

# A Computational Pipeline for High-Throughput Discovery of *cis*-Regulatory Noncoding RNA in Prokaryotes

Zizhen Yao<sup>1\*</sup>, Jeffrey Barrick<sup>2‡</sup>, Zasha Weinberg<sup>3</sup>, Shane Neph<sup>1,4</sup>, Ronald Breaker<sup>2,3,5</sup>, Martin Tompa<sup>1,4</sup>, Walter L. Ruzzo<sup>1,4</sup>

**1** Department of Computer Science and Engineering, University of Washington, Seattle, Washington, United States of America, **2** Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, United States of America, **3** Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, Connecticut, United States of America, **4** Department of Genome Sciences, University of Washington, Seattle, Washington, United States of America, **5** Howard Hughes Medical Institute, Yale University, New Haven, Connecticut, United States of America

**Noncoding RNAs (ncRNAs) are important functional RNAs that do not code for proteins. We present a highly efficient computational pipeline for discovering *cis*-regulatory ncRNA motifs de novo. The pipeline differs from previous methods in that it is structure-oriented, does not require a multiple-sequence alignment as input, and is capable of detecting RNA motifs with low sequence conservation. We also integrate RNA motif prediction with RNA homolog search, which improves the quality of the RNA motifs significantly. Here, we report the results of applying this pipeline to Firmicute bacteria. Our top-ranking motifs include most known Firmicute elements found in the RNA family database (Rfam). Comparing our motif models with Rfam's hand-curated motif models, we achieve high accuracy in both membership prediction and base-pair-level secondary structure prediction (at least 75% average sensitivity and specificity on both tasks). Of the ncRNA candidates not in Rfam, we find compelling evidence that some of them are functional, and analyze several potential ribosomal protein leaders in depth.**

Citation: Yao Z, Barrick J, Weinberg Z, Neph S, Breaker R, et al. (2007) A computational pipeline for high-throughput discovery of *cis*-regulatory noncoding RNA in prokaryotes. *PLoS Comput Biol* 3(7): e126. doi:10.1371/journal.pcbi.0030126

## Introduction

Recent discoveries of novel noncoding RNAs (ncRNAs) such as microRNAs and riboswitches suggest that ncRNAs have important and diverse functional and regulatory roles that impact gene transcription, translation, localization, replication, and degradation [1–3]. In the last few years, several groups have performed genome-scale computational ncRNA predictions based on comparative genomic analysis. In particular, Barrick et al. [4] used a pairwise, BLAST-based approach to discover novel riboswitch candidates in bacterial genomes, many of which now have been experimentally verified. Similar studies have been conducted in various bacterial groups [5–8]. More recent work has extended these searches to eukaryotes [9–13], discovering a large number of known microRNAs while producing thousands of novel ncRNA candidates. With some exceptions, such as [4] and [13], these approaches follow a similar paradigm, which is to search for conserved secondary structures on multiple-sequence alignments that are constructed based on sequence similarity alone. Typically, these schemes use measures such as mutual information between pairs of alignment columns to signal base-paired regions. However, the signals such methods seek, namely compensatory base-pair mutations, are exactly the signals that may cause sequence-based alignment methods to misalign, or alternatively refuse to align, homologous ncRNA sequences. Even local misalignments may weaken this key structural signal, making the methods sensitive to alignment quality, which is especially problematic on diverged sequences.

In this paper, we present a novel structure-oriented

computational pipeline for genome-scale prediction of *cis*-regulatory ncRNAs. It exploits, but does not require, sequence conservation. The pipeline differs from previous methods in three respects. First, it searches in unaligned upstream sequences of homologous genes, instead of well-aligned regions constructed by sequence-based methods. Second, we predict RNA motifs in unaligned sequences using a tool called CMfinder [14], which is very sensitive to RNA motifs with low sequence conservation, and robust to inclusion of long flanking regions or unrelated sequences. Finally, we integrate RNA motif prediction with RNA homology search. For every predicted motif, we scan a genome database for more homologs, which are then used to refine the model. This iterative process improves the model and expands the motif families automatically.

In this study, we apply this pipeline to discover ncRNA elements in prokaryotes. We chose prokaryotes mainly because of the large number of fully sequenced genomes

**Editor:** Gary Stormo, Washington University in St. Louis, United States of America

**Received:** February 5, 2007; **Accepted:** May 17, 2007; **Published:** July 27, 2007

**Copyright:** © 2007 Yao et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Abbreviations:** CDD, conserved domain database; CM, covariance model; ncRNA, noncoding RNA; Rfam, the RNA family database; r-protein, ribosomal protein

\* To whom correspondence should be addressed. E-mail: yzizhen@cs.washington.edu

‡ Current address: Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, Michigan, United States of America

## Author Summary

For decades, scientists believed that, with a few key exceptions, RNA played a secondary role in the cell. Recent discoveries have sharply revised this simple picture, revealing widespread, diverse, and surprisingly sophisticated roles for RNA. For example, many bacteria use RNA elements called “riboswitches” to switch various gene activities on or off in response to extremely sensitive detection of specific molecules. Discovery of new functional RNA elements remains a very challenging task, both computationally and experimentally. It is computationally difficult largely because of the importance of an RNA molecule’s 3-D structure, and the fact that molecules with very different nucleotide sequences can fold into the same shape. In this paper, we propose a computational procedure, based on comparing the genomes of multiple bacteria, for discovery of novel RNAs. Unlike most previous approaches, ours does not require a letter-by-letter alignment of these diverse genomes, making it more applicable to RNA elements whose structure, but not nucleotide sequence, has been preserved through evolution. In an extensive test on the Firmicutes, a bacterial phylum containing well-studied organisms such as *Bacillus subtilis* and important pathogens such as anthrax, we recover most known noncoding RNA elements, as well as making many novel predictions.

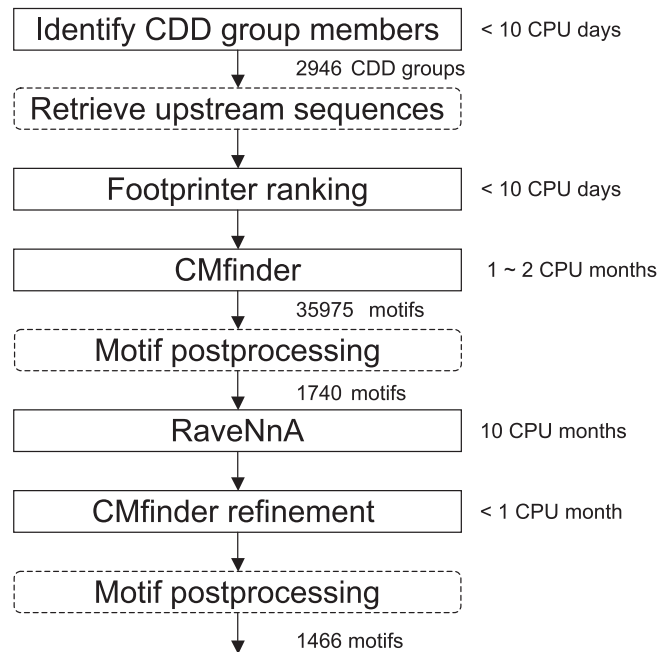
and the great sequence divergence among the species, which can be well-exploited by our approach.

Our approach has two key advantages. First, it is efficient and highly automated. Earlier steps are more computationally efficient than later steps, and we can apply filters between steps so that poor candidates are eliminated from subsequent analysis. Thus, even though we use some computationally expensive algorithms, the pipeline is scalable to larger problems. Besides providing RNA motif prediction, the pipeline also integrates gene context and functional analysis, which facilitates manual biological evaluation. Second, this pipeline is highly accurate in finding prokaryotic ncRNAs, especially RNA *cis*-regulatory elements.

To demonstrate the performance of this approach, we report our search results in Firmicutes, a Gram-positive bacterial division that includes *Bacillus subtilis*, a relatively well-studied model organism with many known ncRNAs. The method exhibits low false-positive rates on negative controls (permuted alignments), and low false-negative rates on known Firmicute ncRNAs. The RNA family database (Rfam) [15], a partially hand-curated database of noncoding RNAs, includes 13 ncRNA families categorized as *cis*-regulatory elements with representatives in *B. subtilis*. Of these, 11 are included among our top 50 predictions and a 12th appears somewhat lower in our ranking. Two other Rfam families are also represented among our top 50 predictions. In addition, both the secondary structure prediction and identified family members are in excellent agreement with Rfam annotation. For the 14 Rfam families mentioned above, we achieved 91% specificity and 84% sensitivity on average in identifying family members, and 77% specificity and 75% sensitivity in secondary structure prediction. Many promising novel ncRNA candidates were also discovered and are discussed below.

## Results

In outline, our pipeline consists of the following major steps. (See Figure 1, Materials and Methods, and the online

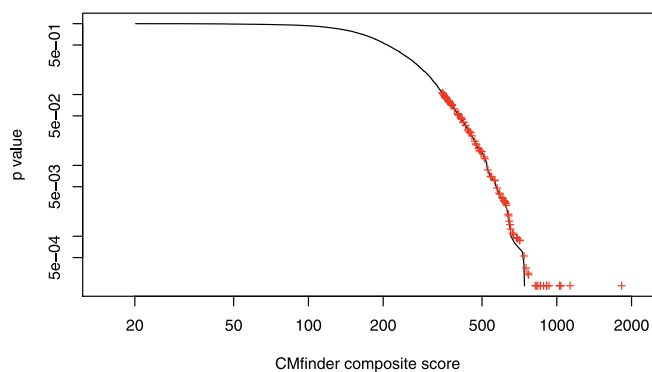


**Figure 1.** Pipeline Flowchart

The boxes with solid lines indicate steps involving intensive computation (approximate running time is specified next to each). Other intermediate steps are specified in the boxes with dashed lines.

doi:10.1371/journal.pcbi.0030126.g001

supplement at <http://bio.cs.washington.edu/supplements/yzizhen/pipeline> for more details.) First, we used the National Center for Biotechnology Information’s (NCBI’s) Conserved Domain Database (CDD) [16] to identify homologous gene sets. For each gene, we collected its 5’ upstream sequence. We call the set of 5’ sequences associated with one CDD group a dataset. *cis*-Regulatory elements are often conserved within such groups. Second, we applied FootPrinter [17], a DNA phylogenetic footprinting tool, to select datasets that are likely to host ncRNAs. In our experience, functional RNAs such as riboswitches often show low overall sequence conservation, but contain interspersed patches where conservation is high. FootPrinter is very effective at highlighting the latter regions. Third, we used CMfinder to infer RNA motifs in each unaligned sequence dataset. CMfinder is a structure-oriented local alignment tool that is robust to varying sequence conservation and length of extraneous flanking regions. We postprocessed motifs to identify distinct motifs corresponding to different RNA elements by removing poor and redundant motifs and clustering the rest based on overlap. Fourth, we used RAVE<sub>NNA</sub> [18–20] to find additional motif instances by scanning the prokaryotic genome database. Riboswitches, for example, often regulate multiple operons that contribute to a single pathway, but no single CDD domain will be common to all of these operons. Thus, the search step was a powerful adjunct to the motif discovery process. These newly discovered motif members were incorporated into a refined motif model, again using CMfinder, and in some cases the search and motif refinement steps were repeated. Motif postprocessing was also repeated after the search/refinement steps. Both CMfinder and RAVE<sub>NNA</sub> rely on the Infernal covariance model software package [21] for RNA motif modeling and search. Finally, we



**Figure 2.** The Empirical  $p$ -Value Distribution Based on the Permutation Test

The black curve shows the complementary cumulative distribution function for the composite scores on randomized datasets (i.e., for each score, the fraction of permuted alignments exceeding that score). The red pluses show the  $p$ -values for the composite scores of the motifs in the original (unpermuted) datasets. All  $p$ -values are greater than or equal to  $2 \times 10^{-4}$  as there are only 5,000 samples in the background distribution.

doi:10.1371/journal.pcbi.0030126.g002

performed gene context analysis and literature searches (manually) for the top-ranking motifs.

We included 44 completely sequenced Firmicute species (see the online supplement at <http://bio.cs.washington.edu/supplements/yzizhen/pipeline>) and 2,946 CDD groups in this study. For each of the three main steps—FootPrinter, CMfinder, and RAVENNA-based refinement—we produced scores to determine which candidates were worthy of continuing analysis. For evaluation purposes, we recorded the scores of candidates at each step, but eliminated none; in the future, we will use them as filters.

The initial CMfinder step produced 35,975 motifs in total. Motif postprocessing reduced this to 1,740 motifs grouped into 1,050 clusters. After RAVENNA-based refinement, more motifs were identified as redundant and removed. A total of 1,466 motifs remained, grouped into 1,060 clusters. (A few of the original clusters were subdivided based on divergent search results.) The full list of candidates is available in the online supplement at <http://bio.cs.washington.edu/supplements/yzizhen/pipeline>.

### Negative Controls: Permuted Alignments

To evaluate how many of our top candidates could have arisen by chance, we performed a randomized control experiment. We first computed CLUSTALW alignments of the 100 sequence datasets having the highest motif scores (before the RAVENNA scan). We then randomly permuted the alignments 50 times, maintaining the approximate gap pattern (see the online supplement at <http://bio.cs.washington.edu/supplements/yzizhen/pipeline>). After degapping each permuted alignment (treating it as a set of unaligned sequences), we applied CMfinder, retaining the top-ranking motif from each randomized dataset. We used this collection of 5,000 motifs to estimate the background score distribution, and to infer  $p$ -values for predicted motifs in the original datasets. Results are shown in Figure 2. By this measure, all 100 top-scoring motifs have  $p$ -values less than 0.1, with the median at 0.016. In addition, 73 of the 100

candidates in the original dataset score higher than all motifs in the corresponding randomized datasets.

Note that this estimation of  $p$ -values is imperfect. In particular, with the scoring scheme we used, datasets containing phylogenetically close sequences tend to score well in comparison to more diverged sets, because permuting the CLUSTALW alignments preserves their sequence conservation. (Independently permuting individual sequences instead of alignments would be less realistic, since in practice *cis*-regulatory RNA motifs are often embedded in regions exhibiting some sequence conservation for other reasons.) Although imperfect, the significance of real motifs tends to be underestimated by this method.

### Positive Controls: Discovering Known Rfam Families

To roughly assess the sensitivity with which the method discovers true ncRNAs, we looked at its recovery of known Rfam (version 7.0) families. We masked matches to Rfam's tRNA and rRNA models, since otherwise these widespread, strong motifs might hide nearby, weaker, but still interesting ncRNA structures. Other Rfam families were not masked and serve as a positive control for our methods. Table 1 shows the distribution of known Rfam families in our candidate list, together with their ranks after running FootPrinter, CMfinder, and RAVENNA. We used the refined motifs as the final output.

According to Rfam, *B. subtilis* contains members of 21 families, categorized into 13 *cis*-regulatory families, one intron element, and seven RNA gene families. We masked tRNAs and rRNAs (four of the seven gene families). Of the 17 remaining families, 13 appear within our top 50 candidates: 11 *cis*-regulatory families present in *B. subtilis*, together with two of the gene families (RNaseP\_bact\_b and SRP\_bact). The four families not represented are two *cis*-regulatory elements (*ykkC-ykkD* and *ydaO-yuaA*), one RNA gene (tmRNA), and one intron element (Intron\_gpI). The exclusion of Intron\_gpI is not surprising, as we did not search intragenic regions. The *ydaO-yuaA* motif escaped detection because it is present in only three of the 68 sequences in its CDD group. The *ykkC-ykkD* and tmRNA motifs, although not among our top 50, would still have been ranked high enough to be discovered in a blind test. Note that, although our computational pipeline is oriented toward discovery of *cis*-regulatory elements, we sometimes find RNA genes such as RNaseP, SRP, and tmRNA because they happen to be conserved in synteny. We also found a partial tRNA motif, not masked since parts of the tRNA lie outside of the collected upstream sequences.

We can potentially filter the candidates at each step to scale this pipeline for larger genomes. In particular, we could have applied CMfinder to only the top half of the datasets according to FootPrinter, and performed genome scans on only the top 500 motifs, without missing any real Rfam families as listed in Table 1. On average, it takes FootPrinter less than 1 min, and CMfinder 10 min to process each dataset, while it takes RAVENNA 4.8 h to scan each motif. We could save considerable computation time by running expensive algorithms only on good candidates.

As shown in Table 1, the ranks for most known ncRNAs improve at each successive step of the pipeline, as more supporting evidence is found. Starting from FootPrinter motifs, CMfinder improves the alignment and identifies consensus secondary structure, while genome scans locate

**Table 1.** Motifs That Correspond to Rfam Families

Rank	Score			Number		Conserved Domain Database			Rfam
	RAV	CMF	FP	RAV	CMF	ID	Gene	Description	
0	43	107	3,400	367	11	9904	IlvB	Thiamine pyrophosphate-requiring enzymes	RF00230 T-box
1	10	344	3,115	96	22	13174	COG3859	Predicted membrane protein	RF00059 THI
2	77	1284	2,376	112	6	11125	MetH	Methionine synthase I specific DNA methylase	RF00162 S_box
3	0	5	2,327	30	26	9991	COG0116	Predicted N6-adenine-specific DNA methylase	RF00011 RNaseP_bact_b
4	6	66	2,228	49	18	4383	DHBP	3,4-dihydroxy-2-butanone 4-phosphate synthase	RF00050 RFN
7	145	952	1,429	51	7	10390	GuaA	GMP synthase	RF00167 Purine
8	17	108	1,322	29	13	10732	GcvP	Glycine cleavage system protein P	RF00504 Glycine
9	37	749	1,235	28	7	24631	DUF149	Uncharacterised BCR, YbaB family COG0718	RF00169 SRP_bact
10	123	1358	1,222	36	6	10986	CbiB	Cobalamin biosynthesis protein CobD/CbiB	RF00174 Cobalamin
20	137	1133	899	32	7	9895	LysA	Diaminopimelate decarboxylase	RF00168 Lysine
21	36	141	896	22	10	10727	TerC	Membrane protein TerC	RF00080 <i>yybP-ykoY</i>
39	202	684	664	25	5	11945	MgtE	Mg/Co/Ni transporter MgtE	RF00380 <i>ykoK</i>
40	26	74	645	19	18	10323	GlmS	Glucosamine 6-phosphate synthetase	RF00234 <i>glmS</i>
53	208	192	561	21	5	10892	OpuBB	ABC-type proline/glycine betaine transport systems	RF00005 tRNA <sup>a</sup>
122	99	239	413	10	7	11784	EmrE	Membrane transporters of cations and cationic drug	RF00442 <i>ykkC-yxkD</i>
255	392	281	268	8	6	10272	COG0398	Uncharacterized conserved protein	RF00023 tmRNA

The first three columns show ranks for refined motif clusters after genome scans (RAV), CMfinder motifs before genome scans (CMF), and FootPrinter results (FP). We used the same ranking scheme for RAV and CMF. Score: CMfinder composite motif score (after refinement). Number: the number of motif instances before (CMF) and after (RAV) genome scan. Conserved Domain Database: PSSM-ID (accession), gene name, and description. Rfam: Rfam accession and family name. The genome scan here refers to the “mini” scan (rather than full scan; see Materials and Methods).

<sup>a</sup>A few tRNAs partially outside the limits of the collected upstream regions evaded our masking procedure.

doi:10.1371/journal.pcbi.0030126.t001

many more motif instances, typically providing still better alignments and additional clues to their functions.

To measure the quality of our automatically constructed motif models, we compared them with Rfam alignments for the same families. Rfam’s covariance models are built from hand-curated “seed” alignments/structure annotations. These in turn are used to build Rfam’s “full” alignments by automatically searching RFAMSEQ (<http://www.sanger.ac.uk/Software/Rfam/ftp.shtml>), a high-quality, nonredundant subset of EMBL (<http://www.ebi.ac.uk/embl>), and automatically aligning all hits.

For the 14 Rfam families in Table 1 for which we found good matching motifs, we selected the top two motifs from each family, and performed full-genome scans on RFAMSEQ, the same sequence database used to construct the Rfam full alignment. To reduce computation time, we did not scan eukaryote genomes, and the Rfam hits from these genomes were excluded from the following analysis. (This treatment affects only a few eukaryotic Cobalamin and Lysine hits, all believed to be Rfam errors or bacterial contamination in the genome sequences, plus a few THI hits, which are real.) For each motif, we selected scan hits at an E-value cutoff of 100, reconstructed the motif alignments using CMfinder, and removed the low-scoring instances (<20 bits). We compared these predicted motifs to corresponding Rfam full alignments, which serve as the gold standard in this test. Table 2 shows the accuracy of our motifs in membership prediction, motif coverage, and secondary structure prediction. Secondary structures were compared at the base-pair level, and only the base pairs with at least one end falling into the overlapped regions are counted. For both predicted motifs and Rfam full alignments, we removed noncanonical base pairs from each sequence. Of the two motifs chosen for each family, we report the one with better results.

For membership prediction, we achieved an average of

84% sensitivity and 91% specificity. The overlapped regions between predicted motif members and corresponding Rfam members account for 81% of the length of the predicted members, and 82% of the length of Rfam members. In the overlapped regions, the secondary structure prediction has 75% sensitivity and 77% specificity. These results suggest our predicted motif models are very accurate compared with Rfam models, which are learned from the hand-curated seed alignments.

For many riboswitch families, the main differences between our motif models and Rfam models are located in boundary regions. Our predicted motifs tend to include the transcription terminator (if present), which is a stable hairpin followed by a stretch of U’s (e.g., Lysine, S\_box, T-box). Although transcription terminators are functionally important, the Rfam riboswitch models do not include them. On the other hand, CMfinder tends to miss the closing helix of large multiloop structures (e.g., Cobalamin, *ykoK*). Most other differences are local perturbations such as small shifts or extra base pairs.

As shown in Table 2, we achieved more than 80% membership sensitivity for all families except *yybP-ykoY*, Glycine, and Cobalamin. The predicted *yybP-ykoY* motif differs from Rfam’s motif mainly at the multiloop closing helix. Cobalamin and Glycine are two riboswitches with poor sequence conservation (46% and 51% average sequence identity, respectively). While our motifs from the initial full-genome scan may be too specific, sensitivity increases significantly with only a small loss in specificity after another iteration of RAVENNA scan and refinement (unpublished data).

For *ykkC-yxkD* and T-box, we predicted more members than Rfam. The predicted *ykkC-yxkD* motif includes the transcription terminator, which caused false positives in our full-genome scans. These false positives, however, all have

**Table 2.** Motif Prediction Accuracy Compared with Rfam

RNA Family Database	Membership			Overlap			Structure		
	Number	Sn	Sp	nt	Sn	Sp	bp	Sn	Sp
RF00174 Cobalamin	183	0.74 <sup>a</sup>	0.97	152	0.75	0.85	20	0.60	0.77
RF00504 Glycine	92	0.56 <sup>a</sup>	0.96	94	0.94	0.68	17	0.84	0.82
RF00234 glmS	34	0.92	1.00	100	0.54	1.00	27	0.96	0.97
RF00168 Lysine	80	0.82	0.98	111	0.61	0.68	26	0.76	0.87
RF00167 Purine	86	0.86	0.93	83	0.83	0.55	17	0.90	0.95
RF00050 RFN	133	0.98	0.99	139	0.96	1.00	12	0.66	0.65
RF00011 RNaseP_bact_b	144	0.99	0.99	194	0.53	1.00	38	0.72	0.78
RF00162 S_box	208	0.95	0.97	110	1.00	0.69	23	0.91	0.78
RF00169 SRP_bact	177	0.92	0.95	99	1.00	0.65	25	0.89	0.81
RF00230 T-box	453	0.96	0.61	187	0.77	1.00	5	0.32	0.38
RF00059 THI	326	0.89	1.00	99	0.91	0.69	13	0.56	0.74
RF00442 <i>ykkC-ykkD</i>	19	0.90	0.53	99	0.94	0.81	18	0.94	0.68
RF00380 <i>ykoK</i>	49	0.92	1.00	125	0.75	1.00	27	0.80	0.95
RF00080 <i>yybP-ykoY</i>	41	0.32	0.89	100	0.78	0.90	18	0.63	0.66
Mean	145	0.84	0.91	121	0.81	0.82	21	0.75	0.77
Median	113	0.91	0.97	105	0.81	0.83	19	0.78	0.78

All comparisons are to the prokaryotic subset of Rfam full alignments. Membership: the number of sequences in the overlap between our predictions and Rfam's (number), the sensitivity (Sn), and specificity (Sp) of our membership predictions. Overlap: the average length of overlap between our predictions and Rfam's (nt), the fractional lengths of the overlapped region in Rfam's predictions (Sn) and in ours (Sp). Structure: the average number of correctly predicted canonical base pairs (in overlapped regions) in the secondary structure (bp), and the sensitivity (Sn) and specificity (Sp) of our predictions.

<sup>a</sup>After another iteration of RaveNNA scan and refinement, the membership sensitivities of Glycine and Cobalamin increased to 76% and 98%, respectively, while the specificity of Glycine remained the same, and the specificity of Cobalamin dropped to 84%.

doi:10.1371/journal.pcbi.0030126.t002

much less significant E-values than the true positives, and hence are relatively easy to eliminate by inspection. In contrast, for T-box we believe most “false positives” (with respect to Rfam 7.0) are actually real. Out of 291 members not included in the Rfam full alignment, 127 are upstream of and on the same strand as aminoacyl-tRNA synthetase genes, where most T-box leaders are found, and the others are largely in poorly annotated regions.

### Motifs Not in Rfam

We examined the best-scoring motif (see RNA motif discovery in Materials and Methods and the online supplement at <http://bio.cs.washington.edu/supplements/yzizhen/pipeline> for details of the motif-scoring function) in each of the top 200 motif clusters. Of these 200 motifs, 116 were deemed unlikely to represent novel ncRNAs: they have covariance model scores less than 40 bits, single hairpin structures, and most were shorter than 30 nucleotides. (Many of these 116 are nevertheless biologically relevant. Many correspond to transcription terminators of upstream genes, and others contain known inverted repeat motifs targeted by DNA binding proteins.) Of 84 remaining motifs, 20 correspond to Rfam families, and 11 to hypothetical transposons. The remaining 53 are candidates for novel ncRNAs. Literature review suggests that many of these candidates are functional. We manually removed the redundant candidates with the same functional roles (for details, see Manual inspection and ribosomal protein leader analysis in Materials and Methods), and present the rest in Table 3.

**Annotated motifs.** Several candidates turn out to be known regulatory elements that have been described previously in the literature, including the following.

*PyrR attenuator.* Upstream of CDD 28178, we predicted a PyrR RNA binding site [22], which regulates *pyr* operon transcription by switching between alternative antitermina-

tor versus anti-antiterminator plus terminator structures. The motif we predicted corresponds to the anti-antiterminator plus terminator structure, which favors transcription termination. It includes 69 instances in 31 Firmicute species, with two copies per species on average: one copy upstream of the *pyrP* or *pyrR* gene, and one copy upstream of *pyrB*. (Note that the recently released Rfam 8.0 has added a PyrR motif, RF00515, which appears to be in good agreement with our prediction.)

*6S.* This ncRNA binds to  $\sigma^{70}$  RNA polymerase holoenzyme to globally regulate gene expression in response to the shift from exponential growth to stationary phase. Although 6S has been known in *Escherichia coli* and close relatives for more than 35 years [23], the corresponding Rfam model (RF00013 6S/SsrS RNA) is confined to  $\gamma$ -proteobacteria, and its Firmicute homologs were only identified recently by experimental [24] and computational [25] means; see also [26]. We have discovered 6S in Firmicutes independently in this study. The motif we predict is a partial 6S that includes the most conserved core. (Note that the revised 6S model in Rfam 8.0 now contains Firmicute instances.)

*Inverted repeats.* It is difficult to determine whether a motif with inverted repeats functions at the DNA or RNA level without considering its genomic context. Based on the literature, three single-hairpin inverted repeat motifs in Table 3 appear to be known DNA binding sites for regulatory proteins: HrcA binding sites (rank 44), BlaI/MecI binding sites (rank 140), and hypothetical CadC binding sites (rank 50). (All three are longer and had significantly higher covariance model scores than the 116 removed inverted repeats mentioned above.)

**Novel ncRNA candidates: Ribosomal protein leaders.** To demonstrate how CMfinder predictions can accelerate the discovery and characterization of new RNA motifs, we present a detailed analysis of two conserved mRNA leader

**Table 3.** High-Ranking Motifs Not Found in Rfam

Rank	Number	CDD	Conserved Domain Database Description	Annotation
6	69	28178	DHOase IIa: dihydroorotase	PyrR attenuator [22]
15	33	10097	RplL: ribosomal protein L7/L1	L10 r-protein leader; see online supplement at <a href="http://bio.cs.washington.edu/supplements/yzizhen/pipeline">http://bio.cs.washington.edu/supplements/yzizhen/pipeline</a>
19	36	10234	RpsF: ribosomal protein S6	S6 r-protein leader
22	32	10897	COG1179: dinucleotide-utilizing enzymes	6S RNA [25]
27	27	9926	RpsJ: ribosomal protein S10	S10 r-protein leader; see online supplement at <a href="http://bio.cs.washington.edu/supplements/yzizhen/pipeline">http://bio.cs.washington.edu/supplements/yzizhen/pipeline</a>
29	11	15150	Resolvase: N-terminal domain	
31	31	10164	InfC: translation initiation factor 3	IF-3 r-protein leader; see online supplement at <a href="http://bio.cs.washington.edu/supplements/yzizhen/pipeline">http://bio.cs.washington.edu/supplements/yzizhen/pipeline</a>
41	26	10393	RpsD: ribosomal protein S4 and related proteins	S4 r-protein leader; see online supplement at <a href="http://bio.cs.washington.edu/supplements/yzizhen/pipeline">http://bio.cs.washington.edu/supplements/yzizhen/pipeline</a> ; [30]
44	30	10332	GroL: chaperonin GroEL	HrcA DNA binding site [46]
46	33	25629	Ribosomal L21p: ribosomal prokaryotic L21 protein	L21 r-protein leader; see online supplement at <a href="http://bio.cs.washington.edu/supplements/yzizhen/pipeline">http://bio.cs.washington.edu/supplements/yzizhen/pipeline</a>
50	11	5638	Cad: cadmium resistance transporter	[47]
51	19	9965	RplB: ribosomal protein L2	S10 r-protein leader
55	7	26270	RNA pol Rpb2 1: RNA polymerase beta subunit	
69	9	13148	COG3830: ACT domain-containing protein	
72	28	4174	Ribosomal S2: ribosomal protein S2	S2 r-protein leader
74	9	9924	RpsG: ribosomal protein S7	S12 r-protein leader
86	6	12328	COG2984: ABC-type uncharacterized transport system	
88	19	24072	CtsR: Firmicutes transcriptional repressor of class III	CtsR DNA binding site [48]
100	21	23019	Formyl trans N: Formyl transferase	
103	8	9916	PurE: Phosphoribosylcarboxyaminoimidazole	
117	5	13411	COG4129: predicted membrane protein	
120	10	10075	RplO: ribosomal protein L15	L15 r-protein leader
121	9	10132	RpmJ: ribosomal protein L36	IF-1 r-protein leader
129	4	23962	Cna B: Cna protein B-type domain	
130	9	25424	Ribosomal S12: ribosomal protein S12	S12 r-protein leader
131	9	16769	Ribosomal L4: ribosomal protein L4/L1 family	L3 r-protein leader
136	7	10610	COG0742: N6-adenine-specific methylase	ylbH putative RNA motif [4]
140	12	8892	Pencillinase R: penicillinase repressor	Blal, MecI DNA binding site [49]
157	25	24415	Ribosomal S9: ribosomal protein S9/S16	L13 r-protein leader; Figure 3
160	27	1790	Ribosomal L19: ribosomal protein L19	L19 r-protein leader; Figure 2
164	6	9932	GapA: glyceraldehyde-3-phosphate dehydrogenase/erythrose	
174	8	13849	COG4708: predicted membrane protein	
176	7	10199	COG0325: predicted enzyme with a TIM-barrel fold	
182	9	10207	RpmF: ribosomal protein L32	L32 r-protein leader
187	11	27850	LDH: L-lactate dehydrogenases	
190	11	10094	CspR: predicted rRNA methylase	
194	9	10353	FusA: translation elongation factors	EF-G r-protein leader

For each motif, we give its rank (after refinement), number of sequences containing the motif, the Conserved Domain Database ID, gene name and description of the dataset from which it was found, and notes. Two of the ribosomal protein autoregulatory motifs are described more fully in the Results section, and five others are described in the online supplement at <http://bio.cs.washington.edu/supplements/yzizhen/pipeline>. doi:10.1371/journal.pcbi.0030126.t003

structures that most likely are involved in autoregulation of L19 and L13–S9 ribosomal protein expression. Five additional presumed ribosomal autoregulatory motifs are presented in the online supplement at <http://bio.cs.washington.edu/supplements/yzizhen/pipeline>.

Many ribosomal protein (r-protein) operons regulate their own expression in *E. coli* [27,28]. Once enough of a specific r-protein encoded by an operon has been produced (i.e., all of its rRNA binding sites are saturated), excess copies of the protein bind to the 5' untranslated leader region of its mRNA and induce structural changes that compete with ribosome binding or stall initiating ribosome complexes. This general repression mechanism appears to apply to many r-protein operons, but the specific RNA structures recognized by orthologous r-proteins are generally not conserved between *E. coli* and other bacterial groups.

For example, the S15 mRNA leaders from *E. coli*, *Geobacillus stearothermophilus*, and *Thermus thermophilus* assume different, apparently unrelated RNA structures that all seem to mimic the same rRNA binding site [29]. Similarly, the mRNA binding site of S4 differs between *E. coli* and *Bacillus* species [30].

Within a bacterial division, the same regulatory structure may be used in many species. Thus, an mRNA leader structure recognized by L4 is conserved in many, but not all,  $\gamma$ -proteobacteria [31]. Our comparative analysis using CMfinder is well-suited to recognize r-protein mRNA leader motifs conserved at this taxonomic level. Indeed, it detects the only two r-proteins leader structures that have currently been characterized in Firmicutes (S4 and S15). However, the structure predicted for S4 leaders by CMfinder agrees only partially with a previous phylogenetic analysis of this element

based on fewer, exclusively *Bacillus*, species [30]. After manually examining the regions aligned by CMfinder, we predict a consensus structure that is close to the Grundy and Henkin model [30] but has a different pseudoknot (see the online supplement at <http://bio.cs.washington.edu/supplements/yzizhen/pipeline>). The relatively poor performance of CMfinder on the S4 leaders may be partly due to the clustering of a subfamily of *Lactobacillus* sequences with a slightly different consensus structure from the *Bacillus* sequences. CMfinder performed better on the S15 leader (rank 842), accurately predicting the location and extent of the largest helix-2 feature [32]. Here, it misses only the small adjacent helix-3, and an additional stem that overlaps the open reading frame.

CMfinder also predicts a novel regulatory RNA structure upstream of L19, encoded by the *rplS* gene, in *Bacilli*, *Lactobacilli*, *Clostridia*, and *Fusobacteria* species (Table 3, rank 160). In *E. coli*, L19 is expressed as the last of four genes from a polycistronic mRNA [33]. A similar gene order is conserved in some Firmicutes (approximately two-thirds of those with the RNA motif), and there is not an intrinsic transcription terminator between the orthologous upstream *trmD* gene and *rplS* in *B. subtilis*. However, the intergenic distance between *trmD* and *rplS* is typically greater than 100 base pairs in Firmicutes (142 nt in *B. subtilis*) compared with only 41 nt between *trmD* and *rplS* in *E. coli*. Putative promoter  $-35$  and  $-10$  hexamers occur within this intergenic region upstream of each predicted RNA structure (Figure 3A), suggesting that L19 is expressed as a separate transcriptional unit from the upstream genes in Firmicutes.

The putative L19 autoregulatory mRNA structure is a small bulged hairpin (Figure 3B). The length of the terminal P2 stem-loop varies, but the outer P1 helix always has exactly eight base pairs. Most primary sequence conservation occurs in the asymmetric internal bulge and P1 stem. The original CMfinder results include some nonconserved sequences and a spurious stem-loop upstream that are not preserved in all examples. Within the conserved region, CMfinder identifies most of the pairing predicted in our manually refined model.

This RNA structure is always found close to the ribosome binding site (RBS) of the L19 open reading frame. If it is involved in typical r-protein autoregulation, then L19 binding might stabilize an alternate paired conformation wherein the 5' side of P1 sequesters the RBS to repress gene expression (Figure 3C). Alternately, the predicted P1 stem might only be stable in the presence of L19, and when it forms, its proximity to the open reading frame might prevent translation initiation. Ribosomal protein L19 binds to the large rRNA subunit at the 50S–30S interface. We were unable to identify any homology between the predicted mRNA leader structure and its 23S rRNA binding site in the *E. coli* ribosome [34], or homologous positions in the *B. subtilis* ribosome [35], that might suggest a simple regulatory model. It is possible that the predicted regulatory hairpin mimics the structure of the rRNA binding site, or participates in a more complex regulatory mechanism.

CMfinder predicts a second novel RNA structure (Figure 4) upstream of the L13–S9 operon, encoded by the *rplM* and *rpsI* genes, in *Bacilli* and *Lactobacilli* species (Table 3, rank 157). There is a strong, near-consensus promoter directly upstream of this motif that defines a conserved transcription start site. The L13–S9 structure is also a bulged hairpin, but it is larger

than the L19 motif. There is striking conservation of seven loop nucleotides (CCCCGGA) that are identical in all sequences. Additional conservation occurs in the bulge and within the P1 helix. CMfinder correctly predicts the P2 helix in this manually revised model, and it also identifies the core base pairs in the P1 helix, except in cases where an inserted stem loop occurs in the 3' side of the bulge.

S9 is a secondary small subunit binding protein, requiring prior S7 binding to associate with 16S rRNA [36]. Most r-proteins involved in autoregulation are primary binding proteins that can bind directly to rRNA, so it seems most likely that L13, a protein that binds to 23S rRNA early in large subunit assembly, recognizes this leader structure. Here again, we were unable to identify any conservation between the rRNA contact sites of L13 and *E. coli* 23S rRNA or the corresponding sites in *B. subtilis* 23S that suggest a regulatory model. There is sometimes a significant distance between the putative regulatory RNA structure and the open reading frame. Alternate pairings between the U-rich 5' side of P1 and a region overlapping the start codon can be devised for many sequences, so it is possible that this alternate conformation is enforced by L13 or S9 binding to the mRNA leader to prevent translation.

In the online supplement at <http://bio.cs.washington.edu/supplements/yzizhen/pipeline>, we present the full manually refined structural alignments for the above two motifs plus five additional putative r-protein leader regulatory motifs: IF-3, L10, L21, S4, and S10. Based on our experiences with these putative mRNA leader structures, it should be straightforward to define many more candidates for r-protein autoregulatory structures in other bacterial groups with our pipeline. Such studies could illuminate how this form of regulation has been modified and preserved during evolution and would make genomic annotation of noncoding RNAs more comprehensive. Five of these seven putative regulatory RNA elements are now included in Rfam 8.0 (see Accession Numbers).

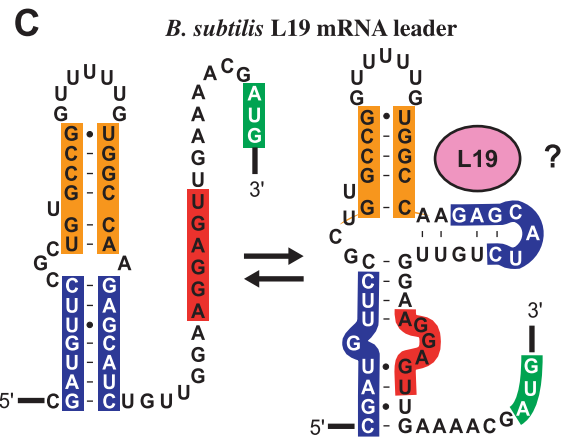
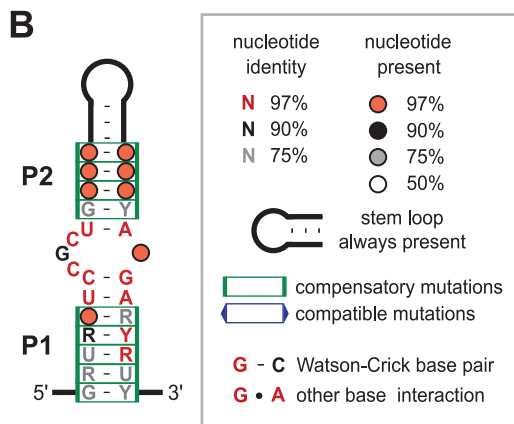
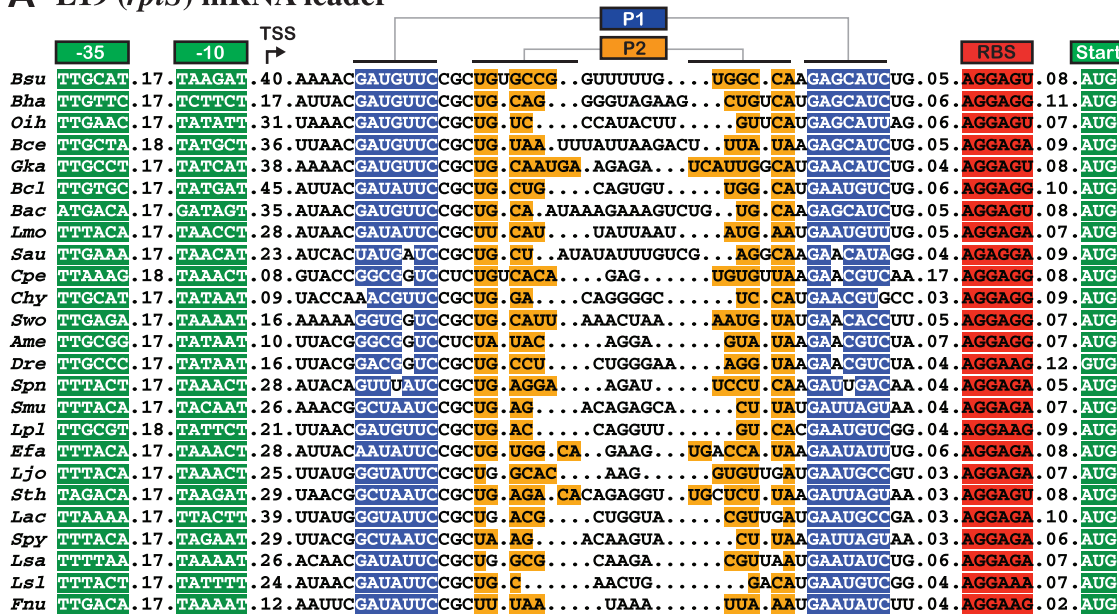
## Discussion

In this study, we have presented a method for automatically finding *cis*-regulatory RNA motifs in prokaryotes. In a careful test with available sequenced Firmicutes, the method exhibited excellent rejection of negative controls (randomly permuted alignments) and excellent recovery of known, experimentally validated ncRNAs, including most riboswitches known in this bacterial group, as well as RNA elements such as 6S that have only recently been recognized there. Careful inspection and refinement of several novel motifs in ribosomal protein leaders provides compelling evidence that they are indeed conserved structures involved in regulation of these important operons.

In addition, our computational pipeline found dozens of other good RNA motifs that constitute strong candidates for novel functional elements, consistent with the increasing appreciation of the importance of RNA in all living organisms. Finally, our method is sufficiently scalable to be applied to all sequenced prokaryotes. We are in the process of doing so, and preliminary results include several novel riboswitch candidates.

We attribute the power of this pipeline to two key characteristics—a relaxation of the constraints on sequence

### A L19 (*rplS*) mRNA leader



**Figure 3.** Putative Autoregulatory Structure in L19 mRNA Leaders

(A) Sequence alignment of a conserved RNA structure found in the 5' UTR of Firmicute *rplS* genes. Possible promoter –35 and –10 boxes in genomic DNA are shown, followed by the putative mRNA leader with the predicted secondary structures (P1 and P2), ribosome binding sites, and start codons highlighted. Numbers represent inserted nucleotides that are not shown. The examples shown are representative of 34 total sequences in the complete alignment, available in the online supplement at <http://bio.cs.washington.edu/supplements/yzizhen/pipeline>. Species abbreviations: *Ame*, *Alkaliphilus metalliredigens*; *Bac*, *Bacillus* sp. NRRL; *Bce*, *Bacillus cereus*; *Bcl*, *Bacillus clausii*; *Bha*, *Bacillus halodurans*; *Bsu*, *Bacillus subtilis*; *Chy*, *Carboxydotherrnus hydrogenoformans*; *Cpe*, *Clostridium perfringens*; *Dre*, *Desulfotomaculum reducens*; *Efa*, *Enterococcus faecalis*; *Fnu*, *Fusobacterium nucleatum*; *Gka*, *Geobacillus kaustophilus*; *Lac*, *Lactobacillus acidophilus*; *Ljo*, *Lactobacillus johnsonii*; *Lmo*, *Listeria monocytogenes*; *Lpl*, *Lactobacillus plantarum*; *Lsa*, *Lactobacillus sakei*; *Lsl*, *Lactobacillus salivarius*; *Oih*, *Oceanobacillus ihyensensis*; *Sau*, *Staphylococcus aureus*; *Smu*, *Streptococcus mutans*; *Spn*, *Streptococcus pneumoniae*; *Spy*, *Streptococcus pyogenes*; *Sth*, *Streptococcus thermophilus*; *Swo*, *Syntrophomonas wolfei*.

(B) Consensus sequence and secondary structure. Pairs supported by compensatory (when both bases in a pair mutate between sequences in the alignment) and compatible (when only one base mutates but pairing is preserved, e.g., G-C to G-U) are boxed.

(C) Structural model of the *B. subtilis* L19 mRNA leader, showing a possible alternate structure that could be stabilized by L19 binding to repress translation.

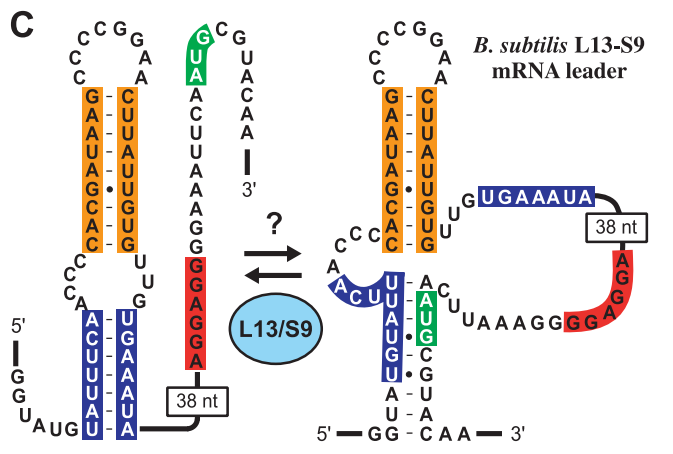
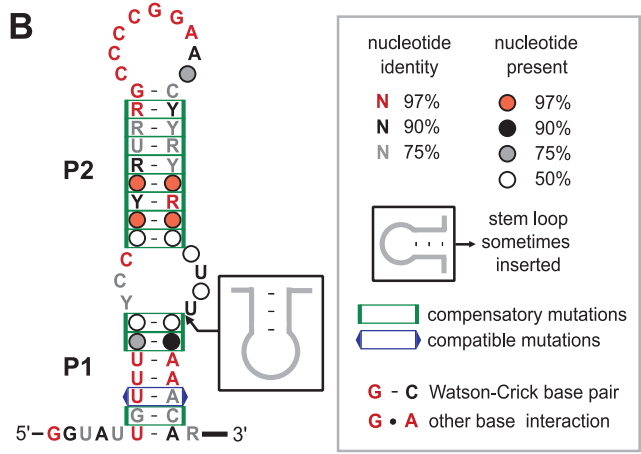
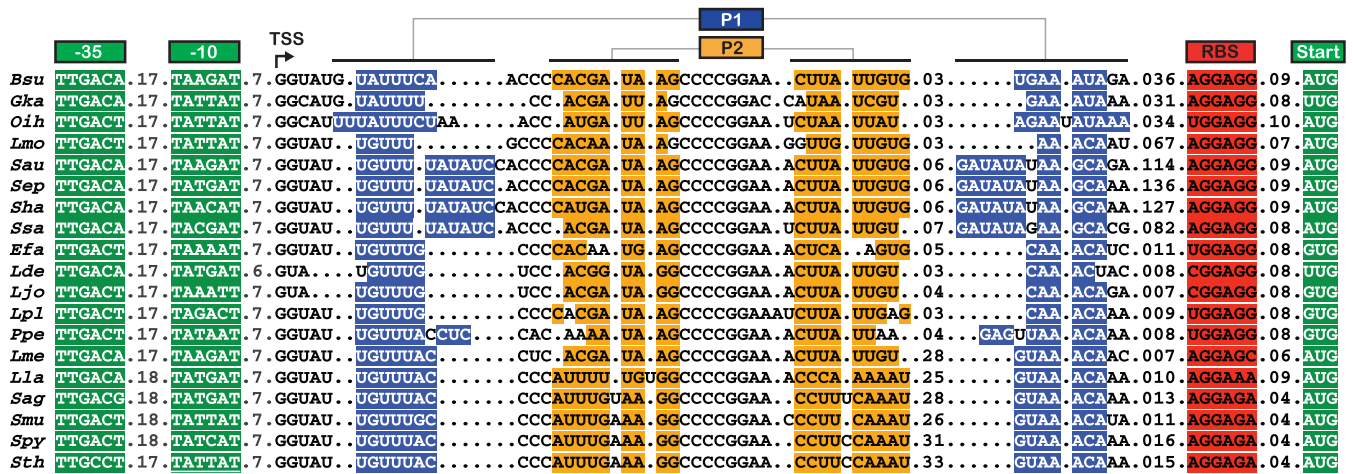
doi:10.1371/journal.pcbi.0030126.g003

conservation imposed by most previous methods, and integration of motif inference with genome-scale search. Our method performs motif inference on regions that are not defined by sequence conservation: we search unaligned sequences upstream of homologous genes, instead of multiple-sequence alignments constructed by sequence comparison tools. In addition, both the RNA motif-finding algorithm CMfinder and the RNA homology search algorithms RAVENNA/Infernal exploit structural information. Sequence conservation can be used as well, but is not required. Finally,

automatic refinement of motifs to incorporate genome-scale search results has proven to be a powerful component of the pipeline (as in other contexts, such as PSI-BLAST [37]). The integration of these tools enables us to discover RNA motifs with low sequence conservation, and to expand the motif family with remote homologs. For example, the predicted motif for the Glycine Riboswitch has only 35% average pairwise sequence similarity. Remote RNA homologs with appropriate gene context are particularly important, as they



### A L13-S9 (*rplM-rpsI*) mRNA leader



**Figure 4.** Putative Autoregulatory Structure in L13-S9 mRNA Leaders  
 (A) Sequence alignment of a conserved RNA structure found in the 5' UTR of Firmicute *rplM-rpsI* operons. The examples shown are representative of 27 total sequences in the complete alignment, available in the online supplement at <http://bio.cs.washington.edu/supplements/yzizhen/pipeline>. Details are as in the legend for Figure 3 with additional species abbreviations: *Lde*, *Lactobacillus delbrueckii*; *Lla*, *Lactococcus lactis*; *Lme*, *Leuconostoc mesenteroides*; *Ppe*, *Pediococcus pentosaceus*; *Sag*, *Streptococcus agalactiae*; *Sep*, *Staphylococcus epidermidis*; *Sha*, *Staphylococcus haemolyticus*; *Ssa*, *Staphylococcus saprophyticus*.  
 (B) Consensus sequence and secondary structure.  
 (C) Structural model of the *B. subtilis* L13-S9 mRNA leader, showing a possible alternate structure that could be stabilized by L13 or S9 binding to repress translation.  
 doi:10.1371/journal.pcbi.0030126.g004

are the strongest evidence, short of experiments, that a motif is functional, as well as providing clues to that function.

Future work will seek to strengthen this pipeline by improved exploitation of phylogeny and by an improved scoring system.

Phylogeny is crucial in all comparative genome analysis, without which the concept of conservation is meaningless. It is important in our work because the sequences upon which motif inference is performed are not evolutionarily equidistant, and the significance of conserved nucleotides and compensatory mutations are distance-dependent. Building on the classic phylogenetic likelihood model of Felsenstein [38], Pfold [39] and Evofold [12] use an RNA-oriented phylogenetic model to select from a given multiple-sequence alignment the regions that fit the structural model best. Unfortunately, in our application, neither an alignment nor an evolutionary tree is initially available, and, for our application, use of the corresponding species tree is

inadequate in the common case when there are multiple sequences per species. Incorporating phylogeny into motif search is another challenge.

We would also like to improve our scoring scheme. As predicted motifs are subject to expensive manual evaluation and experiments, automatic candidate evaluation to guide resource investment is critical. Our current composite scoring system attempts to discriminate among potential RNA motifs by considering a set of features, including species distributions, structure stabilities, motif sizes, and local sequence conservation patterns. While we can easily recognize motifs that are significant in all these aspects, it is more difficult to order those that are only good by some, but not all, criteria. We have tried to combine the features automatically using machine-learning algorithms such as support vector and logistic regression. However, due to the heterogeneity of the features and limitations of available training data, the results were not as good as our handcrafted

composite scoring function. One particular issue is that many of our top-scoring motifs are short single hairpins. They score well because they are widespread, structurally stable, and contain limited but clear sequence conservation. Although short motifs can be functionally important, many do not contain sufficient signal for genome scale homology scans, resulting in false positives that degrade the motif. Other complications include transposons, transcription terminators, DNA-protein binding sites, RNA-polymerase and RNA-ribosome binding sites, etc. The key challenge here is to design a metric that is correctly normalized across various known features and various types of ncRNAs with different sizes, structures, and phylogenetic divergence.

These opportunities for improvement notwithstanding, the approach described in this study has proven itself to be highly effective in discovering noncoding RNA elements in prokaryotes, and promises more discoveries to come.

## Materials and Methods

**Genome sequence and protein homolog data.** We obtained genome sequences from 67 fully sequenced Firmicute species from the NCBI microbial database (RefSeq [40] release 14, 20 November 2005). We first collected amino acid sequences from all annotated protein-coding genes in these species, and categorized them based on NCBI's CDD (version 2.05) [16]. The CDD domain models are curated from various resources, including Pfam, SMART, and COG. In the NCBI microbial database, 92% of all functionally annotated proteins (i.e., with nonhypothetical description field) are assigned to at least one CDD group, as are 32% of “hypothetical” proteins. By definition, all members of a CDD group contain a conserved domain in their protein sequences. A group typically includes both orthologs and paralogs. We assigned proteins to a CDD group using “rpsblast” from the NCBI BLAST package (<http://www.ncbi.nlm.nih.gov/BLAST>), with an E-value cutoff threshold of 0.01. To reduce redundancy, we removed near-duplicate genomes from analysis. To do this, we created a vector for each complete genome, whose *i*th component was the number of predicted occurrences of the *i*th conserved domain in that genome. We normalized these vectors to have unit (Euclidean) length, and measured their similarity in terms of the projection of one CDD vector onto another (i.e., the dot product between them). Beginning with records assigned the lowest accession numbers, we then assembled a set of genomes by accepting each subsequent genome only when its similarity index with all selected datasets was less than 0.95. After removing redundancy in this way, 44 complete genomes remained for processing in subsequent steps.

We removed CDD groups that contained too few members (four or less), since motif discovery is unreliable on such small groups. We also removed 145 groups with too many members (70 or more), since motif discovery is expensive on such large groups.

**Collecting upstream sequences.** For each gene in a CDD group, we collected a few hundred nucleotides upstream of its start codon, which typically includes both 5' UTR and promoter sequences. The prevalence of operons in prokaryotic genomes complicates the extraction of the regulatory regions, as the desired regulatory region may be upstream of the entire operon rather than immediately upstream of the selected gene. To handle this complication in a conservative manner, we extracted the noncoding sequences upstream of the gene and upstream of its plausible operon using MicroFootPrinter [41]. Specifically, if the next coding region upstream is in the same orientation and fewer than 100 nucleotides upstream, this short intergenic sequence is included in our sequence dataset, and the same procedure is applied to the upstream gene. This process continues until interrupted either by a coding region in the opposite orientation or an intergenic region longer than 100 nucleotides. Finally, up to 600 nucleotides of the last intergenic region are included in the sequence dataset. After collecting the upstream sequences, we removed redundant sequences (95% sequence identity across 80% of the sequence according to BLAST), and masked regions that match tRNA or rRNA models in the Rfam database.

**Ranking using FootPrinter.** FootPrinter [17] identifies conserved sequence motifs in a set of unaligned homologous sequences using phylogenetic analysis. We scored each FootPrinter motif by the

number of motif instances minus the corresponding parsimony score, and scored each dataset as the sum of its top 30 motif scores. The resulting scores are used to rank all datasets. This ranking is performed by MicroFootPrinter [41], a front end to FootPrinter [17].

**RNA motif discovery.** We used CMfinder version 0.2 [14] for RNA motif prediction in unaligned sequences. For each dataset, we produced up to five single stem-loop motifs, five double stem-loop motifs, and used CMfinder heuristics to combine the motifs into more complicated structures if possible.

At various subsequent points, we ranked all CMfinder motifs using a heuristic scoring function that favors motifs with instances in diverged species, stable secondary structure, and local sequence conservation. We used local sequence conservation to discriminate trustworthy alignments with reliable anchors from purely structural motifs (e.g., alignments of single hairpins) that could easily arise by chance, while penalizing global sequence conservation, as highly similar sequences are more likely to be conserved by selection pressure on primary sequence than on structure. We refer to these scores as *composite scores*. The details of the scoring function are described in the online supplement at <http://bio.cs.washington.edu/supplements/yzizhen/pipeline>.

**Motif postprocessing.** Next, we filtered the motif set to remove poor motifs and combine redundant ones.

Operationally, a “motif” is a covariance model (CM), and a “motif instance” is a sequence that matches the CM with a score above a specified threshold. For each motif, we removed instances with CM score less than ten bits, and removed all but one copy of completely identical instances. Then, we ranked the motifs by composite scores, as outlined above and detailed in the online supplement at <http://bio.cs.washington.edu/supplements/yzizhen/pipeline>. We further removed motifs with at most four instances and pairwise similarity greater than 0.95, and motifs with composite scores below 50. Afterwards, we selected up to four motifs for each dataset, selected in decreasing score order so that the lower ranking motifs do not overlap significantly with any higher ranking selected motif. By our definition, motif A overlapped significantly with another motif B if the number of nonoverlapping instances of A was less than 30% of the number of overlapping instances, and the average length of the nonoverlapping regions in the overlapped instances of A was less than half of the average length of the overlapped regions. Next, we removed redundant motifs from different datasets. We called motif A redundant with motif B if A overlapped significantly with B and the number of its predicted bases pairs not in B was less than 30% of the number of its base pairs in B. If A and B are redundant with each other, we chose the higher-ranking motif.

Finally, we clustered overlapping motifs as follows. We identified the overlap between motifs according to the genomic coordinates of their instances. One motif was grouped with another if at least half of its instances overlapped, and the overlapped regions are longer than half of the motif length. The motifs were clustered progressively, with high-ranking motifs processed first. We ranked clusters based on their highest-scoring motifs.

**Genome scans for RNA motif homologs.** One of the key strengths of our method is its integration of motif discovery with motif search. Motif discovery is focused on groups of orthologs defined by common CDD membership, since such groups seem likely to be enriched for common *cis*-regulatory elements. However, many *cis*-regulatory elements such as riboswitches will be found near a variety of operons involved in a coherent pathway, which may *not* share a common CDD group. Hence, genome-scale search for additional motif instances is an important component of our approach. Additional instances allow us to construct more accurate motif models, as well as giving insight into potential biological roles for the elements.

Given RNA motifs produced by CMfinder, we searched for additional instances using Infernal CMs [21] accelerated with the ML-heuristic filter [20] implemented in RAVENNA 0.2f. For reasons of speed, two levels of search were used. The initial search database was derived from all 75 finished Firmicute genomes in RefSeq17 (30 April 2006) [40], a total of approximately 200 million nucleotides. Based on sequence annotations, we extracted only intergenic regions for searching, but extended each by 50 nucleotides in each direction to account for common errors in protein-coding gene annotations. The resulting database contained approximately 34 million nucleotides. This small database made it feasible to perform searches for all motifs (averaging 4.8 CPU h per motif), and reduced false positives when compared with the full-genome database. After motif refinement (incorporating hits from this “mini” scan), we performed “full” scans with selected motifs. Full scans examined the prokaryotic subset of the 8 GB RFAMSEQ dataset (version 7.0, March 2005), a total of approximately 900 MB. In particular, comparisons to Rfam (e.g.,

Table 2) were based on full scans, since Rfam full alignments are also derived from scans of RFAMSEQ. For model refinement, we ran CMfinder on all hits with RAVE-NNA E-values < 10. E-values were calculated as in [42]. The necessary extreme value distribution calculations dominate the run times for mini-scans, but not for full scans. The refined motif set is again postprocessed and ranked as described above.

**Identifying known Rfam motifs.** To find which of our predicted motifs were already known, we compared them against the Rfam database. Specifically, we BLASTed our motif instances against Rfam full family members (produced by scanning Rfam covariance models on the RFAMSEQ genomic database; see [15]). For BLAST, we used a word size 12, and selected the hits with length greater than 30 nt, E-value < 10, and sequence identity exceeding 90%. These permissive BLAST thresholds resulted in a few isolated hits that we believe to be false positives. These motifs match fragments, each of about 30 bases, of the Rfam RNA-OUT, Intron-gpII, QaRNA, and RNaseP\_bact\_a families. In general, they are too short, weak, and/or isolated to be compelling, in sharp contrast to the matches reported in Table 1.

**Manual inspection and ribosomal protein leader analysis.** The genomic contexts of the refined motif instances were drawn using the Bio:Graphics modules of BioPerl [43]. For the ribosomal motifs, CMfinder structural alignments were trimmed to relevant regions and manually revised before conducting standard genome scans against the microbial subset of the RefSeq17 database. Hits with the correct genomic context were aligned according to the starting covariance model and manually revised once more to create final sequence alignments (available in the online supplement at <http://bio.cs.washington.edu/supplements/yzizhen/pipeline>). The Neural Network Promoter Prediction program [44] (version 2.2) was used to predict putative transcription start sites, and programs from the Vienna RNA package [45] were used to examine possible regulatory conformations.

**Online supplement.** Additional datasets and technical details are available at <http://bio.cs.washington.edu/supplements/yzizhen/pipeline>.

## References

- Hentze M, Kuhn L (1996) Molecular control of vertebrate iron metabolism: mRNA-based regulatory circuits operated by iron, nitric oxide, and oxidative stress. *Proc Natl Acad Sci U S A* 93: 8175–8182.
- Conne B, Stutz A, Vassalli J (2000) The 3' untranslated region of messenger RNA: A molecular "hotspot" for pathology? *Nat Med* 6: 637–641.
- Mandal M, Boese B, Barrick J, Winkler W, Breaker R (2003) Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. *Cell* 113: 577–586.
- Barrick JE, Corbino KA, Winkler WC, Nahvi A, Mandal M, et al. (2004) New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc Natl Acad Sci U S A* 101: 6421–6426.
- Axmann IM, Kensch P, Vogel J, Kohl S, Herzel H, et al. (2005) Identification of cyanobacterial non-coding RNAs by comparative genome analysis. *Genome Biol* 6: R73.
- Seliverstov A, Putzer H, Gelfand M, Lyubetsky V (2005) Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria. *BMC Microbiology* 5: 54.
- Corbino K, Barrick J, Lim J, Welz R, Tucker B, et al. (2005) Evidence for a second class of S-adenosylmethionine riboswitches and other regulatory RNA motifs in alpha-proteobacteria. *Genome Biol* 6: R70.
- Puerta-Fernandez E, Barrick J, Roth A, Breaker R (2006) Identification of a large noncoding RNA in extremophilic eubacteria. *Proc Natl Acad Sci U S A* 103: 19490–19495.
- McCutcheon JP, Eddy SR (2003) Computational identification of non-coding RNAs in *Saccharomyces cerevisiae* by comparative genomics. *Nucleic Acids Res* 31: 4119–4128.
- Coventry A, Kleitman DJ, Berger B (2004) MSARE: Multiple sequence alignments for statistical detection of RNA secondary structure. *Proc Natl Acad Sci U S A* 101: 12102–12107.
- Washietl S, Hofacker IL, Lukasser M, Huttenhofer A, Stadler PF (2005) Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nature Biotechnol* 23: 1383–1390.
- Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, et al. (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comp Biol* 2: e33.
- Torarinsson E, Sawera M, Havgaard JH, Fredholm M, Gorodkin J (2006) Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res* 16: 885–889. (Erratum [2006] *Genome Research* 16: 1439.)

## Supporting Information

### Text S1. r-Protein Leader

Structural motifs and annotations for predicted ribosomal protein leaders.

Found at doi:10.1371/journal.pcbi.0030126.sd001 (20 KB TAR).

### Text S2. Appendix

Additional technical details.

Found at doi:10.1371/journal.pcbi.0030126.sd002 (16 KB PDF).

### Accession Numbers

Five of the ribosomal protein leaders discussed in the Results section appear in Rfam release 8.0 (<http://www.sanger.ac.uk/Software/Rfam>), with the following accession numbers: L10 r-protein leader (RF00557), L13 r-protein leader (RF00555), L19 r-protein leader (RF00556), L20 (IF-3) r-protein leader (RF00558), L21 r-protein leader (RF00559).

## Acknowledgments

We thank Ryan H. Moy for help with the ribosomal protein leader analysis and the Yale Life Sciences High Performance Computing Center for great technical assistance and computing resources. JEB was a Howard Hughes Medical Institute predoctoral fellow.

**Author contributions.** ZY, JB, ZW, RB, MT, and WLR conceived and designed the experiments. ZY and ZW performed the experiments. ZY and JB analyzed the data. SN contributed reagents/materials/analysis tools. ZY, JB, and WLR wrote the paper.

**Funding.** Computing resources partially supported by National Institutes of Health (NIH) grant RR19895. Additional support provided by National Science Foundation grant DBI-0218798 and NIH grants R01 HG02602 and R33 DK07027.

**Competing interests.** The authors have declared that no competing interests exist.

- Yao Z, Weinberg Z, Ruzzo WL (2006) CMfinder—A covariance model based RNA motif finding algorithm. *Bioinformatics* 22: 445–452.
- Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR (2003) Rfam: An RNA family database. *Nucleic Acids Res* 31: 439–441.
- Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, et al. (2005) CDD: A Conserved Domain Database for protein classification. *Nucleic Acids Res* 33: D192–D196.
- Blanchette M, Tompa M (2003) FootPrinter: A program designed for phylogenetic footprinting. *Nucleic Acids Res* 31: 3840–3842.
- Weinberg Z, Ruzzo WL (2004) Faster genome annotation of non-coding RNA families without loss of accuracy. In: RECOMB04, Proceedings of the Eighth Annual International Conference on Computational Molecular Biology, pp. 243–251. ACM Press. Available: <http://doi.acm.org/10.1145/974614.974647>. Accessed 29 May 2007.
- Weinberg Z, Ruzzo WL (2004) Exploiting conserved structure for faster annotation of non-coding RNAs without loss of accuracy. *Bioinformatics* 20: i334–341.
- Weinberg Z, Ruzzo WL (2006) Sequence-based heuristics for faster annotation of non-coding RNA families. *Bioinformatics* 22: 35–39.
- Eddy SR (2003–2006) Infernal User's Guide. Available: <ftp://selab.janelia.org/pub/software/infernal/Userguide.pdf>. Accessed 29 May 2007.
- Lu Y, Turner RJ, Switzer RL (1996) Function of RNA secondary structures in transcriptional attenuation of the *Bacillus subtilis pyr* operon. *Proc Natl Acad Sci U S A* 93: 14462–14467.
- Brownlee GG (1971) Sequence of 6S RNA of *E. coli*. *Nat New Biol* 229: 147–149.
- Trotochaud AE, Wasserman KM (2005) A highly conserved 6S RNA structure is required for regulation of transcription. *Nat Struct Mol Biol* 12: 313–319.
- Barrick JE, Sudarsan N, Weinberg Z, Ruzzo WL, Breaker RR (2005) 6S RNA is a widespread regulator of eubacterial RNA polymerase that resembles an open promoter. *RNA* 11: 774–784.
- Willkomm DK, Minnerup J, Huttenhofer A, Hartmann RK (2005) Experimental RNomics in *Aquifex aeolicus*: Identification of small non-coding RNAs and the putative 6S RNA homolog. *Nucleic Acids Res* 33: 1949–1960.
- Zengel JM, Lindahl L (1994) Diverse mechanisms for regulating ribosomal protein synthesis in *Escherichia coli*. *Prog Nucleic Acid Res Mol Biol* 47: 331–370.
- Kenner J, Nomura M (1996) Regulation of ribosome synthesis. In: Neidhardt FC, Ingraham JL, Curtiss RC III, editors. *Escherichia Coli and*

- Salmonella: Cellular and molecular biology. Washington (D.C.): ASM Press. pp. 1417–1431.
29. Springer M, Portier C (2003) More than one way to skin a cat: Translational autoregulation by ribosomal protein S15. *Nat Struct Biol* 10: 420–422.
  30. Grundy FJ, Henkin TM (1992) Characterization of the *Bacillus subtilis rpsD* regulatory target site. *J Bacteriol* 174: 6763–6770.
  31. Allen T, Shen P, Samsel L, Liu R, Lindahl L, et al. (1999) Phylogenetic analysis of L4-mediated autogenous control of the S10 ribosomal protein operon. *J Bacteriol* 181: 6124–6132.
  32. Scott LG, Williamson JR (2001) Interaction of the *Bacillus stearothermophilus* ribosomal protein S15 with its 5′-translational operator mRNA. *J Mol Biol* 314: 413–422.
  33. Wikstrom PM, Lind LK, Berg DE, Björk GR (1992) Importance of mRNA folding and start codon accessibility in the expression of genes in a ribosomal protein operon of *Escherichia coli*. *J Mol Biol* 224: 949–966.
  34. Schuwirth BS, Borovinskaya MA, Hau CW, Zhang W, Vila-Sanjurjo A, et al. (2005) Structures of the bacterial ribosome at 3.5 Å resolution. *Science* 310: 827–834.
  35. Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, et al. (2002) The comparative RNA web (CRW) site: An online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* 3: 2.
  36. Powers T, Changchien LM, Craven GR, Noller HF (1988) Probing the assembly of the 3′ major domain of 16S ribosomal RNA. Quaternary interactions involving ribosomal proteins S7, S9 and S19. *J Mol Biol* 200: 309–319.
  37. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25: 3389–3402.
  38. Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol* 17: 368–376.
  39. Knudsen B, Hein J (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res* 31: 3423–3428.
  40. Pruitt K, Tatusova T, Maglott D (2005) NCBI reference sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 33: 501–504.
  41. Neph S, Tompa M (2006) MicroFootPrinter: A tool for phylogenetic footprinting in prokaryotic genomes. *Nucleic Acids Res* 34: W366–W368.
  42. Klein RJ, Eddy SR (2003) RSEARCH: Finding homologs of single structured RNA sequences. *BMC Bioinformatics* 4: 44.
  43. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* 12: 1611–1618.
  44. Reese MG (2001) Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. *Comput Chem* 26: 51–56.
  45. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, et al. (1994) Fast folding and comparison of RNA secondary structure. *Chemical Monthly* 125: 167–188.
  46. Narberhaus F (1999) Negative regulation of bacterial heat shock genes. *Mol Microbiol* 31: 1–8.
  47. Endo G, Silver S (1995) CadC, the transcriptional regulatory protein of the cadmium resistance system of *Staphylococcus aureus* plasmid p1258. *J Bacteriol* 177: 4437–4441.
  48. Derre I, Rapoport G, Msadek T (1999) CtsR, a novel regulator of stress and heat shock response, controls *clp* and molecular chaperone gene expression in Gram-positive bacteria. *Mol Microbiol* 31: 117–131.
  49. Sharma VK, Hackbarth CJ, Dickinson TM, Archer GL (1998) Interaction of native and mutant MecI repressors with sequences that regulate *mecA*, the gene encoding penicillin binding protein 2a in methicillin-resistant staphylococci. *J Bacteriol* 180: 2160–2166.