

The suboptimal structures find the optimal RNAs: homology search for bacterial non-coding RNAs using suboptimal RNA structures

Josef Pánek^{1,*}, Libor Krásný², Jan Bobek^{1,3}, Edita Ježková^{1,3}, Jana Korelusová² and Jiří Vohradský^{1,*}

¹Laboratory of Bioinformatics, ²Laboratory of Molecular Genetics of Bacteria, Institute of Microbiology, Academy of Sciences of the Czech Republic, Vídeňská 1073, 14220 Prague and ³Institute of Immunology and Microbiology of the First Faculty of Medicine, Charles University in Prague and General Teaching Hospital, Studničkova 7, 128 00 Prague 2, Czech Republic

Received September 7, 2010; Revised October 25, 2010; Accepted November 3, 2010

ABSTRACT

Non-coding RNAs (ncRNAs) are regulatory molecules encoded in the intergenic or intragenic regions of the genome. In prokaryotes, biocomputational identification of homologs of known ncRNAs in other species often fails due to weakly evolutionarily conserved sequences, structures, synteny and genome localization, except in the case of evolutionarily closely related species. To eliminate results from weak conservation, we focused on RNA structure, which is the most conserved ncRNA property. Analysis of the structure of one of the few well-studied bacterial ncRNAs, 6S RNA, demonstrated that unlike optimal and consensus structures, suboptimal structures are capable of capturing RNA homology even in divergent bacterial species. A computational procedure for the identification of homologous ncRNAs using suboptimal structures was created. The suggested procedure was applied to strongly divergent bacterial species and was capable of identifying homologous ncRNAs.

INTRODUCTION

Non-coding RNAs (ncRNAs) control a variety of cellular processes in both prokaryotic and eukaryotic species. In prokaryotes, ncRNAs can affect transcription by interacting with RNA polymerase (1), act as post-transcriptional regulators by interacting with mRNAs, act as *cis* acting transcriptional regulators (riboswitches) (2), or interact with and modulate the activities of cellular proteins (3). The proportion of non-coding regions, where ncRNA genes are most frequently located, increases with the

complexity of the organism. These regions form about 98.5% of the human genome (4) and typically <20% of bacterial genomes. Although the identification of bacterial ncRNAs is of prime importance, only a minor fraction of all potential ncRNAs in bacteria have been identified so far. Experimental detection and identification of bacterial ncRNAs can be expensive and time consuming. Therefore, relatively cheap and fast computational identification of bacterial ncRNAs has become the first method of choice.

However, computational searches for bacterial ncRNAs are severely limited by weak conservation of ncRNA properties. Generally, computational searches are based on comparisons of sequences and optimal secondary RNA structures (5), which both display limited similarity, especially between divergent bacterial species. The broad limitation of the computational searches is obvious from the Rfam database (6). This database contains approximately 150 different bacterial ncRNAs, which were experimentally identified or computationally predicted. The majority of them were computationally identified only in species closely related to the species in which they were originally found.

There are four physiologically well-studied bacterial ncRNAs: M1, tm, 4.5S and 6S RNAs. The first three have been broadly identified in bacteria, whereas the identification of 6S RNA was much more limited. 6S RNA regulates protein transcription via interaction with the σ factor–RNA polymerase complex. The interaction is caused by the 6S RNA structure as it resembles an open promoter DNA region (1).

Computational identification of 6S RNA was limited to groups of related species. According to Rfam 10.0, 1700 6S RNAs are known. The Rfam seed alignment contains 154 6S RNAs that comprise a few groups of related species: Gram-negative (G⁻), mostly γ , proteobacteria

*To whom correspondence should be addressed. Tel: +420 296442190; Fax: +420 296442389; Email: panek@biomed.cas.cz
Correspondence may also be addressed to Jiří Vohradský. Tel: +420 296442190; Fax: +420 296442389; Email: vohr@biomed.cas.cz

(92 sequences), cyanobacteria (15 sequences), and Gram-positive (G⁺), low G+C bacteria of *Bacillus*, *Staphylococcus* and *Streptococcus* (38 sequences). Examples of 6S RNAs that had to be identified experimentally, even in related species, can be found, e.g. *B. subtilis* 6S-2 RNA (7–9), *Bordetella pertussis* 6S RNA (9). A comprehensive computational identification of 6S RNA found approximately 100 6S RNAs homologs in eubacteria (10), of which most were in evolutionarily related species. Of distantly related to species with known 6S RNA, only *Symbiobacterium thermophilum* 6S RNA was reported (10). 6S RNA has not been identified in almost all G⁺ high G+C species, in most G⁺ low G+C species and numerous G[–] species.

Here, we set out to identify 6S RNA in divergent bacterial species using structural similarity. Our computational identification of 6S RNA is based on structures with higher than minimal free energy (FE) (suboptimal structures). We demonstrate that the suboptimal structures better capture the functional properties of the 6S RNA molecule than any other genomic or structural features. For the example of species distantly related to already known 6S RNAs—*Streptomyces* and *Mycobacterium*—we show that such an approach can identify homologous ncRNAs in related and divergent bacterial species.

MATERIALS AND METHODS

Data and similarity measures

Genome sequences and genomic annotations were imported from the NCBI ftp site (<ftp://ftp.ncbi.nih.gov/genomes/Bacteria>). Rho-independent terminators were predicted by TransTermHP (11) with default parameters. Sequence comparisons were performed by BLAST. BLAST parameters were optimized for cross-species exploration (-r 1 -q 1 -G 1 -E 2 -W 9 -F 'm D' -U) (12). RNAdistance for structure comparisons used structure tree comparison and returned the edit distance as a pairwise structural similarity score. The smaller the value of the score the higher the similarity; a value of zero indicates identical structures.

Bacterial strains, media and northern blot analysis

Streptomyces coelicolor A3(2) M145 cells were grown in NMMP medium. At indicated time points, cells were harvested and RNA was isolated as described in (13). UV spectroscopy and agarose gel electrophoresis were used to assess the quantity and quality of total RNA samples. Total RNA samples were separated on 6% denaturing polyacrylamide gels and transferred to nylon membranes (BioRad) using a Trans-Blot semi-dry transfer cell (BioRad) (25 V, 4°C over night). Membranes were UV-cross linked. The membranes were hybridized with 5'-end-labeled oligonucleotides, which corresponded to the internal part of the 6S RNA, overnight at 42°C in ULTRAhybridization buffer (Ambion). The detection and quantification of signals were conducted using a phosphorimager (BioRad).

The *Mycobacterium smegmatis* cells were grown at 37°C in Middlebrook 7H9 medium supplemented with 0.05% Tween-80. For experiments with this organism, the *M. smegmatis* mc155 strain was used (a kind gift from Dr J. Weiser from the Institute of Microbiology in Prague, Czech Republic). RNA extractions were carried out as described in (14). Time points were taken from the exponential phase (OD₆₀₀ ~0.4), from the entry into the stationary phase (OD₆₀₀ ~1.7), and from 2 h into the stationary phase (OD₆₀₀ ~3.4). Gel electrophoresis, northern blotting and hybridization were performed as described in (13). Briefly, 4 µg of total RNA was loaded per lane onto 7% polyacrylamide gels and transferred to Amersham Hybond-N membranes. Probes were 5' ³²P-labeled oligonucleotides (Table 3), and signals were visualized by PhosphorImaging (BioRad).

Bacillus subtilis 168 and *Escherichia coli* DH5α were grown in LB medium.

Western blotting. *Mycobacterium smegmatis* mc155, *Bacillus subtilis* 168 and *Escherichia coli* DH5α cells from the stationary phase of growth (~2 h after exiting the exponential phase) were homogenized by sonication (sonicator Hielscher UP200S) in lysis buffer [20 mM Tris (pH 8.0), 150 mM KCl, 1 mM MgCl₂ and 1 mM DTT]. Equal protein amounts (20 µg) of cell lysates were electrophoresed through a reducing SDS PAGE (NuPAGE[®] 4–12% Bis-Tris Gel, Carlsbad, CA, USA) and electroblotted onto a Protran BA 85 cellulosenitrat (E) (Schleicher & Schuell, Dassel, Germany). The membrane was blocked with 5% milk and incubated with the mouse monoclonal antibody to RNA polymerase sigma 70 for 2 h (1/1000; Abcam, Cambridge, UK). The membrane was washed with TBST buffer [10 mM Tris, 150 mM NaCl, 0.05% Tween-20, (pH 7.4)] and treated with horseradish peroxidase (HRP)-linked goat-anti-mouse IgG (Fc specific)-Peroxidase antibodies (1 h, 1/1000; Sigma-Aldrich). Subsequently, the blot was incubated for 2 min with SuperSignal[®] West Pico Chemiluminescent substrate (Thermo scientific, Rockford, IL, USA), exposed on film and developed.

Immunoprecipitation. *Mycobacterium smegmatis* mc155 and *B. subtilis* 168 cells (15 ml of each) from stationary phase (~2 h after exiting the exponential phase) were pelleted, resuspended in 0.8 ml of lysis buffer (see western blotting) and sonicated. Each lysate (0.2 ml) was incubated for 3 h with 4 µl of Mouse monoclonal [2G10] antibody to RNA polymerase sigma 70 (1/50; Abcam, Cambridge, UK). Subsequently, 20 µl of Dynabeads Protein A (Invitrogen) was added, incubated for 2 h, washed five times with lysis buffer and electrophoresed on a 7% polyacrylamide gel along with 4 µg of total RNA from each organism. After northern blotting, the blot was divided in two, with each part containing total RNA and immunoprecipitation from one organism. Subsequently, hybridization was carried out with 5'-labeled probes against *B. subtilis* 6S RNA (positive control) and *M. smegmatis* Ms1 (for probe sequence see Table 4).

RESULTS

Sequence and structure similarity of 6S RNAs

To address how sequence similarity corresponds to structural similarity in the case of 6S RNA, 147 6S RNA sequences from 147 bacterial species were downloaded from the Rfam 10.0 database (15). Only unique bacterial sequences were used. Minimum free energy (MFE) structures of the 6S RNAs were generated using UNAFold (16). To make a comparison between the sequence and structure similarity, similarity scores for sequences and structures were computed for all selected 6S RNAs. The pair-wise sequence similarities were computed using BLAST (17), and structural similarities for MFE structures were computed by means of RNAdistance (18). The similarity scores formed two distance matrices (one for sequence similarity and one for structural similarity). Each matrix was used as an input to a hierarchical clustering algorithm forming two clustering trees (Figure 1) one for sequence similarity (clusters 1–5) and one for structural similarity (clusters I–V). The lines between the trees of Figure 1 connect clusters, including single 6S RNAs. The width of each line is proportional to the number of 6S RNAs included in the connected clusters. The lines indicate that similar sequences did not have similar structures and *vice versa*: lines connect single sequence clusters with multiple structure clusters, and single structure clusters make connections to multiple sequence clusters. For example, cluster 5, which included sequences of 6S RNA of *E. coli* and other G– relatives, is connected with all structural clusters except for cluster II. Similar results were obtained for other 6S RNAs. Such structure/sequence dissimilarity can be observed even for closely related species. This point can be demonstrated by 6S RNAs of *E. coli*, *Bordetella paraptentis* and *Vibrio vulnificus* (Figure 2), which are all closely related G– bacteria. All three species were found in different clusters (clusters I, III and IV in Figure 1), which indicates their mutual structural dissimilarity, while their sequences were similar and clustered to a single cluster 5 (Figure 1).

Similarity of suboptimal 6S RNA structures

In the search for a better structural model for 6S RNA than the MFE one, we focused on suboptimal structures, i.e. the structures with free energies (FEs) higher than the MFE. Several 6S RNAs with dissimilar optimal and similar suboptimal structures were found. Examples of 6S RNAs from *V. vulnificus* and *B. paraptentis* (Figure 3, cf. with Figure 2) with 6S RNA-like structures had the 19th and fifth lowest FEs, respectively. These examples encouraged us to test whether suboptimal structures can characterize 6S RNA. For this purpose, we computed suboptimal secondary structures for all 147 6S RNAs with FEs within 10% of the MFE, but no more than 75 structures for a single RNA were generated. We hypothesized that when clustered, unlike in the case of optimal structures, there should exist a cluster with 6S RNA-like suboptimal structures, which would contain most of the 6S RNA species. In total, we obtained 3452

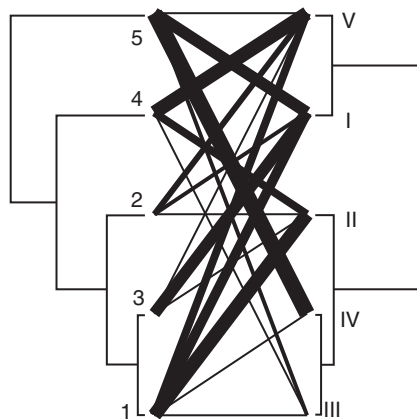


Figure 1. A comparison of the sequence and structural similarity of 6S RNAs by optimal structures. Using both sequence similarity (leftmost tree) and similarity of optimal structures (rightmost tree), 147 6S RNAs were clustered. The similarity scores were pair-wise BLAST *E*-values for sequences and RNAdistance scores for structures. A hierarchical clustering algorithm with 'ward' linkage was used for both trees. Lines between trees connect the positions of sequences and structures of single 6S RNAs. The width of each line is proportional to the number of 6S RNAs included in the line. For proportions, refer to the size of the clusters in Table 1.

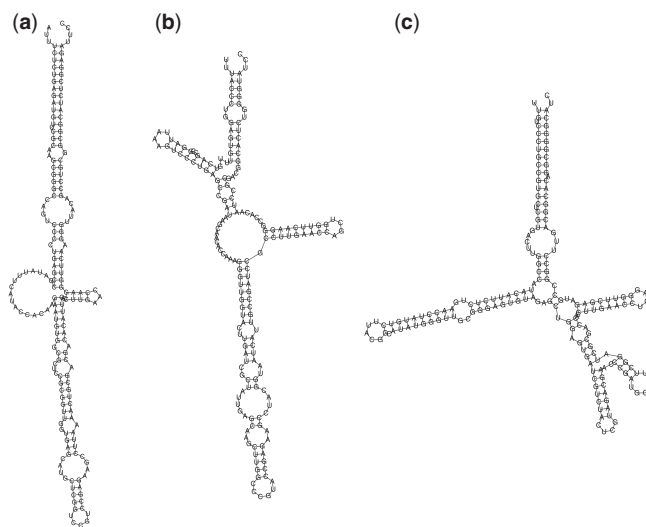


Figure 2. The optimal 6S RNA structures of *E. coli* (a), *V. vulnificus* (b) and *B. paraptentis* (c).

structures—23 structures on average for a single 6S RNA. The similarity matrix was computed for the 3452 structures by the RNAdistance program and 6S RNAs were clustered into five clusters (Table 1). Because each bacterial species was represented by an average of 23 structures, we had to find a way to assign a position in the clustering tree to a specific species. For this purpose, the similarity between the 6S RNA consensus structure (copied from the Rfam database 10.0) and each suboptimal structure of the individual species was computed. The suboptimal structure of a given species most similar to the consensus structure was then identified in the tree, and this position was labeled as representative for the given species. This procedure was repeated for all 147 individual 6S RNAs.

Unfortunately, this approach failed because only 90 6S RNAs (out of 147, 61%) could be labeled (Table 1). Other 6S RNAs exhibited structural dissimilarity of their suboptimal structures to the consensus structure, which was higher than the obvious similarity threshold of 80. This result implies that the consensus structure does not represent all 147 6S RNA. A consensus structure is derived from multiple sequence alignments, i.e. it is based on sequence similarity. We noted above that the sequence and structural similarity do not match each other; therefore, the consensus structure determined from the consensus sequence only represents a limited set of 6S RNAs, as we found here. Another more representative structural template has to be found.

Representative structure template search

We expected that a structural template could be found among the 3452 alternative suboptimal structures. Such

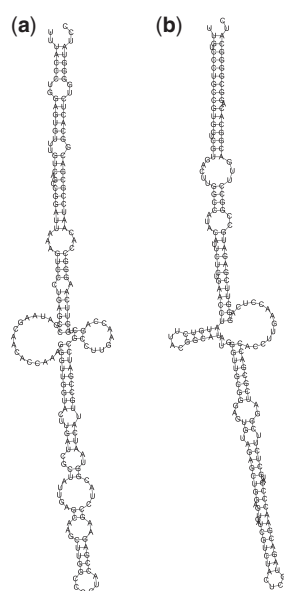


Figure 3. The suboptimal 6S RNA structures of *V. vulnificus* with the 19th lowest FE (a), and *B. parapatensis* with the fifth lowest FE (b).

a structure has to exhibit the highest similarity with suboptimal structures of most of the 147 individual 6S RNA. To find the template, similarity scores for each of the 3452 suboptimal structures to all remaining suboptimal structures were computed. Out of these, those having a score lower than the similarity threshold of 80 were excluded (the lower dissimilarity, the lower score), which prevented a bias towards dissimilar structures. Each of the 3452 suboptimal structures was considered as a potential template. For each of the templates, the following procedure was performed:

```

Select potential template i
  suma = 0
  select 6S RNA j
    select all suboptimal structures of 6S RNA j with
      similarity to template i above
      threshold (similarity score  $\leq 80$ )
    find suboptimal structure with lowest score b or
      b = 0 if not found
    if b  $\neq 0$ , save b, suma = suma + 1
  next 6S RNA j + 1
  meanb = average(saved b)
next template i + 1
  
```

This process was repeated for all 3452 potential templates, and for each of them, two values were recorded: the number of 6S RNAs with a non-zero scores (*suma*) and the average of recorded non-zero scores for the potential template (*meanb*). The template was selected as the one having highest *suma* and lowest *meanb*. This procedure selected a template that was most represented among all 147 6S RNAs and had the highest similarity to at least one of the suboptimal structures of each 6S RNA. The number and average structural similarity of the four best-scoring structures, and for comparison, the consensus 6S RNA structure are shown in Figure 4. The template with the largest number of best-scoring structures (140 out of 147) was the structure with the third lowest FE of *Synechococcus* sp. WH 7803, i.e. the structure characterized 140 out of 147 known 6S RNAs (Table 1). For comparison, the consensus structure characterized only 90 6S RNAs (Table 1),

Table 1. Clustering of suboptimal 6S RNA structures

Cluster	Total no. of optimal/suboptimal structures	Total no. of 6S RNAs ^a	No. of 6S RNAs by consensus structure ^b	Structural similarity by consensus structure ^c	No. of 6S RNAs by suboptimal structure ^d	Structural similarity by suboptimal structure ^e
I	1470	140	74	70	125	62
II	512	67	10	71	10	59
III	789	94	0	—	0	—
IV	628	57	6	72	5	69
V	53	5	0	—	0	—

^aThe number of 6S RNAs with at least one structure either optimal or suboptimal in a cluster.

^bThe number of 6S RNAs with the best scoring optimal/suboptimal structure to the consensus 6S RNA structure in a cluster.

^cThe mean of the pair-wise RNAdistance scores of best-scoring suboptimal structures to consensus 6S RNA structure.

^dThe number of 6S RNAs with the best-scoring optimal/suboptimal structure to the structure of *Synechococcus* sp. WH 7803 6S RNA with the third lowest FE in a cluster.

^eThe mean pair-wise RNAdistance scores of the best-scoring suboptimal structures to the structure of *Synechococcus* sp. WH 7803 6S RNA with the third lowest FE.

as shown above. The distribution of characterized 6S RNAs in structural clusters for the *Synechococcus* template and the consensus 6S RNA structure are shown in Table 1.

Identification of 6S RNAs in suboptimal structure clusters using new structural template

Using the structure with the third lowest FE of *Synechococcus* sp. WH 7803 as a template, a similarity matrix of RNAdistance scores was computed between the template and all 3452 individual suboptimal structures. The positions of individual 6S RNA species in the

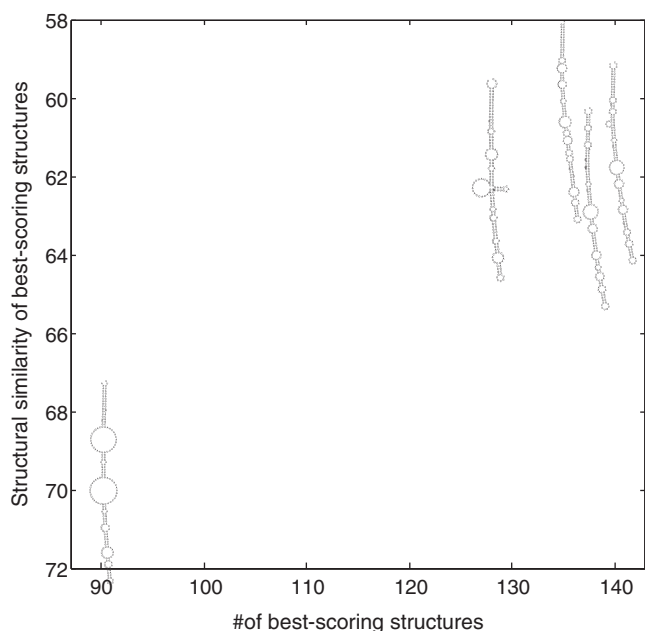


Figure 4. Search for suboptimal 6S RNAs structural templates. *X*-axis shows number of 6S RNAs represented by a template, *y*-axis shows structural similarity of the best-scoring structures to a template. Four best templates and 6S RNA consensus structure are shown. The templates are, represented by thumbnails centered at positions given by values on *x*- and *y*-axes. The templates are as follows (listed from the left to right and from bottom to the top of the figure): consensus 6S RNA structure (Rfam), *Synechococcus* sp. WH 8102 with fifth lowest FE, *E. coli* optimal structure, *Synechococcus* sp. WH 7803 with third lowest FE, *C. watsonii* with 16th lowest FE.

Table 2. Predicted *Streptomyces* ncRNAs

Name	Organism	Genome locus	FE - MFE ^a	Structure similarity to template ^b	Synteny ^c
Sc1	<i>Streptomyces coelicolor</i>	6370627..6370817	-3.4	74	DNA topoisomerase IV subunit B (3e-165)/serine protease (0)
Sc2	<i>Streptomyces coelicolor</i>	3934820..3934630	-4.4	78	hypothetical protein, oxidoreductase (0)/morphological differentiation-associated protein (0)

^aThe FE of a suboptimal structure minus MFE.

^bThe RNAdistance score, which is the lowest of three scores to optimal template structures from *E. coli*, *Synechococcus* sp. WH 7803 with the third lowest FE and *C. watsonii* with the 16th lowest FE are shown.

^cThe left/right flanking genes. The compared species are *S. coelicolor* and *S. avermitilis*. Two names used for a single flanking gene indicate different annotations for the compared species. The BLAST *E*-values of the sequence similarity are in parenthesis.

clusters of suboptimal structures were defined as the position of the best scoring suboptimal structure of the given 6S RNA. Clustering of the suboptimal structures identified five major clusters (I–V, Figure 5). Individual 6S RNAs in clusters I–V were connected with corresponding 6S RNAs in the previously computed tree of sequence similarity (the same as in Figure 1). The width of the connecting lines in Figure 5 is proportional to the number of 6S RNAs connecting the structural and sequence clusters. It can be seen that unlike in Figure 1, where the optimal structures were used, the suboptimal structures of different sequences are grouped in cluster I.

Cluster I contained 1470 suboptimal structures (out of a total of 3452, 42%) with various FEs that corresponded to 140 (95%) out of 147 6S RNAs. Almost half of all suboptimal structures were similar to each other and were clustered into a single cluster. The existence of such a cluster indicates that a 6S RNA structural template

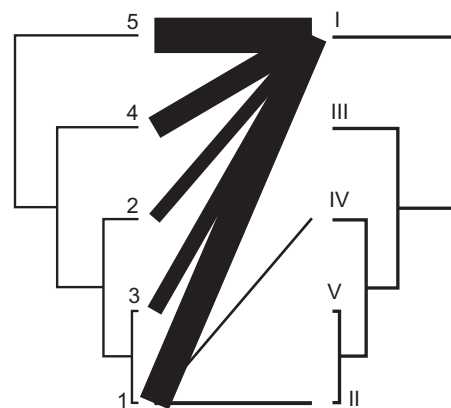


Figure 5. A comparison of the sequence and structural similarity of 6S RNAs using suboptimal structures. Using both sequence similarity (leftmost tree) and similarity of optimal structures (rightmost tree), 147 6S RNAs were clustered. The similarity scores were pair-wise BLAST *E*-values for sequences and RNAdistance scores of the best-scoring optimal/suboptimal structures to the *Synechococcus* sp. WH 7803 structure with the third lowest FE. A hierarchical clustering algorithm with 'ward' linkage was used for both trees. Lines between trees connect the positions of sequences and best-scoring structures of single 6S RNAs. The width of each line is proportional to numbers of 6S RNAs included in the line. For proportions, refer to the size of the clusters in Table 2.

characteristic for dissimilar sequences and different bacterial species exists. It also shows that the structural template does not exist among optimal structures. The comparison demonstrated that suboptimal structures could characterize 6S RNA, i.e. they followed the homology of 6S RNAs via structural similarity regardless of the level of sequence similarity.

A 6S RNA homology search in *Mycobacterium* and *Streptomyces* using suboptimal structures

To test the hypothesis that our structural templates can identify physiological structurally similar ncRNAs in divergent bacteria without the use of sequence similarity, we applied it to the homology search in *Streptomyces* and *Mycobacterium*, G+ high G+C bacteria, which are evolutionary distant from others. The 6S RNA in these species were neither identified, nor experimentally proven. Only in our previous bioinformatic search (19) was some indication of its existence in *S. coelicolor* given.

Because ncRNAs should be conserved in closely related species, we used the intergenic regions (IGR) of *S. coelicolor* sequences, which are similar to those of *S. avermitilis*, for the search. The similarity was measured by BLAST, and sequences with $E\text{-value} \leq 1 \times 10^{-3}$ were used. The sequences produced 768 ncRNA candidate genes, which were 185- to 200-nt long, flanked by a predicted terminator within 50 nt either upstream or downstream of the potential 3'-ends of candidate sequences. Up to 75 optimal/suboptimal structures within 10% of MFE were generated for a single ncRNA candidate sequence, which gave a total of 28 028 optimal/suboptimal structures. The structures were matched to structural templates, which are mentioned below, to identify structurally similar 6S RNA candidates.

The three best structural templates from the search described above (Figure 4), the *E. coli* 6S RNA optimal structure, the *Synechococcus* sp. WH 7803 6S RNA structure with the third lowest FE and the *Crocospaera watsonii* 6S RNA structure with the 16th lowest FE, were used. Structural similarity was computed using the RNAdistance program. The candidate structures were each matched to all of the templates producing three scores, and the best out of the three scores ≤ 80 identified structures of similar ncRNAs. The RNAdistance score threshold 80 was used to filter out pairs of dissimilar structures. The use of three templates delivered structural variability that increased the chance of retrieving good structural hits among suboptimal structures. The matching procedure identified 177 suboptimal structures of eight candidate ncRNAs in *S. coelicolor*. Two sequence candidates (Sc1 and Sc2) were found to be conserved in at least five other *Streptomyces* species (with BLAST $E\text{-values} \leq 1 \times 10^{-20}$), with conserved synteny and similar structures in conserved IGRs in *S. avermitilis* (Table 2 and Figure 6). The expression of these candidates was confirmed experimentally under standard growth conditions (Table 3). Sc2 was found to interact with RNA polymerase in a complex with HrdB, the *Streptomyces* housekeeping sigma factor, which is a presumption of the functionality of 6S RNA (K. Mikulík, paper in

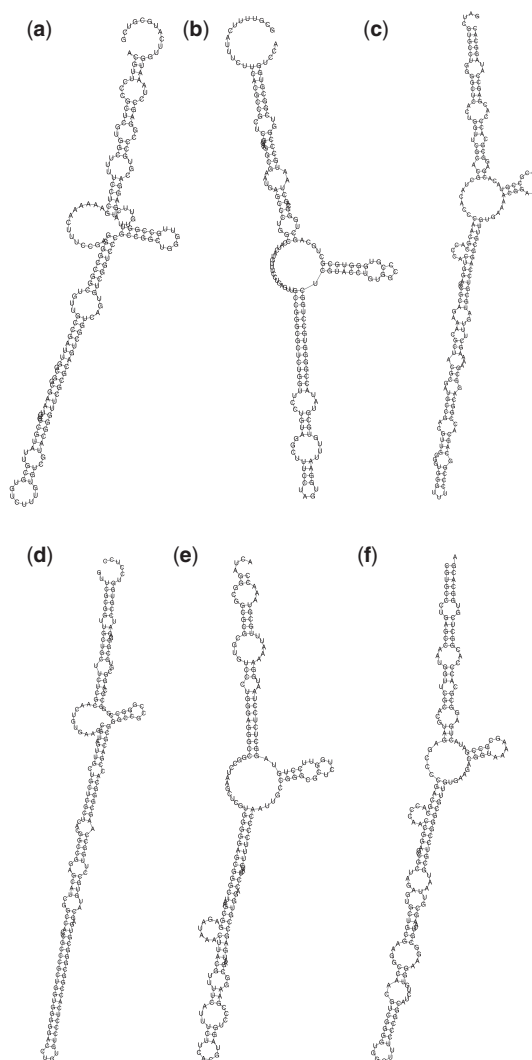




Figure 6. 6S RNA candidates. (a) *S. coelicolor* Sc1, (b) *S. coelicolor* Sc2, (c) *M. smegmatis* Ms1, (d) *S. avermitilis* predicted homolog to Sc1, (e) *S. avermitilis* predicted homolog to Sc2, (f) *M. avium paratuberculosis* predicted homolog to Ms1.

Table 3. Expression of the *S. coelicolor* candidate ncRNAs by northern blot hybridization

ncRNA candidate	Expression ^a					Probe
	g3h	3d	6d			
Sc1						AGTCCTTTGTAC TACCGTCCC GAGTAAG
						
Sc2	40h	2d	3d	6d	g5.5h	AATGTGCCGTAT TGCGTGT
						

^aExpression after 40 h (40h); 2, 3 and 6 days (2d, 3d and 6d, respectively) and after 5, 5 and 3 h during germination (g5, 5h and g3h, respectively) was tested.

preparation). *Scl* was found to be adjacent to *hrdb* gene indicating their coordinated expression.

In *Mycobacterium*, ncRNAs structurally similar to 6S RNA were identified in *M. smegmatis*. IGR sequences conserved in *M. smegmatis* and *M. avium paratuberculosis* were used (BLAST *E*-values $\leq 1 \times 10^{-3}$). Under the same conditions and using same parameters as for *Streptomyces*, 190 candidate terminated sequences were identified, which gave 5130 suboptimal structures. By matching to the structural templates, 86 similar suboptimal structures of four candidate ncRNAs were identified (called Ms1–4). However, none of them passed the conservation test used for *Streptomyces*: either sequences, synteny or structures were not conserved in other *Mycobacterium* species. The best of these candidates was Ms1, which had a sequence and structure broadly conserved in other *Mycobacterium* species (as in *M. avium paratuberculosis*, see Figure 6c and f). For this candidate, only synteny was not conserved, but in a rather interesting way: flanking genes were annotated differently in *M. smegmatis* and *M. avium paratuberculosis* although strong sequence similarity was detected (BLAST *E*-values = 0). Among the *Mycobacterium* species, in which Ms1 was found to be conserved by sequence



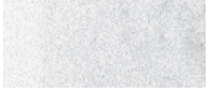

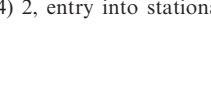
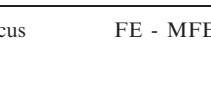
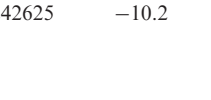
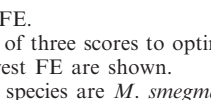
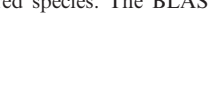

and structure, only the flanking genes of *M. smegmatis* annotation differed.

As none of the four *M. smegmatis* candidates passed fully the conservation criteria, the expression of all of them was tested experimentally under three growth phases (the exponential phase, the entry into stationary phase and the stationary phase, Table 4). One of the four, Ms1, was found to be expressed in all phases. Its characteristics are shown in Table 5. Ms1 produced two transcripts ~280- and 230-nt long (Figure 7a). The expression maximum of both transcripts was in the stationary phase. During all three tested growth phases, the longer transcript prevailed. However, interaction of Ms1 with sigA–RNA polymerase complex was not confirmed experimentally (Figure 7b), suggesting that Ms1 is not 6S RNA, but a novel *Mycobacterium* ncRNA (Figure 6c and f).

DISCUSSION

Using an example ncRNA (6S RNA), suboptimal RNA structures were demonstrated here to be a property capable of improving the identification of bacterial ncRNAs. About 150 bacterial ncRNAs have been

Table 4. Expression of *M. smegmatis* candidate ncRNAs by northern blot hybridization

ncRNA candidate	1	2	3 ^a	Probe
Ms1				GTCGTGGCCGTCGCTTTTCGAAACTACGC
Ms2				CGGGTCACAGCCCAACGTAAGTGCCTCAAC
Ms3				AAGACTTCGACGTGCGCGACCACCGCAAAC
Ms4				CCAAACCCCCACACCACCGGTTTCGTAAC

^a1, exponential phase (OD600 ~0.4) 2, entry into stationary phase (OD600 ~1.7) 3, 2 h into stationary phase (OD600 ~3.4)

Table 5. Predicted *Mycobacterium* ncRNA

Name	Organism	Genome locus	FE - MFE ^a	Structure similarity to template ^b	Synteny ^c
Ms1	<i>Mycobacterium smegmatis</i>	6242435..6242625	-10.2	74	morphological differentiation-associated protein, hypothetical protein (0)/transcriptional regulator, IclR family protein, HAD-superfamily protein subfamily protein IB hydrolase (0) ⁵

^aThe FE of a suboptimal structure minus MFE.

^bThe RNAdistance score, which is the lowest of three scores to optimal template structures from *E. coli*, *Synechococcus* sp. WH 7803 with the third lowest FE and *C. watsonii* with the 16th lowest FE are shown.

^cThe left/right flanking genes. The compared species are *M. smegmatis* and *M. avium paratuberculosis*. Two names used for a single flanking gene indicate different annotations for the compared species. The BLAST *E*-values of the sequence similarity are in parenthesis.

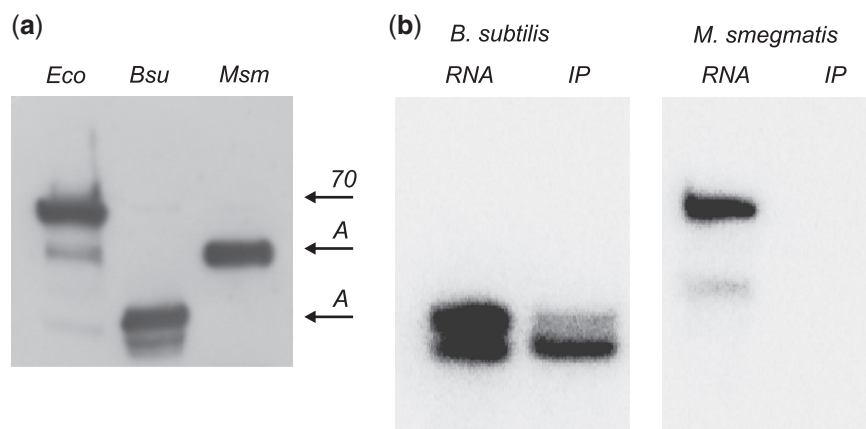


Figure 7. Immunoprecipitation of 6S RNA. (A) Western blot of total protein from *E. coli*, *B. subtilis* and *M. smegmatis* probed with monoclonal antibody against σ^{70} from *E. coli*. The antibody reacts with the main housekeeping sigma factors from all three organisms. (B) Northern blot of total RNA (RNA) and the products of immunoprecipitation experiments (IP). The *B. subtilis* blot (positive control) was probed with an oligonucleotide against 6S RNA demonstrating the presence of this ncRNA on *B. subtilis* RNAP containing σ^A . The *M. smegmatis* blot was probed with an oligonucleotide against Msl RNA, demonstrating that it is likely not a real 6S RNA.

identified so far; however, most were identified in closely related bacterial species and are waiting for identification in other species. We showed that ncRNA identification can be substantially improved using suboptimal RNA structures.

It was demonstrated here that structures with the lowest FE did not characterize the native structure of our example ncRNA, which is functional in the cell. We showed that the best characterizing structure appeared among models with higher than minimal FE, i.e. the suboptimal structures. The FEs of suboptimal structures used throughout our analysis were still quite similar to the minimal one (a 10% limit was used); therefore, their inclusion in the analysis was quite logical.

We demonstrated that to find 6S RNAs, and most probably other ncRNAs, in the intergenic regions of the genome, it was necessary to identify a structural template consistent with the known functional properties of the given ncRNA. At the same time, the template had to be common for most of the known ncRNAs of the same type. Such a template, as shown here for 6S RNA, can be identified among the suboptimal structures.

The presented search through suboptimal structures indicated that more than one structural template that would characterize a single ncRNA may exist. To follow the homology of ncRNAs across bacterial species, more structural models for a single RNA may be required. If you look at Figure 4, the four rightmost structures are very similar in principle, and any one of them can serve as a template. Indeed, using all of them sequentially as templates, we were able to identify 6S RNA in divergent species of *Streptomyces*. Such a procedure effectively overcomes the lack of specificity of consensus ncRNA structures inferred using sequence similarity.

Suboptimal structures of whole ncRNA molecules were used. Within a range of FE, these structures were all matched to each other for all analyzed ncRNAs, and by this computationally straightforward procedure, we demonstrated that 23 suboptimal structures for a single RNA were enough to increase the number of similar

structures of homologous ncRNAs threefold in comparison to optimal structures (from 32 to 95%).

The structural similarity does not automatically imply ncRNA (or even RNA) homology. It means that using the homology search one can also find ncRNAs that are not homologous (as *M. smegmatis* Msl gene, see paragraph 'A 6S RNA homology search in *Mycobacterium* and *Streptomyces* using suboptimal structures' in 'Results' section). Computational biology is not able to give an ultimate answer so far, but is only able to substantially narrow the number of samples requiring wet-lab experiments. By our opinion—and as it is demonstrated by the presented results—the only ultimate proof of ncRNA homology is the wet-lab experiment that unfortunately never is easy and simple.

The matching of suboptimal structures used here principally differs from earlier uses of suboptimal structures for increasing the accuracy of modeling RNA helices (20) and for increasing the accuracy of ncRNAs prediction without a need for homology (21). These papers use suboptimal structure to predict sequence regions that may contain unknown ncRNAs. We do not predict a specific sequence but rather identify ncRNAs from a pool of candidate sequences. Therefore, our goal is to find structural similarity that could reflect functional similarity. The presented results indicate that suboptimal structures better capture the ncRNA homology than other known ncRNA properties and that the use of exclusively optimal structures in computational searches may preclude successful identification of ncRNAs.

CONCLUSIONS

We present this analysis of 'wrong' (suboptimal) instead of the usual analyses of 'right' (optimal) to demonstrate that from a biological point of view, suboptimal structures can be more natural than optimal ones. This notion is demonstrated here through the ability of suboptimal structures to predict ncRNAs. The suboptimal structures

substantially improved the identification of bacterial ncRNAs represented here by 6S RNA.

Although the paper deals with bacterial ncRNAs, a similar approach can also be applied to the identification of eukaryotic ncRNAs. The analysis of all databased 6S RNAs allowed us to identify general principles of homology shared among different bacterial species and allowed the implementation of the presented method for identification of new potential 6S RNAs. The same procedure would have to be adopted for the eukaryotic ncRNAs and the presented algorithmic pipeline modified according to the found principles. The general principle of the conservation of biologically active ncRNA structure among the computed suboptimal structures should be conserved also for the eukaryotic ncRNAs.

Unlike optimal and consensus structures, sequence similarity and genomic properties, suboptimal structures have not been commonly used in computational searches. Optimal structures, which are used commonly, were shown here to be dissimilar, even for 6S RNAs of related species, whereas suboptimal structures effectively demonstrated expected similarity. This result indicates that the first precondition of improvement in the computational identification of bacterial ncRNAs requires replacement of the commonly used sequence similarity, predicted genomic properties and exclusive use of optimal and/or consensus structures by the use of more natural and functionally relevant suboptimal structures. The suboptimal structures apparently better capture the structural properties of the ncRNA molecule. As more structural models need to be computationally compared than when using optimal structures, the demand for computational power increases when using suboptimal structures. Our results indicate that the biologically relevant biocomputational searches for ncRNAs are not that cheap and rapid, as has been generally thought, but require the extensive computation of many structural candidates, among which those relevant to the given ncRNA are found.

FUNDING

Czech Science Foundation (grant No. 303/09/0475, 310/07/1009 P302/10/0468 to J.P., J.V. and J.B., respectively); Institutional Research Concept, Ministry of Education, Youth and Physical Culture of the Czech Republic (AV0Z50200510 and 2B06065); TRIOS (to L.K.). Funding for open access charge: Czech Science Foundation (Grant No. 303/09/0475).

Conflict of interest statement. None declared.

REFERENCES

1. Wassarman, K.M. and Storz, G. (2000) 6S RNA regulates E. coli RNA polymerase activity. *Cell*, **101**, 613–623.

2. Tucker, B.J. and Breaker, R.R. (2005) Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol.*, **15**, 342–348.
3. Vogel, J. and Sharma, C.M. (2005) How to find small non-coding RNAs in bacteria. *Biol. Chem.*, **386**, 1219–1238.
4. Wolfsberg, T.G., McEntyre, J. and Schuler, G.D. (2001) Guide to the draft human genome. *Nature*, **409**, 824–826.
5. Gardner, P.P. and Giegerich, R. (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, **5**, 140.
6. Gardner, P.P., Daub, J., Tate, J.G., Nawrocki, E.P., Kolbe, D.L., Lindgreen, S., Wilkinson, A.C., Finn, R.D., Griffiths-Jones, S., Eddy, S.R. *et al.* (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res.*, **37**, D136–D140.
7. Ando, Y., Asari, S., Suzuma, S., Yamane, K. and Nakamura, K. (2002) Expression of a small RNA, BS203 RNA, from the *yocI-yocJ* intergenic region of *Bacillus subtilis* genome. *FEMS Microbiol. Lett.*, **207**, 29–33.
8. Suzuma, S., Asari, S., Bunai, K., Yoshino, K., Ando, Y., Kakeshita, H., Fujita, M., Nakamura, K. and Yamane, K. (2002) Identification and characterization of novel small RNAs in the *aspS-yrvM* intergenic region of the *Bacillus subtilis* genome. *Microbiology*, **148**, 2591–2598.
9. Trotochaud, A.E. and Wassarman, K.M. (2005) A highly conserved 6S RNA structure is required for regulation of transcription. *Nat. Struct. Mol. Biol.*, **12**, 313–319.
10. Barrick, J.E., Sudarsan, N., Weinberg, Z., Ruzzo, W.L. and Breaker, R.R. (2005) 6S RNA is a widespread regulator of eubacterial RNA polymerase that resembles an open promoter. *RNA*, **11**, 774–784.
11. Kingsford, C.L., Ayanbule, K. and Salzberg, S.L. (2007) Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol.*, **8**, R22.
12. Korf, I., Yandell, M. and Bedell, J. (2003) *BLAST*. O'Reilly & Associates, Sebastopol.
13. Hopwood, D.A., Bibb, M.J., Chater, K.F., Kieser, T., Bruton, C.J., Kieser, H.M., Lydiate, D.J., Smith, C.P., Ward, J.M. and Schrempf, H. (1985) *Genetic Manipulation of Streptomyces – A Laboratory Manual*. The John Innes Foundation, Norwich.
14. Krasny, L., Tiserova, H., Jonak, J., Rejman, D. and Sanderova, H. (2008) The identity of the transcription +1 position is crucial for changes in gene expression in response to amino acid starvation in *Bacillus subtilis*. *Mol. Microbiol.*, **69**, 42–54.
15. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. and Eddy, S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.
16. Markham, N.R. and Zuker, M. (2005) DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res.*, **33**, W577–W581.
17. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
18. Hofacker, I.L. (2004) RNA secondary structure analysis using the Vienna RNA package. *Curr. Protoc. Bioinformatics*, **Chapter 12**, Unit 12.12.
19. Panek, J., Bobek, J., Mikulik, K., Basler, M. and Vohradsky, J. (2008) Biocomputational prediction of small non-coding RNAs in *Streptomyces*. *BMC Genomics*, **9**, 217.
20. Jaeger, J.A., Turner, D.H. and Zuker, M. (1989) Improved predictions of secondary structures for RNA. *Proc. Natl Acad. Sci. USA*, **86**, 7706–7710.
21. Tran, T.T., Zhou, F., Marshburn, S., Stead, M., Kushner, S.R. and Xu, Y. (2009) De novo computational prediction of non-coding RNA genes in prokaryotic genomes. *Bioinformatics*, **25**, 2897–2905.