Computational Identification of Small RNAs in Clostridium acetobutylicum and Prediction of mRNA Targets

Yili Chen, Dinesh Indurthi, Shawn Jones, Eleftherios Terry Papoutsakis
University of Delaware, DE, USA

ABSTRACT

Small non-coding bacterial RNAs (sRNAs) have been found in genomes of many model organisms. Many studies show that sRNAs play important regulatory roles in a variety of cellular processes in bacteria. Clostridium acetobutylicum is a gram-positive, rodshaped anaerobe that produces acetone, butanol and ethanol through fermentation of a variety of carbon sources. It regained interest for potential use in vehicle biofuel production. However, the transcriptional regulation of *C. acetobutylicum* has not been well understood and sRNA regulation is ignored in previous studies. We predicted sRNAs and their mRNA targets in *C. acetobutylicum* ATCC 824 with various computational approaches. The non-coding sRNAs were predicted in the intergenic regions of *C. acetobutylicum* ATCC 824 genome using an integrated computational method, sRNAPredict2. The prediction was followed by Q-RT-PCR and Northern blot validation. The mRNA targets of the validated sRNAs were then predicted by searching in the genome for strong sRNA-mRNA duplexes based on sequence match and the hybrid profile prediction. In summary, 133 sRNAs were predicted, 117 on the chromosome and 16 on the plasmid. Experiments verified the expression of 7 out of 15 randomly selected putative sRNAs. The study identified a group of highly conserved sRNAs that are associated with 16S Ribosomal RNA in genomic location. The high expression level of these sRNAs suggests their potentially important regulation function in *C. acetobutylicum*.

INTRODUCTION

Small non-coding bacterial RNAs (sRNAs) play important regulatory roles in a variety of cellular processes. They are typically 50–500 nucleotides in length and nearly all sRNA species identified to date are encoded in intergenic regions (IGRs). These functional RNA molecules normally do not possess a protein-coding function. Most of them act as post-transcriptional regulators by interacting with specific mRNA targets, modulating target stability and/or translation initiation. In the past few years, new experimental strategies and computational methods have been developed demonstrating that the number of sRNAs in genomes of model organisms is much higher than was previously anticipated.

Since the discovery of first set of sRNAs in E coli by accident (1), several genome-wide methods for sRNA discovery have already been developed. Many studies combined computational searches with experimental validation of selected candidates (2,3). Using a comparative genomic screen approach, Rivas et al (3) predicted 275 sRNAs in E. coli and > 11 out of the 49 tested candidates were experimentally verified. With the availability of the increasing number of bacterial genome sequences, such strategy successfully discovered many sRNAs not only in E coli but also in many other bacteria (4-9). For example, Livny et al. (8) developed a computer program, sRNAPredict, and identified 32 putative intergenic sRNAs in V. cholera, among which 6 were verified. In addition to the identification of sRNA encoding genes, some researchers also tried to identify sRNA targets. Beside the classical genetic

approaches or microarray-based target screening, much of the recent success in identifying sRNAs targets has come from the bioinformatics aided predictions (5,10-12).

Clostridium acetobutylicum (*C. acetobutylicum*) is a gram-positive, rodshaped anaerobe that produces acetone, butanol, ethanol through fermentation of a variety of carbon sources. It recently regained interest for potential use in vehicle biofuel production. Although *C. acetobutylicum* has been studied for decades (13-16), little is known about sRNAs in this microorganism. So far, there are 23 cis-regulating riboswitches reported in the Rfam database (17), but only three sRNA (tmRNA, SRP bact, 6S) were reported. This number is much smaller than E coli. For example, in E coli K12 MG1655, 42 sRNA has been reported in Rfam database.

In this study we did a computational prediction of sRNAs in *C. acetobutylicum* ATCC 824 followed by experimental validation. The sRNAs was predicted by using multiple genetic characteristics commonly associated with sRNA-encoding genes. Expression of the predicted sRNAs was examined by Q-RT-PCR and Northern blot. The study also suggested a few mRNA targets predicted with bioinformatics approach.

MATERIALS AND METHODS

1. Computational Prediction of Intergenic sRNA

sRNAPredict is a computational tool which uses coordinate-based algorithms to integrate the respective positions of individual predictive features of sRNAs and predict putative intergenic sRNAs (8). In our study the second version of this program, sRNAPredict2, was used and the predictive features were prepared as described in Livny's article (18).

The genomic sequences and genome annotations of *Clostridium acetobutylicum* ATCC 824 and its partners, *Clostridium beijerinckii*, *Clostridium_botulinum_A_ATCC_19397* and *Clostridium_perfringens_ATCC_13124*, were downloaded from NCBI. The tRNAs, rRNAs, previously annotated sRNAs and roboswitches were downloaded from Rfam database (17). The intergenic conserved regions between *Clostridium acetobutylicum* ATCC 824 and each of the other strains were identified with WU BLAST 2.0 (19). An E-value cut-off of $1x10^{-10}$ was applied. The putative intergenic rho-independent transcription terminators were predicted with TransTerm (20) and RNAMotif (21). The putative terminators should be no more than 20 nt downstream of the 3' end of the conserved IGRs and with confidence 96% or higher in the prediction using TransTerm. The sequence regions that likely represent conservation of RNA secondary structure were predicted with QRNA (3). A window size of 100 and a slide position of 50 were used when running the QRNA prediction.

All the predictive features were then fed into sRNAPredict2. We set the program parameters as such it only searched for sRNA sequences of 50-550 nt in the intergenic regions of *Clostridium acetobutylicum* ATCC 824 genome.

2. Computational Prediction of sRNA Targets in *C. acetobutylicum*

The reverse and complement sRNA sequence was blasted against the *C. acetobutylicum* ATCC 824 genome. The blastn parameters were set follows: Match/Mismatch = (1,-1); Gap Costs = (Existence: 2, Extension: 1); Expect threshold = 10; Word size = 7. This allows relatively low similar alignments to be found by Blast. The hybrid profiles of selected sRNA and their mRNA target candidates were then predicted with UNAFold (22,23).

3. Validation of sRNA Candidates

Bacterial strains, growth conditions, and maintenance. *C. acetobutylicum* ATCC 824 (Manassas, VA) was used as the WT strain in this study. Strains were stored at -85°C in clostridial growth medium (CGM) (24) containing 15% glycerol and revived by plating onto 2xYTG (16 g/l tryptone, 10 g/l yeast extract, 4 g/l NaCl, 5 g/l glucose, and 15 g/l agar, pH 5.8) agar-solidified plates under anaerobic conditions at 37°C. Single colonies at least 5 days old were transferred to tubes with 10 ml CGM supplemented with 80 g/l of glucose, buffered with 30 mM acetate, and adjusted to pH 7.0. The tubes were then heat shocked at 80°C for 10 min and transferred to an anaerobic incubator at 37°C.

Fermentations. An initial culture of 250 ml of CGM was inoculated from a tube grown to an A_{600} of 0.6-0.8. This initial culture was grown to an A_{600} of ~0.6 and then used to inoculate twelve subcultures of 50 ml each with a 10% inoculum. Four of these subcultures were allowed to grow unstressed, four were stressed with butyrate, and four were stressed with butanol. For the butyrate stress, 175 µl of butyric acid (Riedel-de Haën) was added at an A_{600} of 0.8, and for the butanol stress, 250 µl of 1-butanol (Fisher) was added at an A_{600} of 0.8.

RNA sampling, isolation, and cDNA generation. Samples were collected by centrifuging 3 to 10 ml of culture at 5,000xg for 10 min, 4°C and storing the cell pellet at -85°C. For the unstressed cultures, samples were taken at 6, 12, 18, and 30 hours. For the stressed cultures, samples were taken at 30 min and 1 hour after stress. RNA was isolated according to (25) with the following modification. After individual cell pellets were resuspended in 220 µl of SET buffer with lysozyme (20 mg/ml, Sigma) and proteinase K (4.55 U/ml, Roche), all cell pellets were combined and mixed together. 32 aliquots of 220 µl were then processed as described in (25). To create two pools of RNA, 25 µl of purified total RNA from each processed aliquot was combined together to give ~800 µl. For cDNA generation, the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems) was used according to manufacturer's instructions. For each reaction, 2 µg of total RNA was reverse transcribed using random primers, and then diluted to 20 ng/µl.

Quantitative Reverse Transcription Polymerase Chain Reaction (Q-RT-PCR). For each Q-RT-PCR reaction, 1 µl of the reverse transcription reaction was mixed with 1 µM of each gene/sRNA specific primer (Table 1), SYBR Green PCR Master Mix (Applied Biosystems), and nuclease-free water up to 25 µl. Samples were run on an iCycler iQ5 Real-Time PCR Detection System (Bio-Rad Laboratories) following the manufacturer's instructions. Samples were run in triplicate on the same 96 well plate with three no template controls (in which instead of 1 µl of the reverse transcription reaction, 1 µl of nuclease-free water is added). Following amplification, a melt curve was performed from 55°C to 95°C at 0.5°C/10 sec, to ensure only one product was generated.

Table 1. Primers for Q-RT-PCR Verification

Gene/sRNA	Forward Primer	Reverse primer
6S 1	GAACCTACAGTTCAAACAAGGGAG	ACGATAGGTGGGTGTCCTCA
6S 3	CGCCAAGCTCTTATCTTGAACCTACAG	TCCAAAATACCGCTGCTCTT
CAC0428	AACTTCTAGTATGGCAGCTTT	AGCGTTGCAAGTACGGCTAT
CAC0681	AGTCGTATGCTCAGTTCCTGTTGA	TCCGTCTCCAATTTTTCCTG
CAC1094	GAAGATGACGACCATCCATTTGGC	GCGGCAAACATTATGAGTATTGCTTCA
CAC1322	GGAGCAGGTGTTATTGGATGCTCA	GCAGAATTTGCCTTGCTTGTTCCC
CAC2139	GATCCTAGTACAGTTAAATTCACAGC	TTGTGGTCTCCGTTTTGA
CAC2179	GCGAATATTTCGAATCCTGTGGAC	CAAAGGCTGTGGTGCTTCTTTAGG

CAC2614	TACAGAGCATGGAAGGCTATGGCT	ACCAACCCCTCGAAGTCTTT
CAC2957	CAGGTTGTAGCAGATGCAATGATGGT	CCTGCTTCAACGACTGGAAT
CAC3313	GGAGAAGAAAGAGAGGTTATCGCT	GACAAAGAGAATCCCATTTGGGAAGAC
CAP0060	GAACACATATCCAAGAGCACACACA	TGCGGGAAAGAATTTCAAAC
ch1	CAATCCGCTGTAGCAGGGTTGAAT	TTCACTTACCGCTGCTTCCT
ch6	GGACTTAATATACATGACGTAGAATC	CACCCTTTTCAAGCCATTTT
ch10	GTTCGCTGCACTAGACAGCTTAAT	CATGTTGTCACGTGAGTGCAGCAA
ch21	GAGGGTATCTAAGCTAACGACAAGAG	TCCTTCGTCTGGACTTGCTT
ch23	TCTTGATTAACTCCTCATTGACTT	TTTCCTGCATAAGTACAACAAAAA
ch25	TTCCGGGTAGCATCGCTTGAATCT	AGCCGAGTTCTGTATTCGACA
ch28	GAGCATTTCATACATAAGTTCCGTGT	TTGAATGCCCATGACCATAA
ch33	AGTTTGGAAAGGCTATTGATTT	GCTGGGCTGCCATAATAAAT
ch37	TGGGCATTATAATAGCGTCAAAGA	ACGCCACACCTAAACAATCC
ch41	TCATGCTGTAAGTGTGTGC	ACACCCTCTTTACTTATGTATT
ch43	CAGAATCGCTGTATACTGTGTAATGTA	ACCTTTCGCCAAAAGTAGGA
ch52	ATCCTTTGATAAGGAAGAGTAGCC	ATCACACCACCCTCAGCTCT
ch83	AGAGTGGCTTATAGATGTTAGT	TCGCAAATCTATTCCTCTTTCTG
pd9	AGTATCGGGAATACAAAGTCTGAT	CCTCCTGCATAAACCCCTCT
pd14	CTGCAAATAGAAATTAAGTAGGTCT	GGGTGTTACAGCACCCTATTG
SRP_bact	AATTGGGTCCCACGCAACGGAAAT	TCAGATTTATCCACGGCACA

Probe preparation for Northern analysis. Probes were designed for the detection of two predicted sRNA sequences: ch1 and ch25. Probe template was PCR amplified from one of the cDNA pools. The ch1 PCR primers were the same as the ones used for Q-RT-PCR (Table 1), and the ch25 PCR primers were 5'-GGAGTGGCCCGCTCTGCTTCCGGG-3', for the forward primer, and 5'-ACACTAAGCACGAAACCTAGTGTT-3', for the reverse primer. The probes were labeled with [α-^{32}P]dCTP (3,000 Ci/mmol) using the NEBlot Kit from New England Biolabs (Ipswich, MA), following the manufacturer's instructions, and unincorporated [α-^{32}P]dCTP was removed using illustra ProbeQuant G-50 columns (GE Healtchcare, Buckinghamshire, UK), following the manufacturer's instructions.

Northern blot and analysis. Twenty-microgram samples of total RNA were run on a 2.0% MOPS-formaldehyde agarose gels with 0.05 µg/ml ethidium bromide for 2.45 hours at 80 volts. Gels were imaged under UV light to ensure quality of RNA and to check for even loading of all lanes. Gels were rinsed twice with DEPC-treated water, incubated in 10 gel volumes of 10x SSC for 40 min, and transferred to a 0.45 µm pore-sized, positively charged membrane (Roche Applied Sciences, Indianapolis, IN) by capillary action using 10x SSC as a transfer buffer overnight. Membranes were fixed by heating at 80°C for 2 hr. Membranes were initially stained with methylene blue to confirm RNA transfer and destained by washing in 0.2x SSC and 1% SDS for 15 min at room temperature. Membranes were prehybridized with Ultrahyb Ultrasensitive Hybridization Buffer (Ambion, Austin, TX) for 2 hr at 42°C with gentle agitation. Probes were denatured by boiling for 5 min, cooled on ice for 5 min, and then added to prewarmed hybridization buffer. The prehybridization solution was poured off, the probe/hybridization buffer was added, and membranes were incubated at 42°C for 12-16 hr

with gentle agitation. Finally, membranes were washed twice in 2x SSC, 0.1% SDS for 15 min at 42°C, and twice more in 0.1x SSC, 0.1% SDS for 15 min at 42°C.

RESULTS

1. Prediction of sRNAs

SRNAs are located in intergenic regions (IGRs) in the genome, we therefore searched for sRNAs in the 3,277 IGRs of *C. acetobutylicum*. Most sRNAs are conserved only among closely related species and relatively few sRNAs have been identified on the basis of their sequence homology to previously known sRNAs (26). Therefore, we analyzed the conservation of IGRs across *Clostridium acetobutylicum* ATCC 824, *Clostridium beijerinckii*, *Clostridium_botulinum_A_ATCC_19397* and *Clostridium_perfringens_ATCC_13124* with WU-BLAST2. To ensure the significance of the conservation, we set the BLAST E-value cut-off as 1e-10. The sRNAs were searched only in these conserved IGRs. The putative Rho-independent terminator is an important marker of the end of transcripts in bacteria. In this study, it was identified using TransTerm (20) and RNAMotif (27). A rule used in TransTerm prediction was that putative terminators should be no more than 20 nt downstream of the 3' end of the conserved IGRs and with confidence 96% or higher.

All the above features were fed into RNAPredict2, which is a computational tool that uses coordinate-based algorithms to integrate the respective positions of individual predictive features of sRNAs and predict putative intergenic sRNAs (8,28). In total, we predicted 133 sRNA in the genome of *C. acetobutylicum* ATCC 824, of which 117 were on the chromosome and 16 were on the plasmid. In the 133 predicted non-coding RNA on the chromosome, 102 were novel and 31 matched the annotated *cis*-regulatory RNA element or riboswitches on the chromosome in Rfam database (17).

2. Validation of sRNA Prediction.

In order to experimentally verify the predictions, Q-RT-PCR was performed on 15 predicted sRNA sequences, which were largely randomly selected. Their cycle threshold (Ct) values were compared to both positive and negative control sequences to determine whether they are expressed or not, and if expressed, their relative expression level (Figure 1). Three of the positive controls (6S 1, 6S 3, and signal recognition particle (SRP bact)) are noncoding RNA genes identified in almost all sequenced bacteria(29-31), while the remaining four genes (CAC0681, CAC2139, CAC1322, CAC2957) were chosen because they are always expressed during a batch culture but at different expression levels (25). Conversely, the six negative control genes were chosen because they were never above the threshold of expression during a batch culture (25). Though genes below the threshold of expression are assumed to not be expressed, these genes could still be expressed but at very low levels compared to the rest of the transcriptome. The Ct values from three replicates of both pools were averaged together and standard deviations were calculated.

Comparing the Ct values of the two controls, three regions could be identified: genes which are expressed, genes which are either lowly expressed or not expressed, and genes which are not expressed (Figure 1). Though several negative and positive control genes overlap in the middle region, the standard deviations of the negative control genes display more variance than most of the positive control genes.

Of the 15 predicted sRNA tested, 7 were definitively expressed, 5 were either lowly or not expressed, and 3 were not expressed (Figure 1). Three of the sRNA (ch10, ch52, and ch25) had Ct values comparable to or less than SRP bact, which is involved in translation and

targeting of proteins to the cell membrane, indicating possible roles in global cellular function for these sRNA. Alternatively, one of these sRNA, or any of the expressed ones, could be highly upregulated during only one stage of growth or under butyrate or butanol stress, but since a pool of RNA was used, it is impossible to separate out the sRNA which are constantly expressed and only ones expressed under a certain condition. Regardless, with only three sRNA predictions being definitively not expressed, the confidence in the predictions should be quite high, and it should be noted, that the three sRNA not expressed could still be expressed under a certain stress condition not tested.

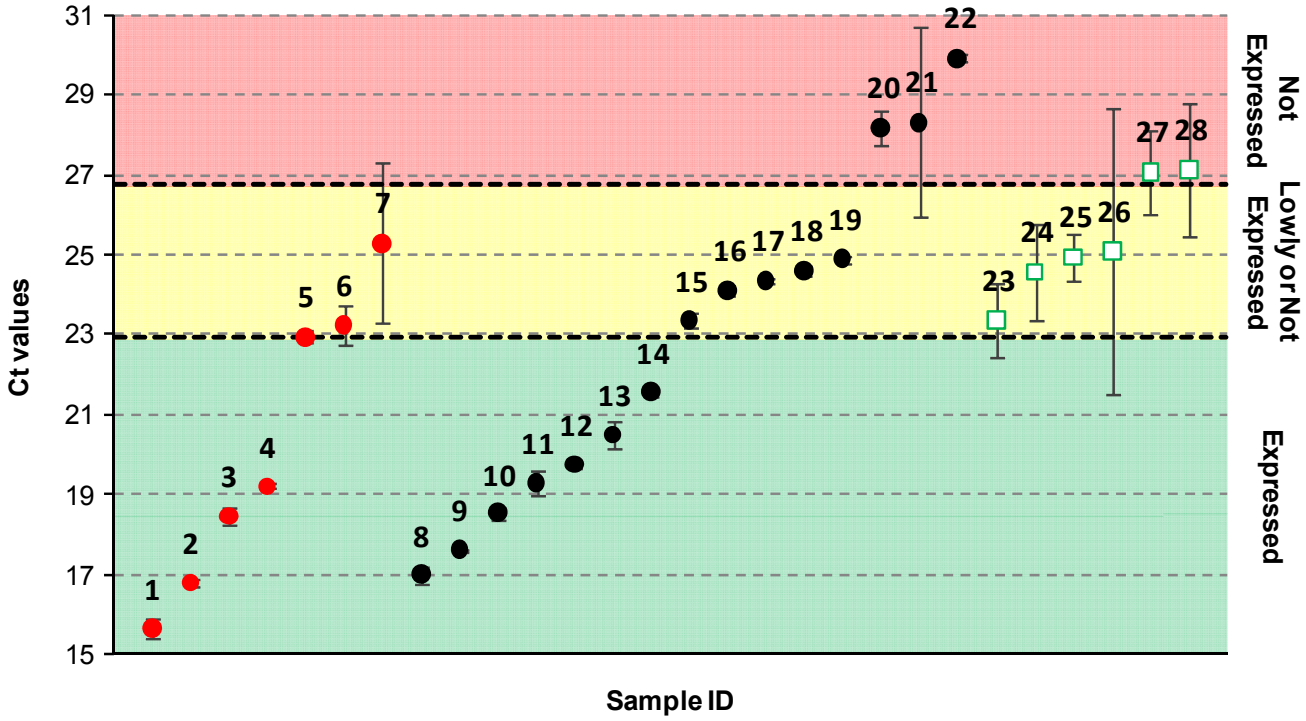


Figure 1. Q-RT-PCR cycle threshold (Ct) values of sRNA and genes tested. sRNA sequences experimentally tested (black circles, ●): ch10 (8), ch52 (9), ch25 (10), ch6 (11), pd9 (12), ch1 (13), ch43 (14), ch21 (15), ch83 (16), ch33 (17), ch28 (18), ch37 (19), ch41 (20), ch23 (21), and pd14 (22). Positive controls (red cirlces, ●): 6S 1 (1), 6S 3 (2), SRP_bact (3), CAC0681 (4), CAC2139 (5), CAC1322 (6), and CAC2957 (7). Negative controls (open green squares, □): CAC3313 (23), CAC0428 (24), CAC1094 (25), CAP0060 (26), CAC2179 (27), and CAC2614 (29). Genes and sRNA samples within the shaded green area are expressed, while those within the shaded yellow and red areas are either lowly expressed or not expressed. The standard deviation between 6 replicate samples is shown by the error bar for each sample.

To further validate the predictions, two sRNA were chosen for Northern blot confirmation: ch1 and ch25. Ch25 was highly expressed on the order of SRP bact, while ch1 was expressed but to a lesser extent. When the two RNA pools were probed, a single band was seen for both ch1 and ch25 (Figure 2). The band for ch1 appears to be around 400 nt (Figure 2, middle panel), which corresponds to the prediction of 314 nt for this sRNA. In contrast, the ch25 band appears to be around 300-400 nt (Figure 2, right panel), but the predicted length was only 138 nt. Thus, the predictions may not necessarily provide the correct length for the sRNA, but did

predict a large enough length for identification. Also, the ch25 band appears more intense than the ch1 band, indicating that ch25 is more highly expressed than ch1, as shown in the Q-RT-PCR.

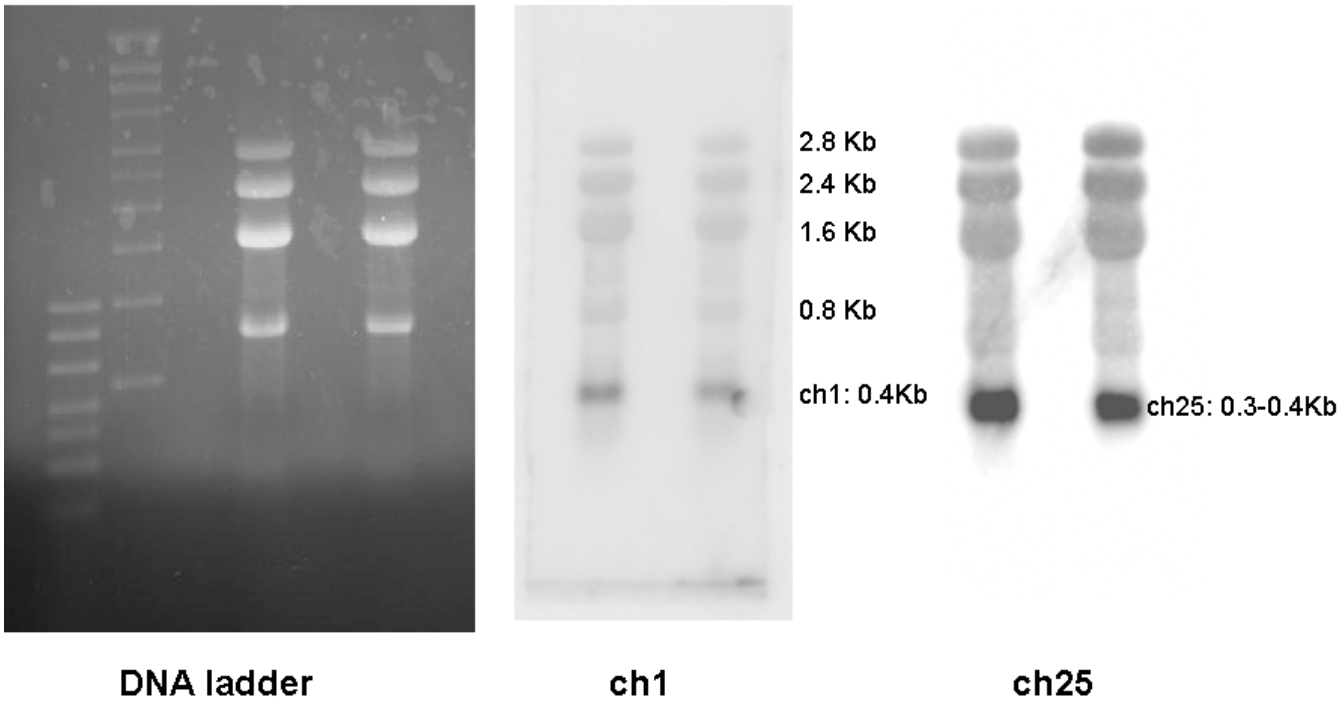


Fig2. Detection of sRNAs of ch1 and ch25 by northern blots blots. **DNA ladder:** 18ug of pooled total RNA were resolved on a 1% denaturing gel and stained with Ethidium bromide (which acts as a loading control). Four prominent rRNA bands shown are 2.8kb (23S), 2.4kb, 1.6kb (16S) and 0.8kb respectively. RNA markers used are 0.1-1kb from Agilent (0.1, 0.2, 0.3, 0.4, 0.6, 0.8 and 1 kb), and 0.5 – 9kb RNA millennium markers from Ambion (0.5, 1, 1.5, 2, 2.5, 3, 4, 5, 6, 9 kb). **ch1:** Northern blots were probed with p32 labeled 145nt ch-1 double stranded DNA. Primers used are the same as the ones Shawn used for Q-RT PCR. Hybridization was performed at 42^{o}C using Ambion's Ultrahyp hybridization buffer for 16-20hrs. Blots were exposed to phosphor screen for 24hrs and imaged using Typhoon imager. **ch25:** Northern blot was probed with p32 labeled 138nt ch-25 double stranded DNA. Primers used for probe synthesis.

Among the tested sRNA candidates, ch10 had the highest expression. Blast of this 360 nt sequence against the *C. acetobutylicum* ATCC 824 genome identified 10 other copies with significant similarity (E < 5e-40). All these 11 copies contain a highly conserved ~130 nt core sequence (Figure 3). It is interesting that all these copies are distributed on the inverse strand of the intergenic regions preceding rRNA-16S ribosomal RNA (16S rRNA) genes. It is not a coincidence that each of the 11 rRNA-16S rRNA genes scattering on *C. acetobutylicum* ATCC 824 genome neighbors with a highly conserved putative sRNA sequence. Q-RT-PCR validated the expression of most of these sequence copies. Five copies (ch10_1, ch10_2, ch10_3, ch10_9 and ch10_11), for which unique primers were able to be designed, showed high expression levels (Ct value < 17). In the other copies, ch10_5, ch10_6, ch10_7 and ch10_8 are identical to each other. Q-RT-PCR also showed high expression of this identical set although it cannot distinguish which one was actually expressed. In the remaining text, this identical sequence set will be named as NRS.

```
Ch10_6     168612   AGTTTGATATAATACAGATGCTAATCTGTATTCCTTTAGCTCATCGCTAAATTTATTTAA  168553
Ch10_7     336180   AGTTTGATATAATACAGATGCTAATCTGTATTCCTTTAGCTCATCGCTAAATTTATTTAA  336121
Ch10_8     341319   AGTTTGATATAATACAGATGCTAATCTGTATTCCTTTAGCTCATCGCTAAATTTATTTAA  341260
Ch10_5     163473   AGTTTGATATAATACAGATGCTAGTCTGTATTCCTTTAGCTCATCGCTAAATTTATTTAA  163414
Ch10_3     158335   AGTTTGATATAATACAGATGCTAGTCTGTATTCCTTTAGCTCATCGCTAAATTTATTTAA  158276
Ch10_1     9709     AGTTTGATATAATACAGATGCTAATCTGTATTCCTTTAGCTCATCGCTAAATTTATTTAA  9650
Ch10_4     254688   AGTTTGATATAATACAGATGCTAATCTGTATTCCTTTAGCTCATCGCTAAATTTATTTAA  254629
Ch10_10    1105276            ATACAGATGCTAATCTGTATTCCTTTAGCTCATCGCTAAATTTATTTAA  1105228
Ch10_9     331042   AGTTTGATATAATACAGATGCTAGTCTGTATTCCTTTAGCTCATCGCTAAATTTATTTAA  330983
Ch10_2     3337462    TTTGATATAATACAGATGCTAATCTGTATTCCTTTAGCTCATCGCTAAATTTATTTAA  3337519
Ch10_11    310729             ATACAGATGCTAATCTGTATTCCTTTAGCTCATCGCTAAATTTATTTAA  310681

Ch10_6     168552   ATAGAATA-AATTCTATTCAAAAGAATTGCTGGTTCTTTAATTCTCTGTTTAATTTTCAA  168494
Ch10_7     336120   ATAGAATA-AATTCTATTCAAAAGAATTGCTGGTTCTTTAATTCTCTGTTTAATTTTCAA  336062
Ch10_8     341259   ATAGAATA-AATTCTATTCAAAAGAATTGCTGGTTCTTTAATTCTCTGTTTAATTTTCAA  341201
Ch10_5     163413   ATAGAATA-AATTCTATTCAAAAGAATTGCTGGTTCTTTAATTCTCTGTTTAATTTTCAA  163355
Ch10_3     158275   ATAGAATA-AATTCTATTCAAAAGAATTGCTGGTTCTTTAATTCTCTGTTTAATTTTCAA  158217
Ch10_1     9649     ATAGAATA-AATTCTATTCAAAAGAATTGCTGGTTCTTTAATTCTCTGTTTAATTTTCAA  9591
Ch10_4     254628   ATAGAATA-AATTCTATTCAAAAGAATTGCTGGTTCTTTAATTCTCTGTTTAATTTTCAA  254570
Ch10_10    1105227  ATAGAATA-AATTCTATTCAAAAGAATTGCTGGTTCTTTAATTCTCTGTTTAATTTTCAA  1105169
Ch10_9     330982   ATAGAATA-AATTCTATTTAAAAGAATTGCTGGTTCTTTAATTCTCTGTTTAATTTTCAA  330924
Ch10_2     3337520  ATAGAATA-AATTCTATTCAAAAGAATTGCTGGTTCTTTAATTCTCTGTTTAATTTTAAA  3337578
Ch10_11    310680   ATAGAATATAATTCTATTCAAAAGAATTGCTGGTTCTTTAATTCTCTGTTTAATTTTCAA  310621

Ch10_6     168493   AGTTCAATTGTTCGCTGCACT-AGACAGCTTATATATAATATCATTTTCAGCAAGAT---  168438
Ch10_7     336061   AGTTCAATTGTTCGCTGCACT-AGACAGCTTATATATAATATCATTTTCAGCAAGAT---  336006
Ch10_8     341200   AGTTCAATTGTTCGCTGCACT-AGACAGCTTATATATAATATCATTTTCAGCAAGAT---  341145
Ch10_5     163354   AGTTCAATTGTTCGCTGCACT-AGACAGCTTATATATAATATCATTTTCAGCAAGAT---  163299
Ch10_3     158216   AGTTCAATTATCCGCTGCACT-AGACAGCTTATACATAATATCATCATCATTTAACT---  158161
Ch10_1     9590     AGTTCAATTGTTCGCTGCACT-AGACAGCTTAATTATAATATCATTTTTATTTAATC---  9535
Ch10_4     254569   AGTTCAATTGTTCGCTGCACT-AGACAGCTTATATAGAATAACATTTTTTTTCTTTG---  254514
Ch10_10    1105168  AGTTCAATTGTTTGCTGCACT-AGACAGCTTATATATATTATCATTTTGAAGAAACT---  1105113
Ch10_9     330923   AGTTCAATTGTTTGCTGCACT-AGACAGCTTAATTATATTATCA----CAACAAATTTAT  330869
Ch10_2     3337579  AGTTCAATTGTTCGCTGCACTCAGACAGCTTAATTATAATAACATCTACATTTAATC---  3337635
Ch10_11    310620   AGTTCAATTGTTCGCTGCACT-AGACAGCTTATACATATTATCATTTTCAACAAGCT---  310565

Ch10_6     168437   -TTGTCAACATCTTT-TAAAAA---CTTTTTTATTTTTAAAAGTAACAAACTTGAGACAA  168383
Ch10_7     336005   -TTGTCAACATCTTT-TAAAAA---CTTTTTTATTTTTAAAAGTAACAAACTTGAGACAA  335951
Ch10_8     341144   -TTGTCAACATCTTT-TAAAAA---CTTTTTTATTTTTAAAAGTAACAAACTTGAGACAA  341090
Ch10_5     163298   -TTGTCAACATCTTT-TAAAAA---CTTTTTTATTTTTAAAAGTAACAAACTTGAGACAA  163244
Ch10_3     158160   -TTGTCAACATCTTT-TA--------TTTAATATTTTAAAAA                     158129
Ch10_1     9534     -TTGTCAACTACTTT-T                                              9520
Ch10_4     254513   -TTGTCAACATCTTT-T                                              254499
Ch10_10    1105112  -TTGTCAACACATTTCTAAAAA---CTTTTTTATTTTTAAAAATTACAAACTTCAATCAA  1105057
Ch10_9     330868   ATTGTCAACATATTT-T                                              330853
Ch10_2     3337636  -TTGTCAACTACTTT-T                                              3337650
Ch10_11    310564   -TTGTCAACATCTTT-TAAAAAACTTTTTTTTATTTTTAAAAGCGACAAACTTGAGACAA  310507

Ch10_6     168382   CGAATACTATCATATCACATTTAATTATAAAGTCAATACATAT-TTTATAATTAT-TTTT  168325
Ch10_7     335950   CGAATACTATCATATCACATTTAATTATAAAGTCAATACATAT-TTTATAATTAT-TTTT  335893
Ch10_8     341089   CGAATACTATCATATCACATTTAATTATAAAGTCAATACATAT-TTTATAATTAT-TTTT  341032
Ch10_5     163243   CGAATACTATCATATCACATTTAATTATAAAGTCAATACATAT-TTTATAATTAT-TTTT  163186
Ch10_10    1105056  CGAATAATATCATACCACATTTATTCATATTGTCAATAAAAAT-TT                1105012
Ch10_11    310506   CGAATACTATCATACCATATTTATTAAGAATGTCAATAGTTTTATATATAAATATCTTTT  310447

Ch10_6     168324   CATGAGGATAATTTCCTCTTGGAGTAGAAAAGTGAATGCTCCATTATTAAAACATTCACT  168265
Ch10_7     335892   CATGAGGATAATTTCCTCTTGGAGTAGAAAAGTGAATGCTCCATTATTAAAACATTCACT  335833
Ch10_8     341031   CATGAGGATAATTTCCTCTTGGAGTAGAAAAGTGAATGCTCCATTATTAAAACATTCACT  340972
Ch10_5     163185   CATGAGGATAATTTCCTCTTGGAGTAGAAAAGTGAATGCTCCATTATTAAAACATTCACT  163126
Ch10_11    310446   CAT                                                           310444

Ch10_6     168264   TCTCT  168260
Ch10_7     335832   TCTCT  335828
Ch10_8     340971   TCTCT  340967
Ch10_5     163125   TCTCT  163121
```

Figure 3. Multiple sequences alignment of ch10 similar copies. The genome coordinates of the sequences in the row were annotated before and after each sequence.

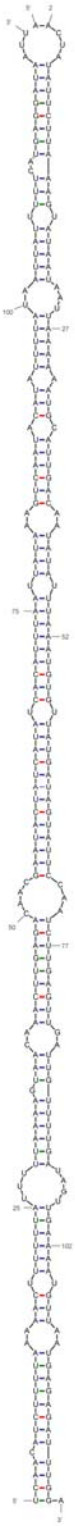
3. Prediction of mRNA targets

Blast approach was applied to search target candidates of ch10 (ch10_1). 212 distinct genomic sequences were matched. Since we were looking for mRNA targets, we were particularly interested in the coding regions of the matched sequences with a reasonable length. In these 212 matched sequences, 41 are within genomic coding regions and have length greater than 34 nt. The hybrid profile of these potential mRNA targets were then predicted with UNAFold (22,23). The highest ranked target gene is listed in Table 2 based on ∆G.

The targets of NRS were predicted in the same way. Figure 4 shows its strongest sRNA–mRNA duplexes hybrid profile (∆G = -64.4 kcal/mol) in the prediction. The genomic region of this hybrid mRNA is in the 3' end of CAC0290 (sensory transduction histidine kinase) spanning 28 nt upstream of the translation stop codon and 95 nt downstream of the annotated ORF. CAC0290 is a gene of 467 nt (328308..329015). It locates downstream of CAC0289 (Response regulator (CheY domain, HTH domain) 328308..329015) and shares the operon with CAC0289. The hybrid region on NRS is from 179 nt to 301 nt. However, this is out of the highly conserved region of the ch10 copies, which is from 56 nt to 161 nt.

Table 2. Target candidates of NRS from Blast search

Gene ID	Start .. End	Gene Description	∆G (KCal/mol)
CAC2775	2908368..2908565	phosphohydrolase	-58.5
CAC1543	1687487..1687731	lactate dehydrogenase	-55.7
CAC0979	1126791..1126986	elongation subunit of DNA-dependent DNA polymerase	-53.2
CAC2786	2915756..2915980	hypothetical protein	-51.1
CAC1185	1336104..1336305	hypothetical protein	-50.2
CAC0356	414863..414987	putative polygalacturonase (pectinase)	-45.9
CAC3548	3744232..3744341	hypothetical protein	-44.2
CAC0946	1086612..1086775	ComE-like protein	-43.5
CAC0823	954046..954181	hypothetical protein	-41.8
CAC0052	61980..62113	hypothetical protein	-39
CAC2653	2766219..2766384	aspartate carbamoyltransferase regulatory subunit	-37.1
CAC1784	1932861..1933048	DNA uptake protein	-36.9
CAC1751	1898514..1898620	chromosome segregation SMC protein, ATPase	-35.9
CAC0691	798275..798380	hypothetical protein	-34.2
CAC2097	2194262..2194370	hypothetical protein	-34
CAC2104	2197935..2198075	general secretion pathway protein F	-33.9
CAC2064	2167897..2167999	purine nucleoside phosphorylase	-33.2
CAC1359	1504816..1504912	xylanase/chitin deacetylase	-32.4
CAC3535	3731416..3731517	Type II restriction enzyme, methylase subunit	-31.7
CAC0243	275283..275429	permease	-31.2



dG = -64.4 dH = -618.6 A-B

Figure 4. Predicted sRNA-mRNA duplexe for NRS. The target sequence spanning the 3'-end of CAC0290 gene (sensory transduction histidine kinase).

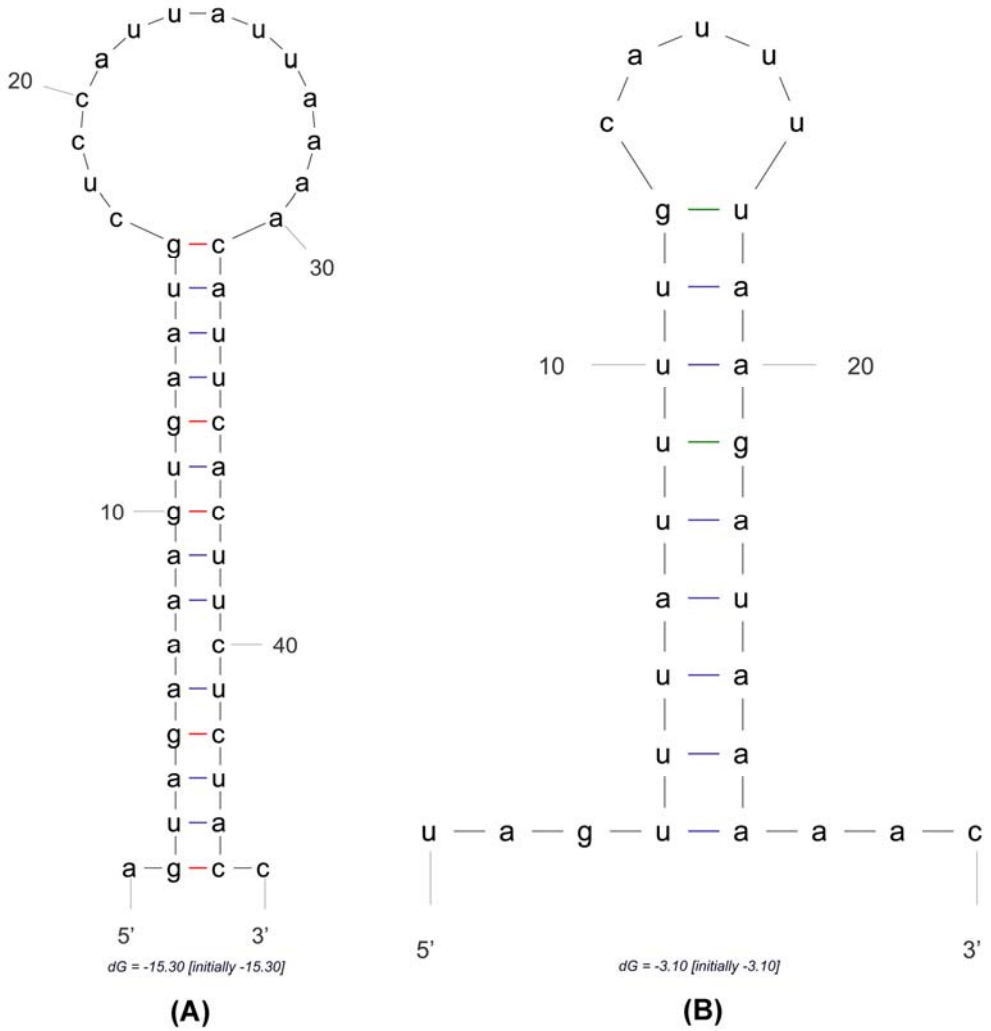


Figure 5. Potential transcript terminator of the validated ch10 sRNA copies. **A.** Loop structure near the end of NRS transcript. The loop sequence is on the inverse strand of the rho-independent terminator of 5S ribosomal RNA, which is upstream of the complementary sequence loci of NRS. **B.** Loop structure near the end of ch10_1 sRNA transcript.

DISCUSSION

This is the first study of genomic identification of sRNAs in the anaerobic bacterium Clostridium acetobutylicum. We predicted 102 novel sRNAs with size ranging from 60 nt to 500 nt. 86 of these sRNAs scatter on the chromosome and the other 16 are on the plasmid. The Q-RT-PCR and Northern blot experiment validated the expression of 7 out 15 randomly selected sRNA candidates. Although Northern blot showed that the prediction did not always capture the correct length of the real sRNA. However, it did predict a large enough length for identification.

The most interesting sRNA candidate is located on the chromosome near the 5' end of a 16S rRNA. This sRNA has a experimental validated high expression level. The sRNA sequence is highly conserved in 11 genomic loci. All these loci are located on the inverse strand of the intergenic regions preceding the 5' end of 16S rRNA genes. Q-RT-PCR verified

the expression of most of these copies. The high expression levels and the consistent association of their genomic location with16S rRNA genes suggest their unknown but important functions in cells.

Despite the high similarity of these sRNA sequences, the computational prediction only caught one of the copies because no rho-independent terminator was found in the other 10 copies. However, it is noticed that NRS sequences occupies the full intergenic region, although on the inverse strand, between the genes coding for 16S rRNA and 5S rRNA. The 3' end of the 5S rRNA transcript forms a rho-independent terminator, so it is possible that the complementary sequence may also forms a weak but functional terminator structure and stop the extension of the sRNA transcription (Figure 5.A). Because this is not a classical strong rho-independent terminator, the prediction method failed to detect it and missed the sRNAs.

A strong terminator structure was also found in the ch10_1 copy, which was identified by the prediction and treated as the end of the transcript. However, Q-RT-PCR data (not included in the paper) shows that only the first half of the predicted sequence was effectively transcribed. The transcribed region is at the 5' end and contains the full conserved core sequence in the alignment (Figure 3). This means that the rho-independent terminator used to find the end of this sRNA is in fact not responsible to stop the transcription. Instead, a weak loop near the end of the transcript was found by the secondary structure prediction (Figure 5.B). But whether this weak loop is functional and determines the end of transcript needs further investigation.

The target prediction suggests a few target genes for ch10 and NRS. Given the features of how sRNA works on target mRNA, Blast is not the most suitable but still an effective approach to search for target candidates. The result provides some evidence for the potential function of the identified sRNAs.

In summary, this computational study gives new perspectives of sRNA activities in clostridium. Most importantly, it identified multiple copies of highly conserved sRNA which are located upstream of 16S rRNA. The function of these sRNAs is unknown, but their consistent genomic locations and high expression levels suggest possibilities of an import regulatory mechanism in clostridium.

REFERENCES

1. Gottesman, S. (2004) The small RNA regulators of Escherichia coli: roles and mechanisms*. *Annu Rev Microbiol*, **58**, 303-328.
2. Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E.G., Margalit, H. and Altuvia, S. (2001) Novel small RNA-encoding genes in the intergenic regions of Escherichia coli. *Curr Biol*, **11**, 941-950.
3. Rivas, E., Klein, R.J., Jones, T.A. and Eddy, S.R. (2001) Computational identification of noncoding RNAs in E. coli by comparative genomics. *Curr Biol*, **11**, 1369-1373.
4. Wassarman, K.M., Repoila, F., Rosenow, C., Storz, G. and Gottesman, S. (2001) Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev*, **15**, 1637-1651.
5. Mandin, P., Repoila, F., Vergassola, M., Geissmann, T. and Cossart, P. (2007) Identification of new noncoding RNAs in Listeria monocytogenes and prediction of mRNA targets. *Nucleic Acids Res*, **35**, 962-974.
6. del Val, C., Rivas, E., Torres-Quesada, O., Toro, N. and Jimenez-Zurdo, J.I. (2007) Identification of differentially expressed small non-coding RNAs in the legume

endosymbiont Sinorhizobium meliloti by comparative genomics. *Mol Microbiol*, **66**, 1080-1091.
7. 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.
8. Livny, J., Fogel, M.A., Davis, B.M. and Waldor, M.K. (2005) sRNAPredict: an integrative computational approach to identify sRNAs in bacterial genomes. *Nucleic Acids Res*, **33**, 4096-4105.
9. Axmann, I.M., Kensche, P., Vogel, J., Kohl, S., Herzel, H. and Hess, W.R. (2005) Identification of cyanobacterial non-coding RNAs by comparative genome analysis. *Genome Biol*, **6**, R73.
10. Tjaden, B. (2008) TargetRNA: a tool for predicting targets of small RNA action in bacteria. *Nucleic Acids Res*.
11. Zhao, Y., Li, H., Hou, Y., Cha, L., Cao, Y., Wang, L., Ying, X. and Li, W. (2008) Construction of two mathematical models for prediction of bacterial sRNA targets. *Biochem Biophys Res Commun*, **372**, 346-350.
12. Tjaden, B., Goodwin, S.S., Opdyke, J.A., Guillier, M., Fu, D.X., Gottesman, S. and Storz, G. (2006) Target prediction for small, noncoding RNAs in bacteria. *Nucleic Acids Res*, **34**, 2791-2802.
13. Alsaker, K.V. and Papoutsakis, E.T. (2005) Transcriptional program of early sporulation and stationary-phase events in Clostridium acetobutylicum. *J Bacteriol*, **187**, 7103-7118.
14. Paredes, C.J., Alsaker, K.V. and Papoutsakis, E.T. (2005) A comparative genomic view of clostridial sporulation and physiology. *Nat Rev Microbiol*, **3**, 969-978.
15. Paredes, C.J., Rigoutsos, I. and Papoutsakis, E.T. (2004) Transcriptional organization of the Clostridium acetobutylicum genome. *Nucleic Acids Res*, **32**, 1973-1981.
16. Tomas, C.A., Alsaker, K.V., Bonarius, H.P., Hendriksen, W.T., Yang, H., Beamish, J.A., Paredes, C.J. and Papoutsakis, E.T. (2003) DNA array-based transcriptional analysis of asporogenous, nonsolventogenic Clostridium acetobutylicum strains SKO1 and M5. *J Bacteriol*, **185**, 4539-4547.
17. Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R. and Bateman, A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*, **33**, D121-124.
18. Livny, J. (2007) Efficient annotation of bacterial genomes for small, noncoding RNAs using the integrative computational tool sRNAPredict2. *Methods Mol Biol*, **395**, 475-488.
19. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389-3402.
20. Ermolaeva, M.D., Khalak, H.G., White, O., Smith, H.O. and Salzberg, S.L. (2000) Prediction of transcription terminators in bacterial genomes. *J Mol Biol*, **301**, 27-33.
21. Macke, T.J., Ecker, D.J., Gutell, R.R., Gautheret, D., Case, D.A. and Sampath, R. (2001) RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res*, **29**, 4724-4735.
22. Markham, N.R. and Zuker, M. (2005) DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res*, **33**, W577-581.
23. Markham, N.R. and Zuker, M. (2008) UNAFold: software for nucleic acid folding and hybridization. *Methods Mol Biol*, **453**, 3-31.
24. Wiesenborn, D.P., Rudolph, F.B. and Papoutsakis, E.T. (1988) Thiolase from Clostridium acetobutylicum ATCC 824 and Its Role in the Synthesis of Acids and Solvents. *Appl Environ Microbiol*, **54**, 2717-2722.

25. Jones, S.W., Paredes, C.J., Tracy, B., Cheng, N., Sillers, R., Senger, R.S. and Papoutsakis, E.T. (2008) The transcriptional program underlying the physiology of clostridial sporulation. *Genome Biol*, **9**, R114.
26. Livny, J. and Waldor, M.K. (2007) Identification of small RNAs in diverse bacterial species. *Curr Opin Microbiol*, **10**, 96-101.
27. Lesnik, E.A., Sampath, R., Levene, H.B., Henderson, T.J., McNeil, J.A. and Ecker, D.J. (2001) Prediction of rho-independent transcriptional terminators in Escherichia coli. *Nucleic Acids Res*, **29**, 3583-3594.
28. Livny, J., Brencic, A., Lory, S. and Waldor, M.K. (2006) Identification of 17 Pseudomonas aeruginosa sRNAs and prediction of sRNA-encoding genes in 10 diverse pathogens using the bioinformatic tool sRNAPredict2. *Nucleic Acids Res*, **34**, 3484-3493.
29. Regalia, M., Rosenblad, M.A. and Samuelsson, T. (2002) Prediction of signal recognition particle RNA genes. *Nucleic acids research*, **30**, 3368-3377.
30. Trotochaud, A.E. and Wassarman, K.M. (2005) A highly conserved 6S RNA structure is required for regulation of transcription. *Nature structural & molecular biology*, **12**, 313-319.
31. 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 (New York, N.Y*, **11**, 774-784.