Journal of Molecular Graphics and Modelling 27 (2009) 770–776

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

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

Computational modeling analyses of RNA secondary structures and phylogenetic inference of evolutionary conserved 5S rRNA in the prokaryotes

Vijai Singh, Pallavi Somvanshi *

Bioinformatics Centre, Biotech Park, Sector-G Jankipuram, Lucknow 226021, Uttar Pradesh, India

ARTICLE INFO

Article history:
Received 4 April 2008
Received in revised form 16 November 2008
Accepted 19 November 2008
Available online 7 January 2009

Keywords:
Bacteria
5S rRNA
Phylogeny
RNA secondary structure
Free energy
Evolution

ABSTRACT

Bacteria are unicellular, ubiquitous microorganisms which grow on soil, acidic hot springs, radioactive wastes, etc. The genome of bacteria constitutes species specific conserved region. The 5S rRNA is one of the most conserved region determined in each bacteria and the size ranges between 110 and 148 bp. On this basis phylogenetic study of 37 bacterial strains was done which results in formation of seven clades and furthermore RNA secondary structure from each clade was made. The lowest free energy (δG) of the 5S rRNA may divulge the most primitive bacteria and slow changes occurs throughout the evolution whereas the higher free energy indicates less stability during the evolution. The RNA secondary structure may provide new insights to understand bacteria evolution and stability.

1. Introduction

Bacteria are unicellular, pleomorphic microorganisms and ubiquitous in every habitat, grows in soil, acidic hot springs, radioactive waste, seawater and deep in earth's crust [1]. There are typically 40 million bacterial cells in a gram of soil and a million bacterial cells in millilitre of fresh water and accumulated the biomass [2]. They play vital role in recycling nutrients and important steps in nutrient cycles which depends on bacteria viz. fixation of nitrogen from the atmosphere. There are 10 times more bacteria present in the human cells when compared with the human body with large numbers of them present on the skin and digestive tract, although, majority of bacteria are become harmless or beneficial by the protective effects of immune system and few are pathogenic which causes infectious diseases, including cholera, syphilis, anthrax, leprosy and bubonic plague. The most common fatal bacterial diseases are respiratory infections; tuberculosis alone kills about 2 million people in a year at sub-Saharan Africa [3].

The cultures of bacteria from all possible habitats present on earth studied, relates to bacterial diversity, diseases, ecological functions and biotechnological applications. The ribosomal operons mainly 16S, 23S and 5S rRNA was proved to be a stable molecular marker for the bacterial identification. The copy number of ribosomal rRNA genes varies from 1 to 15 among bacterial genomes; it is generally believed that all the copies in an organism are identical or nearly identical in nucleotide sequence. The 5S rRNA was present in scattered form in the entire genome of any bacteria. These ribosomal sequences are useful for the phylogenetic study and molecular taxonomy. The copy number of ribosomal operon affects the growth. *Mycobacterium tuberculosis* genome has only 1–2 copy number of ribosomal operon. It is very slow growing bacteria which is the major rationale for growth [4].

Non-coding and structural regulatory motifs of RNA play important roles in gene regulation and other cellular functions. They are characterized by specific secondary structures that are critical to their functions and are conserved in phylogenetically or related functionally. Predicting common RNA secondary structures in multiple aligned sequences remains a challenge in bioinformatics research [5].

RNA structure is less complex than protein structure so, it can be well characterized by identifying commonly occurring location

* Tel.: +91 522 4012076; fax: +91 522 4012081.
E-mail address: psomvanshi@gmail.com (P. Somvanshi).

1093-3263/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2008.11.012

of secondary structural elements. The single stranded RNA base pairs within a single RNA molecule, forming the base pair stem regions. Various bioinformatics tools have been developed for predicting the secondary structure of RNA molecule. Many of these methods attempted to minimize the free energy of folded macromolecule, thus searching for most stable structure [6]. In this study, we deduce the 5S rRNA secondary structure models of 37 bacterial strains. The thermodynamic energy value was also given to further support the stability of strains throughout the evolution.

2. Materials and methods

2.1. Retrieval of genome sequence and identification of conserved domain

The complete 5S rRNA sequences of bacteria strains were retrieved from National Centre for Biotechnology Information microbial genome biology (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi).

2.2. Construction of phylogenetic tree

The 5S rRNA sequences were aligned with CLUSTALX [7]. The computed alignment was manually checked. Pair wise evolutionary distances were computed using Jukes and Cantor algorithm implemented in the MEGA 4 [8] program and a phylogenetic tree was constructed by Neighbor-Joining method.

2.3. RNA secondary structure prediction

Predictions of possible folding of the 5S rRNA bacteria were performed using Mfold. The widely used algorithms for RNA secondary structure prediction, which are based on a search for minimal free energy state [6]. The genetic algorithm (GA) simulates the natural folding pathway which takes place during RNA synthesis. This not only enables new stems be added in the growing RNA chain, but also allows structures to be removed at later stages of the simulation if other pairings are found more favorable. The GA allows prediction of certain tertiary interactions, including RNA pseudo knots. Secondary structure prediction was done individually for each clade in the data set. The minimum free energy was considered from secondary structure of RNA.

3. Results and discussion

The 5S rRNA sequences size ranges between approximately 110–148 bp were used for RNA secondary structure of 37 bacteria. The size and bacterial strains used in this study were given (Table 1). The bacterial strain *Staphylothermus marinus* F1 has the lowest δG −70.40 kcal/mol and *Francisella philomiragia* subsp. *philomiragia* ATCC 25017 has higher δG −27.60 kcal/mol. Thermodynamic values (δG) was observed in 37 different bacterial strains (Table 1) and seven major clades were made (Table 2). The phylogenetic analysis of bacteria (Fig. 1) was obtained and same homology of strain located in same clade. *Enterobacteriaceae* family

Table 1
The 5S rRNA bacterial strain with RNA secondary structure delta G free energy.

S.No.	Bacterial strain	Accession No.	Bases	Free energy kcal/mol
1.	*Streptococcus agalactiae* NEM316	25010075	118	−35.70
2.	*Streptococcus pyogenes* MGAS10394	50913346	117	−35.40
3.	*Staphylothermus marinus* F1	|126464913	115	−70.40
4.	*Staphylococcus saprophyticus* subsp. *saprophyticus* ATCC 15305	73661309	118	−38.60
5	*Shigella dysenteriae* Sd197	82775382	119	−54.40
6.	*Shigella flexneri* 2a str. 2457T	30061571	119	−54.40
7.	*Streptococcus pneumoniae* D39	116515308	116	−30.50
8.	*Streptococcus pneumoniae* TIGR4	118090026	148	−39.90
9.	*Streptococcus sanguinis* SK36	125716887	116	−32.20
10.	*Streptomyces avermitilis* MA-4680,	162960844	126	−49.20
11.	*Vibrio cholerae* O395 chromosome 2	147673035	121	−39.70
12	*Bacillus subtilis* subsp. *subtilis* str. 168	50812173	119	−41.30
13.	*Bacillus cereus* ATCC 14579	30018278	119	−35.60
14.	*Campylobacter jejuni* RM1221	57236892	120	−34.00
15.	*Candidatus Blochmannia floridanus*	33519483	122	−38.80
16.	*Chromohalobacter salexigens* DSM 3043	92112136	116	−44.00
17.	*Chlamydia trachomatis* A/HAR-13	76788711	119	−38.00
18.	*Clostridium botulinum* A str. ATCC 3502	148378011	126	−36.00
19.	*Corynebacterium diphtheriae* NCTC 13129	38232642	121	−43.90
20.	*Coxiella burnetii Dugway* 5J108-111	154705721	117	−41.40
21.	*Deinococcus radiodurans* R1 chromosome 1	15805042	124	−46.40
22.	*Enterococcus faecalis* V583	29374661	116	−36.00
23.	*Escherichia coli* O157:H7 str. Sakai	15829254	120	−54.90
24.	*Haemophilus influenzae* 86-028NP	162960935	115	−44.60
25	*Francisella philomiragia* subsp. *philomiragia* ATCC 25017	167626225	115	−27.60
26.	*Helicobacter pylori* 26695	|15644634	118	−28.80
27.	*Klebsiella pneumoniae* subsp. pneumoniae MGH 78578	152968582	123	−50.20
28.	*Listeria innocua* Clip11262	16799079	110	−36.30
29.	*Legionella pneumophila* str. Paris	54295983	126	−36.50
30.	*Mycobacterium tuberculosis* H37Ra	148659757	123	−41.30
31.	*Mycobacterium tuberculosis* CDC1551	50953765	115	−41.30
32.	*Mycobacterium tuberculosis* H37Rv	57116681	115	−41.30
33.	*Pyrococcus abyssi* GE5	14518450	122	−69.40
34.	*Salmonella enterica* subsp. *enterica* serovar *Paratyphi* A str. ATCC 9150	56412276	122	−54.90
35.	*Salmonella typhimurium* LT2	16763390	122	−54.90
36.	*Shigella boydii* Sb227	82542618	119	−54.40
37.	*Streptococcus pneumoniae* R6	15902044	115	−30.20

Table 2
All these bacterial strains were categorized into the seven clades.

S.No	Clade No	Total bacteria	Bacteria name
1	I	9	*Shigella dysenteriae* Sd197 *Shigella flexneri* 2a str. 2457T *Shigella boydii* Sb 227 *Escherichia coli* O157:H7 str. Sakai, *Salmonella enterica* subsp. enterica serovar Paratyphi A str. ATCC 9150 *Salmonella typhimurium* LT2 *Klebsiella pneumoniae* subsp. pneumoniae MGH 78578 *Canidatus Blochmannia floridanus* *Haemophilus influenzae* 86-028NP
2	II	6	*Chromohalobacter salexigens* DSM 3043 *Coxiella burnetii* Dugway 5J108-111 *Vibrio cholerae* O395 *Francisella philomiragia* subsp. philomiragia ATCC 25017 *Legionella pneumophila* str. Paris *Deinococcus radiodurans* R1 chromosome 1
3	III	5	*Bacillus cereus* ATCC 14579 *Bacillus subtilis* subsp. subtilis str. 168 *Staphylococcus saprophyticus* ATCC 15305 *Clostridium botulinum* ATCC 3502 *Listeria innocua* Clip 11262
4	IV	7	*Enterococcus faecalis* V583 *Streptococcus agalactiae* NEM316 *Streptococcus pyogenes* MGAS10394 *Streptococcus sanguinis* SK36 *Streptococcus pneumoniae* TIGR4 *Streptococcus pneumoniae* D39 *Streptococcus pneumoniae* R6
5	V	1	*Chlamydia trachomatis* A/HAR-13
6	VI	1	*Streptomyces avermitilis* MA-4680
7	VII	8	*Staphylothermus marinus* F1 *Pyrococcus abyssi* GE5 *Campylobacter jejuni* RM1221 *Helicobacter pylori* 26695 *Corynebacterium diphtheriae* NCTC 13129 *Mycobacterium tuberculosis* CDC1551 *Mycobacterium tuberculosis* H37Ra *Mycobacterium tuberculosis* H37Rv

bacteria were observed in separate clade. Several stains of *Streptococcus* realized in the same clade while *Mycobacterium* spp. was seen in another clade.

Seven RNA secondary structures were investigated and have covered entire bacterial strains. The loops structure in RNA secondary structure was destabilized the stability of strains. The RNA secondary structure of bacterial strains was shown (Figs. 2 and 3) and the free energy ranges between −70.40 kcal/mol to −27.60 kcal/mol. The graphical representation of free energy and bacterial strains were shown (Fig. 4). The free energy represents bacterial stability during the course of evolution and environmental selection pressure. The higher free energy signifies that the strains cannot tolerate environmental pressure, due to the formation of multi-loops in the RNA secondary structure and thereby, destabilizes the structure. The 16S–23S intergenic spacer region of *Streptococcus salivarius*, *Streptococcus thermophilus*, and *Lactobacteria lactis* subsp. *cremoris* and the 23S–5S intergenic spacer region 2 of *S. salivarius* and *L. lactis* subsp. *cremoris* were sequenced and compared with the spacer regions 1 and 2 of other *streptococci*. High degree of intraspecific conservation was seen in *S. thermophilus* and *L. lactis*, and same types of sequences were found in *S. salivarius* and *S. thermophilus*. Secondary structure was built to show interaction between the spacer regions 1 and 2 of *S. thermophilus* and *S. salivarius*. The rapid evolution of spacer region 1 in *streptococci* is in part due to insertions and deletions of small RNA stem/loop structures [9]. The genes encode conserved metabolic functions which are polymorphic bears multiple alleles amid different strains [10].

The *E. coli* genome is highly dynamic. The complete sequenced genome of the laboratory K-12 strain and its derivatives served as a vital role in the varied laboratories of the world which proves to be use as an evidence of tremendous plasticity [11]. It was known that K-12 lineage has experienced approximately 200 lateral (horizontal) transfer events from the time it diverged from *Salmonella* about 100 million years ago and 18% of its existing genes were obtained horizontally from other species [12]. The pathogenic strain contains 1.34 million base pairs of lineage-specific DNA which includes 1387 new genes; some of them have been employed in virulence, but in many function was unknown [13].

Secondary structure models exhibited three pairs of 16S rDNA from *E. coli* and *Z. mays* chloroplast ribosome, the 18S rRNA from *S. cerevisiae* and *X. laevis* cytoplasmic ribosome, and the 12S rRNA from human and mouse mitochondrial ribosomes. The secondary structure of ribosomal rDNA has been conserved in the evolution [13]. The Influenza virus encodes conserved non-structural gene which is thermodynamically stable during evolution. The NS1 of Influenza A virus H5N1 strain varied and the phylogeny was reported. The thermodynamic free energy varied between −222.90 and −251.10 kcal/mol of the NS of Influenza virus [14].

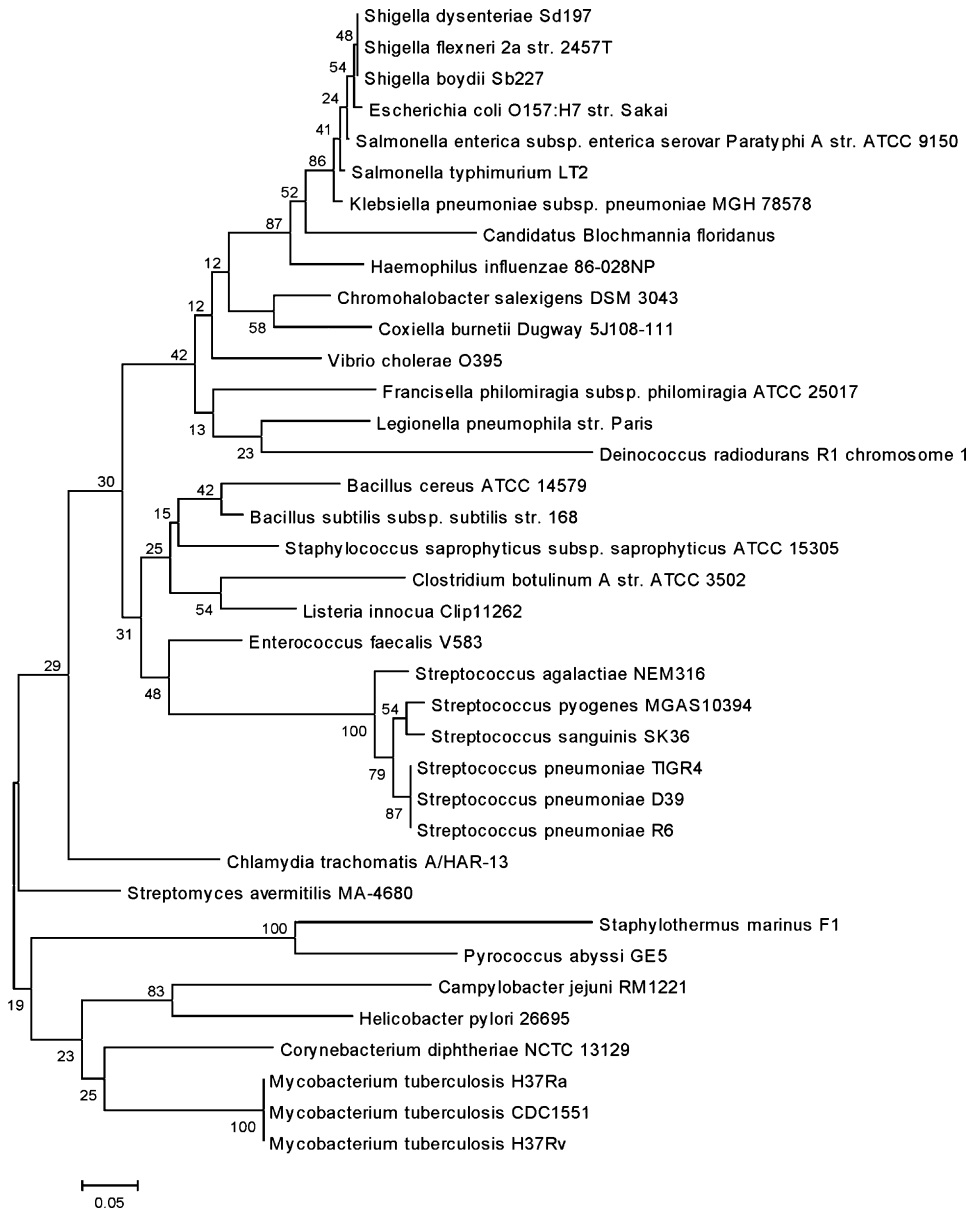

Fig. 1. Phylogenetic tree of 5S rDNA sequences of bacterial strain using NJ method and Jukes and Cantor algorithm with 100 bootstrap values.

Hepatitis C virus (HCV) possesses extensive RNA secondary structure in the core and NS5B-encoding regions of the genome. The analyses of a multiple RNA-folding patterns have predicted by MFOLD to determine the evolutionary conservation in stem–loop structures [15]. Hepatitis G virus (HGV)/GB virus C (GBV-C) causes persistent, non-pathogenic infection in a large proportion of the human population. In this study, structure of the 3′-untranslated region (3′-UTR) of HGV/GBV-C upstream NS5B coding sequence was compared. The investigation of free energies on folding, secondary structure and analysis of covariance between HGV/GBVC genotypes 1–4 were more distantly related HGV/GBV-C chimpanzee variant, the NS5B region contains long stem–loop structures of 38 internally paired nucleotides which were evolutionarily conserved between human and chimpanzee HGV/GBV-C variants [16].

The 5S rRNA gene from *Sphingobium chungbukense* DJ77 was identified. The secondary structure of 199-base-long RNA was proposed. The two-base-long D loop was the shortest among all the known 5S rRNAs. The U19-U64 non-canonical pair in the helix II region was uniquely found in strain DJ77 among sphingomonads [17]. The nucleotide sequence of *Pinus silvestyris* 5S rRNA was determined using two independent methods and compared with other plant 5S rRNAs. It shows more than 90% sequence homology with gymnosperm 5S RNAs. The free energy (δG) analysis of 5S rRNAs from gymnosperms, angiosperms and the other higher plants revealed that the free energy of this ribosomal RNA decreases in evolution [18].

In conclusion, the stability of bacterial strains based on RNA secondary structure and free energy during the evolution was reported. The role of RNA secondary and tertiary structure in governing essential bacterial processes is becoming increasingly obvious. The RNA secondary structure may facilitate further understanding of evolutionary complexity residing within the strains. The free energy of 5S rRNA region might be helpful to predict the primitive and stable bacterial strains in the evolution.

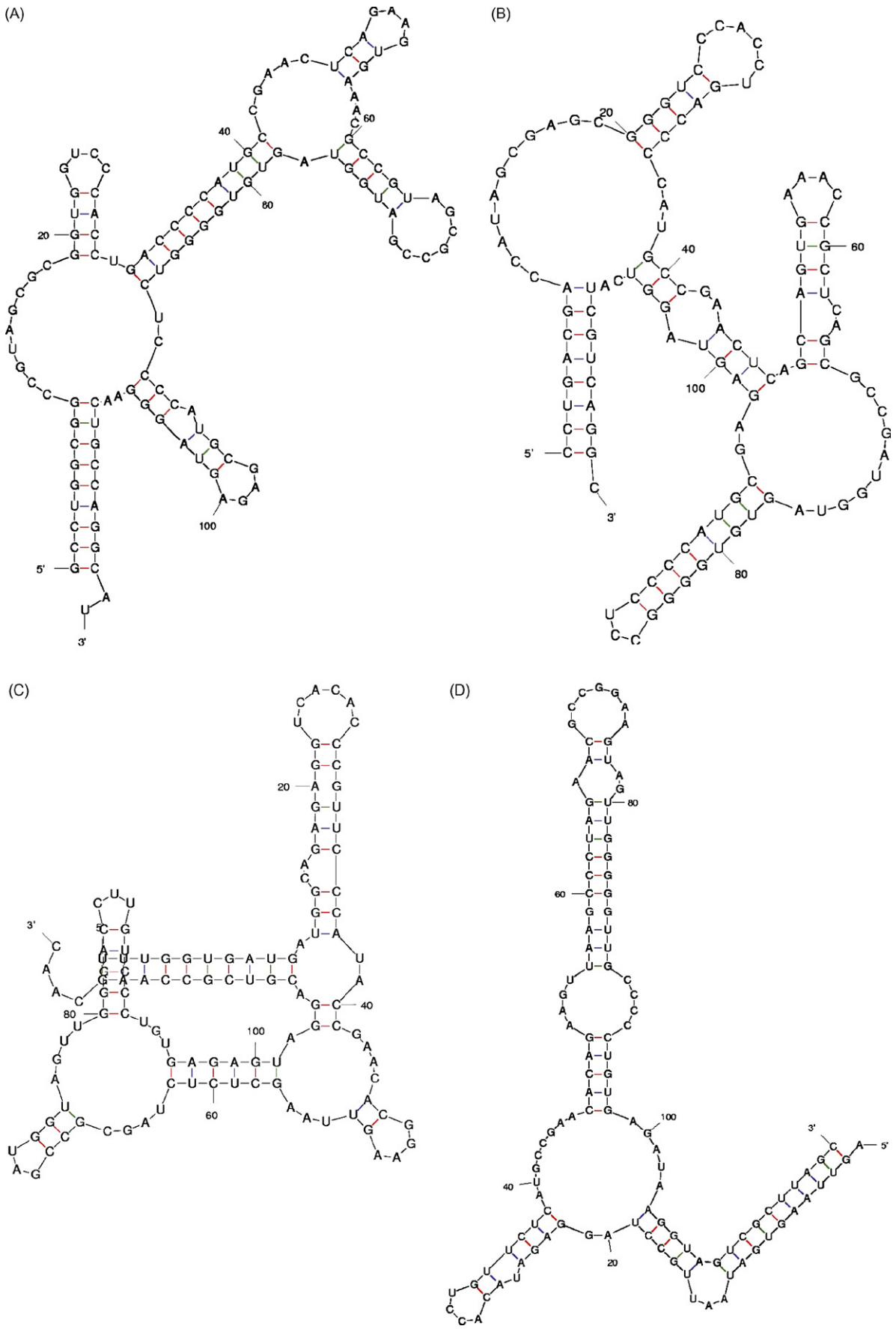

Fig. 2. The 5S rDNA secondary structures of *S. dysenteriae* Sd197, *C. salexigens* DSM 3043 *and B. cereus* ATCC 14579, *S. agalactiae* NEM316 represented for Clade I (2A), II (2B), III (2C) and Clade IV (2D).

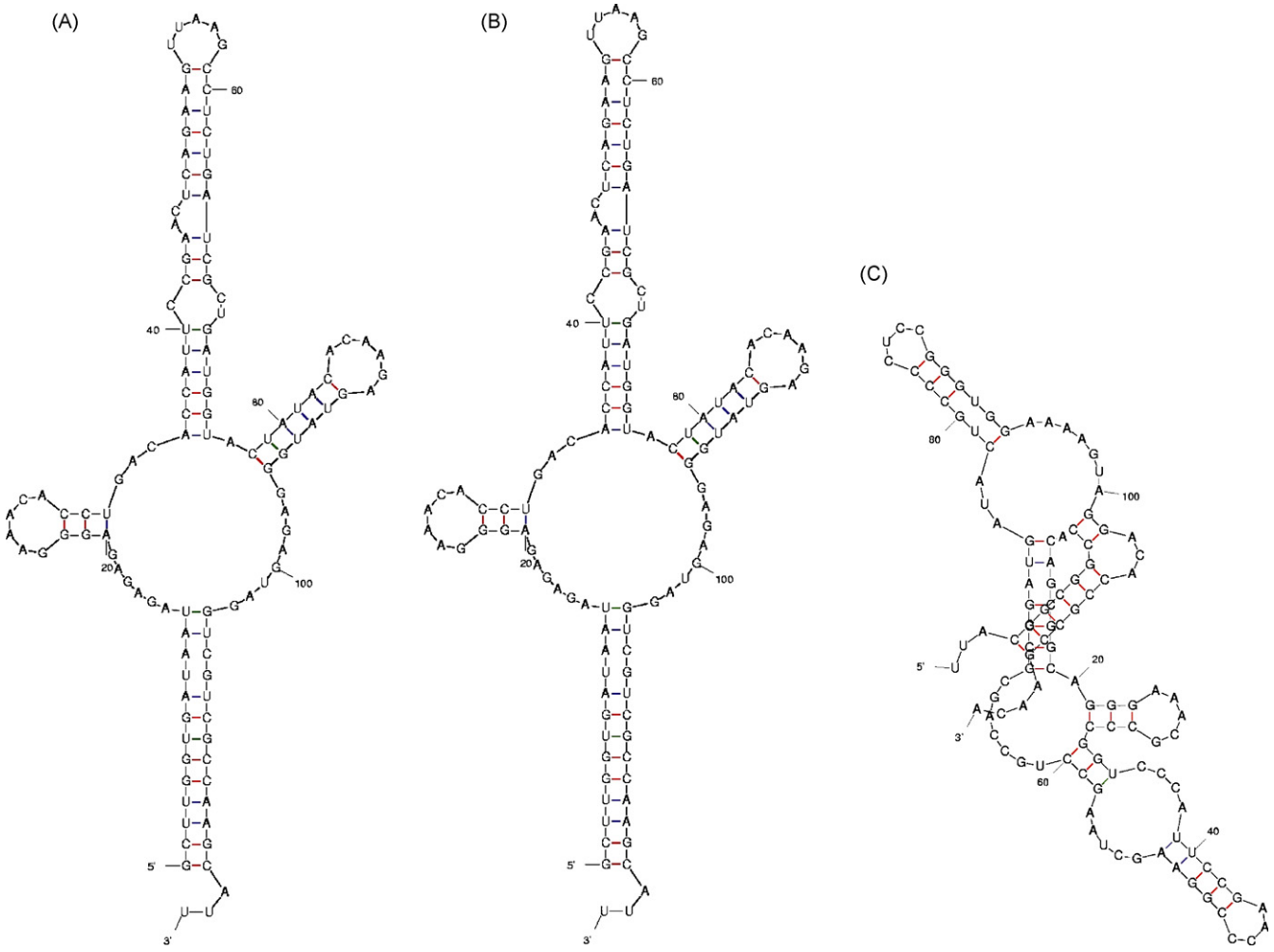

Fig. 3. The 5S rDNA secondary structures of *C. trachomatis* A/HAR-13 *and S. avermitilis* MA-4680 and *M. tuberculosis* H37Rv represented for Clades V (3A), VI (3B) and VII (3C) correspondingly.

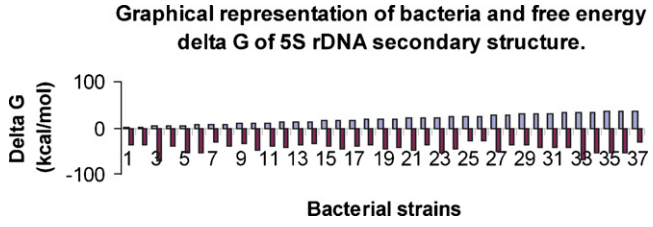

Fig. 4. Graphical representation between different bacterial strains and 5S rDNA secondary structure for free energy. The numbers indicate the strains name. 1: *Streptococcus agalactiae* NEM316, 2: *Streptococcus pyogenes* MGAS10394, 3: *Staphylothermus marinus* F1, 4: *Staphylococcus saprophyticus* subsp. *saprophyticus* ATCC 15305, 5: *Shigella dysenteriae* Sd197, 6: *Shigella flexneri* 2a str. 2457T, 7: *Streptococcus pneumoniae* D39, 8: *Streptococcus pneumoniae* TIGR4, 9: *Streptococcus sanguinis* SK36, 10: *Streptomyces avermitilis* MA-4680, 11: *Vibrio cholerae* O395 chromosome 2, 12: *Bacillus subtilis* subsp. *subtilis* str. 168, 13: *Bacillus cereus* ATCC14579, 14: *Campylobacter jejuni* RM1221, 15: *Candidatus Blochmannia floridanus*,16: *Chromohalobacter salexigens* DSM 3043, 17: *Chlamydia trachomatis* A/HAR-13, 18: *Clostridium botulinum* A str. ATCC 3502, 19: *Corynebacterium diphtheriae* NCTC 13129, 20: *Coxiella burnetii* Dugway 5J108-111, 21: *Deinococcus radiodurans* R1 chromosome 1, 22: *Enterococcus faecalis* V583, 23: *Escherichia coli* O157:H7 str. Sakai, 24: *Haemophilus influenzae* 86-028NP, 25: *Francisella philomiragia* subsp. *philomiragia* ATCC 25017, 26: *Helicobacter pylori* 26695, 27: *Klebsiella pneumoniae* subsp. *pneumoniae* MGH 78578, 28: *Listeria innocua* Clip11262, 29: *Legionella pneumophila* str. Paris, 30: *Mycobacterium tuberculosis* H37Ra, 31: *Mycobacterium tuberculosis* CDC1551, 32: *Mycobacterium tuberculosis* H37Rv, 33: *Pyrococcus abyssi* GE5, 34: *Salmonella enterica* subsp. *enterica* serovar *Paratyphi* A str. ATCC 9150, 35: *Salmonella typhimurium* LT2, 36: *Shigella boydii* Sb227 and 37: *Streptococcus pneumoniae* R6.

References

[1] J. Fredrickson, J. Zachara, D. Balkwill, et al., Geomicrobiology of high-level nuclear waste-contaminated vadose sediments at the Hanford site, Washington State, Appl. Environ. Microbiol. 70 (7) (2004) 4230–4241.
[2] W. Whitman, D. Coleman, W. Wiebe, Prokaryotes: the unseen majority, Proc. Natl. Acad. Sci. U.S.A. 95 (12) (1998) 6578–6583.
[3] C.L. Sears, A dynamic partnership: celebrating our gut flora, Anaerobe 11 (5) (2005) 247–251.
[4] T.B. Taylor, C. Patterson, Y. Hale, W.W. Safranek, J. Clin. Microbiol. 35 (1997) 79–85.
[5] X. Xu, Y. Ji, G.D. Stormo, RNA sampler: a new sampling based algorithm for common RNA secondary structure prediction and structural alignment, Bioinformatics 23 (15) (2007) 1883–1891.
[6] M. Zuker, On finding all suboptimal foldings of an RNA molecule, Science 244 (1989) 48–52.
[7] K. Tamura, J. Dudley, M. Nei, S. Kumar, MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0′, Mol. Biol. Evol. 24 (2007) 1596–1599.
[8] J.D. Thompson, T.J. Gibson, F. Plewniak, F. Jeanmougin, D.G. Higgins, The CLUSTALX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, Nucleic Acids Res. 25 (1997) 4876–4882.
[9] M. Nour, A. Naimi, G. Beck, C. Branlant, 16S–23S and 23S–5S intergenic spacer regions of *Streptococcus thermophilus* and *Streptococcus salivarius*, primary and secondary structure, Curr. Microbiol. 31 (5) (1995) 270–278.
[10] T.S. Whittam, Genetic variation and evolutionary processes in natural populations of *Escherichia coli*, in: F.C. Neidhardt (Ed.), *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology, ASM Press, Washington, DC, USA, 1996 , pp. 2708–2720.
[11] F.R. Blattner, et al., The complete genome sequence of *Escherichia coli* K-12, Science 277 (1997) 1453–1462.
[12] N.T. Perna, et al., Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7, Nature 409 (2001) 529–533.

[13] C. Zwieb, C. Glotz, R. Brimacombe, Secondary structure comparisons between small subunit ribosomal RNA molecules from six different species, Nucleic Acids Res. 9 (15) (1981) 3621–3640.
[14] P. Somvanshi, V. Singh, M. Arshad, Modeling of RNA secondary structure of non structural gene and evolutionary stability of the influenza virus through in silico methods, J. Proteom. Bioinform. 1 (2008) 219–226.
[15] A. Tuplin, D.J. Evans, P. Simmonds, Detailed mapping of RNA secondary structures in core and NS5B-encoding region sequences of hepatitis C virus by RNase cleavage and novel bioinformatics prediction methods, J. Gen. Virol. 85 (2004) 3037–3047.
[16] N.M. Cuceanu, A. Tuplin, P. Simmonds, Evolutionarily conserved RNA secondary structures in coding and non-coding sequences at the 3′- end of the hepatitis Gvirus/GB-virus C genome, J. Gen. Virol. 82 (2001) 713–722.
[17] T.D. Mashkova, M.Z. Barciszewska, A. Joachimiak, M. Nalaskowska, J. Barciszewski, Molecular evolution of plants as deduced from changes in free energy of 5S ribosomal RNAs, Int. J. Biol. Macromol. 12 (4) (1990) 247–250.
[18] H.R. Kwon, Y.C. Kim, Nucleotide sequence and secondary structure of 5S rRNA from *Sphingobium chungbukense* DJ77, J. Microbiol. 45 (1) (2007) 79–82.