG/UA	7%	C
U	UA/UA	3%	D
C	AU/UA	1%	E
U	GC/UA	1%	F
U	GC/UG	1%	G
C	GC/UA	1%	H

each motif in each phylogenetic group (Table 3). For this analysis, we mapped the sequence sets onto a reduced phylogenetic tree that only contained the primary branches in the Archaea (i.e., Crenarchaeota, Euryarchaeota) and the (eu)Bacteria (e.g., Cyanobacteria, Spirochetes, and Proteobacteria). The Purple Bacteria (Proteobacteria) and Gram Positive branches were expanded an additional level (i.e., Alpha subdivision). There are a few key observations from this analysis. First, the most frequent motif—(A), c/gc/ug (at 76%)—occurs in all but two of the major prokaryotic phylogenetic groups. Second, the second most abundant motif—(B), u/au/ua (at 10%)—occurs in the two remaining major phylo-

TABLE 3. Distribution of sequence motifs A–D in bacteria and Archaea.

Sequence motif				Phylogenetic group
A	B	C	D	
				Archaea
41				Crenarchaeota
136				Euryarchaeota
				Bacteria
150	3			Cyanobacteria and chloroplasts
25				Fibrobacter phylum
213				Flexibacter-Cytophaga-Bacteroides phylum
3				Fls.sinusarabici assemblage
33				Fusobacteria and relatives
				Gram-positive phylum
15				Anaerobic Halophiles
395				Bacillus-Lactobacillus-Streptococcus subdivision
22				C. Lituseburensis group
32				C. Purinolyticum group
160				Clostridium and relatives
75				Eubacterium and relatives
803				High G+C subdivision
179	8			Mycoplasma and relatives
44				Sporomusa and relatives
37				Thermoanaerobacter and relatives
41				Green non-sulfur bacteria and relatives
2	4			Green sulfur bacteria
5				Nitrospina subdivision
7	6			Paraphyletic assemblage
17	10	6		Planctomyces and relatives
				Purple bacteria
742	53			Alpha subdivision
	247			Beta subdivision
102				Delta subdivision
2				Env.sar121 group
	70			Epsilon subdivision
149	77	320	160	Gamma subdivision
11				Uncultured magnetotactic bacteria
151	8			Spirochetes and relatives
1				Thermophilic assemblage
5				Thermophilic oxygen reducers
7				Thermotogales

The distribution for the penta-sequence 121/124:237/125:236 is subdivided by phylogenetic group. The four most abundant penta-sequences are shown where A = c/gc/ug; B = u/au/ua; C = u/cg/ua; D = u/ua/ua.

TABLE 4. Summary of structures predicted by ISOPAIR-MCSYM analysis.

Motif	Hydrogen bonding pattern ^a	MC-SYM transformations ^b	Hydrogen bonds formed	Improper bond lengths after minimization?	Most isomorphic structure?
A	121–124	CG_31	C121(N4)–G124(N7)		Yes
	121–124	CG_32	C121(N4)–G124(N7)		
	121–236	CG_32	C121(N4)–G236(N7)	N4–N7 H bond distance	
B	121–124	AU_50	U121(O4)–A124(N6)		Yes
	121–124	AU_52	U121(O4)–A124(N6)		
	121–236	AU_50	U121(O4)–A236(N6)	O4–N6 H bond distance	
C	121–124	CU_101	U121(O4)–C124(N4)		Yes
	121–124	CU_104	U121(O4)–C124(N4)		
	121–236	AU_50	U121(O4)–A236(N6)	O4–N6 H bond distance	
<i>E. coli</i>	121–124	CU_101	U121(O4)–C124(N4)		Yes
	121–124	CU_104	U121(O4)–C124(N4)		
	121–236	AU_50	U121(O4)–A236(N6)	O4–N6 H bond distance	

^aHydrogen-bonding patterns are listed for the base-triple interactions which were consistent with molecular models.

^bThe "Transformation" designations were obtained in MC-SYM version 1.3 (see WWW.IRO.UMONTREAL.CA/~MAJOR/HTML/USERGUIDE.HTML) for the base pairs containing one-hydrogen-bond interactions between the indicated bases.

genetic groups. Third, motifs C and D do not occur exclusively in any of these primary phylogenetic groups; instead they occur in two groups (Bacteria-Planctomyces and Bacteria-Purple Bacteria-Gamma subdivision) that have other motifs (A and B and A, B, and C). Fourth, 23 of the phylogenetic groups have only one motif, six phylogenetic groups have two motifs, one group (Bacteria-Planctomyces and relatives) has three motifs, and one group (Bacteria-Purple Bacteria-Gamma subdivision) has all four motifs. Last, the number of phylogenetic events (times that each of these motifs evolved) is estimated at approximately 10. Here we assume that the most abundant motif—(A), c/gc/ug—is primordial, and thus any motif other than this one has evolved [events have occurred in the Bacteria-Cyanobacteria and Chloroplasts, Bacteria-Fibrobacter, Bacteria-Flexibacter, Bacteria-Gram Positive-Phylum Mycoplasma, Bacteria-Green Sulfur Bacteria, etc. (see Table 3)]. Thus there are 13 such events. However, because some of the primary phylogenetic groups are not necessarily monophyletic sister groups and the actual number of unrelated groups is subject to disagreement, we estimate the number of phylogenetic changes at these five positions to be no less than five and approximately 10.

The RDP Prokaryotic/Chloroplast 16S rRNA alignment was then analyzed to determine the most frequently occurring nucleotide sequences in the (122–129)·(232–239) helix for motifs A, B, and C (Fig. 1). The most frequently occurring sequence for each motif was then used for the molecular modeling experiments. In addition, the *E. coli* sequence (which contains motif C sequence at the base triple) was also used for modeling experiments, as chemical reactivity data was available for it.

Isomorphism and molecular modeling in different sequence motifs

Molecular models for the helix 122 region were constructed using the constraint-satisfaction program MC-SYM that constructs models using nucleotide units with geometries extracted from known structures (Major et al., 1991, 1993; Gautheret et al., 1993). This was done to determine which base-triple interactions could be incorporated into molecular models of the region. In the first exercise, the three most frequent sequence motifs, A, B, and C, as well as the sequence of *E. coli* (Fig. 1) were used for modeling.

MC-SYM scripts to generate models for the sequences shown in Figure 1, A–D, were written with three assumptions. First, all of the base pairs in helix 122 contain normal Watson–Crick or wobble base pairs. Second, the base triples for all the motifs should be isomorphic. Third, the uridine at position 121 in *E. coli* has chemical reactivity (Moazed et al., 1986; P. Wolenzien, unpubl.), indicating that the N3 position of U 121 is not used in the hydrogen bonding in the *E. coli* sequence.

The ISOPAIR program (Gautheret & Gutell, 1997) was used first to determine which conformations would produce isomorphic base triples for the sequences in motifs A, B, C, and *E. coli* (Table 4). All four of the nucleotides in the base pairs 124:237 and 125:236 were considered as possible hydrogen bonding partners for 121 in the ISOPAIR analysis.

The ISOPAIR analysis for a triple occurring between nt 121 and the bp 124:237 revealed that all predicted triples for these nucleotides would occur in the major groove by means of a hydrogen bond between nt 121 and 124. Triples with a hydrogen bond between nt 121

MOST COMMON SEQUENCE FOR MOTIF A	MOST COMMON SEQUENCE FOR MOTIF B	MOST COMMON SEQUENCE FOR MOTIF C	SEQUENCE FOR E. COLI
C 121	U 121	U 121	U 121
239 U - G 122	239 U - A 122	239 U - G 122	239 U - G 122
238 G - C 123	238 G - C 123	238 G - C 123	238 A - U 123
237 C - G 124	237 U - A 124	237 G - C 124	237 G - C 124
236 G - U 125	236 A - U 125	236 A - U 125	236 A - U 125
235 C - G 126	235 G - C 126	235 U - A 126	235 C - G 126
234 C - G 127	234 C - G 127	234 C - G 127	234 C - G 127
A	B	C	D
HYBRID 1 SEQUENCE	HYBRID 2 SEQUENCE	HYBRID 3 SEQUENCE	HYBRID 4 SEQUENCE
C 121	C 121	C 121	C 121
239 U - G 122	239 U - G 122	239 U - G 122	239 U - G 122
238 G - C 123	238 G - C 123	238 G - C 123	238 G - C 123
237 C - G 124	237 C - G 124	237 G - C 124	237 A - U 124
236 C - G 125	236 U - A 125	236 G - U 125	236 G - U 125
235 C - G 126	235 C - G 126	235 C - G 126	235 C - G 126
234 C - G 127	234 C - G 127	234 C - G 127	234 C - G 127
E	F	G	H

FIGURE 1. Sequences for helix 122 region used for modeling. **A–D:** The most common sequence patterns for the three major motifs and *E. coli*. **A:** Motif A (121 = C, 124:237 = G:C, and 125:236 = U:G); **B:** Motif B (121 = U, 124:237 = A:U, and 125:236 = U:A); **C:** Motif C (121 = U, 124:237 = C:G, and 125:236 = U:A); **D:** The sequence for *E. coli* (Brosius et al., 1981), the second most common pattern for motif C. **E–H** contain "hybrid" sequences designed to examine the cause of the strong neighbor effect between the adjacent base pairs 124:237 and 125:236. **E:** Sequence motif A with the 125:236 = U:G base pair substituted with G:C. **F:** Sequence motif A with substitution of 125:236 = A:U. **G:** Sequence motif A with the 124:235 = G:C base pair substituted with C:G. **H:** Sequence motif A with the 124:235 base pair substituted with U:A.

and 237 were eliminated because they would have involved use of the N3 position of nt 121. For motif A, the two isomorphic hydrogen-bonding patterns that were predicted were C-121-N4 to G-124-N7 and C-121-N4 to G-124-O6. The first hydrogen-bonding pattern has the highest degree of isomorphism to the other motifs. For motifs B and C, the hydrogen-bonding patterns, which are isomorphic to the C-121-N4 to G-124-N7 pattern, are, respectively U-121-O4 to A-124-N6 and U-121-O4 to C-124-N4.

The ISOPAIR analysis for a triple occurring between nt 121 and the bp 125:236 revealed that all predicted triples for these nucleotides would occur in the major groove via a hydrogen bond between 121 and 236. Some isomorphic triples were eliminated because they would have involved use of the N3 position of nt 121 in *E. coli*, including one predicted triple between 121 and 125. For motif A, one hydrogen-bonding pattern was predicted for an interaction between 121 and 236: C-121-N4 to G-236-N7. For motifs B and C, the hydrogen-bonding pattern, which is isomorphic to the C-121-N4 to G-236-N7 pattern, is U-121-O4 to A-236-N6.

Even though ISOPAIR only predicted triples occurring between positions 121 and 124 or 121 and 236,

additional MC-SYM scripts were written to attempt to create triples utilizing a hydrogen bond between 121–125 or 121–237. Models were generated for a triple formed by a hydrogen bond between 121 and 125 for motif A, but isomorphic triples for the other motifs could not be identified. It was not possible to generate models for a triple formed by a hydrogen bond between 121 and 237. Additionally, even though no isomorphic base triples were predicted to occur in the minor groove, MC-SYM scripts were written to attempt to generate a hydrogen bond between 121 and each of the 4 nt in their minor groove. Even large amounts of conformational freedom were not sufficient to allow models to be generated in which the hydrogen bond forming the base triple occurred in the minor groove.

Therefore, the only models to be considered were those that contained a base triple formed by a hydrogen bond between 121 and 124 (C-121-N4 to G-124-N7 for motif A) or 121 and 236 (C-121-N4 to G-236-N7 for motif A). The models from MC-SYM that were most consistent with A-form RNA geometry were then subjected to energy minimization. Only the model that utilized a hydrogen bond between 121 and 124 (Fig. 2A) could be minimized to acceptable O3'-P bond lengths

FIGURE 2. Models for the 121(124:237) triple and the 121(125:236) triple in motif A. Both of the molecular models are shown after subjecting them to energy minimization. **A:** Model for the 121(124:237) triple containing a hydrogen bond between the N4 of nt 121 and the N7 of nt 124. **B:** Model for the 121(125:236) triple containing a hydrogen bond between the N4 of nt 121 and the N7 of nt 236. Note that a suitable hydrogen-bond length cannot be attained in the model containing the 121(125:236) triple. The measurements indicated in both panels are between the hydrogen of the hydrogen-bond donor and the heavy atom hydrogen-bond acceptor. Distances between heavy atoms in both panels also indicate an acceptable hydrogen-bond distance (Saenger, 1984) in the structure of **A** but not in **B**.

and hydrogen-bond lengths (Saenger, 1984). The model containing a base triple formed by a hydrogen bond between 121 and 236 could not be minimized to an acceptable hydrogen bond length between 121 and 236 and maintain an acceptable O3'-P bond length between 121 and 122 (Fig. 2B). This analysis was repeated with the same results for all motifs. The models for all motifs for helix 122 using this base triple were superimposed using the backbone and ribose atoms for alignment (Fig. 3). The root mean square (RMS) deviations between models are listed in Table 5.

ISOPAIR was used to analyze the possibility of isogeometric triples in helix 122 for all motifs, A–H, shown in

TABLE 5. RMS deviations of models containing base triples.

Motif	Deviation from ideal A-form RNA structure of same sequence ^a	Deviation of entire model from motif A model ^b	Deviation of base triple from motif A base triple ^b
A	0.74 Å	—	—
B	0.40 Å	1.32 Å	1.29 Å
C	0.61 Å	1.77 Å	1.93 Å
<i>E. coli</i>	0.61 Å	1.80 Å	1.94 Å
D ^c	—	—	—
E	2.72 Å	1.79 Å	2.02 Å

^aRMS deviation obtained for all atoms.

^bRMS deviation obtained for backbone and ribose atoms.

^cIt was not possible to predict structures for motif D that would contain acceptable covalent and hydrogen bond lengths.

Table 2. Isomorphic triples for all of the motifs except motif D could be found. The conformations shown in Figure 2 were predicted for either comparison of the ABC or ABCEFGH motifs. Motif E was modeled as an example of one of the less frequent motifs. An isomorphic structure for the region utilizing a hydrogen bond between 121 and 124 was obtained and the structure containing it could be refined by energy minimization (data not shown). Comparison of the structure for motif E versus motif A indicates a close similarity in the structure of the base triple, but a higher degree of differences in the overall structure (Table 5).

Sequence bias at the base pair adjacent to the base triple interaction

The strong correlation between the sequences at 124:237 and 125:236 was investigated to determine if

FIGURE 3. Superposition of the models for the proposed helix 122 with the base triple 121(125:236). **A:** Superposition of models of helix 122 for motifs A (red), B (green), C (blue), and *E. coli* (yellow). Alignment was done using the backbone and ribose atoms. The atoms are shown only for motif A; the backbones are shown for all four models. **B:** Superposition of the base triple for motifs A, B, C, and *E. coli*. The geometry for the triple of the C (blue) and *E. coli* motifs are nearly identical and share the same molecular structure in this figure. See Table 5 for RMS deviation of the models from one another.

there was a geometrical connection between the bias and the occurrence of the base triple at 121(124:237). When 121 is a C and 124:237 are GC, there are no occurrences of GC or AU in prokaryotes at the base pair 125:236. Hybrid sequences 1 and 2 were created to investigate this bias. In these hybrids, the only change made to the most common sequence for motif A was the substitution of GC for UG at base pair 125:236 (Fig. 1E) and the substitution of AU for UG at base pair 125:236 (Fig. 1F). For hybrid 1 and 2 sequences, models could not be generated for base triples incorporating a hydrogen bond between nt 121 and 124 (results not shown).

Certain base-pair types are also absent at the 124:237 positions. When 121 is a C and 125:236 are UG, there are no CG or UA base pairs at 124:237. To investigate this bias, additional hybrid models, hybrid 3 and hybrid 4 were created. Hybrid 3 (Fig. 1G) is the most common sequence for motif A with base pair 124:237 changed from GC to CG. Hybrid 4 (Fig. 1H) contains a UA base pair at base 124:237. When these hybrid sequences were modeled, it was possible to generate models for a base triple incorporating a hydrogen bond between nt 121 and 124 (Fig. 4). However, these models could not be minimized to acceptable O3'-P and hydrogen-bond lengths (Saenger, 1984). Specifically, the O3'-P bond length between nt 121 and 122 was incompatible with an acceptable bond length for the hydrogen bond between 121 and 124 to form the base triple. Thus, the strong-neighbor effect occurring between base pairs 124:237 and 125:236 can be explained by the requirement for a geometric arrangement needed to allow the formation of the triple at 121(124:237).

FIGURE 4. Models demonstrating the cause of the strong neighbor effect. Both examples shown here have been refined from MC-SYM structures by energy minimization and contain proper backbone bond lengths, anti-base conformations and 3'-endo ribose conformations. **A:** Model for the base triple utilizing a hydrogen bond between C121 and C124 in hybrid 3. The distance between the hydrogen of N4 (C121) and the O2 (C124) exceeds the maximum acceptable length of 2.17 Å. **B:** Model for the base triple utilizing a hydrogen bond between C121 and U124 in hybrid 4. The distance between the O4 (U124) and the hydrogen of N4 (C121) exceeds the maximum acceptable length of 2.17 Å. Hydrogen-bond distances measured between heavy atoms also indicate unacceptably long distances in both cases (Saenger, 1984).

DISCUSSION

Base triples are inherently more difficult to predict than base pairs (see Gautheret et al., 1995). Although the majority of the secondary structure base pairs are conserved in all members of a given RNA type (e.g., tRNA), this is not the situation for base triples. Several base triples in tRNA and group I introns (Michel et al., 1990; Michel & Westhof, 1990) are present in only a subset of their structures (e.g., in the type-2 tRNAs, or in the C1-2 subgroup of group I introns). Second, although similar (base pairing) conformations are only maintained with positional covariations, similar conformations in base triples can be maintained with single, unmatched positional variation (Klug et al., 1974).

The 121(124:237)/(125:236) putative base triple was identified by three comparative sequence analysis approaches. The identification of the base-triple interaction within this sequence was investigated further with ISOPAIR, MC-SYM, and energy minimization to determine the potential for having isomorphic structures and to determine the possibility of forming molecular models. In the present case, it was possible to take advantage of chemical reactivity data that was relevant to the conformations of U121, as U121 was reactive with CMCT (Moazed et al., 1986). The MC-SYM exercises resulted in models in which U121 was hydrogen bonded with either C124 or A236. However, MC-SYM models typically need to be constructed initially with long O3'-P bond lengths, and when energy minimization was performed to determine which models could be refined to acceptable bond lengths, the model with a U121-to-A124 hydrogen bond was the only one in which that could be done.

The interaction that is predicted here involves an extra hydrogen bond between 16S rRNA nt 121 and 124 in the major groove of the helical region. This is a different type of interaction than the recent examples of same-strand near-neighbor interactions in the group I ribozyme (Cate et al., 1996) and in hepatitis delta virus ribozyme (Ferré-D'Amaré et al., 1998) that occur in the minor groove of the helix. However, the base triple interaction in the 16S rRNA at 595(596:644) has been shown to occur in the major groove (Kalurachchi & Nikonowicz, 1998) with the hydrogen bond occurring between nt 595 and 596. There are also several examples in tRNA structures of base triple interactions that involve major groove interactions. Furthermore, we are confident that the predictive modeling program MC-SYM has the capability of constructing minor groove interactions, because another base triple in 16S rRNA between nt 494(440:497) is predicted to contain its extra interaction in the minor groove. Thus, in spite of the notoriety of the narrowness of the RNA major groove in regular helices, there is no clear rule about which groove is used in these types of interactions.

By modeling “hybrid” sequences that utilized these nonoccurring base pairings, we demonstrated that a base triple in which nt 121 is hydrogen bonded to 124 is consistent with the coordinated set of base pairings at 124:237 and 125:236. We conclude that specific base pairings at 124:237 (e.g., CG and UA when 121 is U) and 125:236 (e.g., CG and AU when 121 is U or C) were not allowed due to structural constraints at 121(124:237). The neighbor effect has been widely observed in tRNAs (Gautheret et al. 1995), in Group I introns (Michel & Westhof, 1990), and in 16S and 23S rRNA (R. Gutell, unpubl. data); the work presented here is a demonstration that it is based at least partly on structural constraints.

Finally, by modeling the (122–129)·(232–239) helix with MC-SYM, we demonstrated that the conformations chosen for the triple 121(124:237) in molecules containing the A, B, and C sequence motifs are part of models for the entire helical region that are isomorphic in spite of divergent sequences at many positions. There are two types of exception to this conclusion that occur with the region containing the minor sequence motifs. The first is that we have not been able to determine isomorphic structures for motif D (U(UA/UA)) that occurs in 3% of the sequences. Furthermore the sequence motifs E–H that occur in approximately 1% of our prokaryotic data set can be modeled into base-triple geometries isomorphic with the major motifs A, B, and C. However, the formation of these base triples requires some distortion in the positions of the riboses and bases compared to the normal type-A geometry. The exceptions in motifs D–H occur in a small fraction (approximately 7% total) of the total number of sequences. Thus the idea of strict isomorphism at this region extends to about 93% of the prokaryotic 16S rRNA data set.

In the absence of a base triple in this region, nt 121 would most likely be stacked on nt 122 because of the favorable stacking interactions. However the presence of the base triple involving 121 causes a repositioning and redirection of the phosphate backbone in the region of 121. This may affect the interaction that helix 122 has with adjacent helix (240–242)/(284–286) as well as the trajectory of the single-stranded region 116–121 in the ribosomal subunit.

MATERIALS AND METHODS

Comparative base-triple analysis

Two comparative methods (Gautheret et al., 1995) that identify base triples were used to search for covariations between the known secondary-structure base pairs and the unpaired positions. Method one (chi) calculates the expected and observed frequencies of base triples and their chi-square values for all possible base pair and unpaired nucleotide combinations. The base triple candidates with the highest chi-

squared values are considered possible. Method two (ec and sec) calculates values for the number of pseudophylogenetic events. Because the sequences in our alignments are arranged in phylogenetic order, the sequences that are most closely related are adjacent to one another. The number of coordinated base changes that have occurred throughout the evolution of the RNA under study can be approximated by counting the number of mutual events (or simultaneous changes) in the two (or three) columns (positions) in the alignment. This number is divided by the total number of changes that have occurred at the positions under study. Thus the maximum score of 1 denotes that all of the positional changes are involved in mutual events, while a score of 0.5 signifies that only half of the changes are associated with a mutual event.

More recently a third base-triple covariation algorithm called “covary” has been established (R. Gutell, J. Cannone, V. Schweitzer, unpubl. program; Gutell et al., in prep.). This newest method sums the frequencies of those triples that covary from the most frequent triplet (counting the most frequent triplet). We only consider those triplets where the percentage of pure covariation (covariation without exceptions) from the most frequent triplet is greater than 35% of those triplets that vary from the most frequent triplet; we filter out those triplets with less than 35% covariation. For example, for the (124:237)121 base triple, (GC)C = 74%, (AU)U = 10%, (CG)U = 7%, (UA)U = 3%, (GC)U = 2%, other = 4%. Here the total covariation value = .84, as the only pure triplet covariations are (GC)C and (AU)U, and the filter value is 38% (10/26, where 10 = %(AU)U, 26 = %(AU)U + (CG)U + (UA)U + (GC)U + others). The base pair that covaries best with the unpaired position 121 is 124:237 with the covary score .84, and the unpaired position that covaries best with the 124:237 base pair is 121, with the same .84 covary score. This method will be formally presented elsewhere (R.R. Gutell, S. Subrashchandran, M. Schnare, Y. Du, N. Lin, L. Madabusi, K. Muller, N. Pande, N. Yu, Z. Shang, S. Date, D. Konings, V. Schweiker, B. Weiser, & J. Cannone, in prep.).

For all of these methods, putative base triples are considered possible when the covariation between the base pair and the unpaired nucleotide is mutual (e.g., base pair X covaries best with unpaired nucleotide Y, and Y covaries best with X). These base-triple “mutual best” covariations are then evaluated for other considerations, including weaker covariations among the base pairs in the vicinity of the base-triple base pair [neighbor effect, see Gautheret et al., (1995)], the number of times the covariation occurred in the evolution of the rRNAs, the statistical significance of the triple covariations, and the exceptions (single and double variations). A putative base triple with a mutual best covariation is considered more probable when there are neighbor effects flanking the base pair (of the base triple), and when the base triple candidate is identified as mutual best with all methods—ec (phylogenetic events), chi (statistical methods), and covary (logical method). The putative base triples U121(C124-G237) and/or U121(U125-A236) are the highest-scoring triplet covariations with these three methods.

ISOPAIR

ISOPAIR (Gautheret & Gutell, 1997) was used to determine the hydrogen-bonding patterns that would produce isomor-

phic base triples across the major motifs and the *E. coli* sequence. Additionally, in preparation for molecular modeling of helix 122 for all the motifs and the *E. coli* sequence, the RDP Prokaryotic/Chloroplast 16S rRNA alignment was then searched for the most frequently occurring pattern of nucleotides in helix 122 for motifs A, B, and C.

Molecular modeling

MC-SYM 1.3 scripts (Major et al., 1991) were written to determine which of the base-paired nucleotides (124–237 or 125–236) were capable of forming a hydrogen bond to the single-stranded nt 121. In all MC-SYM scripts, standard global constraints contained in the sample scripts for the program were used to minimize van der Waals' overlaps in the generated models and all scripts were written to generate models that were as close to A-form RNA as possible. Because the formation of the base triple requires an exceptional geometry for nt 121, the O3'-P bonding distances between adjacent nucleotides needed to be initially set at 6.5 Å to allow for searching of conformational space that would result in the formation of models. This is reasonably consistent with the default value of 6 Å used in the ADJACENCY section of the MC-SYM program (<http://www.iro.umontreal.ca/~major/HTML/mcsym.ug.html>). This is well beyond the normal maximum of 1.62 Å for this bond; however, all the O3'-P bond lengths in the models are easily adjusted to the optimal length of 1.58–1.62 Å by energy minimization. For helix 122, very few models (and thus very few conformations) were generated by MC-SYM for each script and therefore clustering of models into families was not required.

Energy minimization

All structures considered isomorphic by the ISOPAIR analysis at the base triple were minimized using Insight II (Molecular Simulations, Inc.). Global constraints were used in the MC-SYM scripts, so van der Waals' overlaps were minimal and the main deviation of the structures generated by MC-SYM from acceptable nucleotide geometries was the O3'-P distances between adjacent nucleotides. Ninety rounds of steepest-descent minimization followed by ten steps of conjugate gradient minimization using the AMBER force field were used. The geometry of the helical regions of the minimized structures (bp 122:239–127:234) were determined by measuring the RMS deviation of the structures from an ideal type-A RNA helix of the same sequence and the geometry of nt 121 was determined by measuring its dihedral angles and bond lengths.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health (NIH) grants GM48207 to R.R.G. and GM43237 to P.W. Robert Cedergren initially suggested the use of MC-SYM in modeling rRNA regions and we gratefully acknowledge his insight into this problem. Francois Major is thanked for his comments and suggestions in the use of MC-SYM, Daniel Gautheret for making the program ISOPAIR available, Vi Schweitzer, Jamie Cannone, Sankarasubramanian Subashchandran for developmental work on the covariation algorithms.

Received March 24, 1999; returned for revision May 5, 1999; revised manuscript received August 7, 1999

REFERENCES

- Ban N, Freeborn B, Nissen P, Penczek P, Grassucci R, Sweet R, Frank J, Moore P, Steitz T. 1998. A 9 Å resolution X-ray crystallographic map of the large ribosomal subunit. *Cell* 93:1105–1115.
- Basavappa R, Sigler PB. 1991. The 3 Å crystal structure of yeast initiator tRNA: Functional implications in initiator/elongator discrimination. *EMBO J* 10:3105–3111.
- Brion P, Westhof E. 1997. Hierarchy and dynamics of RNA folding. *Annu Rev Biophys Biomol Struct* 26:113–137.
- Brosius J, Dull TJ, Sleeter DD, Noller HF. 1981. Gene organization and primary structure of a ribosomal RNA operon from *Escherichia coli*. *J Mol Biol* 148:107–127.
- Cate JH, Gooding AR, Podell E, Zhou K, Golden BL, Kundrot CE, Cech TR, Doudna JA. 1996. Crystal structure of a group I ribozyme domain: Principles of RNA packing. *Science* 273:1678–1685.
- Conn GL, Draper DE, Lattman EE, Gittis AG. 1999. Crystal structure of a conserved ribosomal protein–RNA complex. *Science* 284:1171–1174.
- Conn GL, Gutell RR, Draper DE. 1998. A functional ribosomal RNA tertiary structure involves a base triple interaction. *Biochemistry* 37:11980–11988.
- Correll CC, Freeborn B, Moore PB, Steitz TA. 1997. Metals, motifs, and recognition in the crystal structure of a 5S rRNA domain. *Cell* 91:705–712.
- Ferré-D'Amaré AR, Zhou K, Doudna JA. 1998. Crystal structure of a hepatitis delta virus ribozyme. *Nature* 395:567–574.
- Gautheret D, Damberg SH, Gutell RR. 1995. Identification of base-triples in RNA using comparative sequence analysis. *J Mol Biol* 248:27–43.
- Gautheret D, Gutell RR. 1997. Inferring the conformation of RNA base pairs and triples from patterns of sequence variation. *Nucleic Acids Res* 25:1559–1564.
- Gautheret D, Major F, Cedergren R. 1993. Modeling the three-dimensional structure of RNA using discrete nucleotide conformational sets. *J Mol Biol* 229:1049–1064.
- Golden BL, Gooding AR, Podell ER, Cech TR. 1998. A preorganized active site in the crystal structure of the *Tetrahymena* ribozyme. *Science* 282:259–264.
- Green R, Noller HF. 1997. Ribosomes and translation. *Annu Rev Biochem* 66:679–716.
- Gutell RR. 1996. Comparative sequence analysis and the structure of 16S and 23S rRNA. In: Zimmermann RA, Dahlberg AE, eds. *Ribosomal RNA: Structure, evolution, processing, and function in protein synthesis*. Boca Raton, Florida: CRC Press. pp 111–128.
- Gutell RR, Noller HF, Woese CR. 1986. Higher order structure in ribosomal RNA. *EMBO J* 5:1111–1113.
- Gutell RR, Power A, Hertz G, Putt E, Stormo G. 1992. Identifying constraints on the higher-order structure of RNA: Continued development and application of comparative sequence analysis methods. *Nucleic Acids Res* 20:5785–5795.
- Kalurachchi K, Nikonowicz EP. 1998. NMR structure determination of the binding site for ribosomal protein S8 from *Escherichia coli* 16S rRNA. *J Mol Biol* 280:639–654.
- Kim S-H, Quigley FL, Suddath FL, McPherson A, Sneden D, Rim JJ, Weinzierl J, Rich A. 1973. Three-dimensional structure of yeast phenylalanine transfer RNA: Folding of the polynucleotide chain. *Science* 179:285–288.
- Klug A, Ladner J, Robertus JD. 1974. The structural geometry of co-ordinated base changes in transfer RNA. *J Mol Biol* 89:511–516.
- Leontis NB, Westhof E. 1998. The 5S rRNA loop E: Chemical probing and phylogenetic data versus crystal structure. *RNA* 4:1134–1153.
- Maidak BL, Cole JR, Parker CT Jr, Garrity GM, Larsen N, Li B, Lilburn TG, McCaughey MJ, Olsen GJ, Overbeek R, Pramanik S, Schmidt TM, Tiedje JM, Woese CR. 1999. A new version of the RDP. *Nucleic Acids Res* 27:171–173.
- Major F, Gautheret D, Cedergren R. 1993. Reproducing the three-dimensional structure of a tRNA molecule from structural constraints. *Proc Natl Acad Sci USA* 90:9408–9412.

- Major F, Turcotte M, Gautheret D, Lapalme G, Fillion E, Cedergren R. 1991. The combination of symbolic and numerical computation for three-dimensional modeling of RNA. *Science* 253:1255–1260.
- Michel F, Ellington AD, Couture S, Szostak JW. 1990. Phylogenetic and genetic evidence for base-triples in the catalytic domain of group I introns. *Nature* 347:578–580.
- Michel F, Westhof E. 1990. Modelling of the three-dimensional architecture of Group I catalytic introns base on comparative sequence analysis. *J Mol Biol* 216:585–610.
- Moazed D, Stern S, Noller HF. 1986. Rapid chemical probing of conformation in 16S ribosomal RNA and 30S ribosomal subunits using primer extension. *J Mol Biol* 187:399–416.
- Penczek P, Brassucci RA, Frank J. 1994. The ribosome at improved resolution: New techniques for merging and orientation refinement in 3D cryo-electron microscopy of biological particles. *Ultra-microscopy* 53:251–270.
- Pley HW, Flaherty KM, McKay DB. 1994. Three-dimensional structure of a hammerhead ribozyme. *Nature* 372:68–74.
- Saenger W. 1984. *Principles of nucleic acid structure*. New York: Springer-Verlag.
- Svergun DI, Burkhardt N, Pedersen JS, Koch MHJ, Volkov VV, Kozin MB, Meerwinck W, Stuhmann HB, Diederich G, Nierhaus KH. 1997. Structural analysis of the 70S *E. coli* ribosome and its RNA by solution scattering: I. Invariants and validation of electron microscopic models. *J Mol Biol* 271:588–601.
- Uhlenbeck OC, Pardi A, Feigon J. 1997. RNA structure comes of age. *Cell* 90:833–840.
- van Bohlen K, Makowski I, Hansen HAS, Bartsel H, Berkovitch-Yelin Z, Zaytzev A, Meyer S, Paullre C, Franscheschi F, Yonath A. 1991. Characterization and preliminary attempts for derivatization of crystals of large ribosomal subunits from *Haloarcula marismortui* diffracting to 3 Å resolution. *J Mol Biol* 222:11–15.
- Woese CR, Magrum LJ, Gupta R, Siegel RB, Stahl DA, Kop J, Crawford N, Brosius J, Gutell RR, Hogan JJ, Noller HF. 1980. Secondary structure model for bacterial 16S ribosomal RNA: Phylogenetic, enzymatic and chemical evidence. *Nucleic Acids Res* 8:2275–2293.