Brazilian Journal of Microbiology (2007) 38:573-579
ISSN 1517-8382

THE FUNCTION OF LEXICAL MOTIFS IN THE ORGANIZATION OF THE *ACTINOMYCETES* 5S rRNAS

Sandra M Rodrigues-Subacius[1]; Gabriel Padilla[1]*

[1]Instituto de Ciências Biomédicas, Universidade de São Paulo, São Paulo, SP, Brasil.

Submitted: February 06, 2007; Returned to authors for corrections: April 20, 2007; Approved: July 16, 2007.

ABSTRACT

This work shows results obtained by employing the linguistic method to identify biologically meaningful sites in *Actinomycetes* 5S rRNAs. The approach adopted identifies triplet-words, along the base sequence of 5S rRNA, located mainly at the alpha and beta domains of the 5S secondary structure. There are triplet-words representing universal protein binding sites that include important prokaryote signatures, and sites strategically located in critical regions related to the formation of the 5S ribonucleoproteins (RNP) complex. In those sites, where the GC pressure promoted substitutions, the analysis demonstrates that alterations did not affect their biological significance. Sites formed by GGY (or more rarely GGR), continued to play an important role as ribosomal proteins rpL18 and rpL5 protein receptors. The data suggest that instead of increasing the molecular variability, expected for the diversity in species and habitats occupied for the group, GC pressure functioned as a reducer mechanism for the inter-specific diversity.

Key words: *Actinomycetes*; 5S rRNA; GC pressure, linguistic analysis, lexical motifs

INTRODUCTION

The 5S rRNA molecule has been chosen as a model for several exploratory studies on RNA-protein interactions. First, it is the only known RNA that binds to ribosomal proteins (rp), before its incorporation into the ribosome, facilitating the experimental control of these interactions (29). Thus, the *Eubacteria* 5S ribonucleoproteins (RNP) constitute an independent structural domain that, when integrated into the large ribosomal subunit, exerts a vital function in the A-site activity by mediating the communication between the peptidyl-transferase centre and the EF-G domain, through multiple cross-links with the 23S rRNA in the 50S ribosomal subunit (3). The allosteric intervention in the transmission signal via 5S RNP, not only depends on the 5S RNA molecule, but also on an intricate network of interactions, involving proteins of its own complex (rpL18, rpL5 and rpL25). The integrative and associative nature, of the structural domains and functional ribosomal centres, implies that the 5S+protein complex follows the same organizational principles responsible for the folding pattern of other rRNAs and rRNA-protein interactions (3,22).

Actinomycetes genomes are G+C rich (>72%). The strong displacement from the neutrality content (G+C= 50%) has been attributed to selective forces (19,33) or to mutational bias, named GC pressure, operating on the genome during phylogenetic intra-group evolution (2,17,28,32). The GC pressure effect has been sufficiently intense to increase the GC composition of its genic product (5S rRNA) in *Actinomycetes*, as revealed by the comparison with other Gram-positive species (26,27).

Among *Eubacteria*, *Escherichia coli* 5S rRNAs have been extensively studied using enzymatic and chemical approaches (6), physical techniques (25), structural-functional and genomic analysis (11,13,19), in order to elucidate their molecular structure and binding sites that contributes to the tertiary conformation of the molecule or formation of the 5S RNP. Since little information, concerning *Actinomycetes* 5S rRNA is available in the literature, the results obtained in the present study, based on linguistic analysis, were compared to experimental data obtained using *E. coli* 5S rRNA as a reference molecule. This procedure was useful to test the validity of the theoretical approach adopted here, focused on the identification of sites related to 5S RNA conformation and 5S RNA-protein interaction.

*Corresponding Author. Mailing address: Instituto de Ciências Biomédicas, Universidade de São Paulo, SP, Brasil. E-mail: gpadilla@icb.usp.br

The main objective of the present work was to study the influence of the GC pressure in the *Actinomycetes* 5S rRNA molecules, analysing the appearance of new oligonucleotides (double and triplets), and their influence on several secondary structural domains. Within this context, we adopted a linguistic approach, because it allows the discrimination between short segments (oligonucleotides) which carry some potential biological meaning, for example, recognition signals useful for the intra- and inter-molecular interactions between RNA-RNA and/or RNA-protein, respectively (4,5,20,21).

MATERIAL AND METHODS

The 5S rRNA sequences were obtained from the 5S Ribosomal RNA DataBank, available at http://biobases.ibch.poznan.pl/5Sdata, which contains 536 5S RNA sequences from *Eubacteria* and 61 sequences from *Archaea*. For some species, each record consists of more than one sequence due to intercistronic or intra-specific variability (29). The taxonomic classification was adopted without modification.

5S RNA Model

The 5S rRNA model is from 5S Ribosomal RNA Data Base (Fig. 1). It consists of five helix (I-V) and single stranded elements, i.e., two external loops (C and D) and two internal loops (B and E), and a hinge region (loop A), which connects the three arms of the molecule. This model is based on a multiple sequence alignment of 536 5S RNA from *Eubacteria*. The model is 158 positions long showing high variability because there are only three positions conserved: C at position 61, G at position 62 and U at position 99. For prediction of RNA secondary structure the mfold web server was used as reference. Phylogenetic data suggest quasi-co-linearity between helix II and helix V, maintained by the unusual triple interactions involving the conserved residues U14, G69, and G107 of loop A (Fig. 1). As consequence, the three main structural domains (alpha, beta and gamma) do not interact with each other, leading to spatial conformations for Y-shaped molecule (32). The crystal structure of the large ribosomal subunit from *Haloarcula marismortui* shows interactions of 5S rRNA with ribosomal proteins and 23S rRNAs. There are also interactions with ribosomal proteins L5, L10e, L18, L21 and L30 but only one direct interaction linking 5S rRNA domain γ and helix 38 of 23S ribosomal RNA (1).

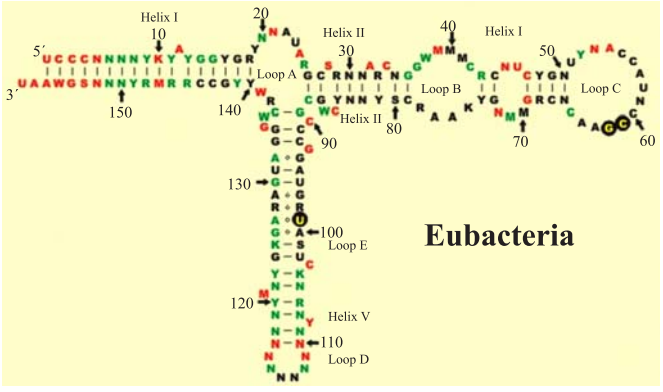

Figure 1. Secondary structure of 5S ribosomal RNA of *Eubacteria*. The numbering of nucleotides corresponds to multiple sequence alignment. Nucleotide symbols are according to IUBMB nomenclature. N:A, C, G or U; Y: C or U; R: A or G; W: A or U; S: G or C; M: A or C; K: G or U (29).

Linguistic Method

The linguistic method is based on a Markovian procedure, which identifies oligonucleotide words through the difference (contrast values, Cv) between the observed frequencies (*f*obs), and the calculated expected value (Exp) (5,21). In this work were recognized as code words only those oligomers, with absolute standard deviation (SD) equal to or greater than 1.00. We adopted this significance level, based on the small size of the 5S sequences, in addition to their high evolutionary conservation. Consequently, the molecule contains a limited number of variable sites, once the 5S gene was under strong mutational restriction (12).

For regression analysis, the average Ks index and contrast values per group were calculated. *Streptomyces* and *Mycobacterium* were not included because their contrast values differ from other groups. Regression analysis was carried out using the Microcal^{Tm}Origin^{Tm} version 4.10 software (Microcal Software, 1996).

RESULTS

Preferred Double Nucleotides

An atypical enrichment in GG dinucleotides (DINGG) was observed in *Actinomycetes* 5S rRNAs, reaching values comparable to those reported for thermophilic *Eubacteria* (28). As consequence the expected proportions of G and C combinations would be elevated (Table 1). The linguistic analysis showed that GC pair was consistently an avoided word (Table 2), and the CC amount exceeded the frequencies theoretically expected; only CC and GU stand out as code word by this analysis.

CC sites and hairpin C

The preference for DiNCC was a common feature among all *Actinomycetes* 5S RNAs, although it was not restricted to this group, some *Firmicutes* species also shared this characteristic. Screening of the CC sites along the *Actinomycetes* 5S structure revealed that approximately 60-75% of these sites are preferentially restricted to hairpin C (Table 3), sometimes encompassing neighbouring positions located on the 5' loop

Table 1. Duplet combinations (CC; GG; CG; GC). Expected and observed values.

	CC		GG		CG		GC	
	Exp	Obs	Exp	Obs	Exp	Obs	Exp	Obs
Coryneform	7,66	11,60	15,22	15,00	10,71	10,00	10,70	8,60
Actinomycetelaes	8,61	12,00	15,90	15,00	11,56	11,00	11,83	9,00
Pseudo-nocardiaceae	9,99	13,83	14,31	14,39	11,92	11,61	12,07	8,94
Frankia	8,84	10,48	15,19	13,55	11,25	9,45	11,70	8,01
Micrococcaceae	9,17	11,60	14,75	15,80	11,59	10,40	11,65	8,60
Aureobacterium	9,15	12,00	14,47	16,00	11,37	11,00	11,65	10,00
Streptomycetes	7,96	10,75	14,05	15,00	10,43	11,00	10,72	8,83
Arthrobacter	10,30	14,00	14,80	16,60	12,13	12,20	12,53	8,20
Clavibacter	9,71	11,00	14,47	15,00	11,71	11,00	12,00	9,00
Actinoplanetes	11,41	16,00	14,35	13,00	12,64	13,00	12,96	11,00
Pimelobacter	10,89	14,00	15,18	14,00	12,71	13,00	13,01	13,00
Mycobacteria	12,65	16,10	10,28	11,30	11,21	11,60	11,53	11,10

Values correspond to the average and standard deviation per *group*. Exp= expected; Obs= observed.

B-side. At least seven highly conserved CC sites were observed in this region, whose arrangement displayed a typical structure, consisting of four C-clusters (≥ CC), three derived from overlapping CC pair. The C-clusters occurred separately from each other in regular intervals of three positions (standard distribution) as observed in *Clavibacteria* (Table 3). The C-cluster II vanishes after the point mutation fixation at target positions C36 and C37 in *Aureobacteria*, *Frankia*, *Clavibacteria* and *Streptomyces*. C-cluster IV is particularly interesting since it is part of a specific-group signature CC48CGGAAGC (Table 3). The conservation of the I and III C-clusters, together with a CAU triplet located in hairpin C, permitted an alternative local alignment showing that non-random arrangement of C-cluster site is a recurrent pattern, which is independent of the eubacterium origin (Table 3). Detectable homologies through this unusual alignment procedure provided some hints about the possible bias in the growth of these clusters. C-cluster size is less variable in *Actinomycetes* 5S RNA, except for some rare species, in which C-cluster III was expanded one position at its 5' end (*M. luteus*, *M. capsulatus* and *Mycobacteria*, Table 3). When *Eubacteria* sequences (*E. coli*, *Bacillae* and *Mycoplasma*) were alignment, the association between growing direction and secondary structural elements became more evident. Clusters located in double strands were extended in the 5' → 3' direction (Table 3).

Table 2. Preferred and avoided double and triplet nucleotides in *Actinomycetes* 5S rRNA.

Actinobacteria 1.0≤std≥1.0	CC (+)	GC (-)	GU (+)	CAU (+)	AGC (+)	GAA (+)	CCG (+)	GGU (+)	GGC (-)	CGU (-)	Specific-group triplets
Coryneform	80%	40%	40%		60%		80%	100%	100%	80%	GAU(+) CAA(+)
Actinomycetales	100%	100%						100%	100%		AAC(+) UUC(+)
Pseudo-nocardiaceae	100%	75%	100%	37%	75%	65%	37%	95%	95%	85%	
Frankia	50%			100%			100%	50%	50%	100%	
Micrococcaceae	75%	50%	25%	100%	75%		50%	50%	50%		
Aureobacterium	100%				100%		100%		100%	100%	
Streptomycetes	83%			100%	92%	75%	85%		100%		UAU(+) UUC(+)
Arthrobacter	100%	80%	80%	80%	80%	80%	60%	80%	100%		UUC(+)
Clavibacter		100%					100%		100%		
Actinoplanetes	100%			100%							UCC(+) GGA(+)
Pimelobacter	100%			100%			100%	100%			CGG(+)
Mycobacteria	90%			50%	70%	100%		70%		90%	UCC(+) CGG(+)

Percentages refer to the number of molecules in each group in which these words are under (-) or over (+) represented.

Table 3. C-clusters and preferred triplet nucleotides located at hairpin C.

5'.....	26	27	28	**[29**	**30**	**31**	**32**	**33**	**34]**	35	36	37	38	39	40	41	42	43	44	45	46	47	**[48**	**49**	**50**	**51**	**52**	**53**	**54**	**55]**	**3'..**
C-clusters						**I**											**III**								**IV**							
Standard Distribution					**C**	**C**	**C**	***G**	***G**	***U**	**C**	**C**	**C**	***A**	***U**	*	**C**	**C**	***G**	***A**	***A**	**C**	**C**	**C**	***G**	***G**	***A**	***A**	***G**	***C**	
Clavibacter					C	C	C	*	*	*	*	*	*	*	*	*	C	C	*	*	*	C	C	C	*	*	*	*	*	*		
Frankia					C	C	C	*	*	*	*	*	*	*	*	*	C	C	*	*	*	C	U	C	*	*	*	*	*	*		
Micrococcus lteus					C	C	C	*	*	*	C	C	C	*	*	*C*	C	C	*	*	*	C	C	C	*	*	*	*	*	*		
																←																
Brevibacteriacasei			C	C	***C***	* G	*G	*U	*	*	*	C	C	*	*	*	C	C	*	*	*	C	C	C	*	*	*	*	*	*		
Micrococcus																																
capsulatus	C	C	*	***C***	C	C	*	*	*	***C***	C	C	*	*	*C*	C	C	*	*	*C*	C	U	C	*	*	*	*	*	*		
																←					←											
Mycobacteria					C	C	C	*	*	*	C	C	C	*	*	*C*	C	C	*	*	*	C	C	C	*	*	*	*	*	*		
																←																
Escherichia coli	**C**	**C**	**C**	*	**C**	**C**	Y	*	*	C	**C**	**C**	**C**	***A***	***U***	*	**C**	**C**	***G**	***A**	***A**	**C**	**U**	**C**	***A**	***G**	***A**	***A**	***G**	***U**	
							→			←																						
Bacillaceae					**C**	**C**	**C**	*	*	*	**C**	**C**	**C**	***A***	***U***	*	**C**	**C**	*	*	*	**C**	**A**	**C**	*	*	*	*	*	*		
Mycoplasmatales					**C**	**C**	Y	*	*	*	**C**	**C**	**C**	***A***	***U***	*	**C**	**C**	*	*	*	**C**	**A**	**C**	*	*	*	*	*	*		

The numbers correspond the positions of the hairpin C, with reference to 5S aligned sequence. Numbers in bold inside the brackets represent the helix bIII (5´ and 3'side, respectively ; more detail see fig. 1); I and III C-*clusters* and the CAU (conserved triplet) were used as consensus segments for the alignment of the EUB (¯G+C) and *E.coli* 5S RNAs; The standard distribution and preferred triplets that occur in this region of the actinomycete 5S rRNAs are showed in third row; The arrows indicate the direction of growing of the C-clusters in helix bIII and loop C.

Preferred Triplet Nucleotides

The code-word definition adopted by the linguistic analysis introduces the notion of word size. The high incidences of CC pair added to GC pressure were responsible for the large quantity of CCG triplets in 5S *Actinomycetes* (aprox. 7 sites per sequence). Moreover, the abundance by itself is not the single criteria to determine the code word of interest (23). The highest proportion of triplet code-words discriminated by this analysis was found in the alpha and beta domains of the secondary structure of *Actinomycetes* as a product of the mutational bias induced by genomic GC pressure (Fig. 1). Helix III was one of the regions most tolerant to G and C fixation. As a consequence, important code-word triplets (CCG, GGU and AGC) were generated by substitutions from A48C, U53C and U34G, in Gram positive *Eubacteria* (↓ G+C), along the stem III, forming a group-specific signatures, such as CCGG34U and C48CGGAAGC53 present in 99% and 85% of the species, respectively.

At least one site of each triplet word discriminated by linguistic analysis occurs in hairpin C. An intense overlap of the triplet word predominates in all extensions of this region encompassing all C-cluster sites, whose limit (3' end), is always defined by the presence of a G (CCCG), except for CCC38AU segment. Three of seven occurrences of CCG overlap the C-clusters in hairpin C (Table 3). Frequently, the G terminus is shared by another favourite triplet word (CCG33GU). Overlapping bases, when filling the central position of two distinct words, generally correspond to universal invariant positions of the molecule. These positions might play an important role as specific interaction sites. It was shown experimentally that L5 ribosomal proteins interact with residues from the CCC38AU segment (24). Positional ambiguity between the RNA and its binding proteins could explain the superposition of CAU into C-cluster II. However, in *Streptomyces* and other species, base substitutions were responsible for the degeneracy of C-cluster, conserving only the invariant C3 position, which remained available to interact with rpL5.

The GAA triplet was not a preferred consensual word for all *Actinomycete* species (Table 2). Nonetheless, the two GAA sites in hairpin C were nearly universal in the *Actinomycetes* 5S RNAs. Invariably, both sites overlapped with another preferred word, either as part of the most important phylogenetic signal of the prokaryotic 5S RNA (CCG44AA) (Fig. 1), or by sharing the A53 invariant position with the AGC preferred triplet in stem III (GGA52A53GC). The A52A53 forms a very stable and well-defined bulge flanked by the guanines (Fig. 1 and Table 3).

We noted that, although DINGC is an avoided word of the *Actinomycetes* dinucleotide vocabulary, the statistical significance of the $^{5'}$RGC$^{3'}$ triplets clearly depend on the previous purine, *i.e.*, A (AGC) or G (GGC) (Table 2). While GGC triplets were under-represented in the 5S *Actinomycetes*, AGC triplets raised as code words related with structural motifs. Along the secondary structure of the molecule (Fig. 1), these residues may be part of loops (loop A) or participate, when adjacent to double strands, in the formation of single (A66GC in the βII helix) or double (A53) bulges. Furthermore, the $^{5'}$GC$^{3'}$/$^{5'}$GC$^{3'}$ complement themselves by secondary folding of the molecule, near to the 5' and 3' ends of helix II adjacent to loop A, thus contributing to the conformation of this hinge region (Fig. 1).

GC pressure and the GGY triplets

The behaviour of the $^{5'}$GGY$^{3'}$ preferred triplet in response to GC pressure is particularly interesting in terms of linguistic analysis. All the sites occupied by this triplet are useful for illustrating specific site changes, dependent on the GC pressure intensity. The GGY triplet is determined by the pyrimidine occurring at the third position. In most actinomycetes GGU is the preferred base combination rather than GGC that tends to be strongly avoided with rare exceptions (Table 2). GGY sites are frequently distributed along the 5S sequence as isolated sites (Fig. 1), overlapping another triplet (CCGGU), or forming an hexamer with two tandem repeats (GGUGGU) present in the helix/loop junctions (Fig. 1). It is noteworthy that GGY triplets (GGC/U) are frequently found inside helix βII (Fig. 1). We observed that exchange of U for C at GGY variable sites is strongly dependent of the GC pressure acting on actinomycete genomes. The site-specific efect of the GC pressure is well illustrated by the action of the G4GYG7GY universal site located in helix I of almost all *Firmicutes*. The first frame is preferentially represented by G4GU in *Eubacteria* (↓G+C), while this triplet prevails in the second module in approximately 80% of the *Actinomycetes*. The preference for GGC in this position is restricted to 17% of the *Actinomycetes* species that commonly present the highest Ks values (*Streptomycetes, Arthrobacter, Clavibacter, Actinoplanets, Pimelobacter*).

Ks Index and GGU/GGC Triplets

The comparison between the contrast values and the Ks index predict an evolutionary relationship between GGU and GGC triplets (26). This index was used as a phenotypic marker of point mutations activities in the 5S gene as reflected by its genic product. Nevertheless, the present data have revealed that the Ks index also seems particularly sensitive for detecting variations in the GC pressure magnitude, especially in those *Eubacteria* species, like *Actinomycetes*, whose genomes are submitted to intense and variable GC mutational pressure.

A strong linear correlation between the average Ks values per group and the contrast values for the GGU and GGC triplets was observed, indicating a sharp dependence on the intensity of the mutational pressure. Furthermore, the frequency of GGU triplets tends to reduce uniformly with the increase of Ks values, while GGC triplets followed the opposite trend (Fig. 1). This implies that G+C input into the 5S gene, induced by the GC pressure, was sufficiently intense to promote an expressive U → C substitution at the 3' end position of the $^{5'}$GGY$^{3'}$ triplet.

Only two actinomycetes groups, *Streptomyces* and *Mycobacterium* did not follow the tendency described above. In *Streptomyces* 5S rRNA, GGC sites are so deleterious that they simply do not occur. On the other hand, in *Mycobacterium* 5S rRNA, as well in two other rare actinomycetes *Actinoplanes* and *Arthrobacter*, GGA rather than GGU is the favourite triplet word. In these species, GGU sites were often replaced by GGA. These changes were especially significant in sites located inside helix βII, sometimes, enclosing loop B (Fig. 1).

DISCUSSION

Actinomycete Contrast-Vocabulary

The most remarkable features of the actinomycete contrast-vocabulary are the inter-specific homogeneity, in addition to the restricted repertoire of triplet words, when compared with Gram-positive eubacteria (↓G+C) and protobacterium vocabularies. The GC pressure exerted on the 5S gene functioned as a mechanism of reducing differences, despite a wide species diversity of the most variable habitats (2).

The strong selection of certain combinations of G+C triplets in the 5S molecule, to detriment of others, should reflect not only their structural and/or functional meaning, but also universal genomic properties reflected by genic products. To illustrate this effect Trifonov and Bettecken (1997) (31) showed that CCG triplet is exceptionally abundant in G+C enriched genomes. In addition, phage genomes also yield a triplet pattern rich in CCG (5). In the present study, this combination was dominant in all *Actinomycetes* 5S rRNAs, while in the *Firmicutes*, the content remained under-represented, with only the universal sites that compose the small helix IV and those contained in the preserved prokaryotic signature loop C.

Triplet Words as Structural and Functional Lexical Motifs

The abundance of some triplet-words, in response to the genomic GC pressure, reflects further fixation of point mutations into variable positions of the 5S molecule. There are two possible explanations: first, the emergence of such triplet-words does not jeopardize the specific intra- or inter-molecule interactions, or second, some of these new sites confer additional advantages on the structure and/or function of the molecule.

Several lines of evidence support these interpretations, for example, the main characteristic of the GGU and GGC sites, refers to the free exchange between the two triplets as a function of the direction and magnitude of the genomic GC pressure. Such performance strongly implies that both triplets (GGU/C) belong to a unique family of functional code-words, which probably includes the GGA/G (GGR) combination. Therefore, these words should represent morphological units of the 5S RNA language that carry a similar biological meaning.

Although experimental data regarding *Actinomycetes* 5S rRNAs are scarce, enzymatic assays carried out with *E. coli* 5S RNA have demonstrated that several positions occupied by G are strongly protected by the rpL18 ribosomal protein (10), including the G24 position that lies inside the string formed by the two juxtaposed GGU sites. Furthermore, the *E. coli* rpL18 may also interact with the carbon-phosphate backbone of G7, inserted into the GG7CGGC hexamer located in the α helix (23).

The evolutionary relationship between GGU/C (GGY) found in the present study suggests that both sites are functionally equivalent. Consequently, the analogous positions occupied by GGU in *Actinomycetes* 5S RNA, such as GG24U (loop B) and the GG7UGGU duplicate site in the α helix, may represent primary recognition targets for actinomycete rpL18. Since repeats can be related to multiple protein-binding sites, inducing an amplification and/or redundancy of their local expression, we propose an expansion of the rpL18 contact region, specifically when GGY triplets are arranged in tandem, as observed for *E. coli* 5S (GGUGG24U) at the helix B/loop B junction.

The universal binding sites for rpL18 ribosomal protein and probably for rpL5, is confirmed by the presence of G 2307GGACAUCAGGAGGU2320 and 2305UCGGAC2310, which GU and GGA, isolated or in tandem in the 23S rRNA fragment, for example interact with the two proteins inside the *E. coli* ribosome, respectively (18,19). It is noteworthy that the GGU triplet is also found in the γ domain, where it occurs inserted into a hexamer composed of two inverted juxtaposed triplets (AUGGUA), which are highly conserved and essential for the interaction of the 5S RNA with the L25 β-barrel (25).

The localization of the code words along the 5S secondary structure, demonstrates that the effect of the GC pressure on the emergence of a group-specific vocabulary has mainly affected sites located in α and β structural domains. The main binding sites of *E. coli* 5S RNA, the rpL5 and rpL18 proteins, are found in these domains with the hinge region conformation being critical for the initial recognition of rpL18 (9, 10, 14, 18, 23, 24, 34). Such a configuration seems to guarantee the conformation stability necessary for recognition. Thus, these AGC sites located at the extremities of the helix play an important role as structural motifs. So far, no specific binding probes for the rpL18 and/or rpL5 proteins involving the AGC residues in 5S rRNAs exist. On the other hand, experimental evidence indicates the importance of the intercalating properties of the unpaired residues, either in keeping the local conformation or in relaxing the internal one (16). Long-range tertiary interaction might explain the high intra- and inter-specific conservation of the impaired sites.

Some triplet-words, discriminated by the linguistic analysis, apparently did not suffer any GC pressure influence, for example CAU. Some studies have adopted an alternative tertiary conformation for the *E. coli* 5S RNA, based on interaction between C38 and U40 and the rpL5 ribosomal protein (14). However, the presence of this triplet in the same 23S fragment, where GGR triplets are found, points to a potential rpL18 protein receptor, which serves as a functional motif analogue to GGR triplets.

5S and 23S RNAs interactions

A large data concerning the 5S-23S rRNA interactions came from studies with *E. coli*. The present study has identifies some group-specific differences, that should imply significant alterations in the conformation and topology of the 5S molecule, either in its free state, or when complexes to ribosomal proteins and/or 23S rRNA. This is the case, for example, for a short base segment that composes loop D. This string apparently exhibits characteristics similar to those found in the tetra loop belonging to the GNRA family. In *E. coli* 5S RNA, loop D contains the U89 position that can be related to inter-nucleotide interaction between 5S-23S rRNA (7,8). The presence of G89, as part of a tetra loop structure indicates that the binding rules involving the *Actinomycetes* 5S and 23S RNAs are group-specific, especially when considering the tetra loops represents important structural motifs of tertiary interactions.

ACKNOWLEDGMENTS

We are grateful to FAPESP for financial support to SS (Proc. 99/06637-1). To Dr. Samuel Pietrokoviski (Weismann Institute of Science-Israel) for providing the Contrast program, Dr. Rafael Santos (IP&D University of Vale do Paraiba) for the help with computational programs, Drs. Beatriz Fernandes and Stephen Blanksby and Mrs. Kerstin Markendorf for suggestions and critical review of the manuscript.

RESUMO

A função dos motivos léxicos na organização do 5S rRNAs de *Actinomicetes*

Neste trabalho são apresentados resultados obtidos empleando o método linguístico para identificar sítios no 5S rRNAs de *actinomicetes* com significado biológico. A abordagem identificou palavras-tripletes, junto com a sequência de bases do 5S rRNAs, localizados principalmente nos domínios alfa e beta da estrutura secundária. Entre eles, existem palavras-tripletes que representam sítios de ligação de proteínas universais, que incluem importantes assinaturas procarióticas, além de sítios estrategicamente colocados em regiões críticas relacionados com a formação do complexo 5S ribonucleoproteína (RNP). Nestes sítios, onde a pressão GC promove substituições, as alterações não afetaram seu significado biológico. Sítios formados por GGY (ou mais raramente GER), jogam um papel importante como receptores de proteínas ribossomicas rpL18 e rpL5. Os dados também sugerem que ao contrário de aumentar a variabilidade molecular, esperada pela diversidade em espécies e habitats ocupados pelo grupo, GC funciona como um mecanismo para diversidade inter-específica.

Palavras-chave: *Actinomicetes*, 5S rRNA, pressão GC, análise linguístico, motivos léxicos

REFERENCES

1. Ban, N.; Nissen, P.; Hansen, J.; Moore, P.B.; Steitz, T.A. (2002). The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, 289, 905-920
2. Bentley, S.D.; Chater, K.F.; Cerdeno-Tarraga, A.M.; Challis, G.L.; Thomson, N.R.; James, K.D.; Harris, D.E.; Quail, M.A.; Kieser, H.; Harper, D.; Bateman, A.; Brown, S.; Chandra, G.; Chen, C.W.; Collins, M.; Cronin, A.; Fraser, A.; Goble, A.; Hidalgo, J.; Hornsby, T.; Howarth, S.; Huang, C.H.; Kieser, T.; Larke, L.; Murphy L.; Oliver, K.; O'Neil, S.; Rabbinowitsch, E.; Rajandream, M.A.; Rutherford, K.; Rutter, S.; Seeger, K.; Saunders, D.; Sharp, S.; Squares, R.; Squares, S.; Taylor, K.; Warren, T.; Wietzorrek, A.; Woodward, J.; Barrell, B.G.; Parkhill, J.; Hopwood, D.A. (2002). Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature*, 417, 141-147.
3. Bogdanov, A.A.; Dontsova, O.; Dokudovskaya, S.S.; Lavrik, I.A. (1995). Structure and function of 5S rRNA in the ribosome. *Biochem. Cell. Biol*., 73, 869-876.
4. Bolshoy, A. (2003). DNA sequence analysis linguistic tools: contrast vocabularies, compositional spectra and linguistic complexity. *Appl. Bioinformatics*, 2, 103-12.
5. Brendel, V.; Beckmann, J.S.; Trifonov, E.N. (1986). Linguistics of nucleotide sequences: Morphology and comparison of vocabularies. *J. Biomolec. Str. Dyn*, 4, 11-21.
6. Brunel, C.; Romby, P.; Westhof, E.; Ehresmann, C.; Ehresmann B. (1991). Three -dimensional model of *Escherichia coli* ribosomal 5S RNA as deduced from structure probing in solution and computer modeling. *J. Mol. Biol*., 221, 293-308.
7. Dokudovskaya, S.; Dontsova, O.; Schpanchenko, O.; Bogdanov, A.A.; Brimacombe, R. (1996). Loop IV of 5S ribosomal RNA has contacts both to domain II and to to domain V of the 23S RNA. *RNA*, 2, 146-152.
8. Dontsova, O.; Tishkov, V.; Dokudovskaya, S.; Bogdanov, A.A.; Doring, T.; Rinke-Appel, J.; Thamm, S.; Greuer, S.; Brimacombe, R. (1994). Stem-loop IV of 5S rRNA lies close to the peptidyltransferase center. *Proc. Natl Acad. Sci.,* USA 91, 4125-4129.
9. Egebjerg, J.; Christiansen, J.; Brown, R.S.; Larsen, N.; Garrett, R.A. (1989). Protein L18 binds primarily at the junctions of helix II and internal loops A and B in *Escherichia coli* 5S rRNA. *J. Mol. Biol*., 206, 651-668.
10. Garrett, R.A.; Noller, H.F. (1979). Identification of kethoxal-reactive sites on the 5S RNA. *J. Mol. Biol.,* 6, 637-648
11. Gu, S.Q.; Jockel, J.; Beinker, P.; Warnecke, J.; Semenkov, Y.P.; Rodnina, M.V.; Wintermeyer, W. (2005). Conformation of 4.5S RNA in the signal recognition particle and on the 30S ribosomal subunit. *RNA*, 11, 1374-84
12. Guimarães, R.C.; Trifonov, E.N.; Lagunez-Otero, J. (1997). Taxonomy of 5S ribosomal RNA by the linguistic technique: Probing with mitochondrial and mammalian sequences, *J. Mol. Evol*., 45, 271-277.
13. Kierzek, E.; Kierzek, R.; Turner, D.H.; Catrina, I.E. (2006). Facilitating RNA structure prediction with microarrays. *Biochemistry*, 45, 581-93.
14. Mueller, F.; Sommer, I.; Baranov, P.; Matadeen, R.; Stoldt, M.; Wohnert, J.; Gorlach, M.; Van Heel, M.; Brimacombe, R. (2000). The 3D arrangement of the 23S and 5S rRNA in the *Escherichia coli* 50S ribosomal subunit based on a Cryo-electron microscopic reconstruction at 7.5 A resolution. *J. Mol. Biol*., 298, 35-59.
15. Naya, H.; Romero, H.; Zavala, A.; Alvarez, B.; Musto, H. (2002). Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. *J. Mol. Evol.*, 55, 260-264.
16. Nazar, R.N. (1991). Higher order structure of the ribosomal 5S RNA. *J. Biol. Chem.*, 266, 4562-4567.
17. Osawa, S.; Jukes, T.H.; Muto, A.; Yamao, F.; Ohama, T.; Andachi, Y. (1987). Role of directional mutation pressure in the evolution of the eubacterial genetic code. Cold Spring Harbor Symp. *Quant. Biol*., LII, 777-789.
18. Osswald, M.; Greuer, B.; Brimacombe, R. (1990). Localization of a series of RNA-protein cross-link sites in the 23S and 5S ribosomal RNA from *Escherichia coli* induced by treatment of 50S subunits with three different bifunctional reagents. *Nucl. Acids Res*., 18, 6755-6760.
19. Ostegaard, P.; Phan, H.; Johansen, L.B.; Egebjerg, J.; Ostegaard, L.; Porse, B.T.; Garrett, R.A. (1998). Assembly of proteins and 5S rRNA to transcripts of the major structural domains of 23S rRNA. *J. Mol. Biol*., 284, 227-240.
20. Pietrokovski, S. (1989). Nucleotide sequence dialects and vocabularies. MSc Thesis, *Weizmann Inst. Sci*., Rehovot, Israel.
21. Pietrokovski, S.; Hirson, J.; Trifonov, E.N. (1990). Linguistic measure of taxonomic and functional relatedness of nucleotide sequences, *J. Biomolec. Str. Dyn*., 7, 1251-1268.
22. Sergiev, P.; Dokudovskaya, S.S.; Romanova, E.; Topin, A.; Bogdanov, A.A.; Brimacombe, R.; Dontosova, O. (1998). The environment of 5S rRNA in the ribosome:cross-links to the GTPase-associated area of 23S rRNA. *Nucl. Acids. Res*., 26, 2519-2525.
23. Shpanchenko, O.V.; Zvereva, M.I.; Dontsova, A.O.; Nierhaus, K.H.; Bogdanov, A.A. (1996). rRNA sugar-phosphate backbone protection in complexes with specific ribosomal proteins. FEBS. *Letter*, 394, 71-75.
24. Speek, M.; Lind, A. (1982). Structural analyses of *E. coli* 5S RNA fragments, their associates and complexes with proteins L18 and L25. *Nucl. Acids Res.*, 10, 947-63.
25. Stoldt, M.; Wohnert, J.; Ohlenschlager, O.; Gorlach, M.; Brown, L. (1999). The NMR structure of the 5S rRNA E-domain-protein L25 complex shows pre-formed and induced recognition. *EMBO J*., 18, 6508-6521.
26. Subacius, S.M.R.; Bussab, W.O. (1998). Purine and pyrimidine composition in 5S rRNA and its mutational significance. *Genetics Mol. Biol*., 21, 255-258.
27. Subacius, S.M.R. (1994). Evolution of the 5S rRNA, PhD Thesis, Institute of Biosciences-University of São Paulo, Brazil.
28. Sueoka, N. (1988). Directional mutation pressure and neutral molecular evolution, *Proc. Natl. Acad. Sci.*, USA 85, 2653-2657.
29. Szymanski, M.; Barciszewska, M.Z.; Erdmann, V.A.; Barciszewski, J. 2002. 5S ribsomal RNA database. *Nucl. Acids. Res*., 30, 176-178
30. Trifonov, E.N.; Bettecken, T. (1997). Sequence fossils, triplet expansion and reconstruction of earliest codons. *Gene*, 205, 1-6.
31. Trifonov, E.N.; Bolshoi, G. (1983). Open and closed 5S ribosomal RNA: the only two universal structures encoded in the nucleotide sequences. *J. Mol. Biol*., 169, 1-13
32. Wada, A.; Suyama, A.; Hanai, R. (1991). Phenomenological theory of GC/AT pressure on DNA base composition. *J. Mol. Evol*., 32, 374-378.
33. Wright, F.; Bibb, M.J. (1992). Codon usage in G+C rich *Streptomyces* genome, *Gene*, 113, 55-65.
34. Zhang, P.; Popieniek, P.; Moore, P.B. (1989). Physical studies of 5S rRNA variants at position 66. *Nucl. Acids. Res*., 17, 8645-8656.