Abstract
Discovering and defining metabolite-binding riboswitches and other structured regulatory RNA motifs in bacteria
Jeffrey Evan Barrick
2006

Riboswitches are noncoding mRNA elements that directly control gene expression in response to changes in cellular conditions. Riboswitches that sense metabolite concentrations fold into complex receptors that bind the target small molecule and concomitantly switch the conformation of nearby RNA structures to alter protein production. To investigate the scope of riboswitch regulation in bacteria, we have used comparative genomics approaches to identify new regulatory RNA elements with conserved secondary structures and genomic contexts, primarily in *Bacillus subtilis* and *Agrobacterium tumefaciens*. Four of these motifs have now been shown to function as riboswitches, including (1) a self-cleaving ribozyme triggered by glucosamine-6-phosphate, (2) a riboswitch that cooperatively binds glycine with two tandem aptamers, (3) a miniature $preQ_1$ riboswitch, and (4) a second class of *S*-adenosylmethionine riboswitches found mainly in α-proteobacteria. Several other regulatory RNA elements that we discovered have complex structures and occur upstream of related genes in diverse bacterial species, but we do not yet know what cellular conditions or compounds these "orphan riboswitches" may sense. We have systematically catalogued examples of ten classes of metabolite-binding riboswitches in new genomic sequences in order to understand large-scale trends in their phylogenetic distributions and mechanisms and to refine their structural models. While searching for new riboswitches, we inadvertently discovered that homologs of *Escherichia coli* 6S RNA are significantly more widespread in bacteria than previously realized and characterized the conserved features of this regulator of RNA polymerase activity. The increasing number of structured RNAs operating with complex regulatory mechanisms in the genomes of contemporary bacteria may reflect unexpected regulatory sophistication in an ancient RNA World.

Discovering and defining metabolite-binding riboswitches and other structured regulatory RNA motifs in bacteria

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

By
Jeffrey Evan Barrick

Dissertation Director: Ronald R. Breaker

December 2006

To my parents

Contents

Figures

Tables

Acknowledgements

Five years is a forever in the blink of an eye. Excavating memories of my graduate life and studies is like digging into the Burgess Shale. What amazingly wonderful happenings and distractions! What fascinating and unexpected friendships! What improbably hopeful monsters! All that remains in my memory are scattered shapes strewn in the newly hardened mud, but I know that this was once a world, vivaciously alive, and that the ephemeral descendents of those instants live on.

I would like to thank all of my MB&B classmates and colleagues, especially Kelly Sheppard, Kyle Friend, Adrian Olivares, Jess Williamson, and Scott Boyle, for the close community we shared, especially during the early years of our studies. It has sustained me and endured due to their immense generosity. I thank the many inhabitants of the Hostel California at 169 Livingston over the years for making my life more interesting, especially my constant co-conspirators Dennis Mishler and Keith Corbino for many an arrant escapade. Other friends are too numerous for me to mention them all: Amanda Solem, Patrick Nyman, Sumit Bora, Rich Wing, ultimate frisbee players in the KBT quad, my fellow biochemistry TAs, and many happy hour companions. I'm particularly touched by Shira Weidenbaum's constant encouragement, warmth, and wit.

So many members of the Breaker Lab have been instrumental in carrying out the work I describe and leveraging my informatics data mining into remarkable experimental finds. I thank Narasimhan Sudarsan who is more ingenious than Nature; Ali Nahvi for turning over Luigi's house and keeping me in the mix; Adam Roth for his keen eye that has pored over reams of search results; Wade Winkler for his inspiring and tireless work; Keith Corbino for keeping the lab in order and programming parts of BLISS; Ryan Moy for bioinformatics work that has followed up on new motifs; Beth Regulski, Kirstin Block, and Ethan Butler for their experiments on 6S RNA; benchmates Noah Gourlie, Tyler

Ames, Jane Kim, Ken Blount; and Ben Boese, Ming Cheah, Smadar Cohen-Chalamish, Jennifer Collins, Gail Emilsson, Elena Puerta Fernandez, Ming Hammond, Inbal Jona, Mark Lee, Kris Link, Jinsoo Lim, Maumita Mandal, Shingo Nakamura, Izabela Puskarz, Brian Tucker, Joy Wang, Rüdi Welz, Ken Wickiser, Lixia Xu, and Maris Zivarts.

I owe others in the Yale research community for their generous help: Matt Cabeen and Michelle Aaron for microbiology advice; Cecilia Guerrier-Takada for technical advice; Junhyong Kim for a memorable rotation; and Don Engelman for teaching and career mentorship. Outside of Yale, I thank Alex Bateman, Sam Griffiths-Jones, and the rest of the Rfam/Pfam team for an unbelievable month working at the Sanger Center near Cambridge. Finally, much of this work would not have been possible without computer programs and puns developed at the University of Washington in Seattle by Zasha Weinberg in the lab of Larry Ruzzo.

I thank my committee members Alanna Schepartz and Lynne Regan for remaining helpful and supportive even after my projects veered unexpectedly from mRNA display into the RNA World. I am grateful to Eric Westhof for donating his time and expertise to improve this document. Finally, I thank Ron Breaker for granting me so many unbelievable opportunities and the independence to indulge in my own interests — I tried my best to find every riboswitch. A Howard Hughes Medical Institute Predoctoral Fellowship supported my five years of graduate study.

"Every one wanted to say so much that no one said anything in particular"

Captains Courageous, Rudyard Kipling

—Jeffrey Barrick

August 2006

Preface

Chapter 1 briefly describes the discovery of the first riboswitch classes and how it motivated us to apply comparative genomics to predict new *cis*-regulatory RNA motifs in bacterial genomes that would be candidates for new riboswitches. I describe the design and implementation of the BLISS database, relate various complications that we encountered during its development, and note other conserved bacterial sequence features that can be confused with novel *cis*-regulatory RNA elements. A final section compares our results to those obtained by other efforts to identify conserved regulatory elements with different approaches.

Chapter 2 provides an overview of the tools and approaches that we use to characterize putative *cis*-regulatory RNA motifs. I describe RNA homology search methods and how information about the genomic contexts of putative matches can be used to collaborate predictions of more diverged motif examples. In this context, I compare the benefits and shortcomings of various computational programs and databases for predicting various facets of riboswitch structure and regulation. I then outline important considerations for manually constructing high-quality RNA sequence alignments and detail how the valuable information gained by characterizing a regulatory RNA element can be shared with the scientific community.

Chapters 3 and 4 present our current knowledge of the regulatory RNA motifs that we discovered and defined in the low G+C Gram-positive soil bacterium *Bacillus subtilis* and the α-proteobacterial plant pathogen *Agrobacterium tumefaciens*, respectively. Four of these elements have subsequently proven to function as riboswitches, and I briefly describe their distinguishing characteristics. I also summarize what is suspected from biochemical and bioinformatic evidence about the genetic

regulons, regulatory mechanisms, and cellular roles of the other "orphan riboswitch" motifs whose function remains a mystery.

Chapter 5 is a systematic survey of widespread riboswitch classes in genomic and environmental sequences. I present the phylogenetic distributions of each riboswitch class and the preferred expression platform mechanisms in different bacterial groups. Then I describe new consensus features, structural motifs, and base interactions found in several riboswitch classes by examining expanded multiple sequence alignments.

Chapter 6 describes the serendipitous discovery, while examining candidate *B. subtilis* motifs, that homologs of *E. coli* 6S RNA exist in most bacterial species. 6S RNA is a different kind of regulatory RNA: an abundant noncoding RNA that represses transcription from certain promoters during nutrient limitation. I describe the conserved structural features of 6S RNA and relate them to its likely role as a mimic of promoter DNA that sequesters σ^{70}-containing RNA polymerase holoenzyme. Then, a section reviews the known properties of 6S RNAs from different species, several of which had been previously studied without realizing that they were structurally homologous RNAs.

Chapter 7 begins by discussing usage of the term "riboswitch" and mentioning other types of natural riboswitches that have been discovered by others. I then describe new approaches that promise to discover even more *cis*-regulatory RNA elements and riboswitches in microbial and eukaryotic genomes. Finally, this chapter closes by considering the intriguing possibility that some riboswitches classes are descended from metabolic ribozymes that existed in an ancient RNA World.

1 Discovering riboswitches in bacterial genomes

1.1 Introduction

In September of 2002 our laboratory reported that a conserved structure in the mRNA leader of the *btuB* gene in *E. coli* was able to directly bind coenzyme B_{12} (adenosylcobalamin) in the complete absence of proteins and autonomously control gene expression from this transcript [202]. This RNA structure was the first example of a "riboswitch", an mRNA element able to proactively fold into a complex structure, bind a small molecule metabolite, and change its conformation in a way that controls gene expression without the involvement of other cellular factors. This result specifically followed up on earlier evidence that the leader regions of this mRNA and an operon of cobalamin biosynthesis genes from *S. typhimurium* (the *cob* operon) both contained a conserved 25 nt B12 box sequence and extensive RNA secondary structure [225] and that adenosylcobalamin inhibited binding of purified ribosomes to the *btuB* transcript [209]. The ability of mRNAs, commonly thought of as passive messengers in the central dogma, to control their own fates was surprising in a way that echoed the earlier discovery of natural ribozymes [43, 79].

In a broader sense, *in vitro* selection experiments that had successfully been used to isolate RNA aptamers that bound to a variety of small molecules from random nucleic acid sequences [94, 321] were harbingers of this new regulatory role for RNA in cells. Indeed, our laboratory had been inspired to search the scientific literature for RNAs that bound to small molecules and controlled gene expression by the ease with which we had created allosteric hammerhead ribozyme biosensors that recognized a variety of small molecules and metals like theophylline, cAMP, FMN, and Co^{2+} [251, 261]. We reasoned that cases where conserved mRNA elements were known to be important for

genetic regulation, but regulatory proteins that mediated the effect had "not yet been identified", would be good candidates for control by natural allosteric RNA sensors.

After the initial report of a riboswitch directly controlling coenzyme B_{12} biosynthesis and transport genes, our research group and other laboratories rapidly described new classes of riboswitches in bacteria that were already latent in the scientific literature. Thus, conserved THI-box structures [190, 234] proved to be riboswitches that recognize thiamine pyrophosphate (TPP) [191, 322]. RFN elements, predicted to adopt regulatory RNA structures in the leaders of genes directing the biosynthesis and transport of riboflavin from bioinformatic analyses [86, 298], were riboswitches that sensed flavin mononucleotide (FMN) [191, 323]. Similarly, S-box RNA structures common in low G+C Gram-positive bacteria [103] were riboswitches that bound the coenzyme *S*-adenosylmethionine (SAM) [185, 326]. More surprisingly, RNA sequences that were at one time thought to encode a regulatory leader peptide upstream of the *lysC* gene in *B. subtilis* [149, 213] proved to adopt a complex structure that directly binds lysine [105, 236, 269]. Finally, the long mRNA leader of a purine transport operon in *B. subtilis* known to be important for *cis*-regulation [48] harbored a guanine-sensing riboswitch [178], and variants of the same conserved RNA structure specifically recognized adenine [179].

Riboswitches were a widespread type of genetic control in bacteria. In total, six riboswitch classes that recognized seven different fundamental metabolites ranging from coenzymes to amino acids to nucleobases had been discovered in less than two years (Figure 1.1). Evidence for each of these riboswitch classes had been encountered piece-meal by researchers primarily interested in understanding regulation of a specific metabolic process or operon. Many of the original publications predated the era of microbial genomics, which began in 1995 with the determination of the complete

Figure 1.1 Riboswitch classes discovered by genetics

The conserved structures of the AdoCbl, TPP, FMN, SAM, guanine, adenine, and lysine riboswitch aptamers are shown alongside the chemical structures of their ligands. Red nucleotides are conserved in at least 80% of the known sequences and unfilled circles represent nucleotides whose identity is not conserved. R and Y represent conservation of purines (A, G) or pyrimidines (C, U), respectively, when a single base is not conserved at the 80% level. Other lines represent less-conserved structures, with allowed insertions of nucleotides and typical feature lengths labeled at certain positions. In the purine consensus structures, the base that determines the specificity of the aptamer through a Watson-Crick base pair to the ligand is marked with a triangle. Two unnatural RNA aptamers isolated by *in vitro* selection that bind to FMN [38] and guanine [144] are shown for comparison to the natural aptamers that bind those compounds. The molecular resolution structure of an aptamer [171] that binds to vitamin B_{12} (cyanocobalamin), which differs only at an axial ligand on the porphyrin from coenzyme B_{12} (adenosylcobalamin), has also been solved [271]. Only the minimal conserved structures are drawn for the three unnatural aptamers. Additional flanking nucleotides were present during selection and characterization of these sequences.

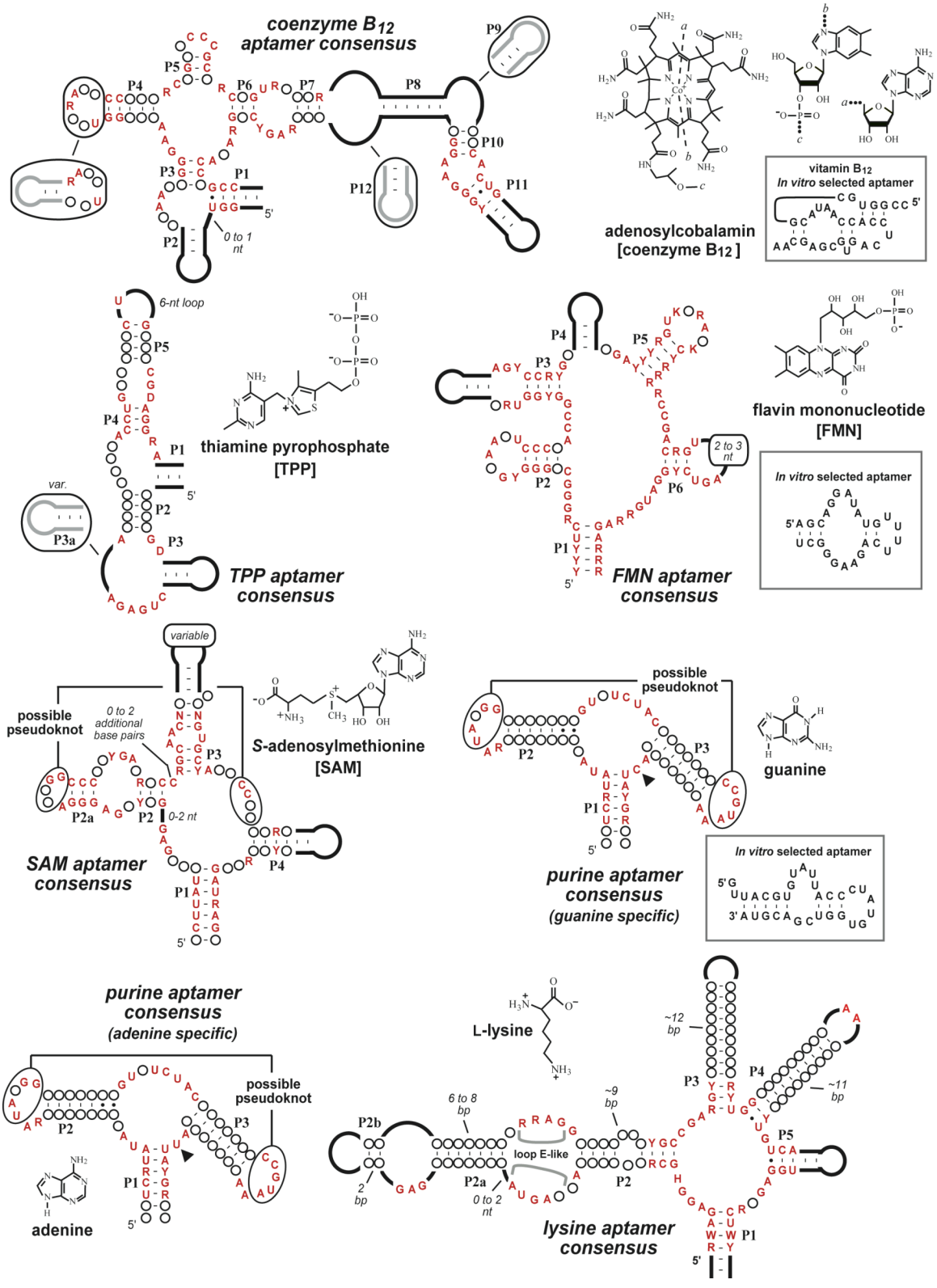


Figure 1.1 Riboswitch classes discovered by genetics

genome sequence of *Haemophilus influenzae* [78], and has matured over the last decade to a point where hundreds of complete microbial genomes have been sequenced. We wondered how prevalent riboswitch regulation was in biology and how many more natural aptamers we could discover in bacteria by designing a targeted comparative approach that mined these new genomic sequences.

Comparative genomics has been used to attribute functions to many suspiciously conserved DNA and RNA sequence motifs in genome sequences. The basic assumption of this approach is that genomic sequences that do not serve any function are unconstrained and will tend to be randomized by mutational drift after sufficient evolutionary time has passed. On the other hand, the essential conservation of functional sequences that contribute to an organism's survival will tend to be preserved in many of the present-day species that have inherited DNA from a common ancestor. To apply this approach, design choices must be made about what sequences to compare and what sort of sequence homology is expected. Integrating the output of multiple computational tools, cross-referencing information from relevant bioinformatic databases, and presenting this rich data in a form that aids comparisons and interpretation will increase the sensitivity and usefulness of predictions for specific downstream experimental applications. A schematic of programs and databases specifically useful for finding *cis*-regulatory RNA motifs in bacteria that are riboswitch candidates is shown in Figure 1.2. The following sections detail the assumptions and computational tools that we used to create two versions of the BLISS (Breaker Lab Intergenic Sequence Server) database, a resource for riboswitch discovery [19, 51].

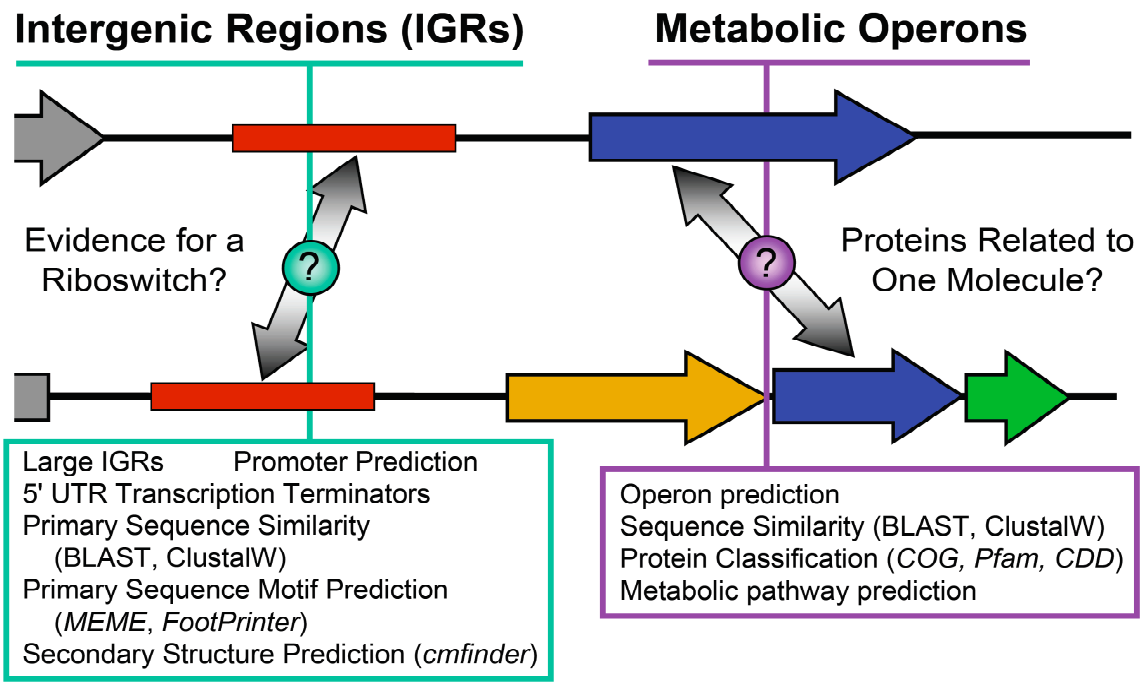


Figure 1.2 Comparative genomics approaches for identifying riboswitches and other structured regulatory RNA elements in bacteria

If conserved sequence motifs or certain intergenic region characteristics commonly exist upstream of genes with a similar function in diverged bacterial species then it is likely that these intergenic regions harbor a regulatory RNA element. Examples of programs and databases relevant for assessing intergenic region sequence similarity and characterizing the functions of protein sequences are shown in parenthesis. Descriptions of many of these tools can be found in Chapters 1 and 2.

1.2 BLISS: a database for riboswitch discovery

We sought to exploit the common features of known riboswitches to create a targeted comparative genomics discovery pipeline capable of identifying new bacterial RNA motifs that might function as riboswitches. We knew that many of the known riboswitch classes were generally more phylogenetically widespread in bacteria than other systems of protein-mediated genetic regulation. We therefore began our analyses with all of the completed microbial genomic sequences available from the RefSeq database at the time [223]. Examples of every riboswitch class were known to occur in the Gram-positive soil bacterium *Bacillus subtilis*, and some riboswitch classes appeared to be restricted to it and other species of low G+C Gram-positive bacteria in the *Bacillus*/*Clostridium* group (Firmicutes). Therefore, we restricted our searches to consider 91 genomes and only compared them to the *B. subtilis* chromosome sequence in the first version of the BLISS database (v1). We relaxed this assumption and conducted an all-versus-all comparison of 116 genomes for the second version of the BLISS database (v2).

Riboswitches are anomalously large noncoding elements for bacterial genomes, and they almost exclusively occur in the 5' untranslated regions (UTRs) of messenger RNAs encoding genes related to the biosynthesis, salvage, transport, or availability of a specific metabolite. Microbes tend to have "lean" genomes where 80-90% of the nucleotides encode proteins. Therefore, an unusually large intergenic region (IGR) adjacent to the beginning of an operon may be evidence that part of its sequence is transcribed and that its mRNA could therefore contain a sizeable 5' UTR regulatory element. Indeed, the median length of the 25 *B. subtilis* IGRs containing known riboswitches was 330 nucleotides compared to only 152 nucleotides for the complete set of 2913 IGRs we analyzed. We disregarded small IGRs with lengths less than 30 nt because they usually indicate that neighboring genes on the same strand are part of the

same transcriptional unit [193, 243], and small IGRs are not likely to harbor complex RNA structures.

Clearly, the possible existence of a large 5' UTR in an IGR is only the first and roughest estimate of whether a riboswitch resides there. The aptamer domains of riboswitches adopt complex structures that precisely recognize metabolites and exhibit sequence conservation typical of structured RNAs. There are short blocks (typically <14 nt) of primary sequence conservation in the context of base-paired stems containing compensatory mutations. These consensus features may be interspersed with inserted RNA structures of highly variable lengths without consensus sequences. While it would theoretically be best to look for conserved RNA features with an algorithm that simultaneously constructs an alignment and predicts a common secondary structure, even heuristic implementations of this general approach [245] for pair-wise alignments are too computationally expensive to compare the hundreds of thousands of IGR sequences in our dataset. Instead, we used BLASTN [7] to compare just the primary sequences of intergenic regions from a query genome — initially only *B. subtilis* (v1), but eventually all genomes (v2) — to all IGRs in the complete set of genomes to locate sequence homology that might be due to a common RNA structure. We added sequence alignments of all IGR matches identified by BLAST to the BLISS database so that they could be examined for evidence of conserved riboswitch-like structures. For known riboswitches, these alignments revealed a collection of base-paired stems that were supported by compensatory nucleotide mutations in many instances.

The other main distinguishing characteristic of microbial riboswitches is that they always occur upstream of genes involved in the biosynthesis, transport, salvage, or utilization of a single small molecule metabolite. In order to evaluate the genomic contexts of conserved intergenic sequences, we systematically predicted COG functional classifications [274, 276] for all gene products in the genomes of the BLISS

database. We then tabulated the frequency with which proteins in various COGs appeared in the putative operons downstream of BLAST matches to each IGR. Once again, this reckoning assumed that genes on the same strand within 30 nt are part of the same transcriptional unit. Adding this information to the BLISS database allowed us to quickly judge the regulatory potential of putative 5' UTR sequence elements.

In most instances, known riboswitches in *B. subtilis* control the formation of Rho-independent transcription terminator structures. Metabolite binding disrupts an alternate conformation with an antiterminator stem, allowing the formation of a characteristic GC-rich stem-loop directly followed by a run of consecutive uracil bases that causes RNA polymerase to prematurely terminate before any open reading frames are transcribed [107, 329]. These regulatory terminators can often be distinguished based on their IGR positions from constitutive terminators which occur at the ends of transcriptional units. Regulatory terminators are located in 5' UTRs between upstream riboswitch aptamer domains and a downstream coding region on the same strand, while constitutive terminators occur shortly after the end of an upstream gene on the same strand. BLISS (v1) incorporates TransTerm [72] predictions of intrinsic terminators in order to identify conserved UTR elements with riboswitch-like regulatory potentials. This prediction was not included in the successor BLISS database (v2) because riboswitches that regulate termination are rare in other bacterial divisions.

A web interface to the BLISS database (http://bliss.biology.yale.edu) allows IGR sequence alignments and associated evidence of riboswitch function to be interactively viewed and annotated. An index web page shows intergenic regions from a specific organism optionally sorted and filtered based on various characteristics of the IGR sequence or the genomic contexts and taxonomic distributions of its BLAST matches (Figure 1.3). Clicking on an intergenic region link brings up a more detailed web page

Figure 1.3 BLISS database organism IGR index web page

At the top of the page is a form with display, sorting, and filtering options for the list of IGR information tables from the selected genome fragment. The three top-ranked IGRs of the complete list are shown when sorting by the maximum number of times a protein in any one COG occurs in the putative operons downstream of its BLAST matches. The third table is labeled as a legend: "detailed view link" – links to a more detailed web page showing the BLAST hit alignment for the IGR (see Figure 1.4); "IGR coordinates" – strand (+/–), start, end in the GenBank record; "length" – number of nt in the IGR, "#BLAST hits" – number of matches to the IGR (number of matches where the downstream gene is on the same strand, i.e. that could be 5' UTRs); "GC%" – average over all BLAST matches of the highest G+C% measured in any 15 nt window for each hit; "annotation panels" – colored banners created from annotation; "annotation link" – link to user-editable TWiki page of annotation for this IGR (See Figure 1.5); "genes downstream of IGR" – genes in the putative operon downstream of this IGR, one gene per line, showing the gene's GenBank protein accession number, name, COG assignments, and description; "microbial groups of BLAST hits" – abbreviated major microbial taxonomic divisions (number of matching IGRs in each group); "species of BLAST hits" – abbreviated microbial species designations, usually the first letter of the genus and first two letters of the species (number of matching IGRs in each species); "COGs downstream of BLAST hits" – a table showing information on the COG classifications of genes occurring in putative operons downstream of all BLAST matches. Each line has the COG id, number of species where that COG occurs in the putative operon downstream of a matching IGR, number of microbial groups that COG occurs in a putative operon downstream of a matching IGR, name of the prototype gene for that COG, and a description of the functions of proteins in the COG.

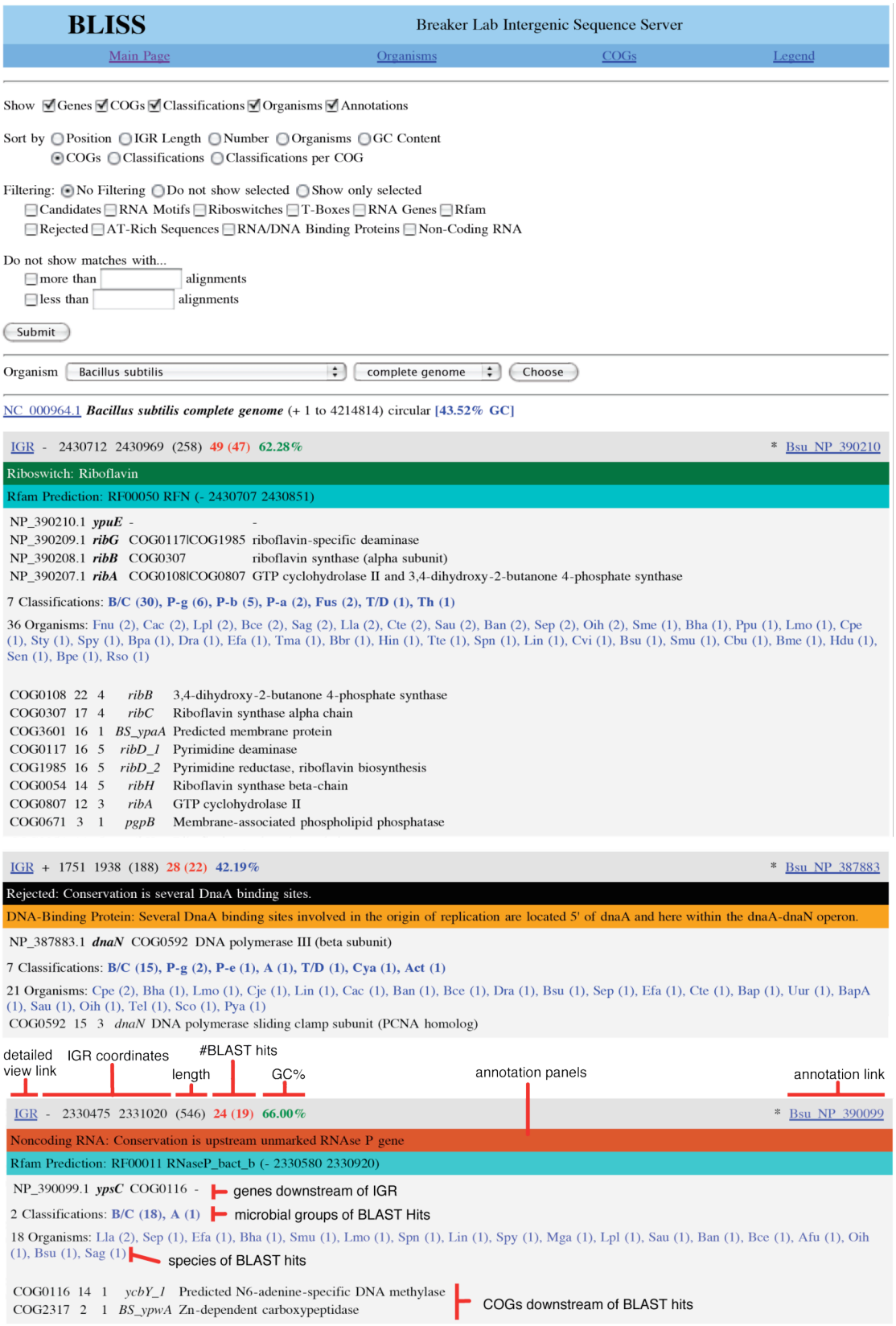


Figure 1.3 BLISS database organism IGR index web page

showing its nucleotide sequence, a table with information about BLAST matches to this IGR, and a multiple sequence alignment of these hits threaded onto the current IGR sequence (Figure 1.4). Clicking on links in the BLAST match table loads the detailed view of the IGR containing the hit, allowing a user to traverse a neighborhood of IGRs connected by BLAST hits to possibly discover more divergent sequence homologies.

A key feature of BLISS is the integration of the open source TWiki collaboration tool to coordinate annotation of known IGR sequence elements by multiple users [279]. TWiki allows freeform HTML web page revision by a community of registered users and supports full version control that records a history of all page edits. BLISS generates a separate TWiki webpage for each intergenic region automatically when a user chooses to add annotation. Keywords added to these pages are recognized by the web interface to prominently display banners with annotation information directly on the sortable IGR index (Figure 1.5). Our lab used this annotation capability to record known features of bacterial genomes such as noncoding RNAs and transcription factor binding sites (see below) as we mined IGR lists from selected organisms for new riboswitch candidates. Since 10-20 users eventually contributed annotation, this system was important for minimizing duplication and ensuring completeness during these efforts.

1.3 Results

We initially examined all intergenic regions with at least five BLAST hits in *Bacillus subtilis* for evidence of riboswitches (v1). Later, we surveyed *Agrobacterium tumefaciens* IGRs and selected IGRs near metabolic genes of special interest (e.g. small molecule "synthases") in other organisms (v2). Some regions of intergenic homology were immediately dismissed as uninteresting because they consisted of low complexity AT-rich DNA regions. This type of conservation is unlikely to form defined RNA structures

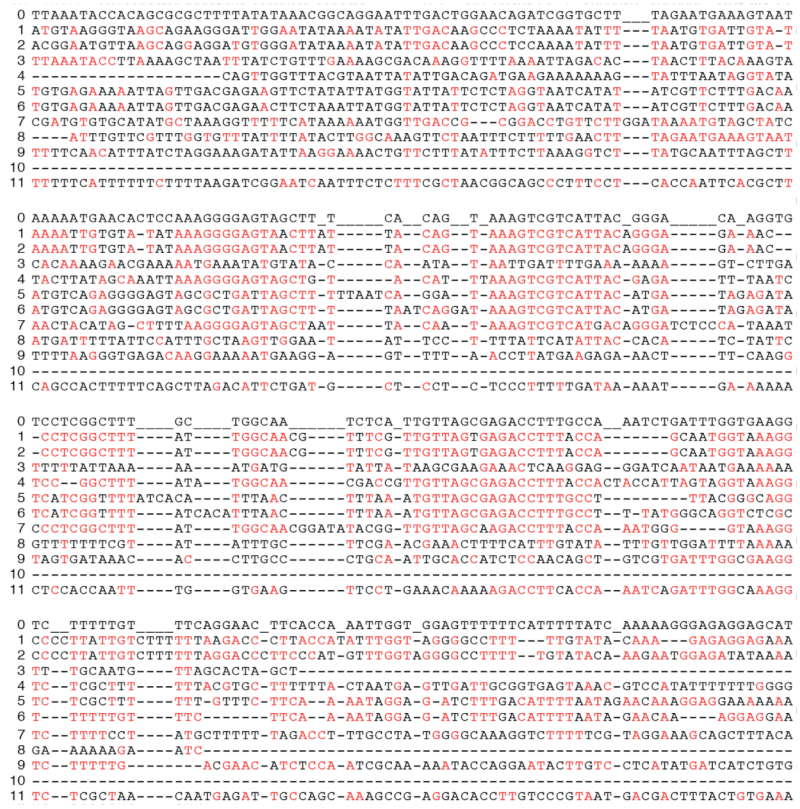
Figure 1.4 BLISS database detailed IGR view web page

The detailed view for the IGR upstream of the *B. subtilis yybP* gene is shown. At the top is the genomic sequence of the IGR. The accompanying "Islands" line is an *ad hoc* representation, where genomic T bases have been replaced by periods and stretches of at least three A bases have been replaced by commas, meant to emphasize regions of sequence with rich RNA structure potential. Rho-independent transcription terminators are also shown in this area when they are present. Below is the statistics table for the IGR as shown on the index page (see Figure 1.3 legend). Following that is a table describing the BLAST matches to the current IGR. Each line has a match index number that links to the detailed view page for that IGR, E-value for the hit, description of the nucleotide record containing the hit, abbreviation of the microbial group of its organism of origin, its GenBank accession number, its coordinates (strand, start, end), and the names of genes and COG classifications of proteins in the putative operon downstream of the matching IGR. Below this table is the alignment of the IGRs containing BLAST matches to the selected IGR. Red nucleotides indicate that a sequence is the same as the selected IGR at that position. Only part of the full alignment is shown for clarity. (Note that BLAST match #10 was to a part of the *B. subtilis* IGR that is not shown.) Finally, the information table for the reverse complement of the current IGR is shown if it could also be a 5' UTR, so that sequence homology that is really functional on the opposite strand can be more readily recognized. Note that only matches 1-2 and 4-7 have homology to the same part of the *B. subtilis* IGR. This results page was used to discover the *yybP* RNA motif (see Section 3.8).

Figure 1.4 BLISS database detailed IGR view (page 1 of 2)

Blast Hits

ID	Expect	Organism	Class	Accession		Start	End	Operon
0	0	*Bacillus subtilis complete genome*	B/C	NC_000964.1	-	4168821	4169252	***yybP [-]***
1	9.5e-12	*Bacillus anthracis complete genome*	B/C	NC_003997.3	+	823855	824206	***BA0809 BA0810 [COG3339 COG0586]***
2	9.8e-12	*Bacillus cereus complete genome*	B/C	NC_004722.1	+	809238	809588	***BC0827 BC0828 [COG3339 COG0586]***
3	6.9e-06	*Fusobacterium nucleatum complete genome*	Fus	NC_003454.1	-	273355	273623	***FN1792 [-]***
4	4.6e-05	*Oceanobacillus iheyensis complete genome*	B/C	NC_004193.1	-	3147335	3147701	***OB3036 [COG0861]***
5	0.001	*Listeria monocytogenes complete genome*	B/C	NC_003210.1	+	1021384	1021664	***lmo0991 [COG0861]***
6	0.001	*Listeria innocua complete genome*	B/C	NC_003212.1	+	1012984	1013263	***lin0990 [COG0861]***
7	0.003	*Bacillus halodurans complete genome*	B/C	NC_002570.1	-	2677723	2678028	***BH2553 [COG0861]***
8	0.003	*Bacteroides thetaiotaomicron complete genome*	Ba	NC_004663.1	-	3188299	3188508	***BT2554 BT2553 [COG1396 -]***
9	0.006	*Streptococcus pyogenes complete genome*	B/C	NC_002737.1	+	1375177	1376325	***No genes***
10	0.007	*Helicobacter hepaticus complete genome*	P-e	NC_004917.1	+	1262219	1262295	***HH1304 [-]***
11	0.009	*Bacillus subtilis complete genome*	B/C	NC_000964.1	+	4168821	4169252	***yybO [COG0477]***



Complementary Intergenic Region

IGR + 4168821 4169252 (432) **12 (6)** **48.00%** ***yybO*** ? Bsu_NP_391937

NP_391937.1 ***yybO*** COG0477 -

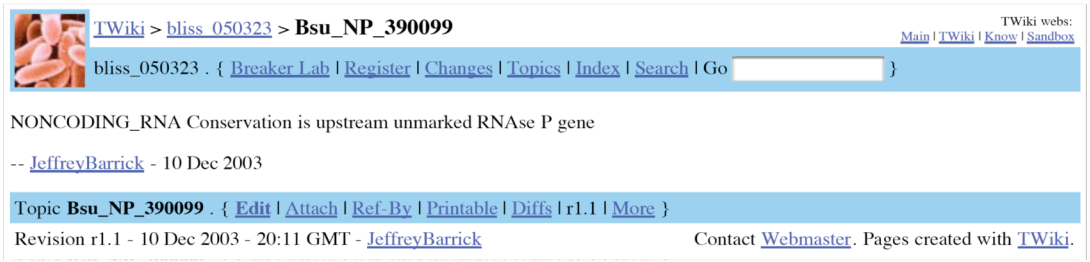
2 Classifications: **B/C (5), Ba (1)**

5 Organisms: Bsu (2), Bce (1), Ban (1), Spy (1), Bth (1)

COG1273 2 1 *BS_ykoV* Uncharacterized conserved protein

Figure 1.4 BLISS database detailed IGR view (page 2 of 2)

A



B

Riboswitch	Known riboswitch sequence	Keyword = RIBOSWITCH
T-Box	Known T-box sequence	Keyword = T_BOX
RNA Motif	Sructured RNA motif of unknown function	Keyword = RNA_MOTIF
RNA-Binding Protein	Known RNA binding site for a protein	Keyword = RNA_PROTEIN
DNA-Binding Protein	Known DNA binding site for a protein	Keyword = DNA_PROTEIN
Noncoding RNA	Known noncoding RNA	Keyword = NONCODING_RNA
Rfam Prediction	Hit to a known RNA element in the Rfam database	(automated)
Rfam Complement Prediction	Hit to the complement of a known RNA element in the Rfam database	(automated)
High Priority Candidate	Candidate for a regulatory RNA motif	Keyword = CANDIDATE_1
Medium Priority Candidate	Candidate for a regulatory RNA motif	Keyword = CANDIDATE_2
Low Priority Candidate	Candidate for a regulatory RNA motif	Keyword = CANDIDATE_3
Rejected	IGR rejected as a riboswitch candidate	Keyword = REJECTED
AT Rich Conservation	Conservation is AT-rich and unlikely to be an RNA structure	Keyword = AT_RICH

Figure 1.5 BLISS database annotation web page

(A) TWiki annotation page for the third intergenic region from Figure 1.3. It displays information on the last user who edited the web page and links to past versions of the page. This intergenic region has been annotated with the NONCODING_RNA keyword and that designation shows up as a colored banner on the information table for this intergenic region. ***(B)*** A full list of recognized annotation keywords and the colored banners that they produce. Rfam prediction banners are automatically added from a pregenerated database table of matches.

because of its degenerate sequence and the poor thermodynamic stability of A–U base pairs. We did not filter out these unstable sequences or low-complexity regions before conducting sequence comparisons, but later added a statistic that reflected the G+C content of BLAST matches to an IGR so that these hits could be flagged (v2).

While scanning IGRs in the BLISS databases for new riboswitch candidates, we encountered many conserved intergenic sequences that are known features of microbial genomes, including (1) repetitive sequences, (2) recognition sites for DNA-binding proteins, (3) leader peptides and misannotated open reading frames, (4) noncoding RNAs, (5) recognition sites for RNA-binding proteins, and (6) other *cis*-regulatory RNA elements. While each of these features is certainly biologically important and interesting in its own right, they were unintended byproducts of our riboswitch search strategy. Therefore, we catalog examples of the irrelevant elements that we encountered below and remark on how they can be recognized and distinguished from structured regulatory RNA motifs that are riboswitch candidates. We primarily draw on our first-hand experiences with *Bacillus*/*Clostridium* species and α-Proteobacteria for this discussion.

1.3.1 Repeat sequences

Mobile DNA sequences that self-propagate within genomes are present in a wide variety of bacterial groups and can be the source of tens to hundreds of copies of very specific intergenic repeat sequences. Families of insertion sequence (IS) elements [177] typically contain a single ORF encoding a transposase enzyme flanked by family-specific DNA sequences, usually consisting of 10-40 nt inverted repeats (IRs), that serve as recombination sites during transposition. When copies of these IRs are found adjacent to transposase genes or pseudogenes, they are easy to identify and ignore. However, transposition events that excise the mobile DNA element often leave behind characteristic sequence scars derived from the IRs, and these sequences are no longer

associated with transposase reading frames. Integrating DNA phages, retroviruses, retrotransposons, and more complex DNA transposon families can also leave behind repetitive sequences as a byproduct of their colonization and replication processes.

Short (≤200 nt) noncoding repeat elements that often contain palindromic regions are also found in many bacterial genomes. Their presence is best characterized in enterobacteria [13]. The *E. coli* genome contains 264 bacterial interspersed mosaic elements (BIMEs) made of combinations of 40-bp REP elements, 32 instances of 56-bp boxC sequences, 25 examples of 127-bp IRU elements, and 6 151-bp RSA sequences. We encountered similar elements, some of which had been previously studied, in α-proteobacteria. Several *Rhizobium* species contain multiple copies of a 108-bp RIME sequence [210], and *Caulobacter crescentus* contains 37 examples of 116-bp CIR elements that are also common in other α-proteobacteria [46]. The evolutionary origins of short palindromic repeat elements are unclear, although all of these examples are believed to be, at least partly, transcribed into RNA sequences.

Retrospectively, choosing *B. subtilis* as our first reference organism was fortunate. Its genome contains very few IS elements and no palindromic repeat families, whereas related species like *Bacillus anthracis* have hundreds of IS scars. Widespread spurious clustering of IGRs due to BLAST hits between repetitive elements renders the current BLISS implementation much less useful for organisms such as *E. coli* and *B. anthracis*. The top IGR in *E. coli* has a staggering 697 BLAST matches, mostly to other IGRs in its own genome or closely related enterobacteria. It is certainly easy to recognize that this intergenic region contains a repetitive element, but it would be a mistake to conclude that it cannot also contain a structured RNA element. Another section of the IGR sequence could be of interest, but the signal from a "modest" 10-20 BLAST hits would be dwarfed by the hundreds of matches to the repeat element.

Repeats could be masked from IGR sequences before BLAST comparisons to solve this problem. The program RepeatMasker is commonly used to filter repeats and low-complexity regions out of eukaryotic genome sequences [258]. The main obstacle to applying it to microbial genomes is that a suitable library of repeat element sequences does not exist. Identification of repeat element families, which are often peculiar to each microbial division and sometimes even confined to specific species, has lagged far behind the pace of microbial genome sequencing. A database of IS elements does exist [257], but it is unlikely to be complete enough to make this filtering generally useful in a collection of over 100 diverse microbial genomes. There is currently an unfilled niche for a computational tool that can be used to efficiently discover IS scars and short palindromic repeats from a microbial genome sequence, though *de novo* methods for predicting repeat families intended for eukaryotic genomes have recently been developed [18, 221]. Surveys of this nature would have practical applications. The high copy number, narrow taxonomic distribution, and rapid evolution of repeat sequences make them useful for strain identification, e.g. fingerprinting *Vibrio cholerae* strains and serotypes based on their IS1004 complements [26].

1.3.2 DNA binding sites for proteins

Our computational screen also detected some conserved DNA elements recognized by regulatory proteins. Transcription factor binding sites are generally quite short (≤17 nt). Their sequences usually consist of two symmetric half-sites separated by 3-5 variable nt that are symmetrically recognized by protein dimers. This sort of palindromic sequence conservation resembles a short RNA hairpin, and we have anecdotally observed that some alignments of protein binding sites even exhibit seemingly "compensatory mutations" where correlated changes in each half-site preserve the overall inverted repeat. For this reason, we generally ignored short BLAST matches consisting of

inverted repeats that did not have flanking conservation or appear to be embedded in more complex RNA-like structures. Many transcriptional repressors are autoregulatory (i.e. they bind to their own promoter regions) so the collection of genes downstream of BLAST matches to an IGR could sometimes be used as extra evidence to disregard their DNA recognition sites as potential riboswitches.

The most prominent DNA binding site in our results was the target sequence for the Fur repressor, an unusually widespread bacterial transcription factor that occurs both in *E. coli* and *B. subtilis* [15]. Fur regulates expression of around 20 operons in *B. subtilis* in response to iron availability by binding to a highly-conserved 19-bp DNA site. This "classical" site has been shown to actually consist of two overlapping 7+1+7 inverted repeats that define an overall site encompassing 21 nt [14]. This unusual mode of conservation may explain why the Fur binding site was so readily detected by BLAST searches. Other conservation that we encountered in *B. subtilis* could be assigned to BlaP [138], FadR [39], LexA [47], CtsR [57, 153], Fnr [228], and HrcA [204] transcription factor binding sites. We also noticed 16-bp Spo0J binding sites involved in chromosome partitioning [170] and DnaA binding sites located near the chromosomal origin that are involved in replication [84].

Several databases of transcription factor binding sites exist. The DBTBS database contains experimentally identified transcription factor binding sites specifically in *Bacillus subtilis* [133]. The PRODORIC database contains weight matrices describing transcription factor binding sites from diverse bacteria and offers a web tool for searching a user-input sequence for regulatory sites [195, 196]. These predictions generally have high false positive rates when applied on a genomic scale due to the limited information present in small transcription factor binding sites. Given the >100 transcription factors present in a typical microbial genome, these databases should not

be considered exhaustive, but they can be useful reference sources for rapidly identifying known DNA binding sites.

1.3.3 Leader peptides and misannotated reading frames

One of the α-proteobacterial motif candidates whose sequence alignments we investigated in depth turned out to be a leader peptide controlling branched chain amino acid synthesis upstream of *ilvB* genes [296]. Generally, cases like this where an entire small ORF is not annotated, can be discriminated from RNA sequence conservation by looking for characteristic mutation patterns. Homologous protein-coding regions will differ mainly at the third wobble positions of codons and by nucleotide insertions occurring in multiples of three that preserve the translational frame. However, leader peptides have very short sequences that are not well-conserved and may be difficult to recognize, It is also helpful to be generally aware that this mechanism usually regulates amino acid biosynthetic genes [151]

We have also encountered systematic start codon misannotations that left the conserved N-terminal portions of related protein sequences within intergenic regions in multiple organisms. These sequences appear to be *cis*-regulatory because they are always "upstream" of the same gene. However, their BLAST matches usually suspiciously extend continuously up to the (incorrect) start codon of each downstream gene, and they can be recognized by finding examples of complete reading frames in protein sequence databases.

1.3.4 Noncoding RNAs

Noncoding RNAs may be mistaken for structured *cis*-regulatory elements when they occur with the same genomic context in related organisms. Early on we recognized that several very promising RNA candidates with clear compensatory mutations and complex

base-paired structures were really RNase P RNAs and the reverse complements of unannotated tRNAs. The annotated 5' and 3' extents of ribosomal RNAs are also sometimes suspect, and rRNAs have highly conserved promoter elements [97] that will appear as intergenic sequence conservation on both strands. After we rediscovered the reverse complements of several known riboswitches (particularly SAM) with BLISS (v1), we added the results table for the reverse complement of an IGR (if it could *also* be a 5' UTR on the other strand) to the detailed IGR view pages in BLISS (v2) so that these misleading organizations could be more readily recognized (Figure 1.4).

Fortunately, there exist excellent tools for finding known RNA elements in genomic sequences. The program tRNAscanSE [172] can be used to accurately locate tRNA sequences, and the Rfam database [100] can be used to scan genomic sequences for over 500 RNA families. Rfam predictions were incorporated into BLISS (v2) and proved useful for removing other widespread noncoding RNAs such as SRP RNA and tmRNA, as well as the reverse complements of known RNA elements, from further consideration. Most of the known classes of riboswitches had even been incorporated into Rfam by this time.

Two motifs that we identified upstream of the *B. subtilis yrvM* and *yocI* genes had been previously reportedly to be stable noncoding RNAs of unknown function [9, 273]. We began to characterize these RNA structures as potential riboswitch candidates and later realized that they are structural homologs of *E. coli* 6S RNA (see Chapter 6).

1.3.5 RNA binding sites for proteins

B. subtilis employs many proteins that sense the availability of a certain metabolite and regulate gene expression by conditionally binding to mRNA leader sequences. These RNA recognition sites can be difficult to distinguish from riboswitches. They have conserved RNA structures, and binding of many of these proteins also regulates

downstream transcription terminators. However, these conserved RNA binding sites tend to be shorter than riboswitches since the RNA is acting as a passive recognition element rather than folding into a complex tertiary structure to act as a molecular sensor itself. Multiple RNA-binding sites for the UMP-sensing PyrR protein [173] and the tryptophan-binding TRAP system [12] are found in *B. subtilis* and related species. Regulatory proteins that sense sugar availability and bind to specific mRNA leader sequences to inhibit premature termination occur in various microbial species, e.g. BglG, SacT/SacY, LicT [241].

We also noticed several examples of sequence conservation upstream of ribosomal protein (r-protein) operons in *B. subtillis* and *A. tumefaciens*. Many *E. coli* r-protein operons are known to be autoregulated. One of the encoded r-proteins binds to the 5' UTR of the operon when it is present in excess of its available rRNA sites. This binding event alters the mRNA leader structure in a way that prevents protein synthesis [143, 331]. This regulatory strategy appears to also be conserved in many other bacterial groups, but the exact mRNA leader structures used for a given r-protein may differ dramatically among lineages. For example, distinct leader mRNA structures are recognized by the *E. coli*, *Bacillus stearothermophilus*, and *Thermus thermophilus* S15 proteins [263], and the mRNA sequences recognized by S4 are different in bacteria allied to *E. coli* and *B. subtilis* [102]. *E. coli* threonyl-tRNA synthase is also known to bind to a conserved structure in its own leader sequence and repress its own expression by competing with ribosome binding [242].

1.3.6 Other structured *cis*-regulatory RNA elements

We also identified two other types of conserved sequences associated with known RNA-based regulatory systems in *B. subtilis*. Conserved T-box elements occur upstream of most aminoacyl-tRNA synthetases and some amino acid biosynthetic genes in low G+C

Gram-positive bacteria [106, 120, 238]. T-boxes operate much like metabolite-binding riboswitches. Binding of a specific, uncharged tRNA species to the T-box element, partly through intermolecular base pairing to the anticodon sequence, induces an alternate antiterminator stem (that disrupts a transcription terminator) to activate gene expression.

The *B. subtilis pyrG* gene is regulated by a novel re-iterative transcription mechanism [187]. When CTP levels are low, RNA polymerase is unable to progress past DNA templating C bases near the beginning of this transcript and slips on three positions templating G residues, adding as many as 7 additional G residues to 5' end of the RNA. Normally an intrinsic terminator halts transcription within the 5' UTR, but this extended G stretch, which can be as long as 10 nt, enforces an alternate antiterminator RNA structure that allows read-through to produce the ORF for this CTP biosynthetic gene.

1.3.7 Riboswitches and riboswitch candidates

Our comparative procedure recovered 6 of the 7 classes of known riboswitches in *B. subtilis*. After eliminating explanations of IGR sequence conservation due to the elements described in the previous sections, several structured regulatory RNA elements remained that were candidates for new natural aptamers functioning in unknown classes of riboswitches. We expanded our initial sequence alignments of these candidates by conducting additional RNA homology searches and predicted improved consensus structures for these motifs with the methods described in Chapter 2. Descriptions of a total of 13 putative *cis*-regulatory RNA structures from *B. subtilis* and *A. tumefaciens* that we characterized in detail are given in Chapters 3 and 4.

1.4 Other comparative methods that discover riboswitches

Other approaches that attempt to identify regulatory motifs in microbial intergenic regions differ from BLISS primarily in their definitions of sequence similarity and the

amount of comparative information from other genomes they include. Several studies that have concentrated on the genome of *B. subtilis* can be directly compared to our results. One method predicts transcription factor binding sites by identifying overrepresented dimer motifs consisting of two sequence words of 4-5 bases separated by 3-30 bases within *B. subtilis* intergenic regions [199]. A second approach identifies PCEs (phylogenetically conserved elements) heuristically extended from exact 3 base sequence matches in pair-wise alignments of *B. subtilis*, *B. halodurans*, and *B. stearothermophilus* IGRs occurring upstream of orthologous genes [278].

BLISS (v1) finds matches from *B. subtilis* IGRs to all IGRs in 91 complete bacterial genomes using BLASTN and then assembles pair-wise alignments to the query IGR into a multiple sequence alignment. Unlike the other methods, it was designed as a tool for manually exploring promising motifs and does not assign detailed statistical scores to conservation. Instead, the interface displays predictions of downstream gene functions and intrinsic transcription terminators alongside sequence statistics and IGR alignments. An integrated tool allows collaborative annotation by users as they examine promising intergenic regions and develop secondary structure models for putative RNA elements.

There is considerable overlap between the regulatory site predictions of these three methods. In BLISS, riboswitches, T-boxes, and Fur protein binding sites dominate the list of *B. subtilis* IGRs with the most aligned sequences,although some other protein binding sites are apparent. The predictions of the dimer motif method [199] are largely complementary. It accurately captures sigma and transcription factor binding sites but does not detect the 19-bp Fur site and only discovers a single motif within the extended sequences of T-boxes. PCE predictions [278] are intermediate between these extremes. This approach detects protein binding sites and covers the conserved portions of several known riboswitches and T-boxes with multiple PCEs. Generally, the use of BLASTN

searches to identify similarity seems to introduce a bias towards finding longer sequence motifs that are more typical of riboswitches than protein binding sites.

After we published our study of *B. subtilis* with BLISS v1 [19] and work on BLISS v2 was underway, another research group reported a very similar computational pipeline that also disproportionately discovered riboswitch-like regulatory motifs [4]. They extracted intergenic regions upstream of proteins assigned to a COG from many bacterial genomes and identified noncoding sequence conservation with the Multiple EM for Motif Elicitation (MEME) tool [16]. The resulting motifs were re-searched against all IGRs to compile regulons (collections of different genes regulated by the same conserved element) with the Motif Alignment and Search Tool (MAST) [17]. The IGRs upstream of these gene sets were then used as input into MEME again, and this cycle was iterated until the predicted motifs and their regulons converged. They recovered an impressive array of biologically meaningful motifs that correspond to T-boxes, DNA- and RNA-binding protein sites, r-protein leaders, and many known riboswitches. Among their candidates for new regulatory elements they specifically reported a "new" high-scoring motif upstream of genes encoding the glycine cleavage system. We had actually identified the same conserved RNA structure (*gcvT*) with BLISS earlier [19], and already demonstrated that it functioned as a glycine-binding riboswitch (see Section 3.3).

MEME is probably a better tool for discovering RNA-like sequence homology than BLAST. It is designed to discover short ungapped nucleotide blocks represented by weight matrices separated by variable linker regions that are not conserved. This model can more accurately represent the conserved core of a natural aptamer than the straight sequence similarity discovered by BLAST where gaps and inserted divergent sequences are always penalized. Another advantage of their approach is that they were able to rank their predictions with p-values representing the probability that a predicted motif was correlated with genes in a certain COG by chance. Later, they created the RibEx web

site [3] that allows a user to search for their MEME motifs in a user input sequence or to display all predictions mapped onto complete genomes with the GeConT web tool [49]. While they tout this web site as a tool for identifying known riboswitches, its real utility is for exploring their motif candidates of unknown function. RNA homology searches using covariance models are far more sensitive and accurate for locating known riboswitch classes than their MAST motif strategy (see Section 2.2). Sensitivity is defined here as the fraction of known riboswitch examples that a search method recovers, and accuracy is the percentage of all its predictions that are truly riboswitches, i.e. not false positives.

Lastly, it is worth mentioning that a simpler comparative genomics technique that was initially intended to discover genes regulated by protein-mediated transcription attenuation mechanisms [151], is equally valid for detecting genes regulated by riboswitches. Here, instead of searching for conserved sequences, one identifies orthologous genes with an unusually high frequency of upstream Rho-independent transcription terminator hairpins positioned for a potential regulatory role. This approach has been used to determine COG functional groups that are correlated with upstream transcription terminators in 26 [158] and 180 [188] complete microbial genomes. We also looked for TransTerm terminator predictions upstream of certain COGs in ~100 microbial genomes (data no shown) and recovered many of the genes now known to be regulated by riboswitches and new RNA motifs discovered using BLISS. It is interesting that the first group specifically discussed conserved terminators arranged upstream of the *B. subtilis yqhI* gene and its orthologs in other organisms. This gene has since been renamed *gcvT*, and their attenuation prediction predated our discovery of the glycine riboswitch. The major downside of this approach is that it does not provide a sequence similarity stepping-stone to begin investigating the nature of the surmised regulation.

1.5 Conclusions

We created the BLISS database to identify new riboswitch candidates based on intergenic sequence homology correlated with conserved genomic contexts. We have successfully converted four conserved elements that we discovered with BLISS into new classes of riboswitches, and several other "orphan riboswitch" motifs that we discovered exhibit hallmarks of metabolite-binding riboswitch function although we have been unable to identify the ligands that they recognize (see Chapters 3 and 4). A number of other computational approaches compare favorably to BLISS and rank the same motifs highly. In all cases, it is necessary to weed out many spurious predictions of RNA sequence conservation that are due to a wide variety of known bacterial genome features to isolate *cis*-regulatory RNA motifs that are likely to function as riboswitches. To our knowledge, BLISS is the only resource that has been productively used as the source of a new experimentally validated riboswitch to date.

Computational platform

The BLISS database (http://bliss.biology.yale.edu) was created with an in-house suite of Perl scripts that runs command line sequence analysis tools, compiles comparative statistics, and populates a MySQL database. Execution of some computational tasks (e.g. BLASTN) was distributed over a small cluster of 4-7 CPUs. The BLISS database web pages are created on-the-fly by Perl scripts that require the CGI module and interface with the database. They are currently hosted on a dual processor G4 Macintosh running Mac OS X Server version 10.4.

Genome sequences

Genome sequences were obtained from the NCBI microbial RefSeq list [223]. This database maintains a nonredundant compilation of GenBank records, but in practice a

sizeable number are from different strains of the same bacterial species, and therefore nearly identical. We hand picked a list of genomes that attempted to minimize the inclusion of near-duplicate sequences that would bias our accounting of BLAST hit results toward these redundant genomes. We considered only IGRs with a length of at least 30 nt. Smaller intergenic regions are unlikely to harbor structured RNA elements and are usually part of the same transcriptional unit when the flanking genes are on the same strand [243] .

Organisms were classified into broad taxonomic groups based on the information in their GenBank records and the Comprehensive Microbial Resource at TIGR [218]. Our three-letter organism abbreviations are derived from the COG database when possible, typically the first letter is from the genus and the next two letters are from the species name. A complete list of sequence accession numbers and organism abbreviations is available via a link from each database's web site.

IGR sequence comparisons

We used NCBI-BLAST (version 2.2.5) to compare intergenic regions from a single reference genome to intergenic regions from all other genomes [7, 8]. The program BLASTN was used with a word size of 7 nucleotides, a gap open penalty of 2, a gap extension penalty of 2, and a nucleotide mismatch penalty of 2 (command line parameters: –W 7 –G 2 –E 2 –q –2). These choices were found to maximize the ratio of true positive matches (hits between two IGRs harboring a riboswitch aptamer) to true negative matches (hits between riboswitch-containing IGRs and IGRs that do *not* contain other examples of riboswitches in the same class) within the *B. subtilis* genome after scanning a panel of different parameter combinations. BLAST hits were symmetrized by taking the higher E-value for each directional query-subject pair of hits between two intergenic regions. IGRs containing bidirectional BLAST hits with E-values ≤ 0.01 were

individually aligned to the complete query IGR sequence using the program SSEARCH of the FASTA package (version 3.4) [214] with a gap opening penalty of 15 (command line parameter: –f –15). Pair-wise alignments to the query alignment were merged into a multiple sequence alignment by introducing additional gaps where necessary.

Gene function predictions

We used the COG database (September 2003) to uniformly assign gene functions to the genomic data sets [275, 276]. Specifically, each annotated protein gene was filtered with the COILS program (version 2.2) [174] and then compared to proteins in the COG database using BLASTPGP [8] with default parameters. Proteins were assigned COGs from these similarity results using the local version of the COGNITOR program [277]. Proteins that are the results of gene fusions are often assigned to multiple COGs. Gene descriptions for each COG are derived from the "whog" file of the database distribution. *E. coli* or *B. subtilis* gene names were assigned to generically identify protein products classified into COGs that contained proteins from these model organisms.

Terminator predictions

For BLISS (v1), Rho-independent transcription terminators were predicted using the software program TransTerm on data sets corresponding to all genomic fragments from each organism [72]. We modified the Perl scripts to ignore distinctions between head-to-tail and tail-to-tail intergenic regions when scoring terminator significance and to not combine confidence values for overlapping terminators on opposite strands. The altered script "smooth_confidence.perl" is available online (http://bliss.biology.yale.edu). We considered terminators with >98% confidence values to be high quality predictions.

Rfam predictions

For BLISS (v2), known noncoding RNA elements were predicted using the Rfam Database (version 6) [99] by searching complete genomic sequences with the provided "rfam_scan.pl" Perl script. Models for ribosomal RNAs and self-splicing introns (Group I and Group II) were omitted to decrease scan times.

IGR annotation

The BLISS database links intergenic regions to the open source TWiki collaboration tool [279]. TWiki allows web pages to be edited by any registered user and supports full version control implemented with RCS to record a history of all page edits. BLISS generates a separate TWiki page indexed according to the source organism's abbreviation and downstream protein accession number (e.g. Bsu_NP_391080) for each intergenic region automatically when a user chooses to add annotation. In addition to allowing freeform HTML annotation, keywords within these files are recognized by the web interface to prominently display information directly on the sortable list of IGR comparative information for each organism.

2 Strategies for defining regulatory RNA motifs

2.1 Introduction

After a putative *cis*-regulatory RNA element has been discovered it can be iteratively improved by identifying new sequence matches, verifying their genomic contexts, and modifying its consensus structure model to incorporate newly discovered variants. It is important to understand the algorithms behind different RNA homology searching methods to employ them most effectively. Similarly, it is helpful to be acquainted with how microbial genomes are annotated to correctly interpret the genomic contexts of possible matches. Finally, a basic knowledge of the typical structures of functional RNAs and how their properties are manifest in sequence alignments is useful for constructing high quality sequence alignments. This chapter is meant to be a brief tutorial that describes tools, databases, and approaches that we have found most useful for defining the structures and regulons of candidate riboswitch elements. It closes by describing the appropriate venues for sharing the valuable information created in this process with the greater scientific community.

2.2 RNA homology search methods

Many computational homology search methods can discover diverged homologs of a known RNA sequence and secondary structures [29]. We concentrate on describing the critical parameters for applying the programs that we have found most useful for identifying new examples of riboswitches and other regulatory motifs, roughly in order of the complexity of their underlying models for representing RNA structures .

2.2.1 BLAST and Smith-Waterman

As described in Section 0, most functional RNAs contain ungapped blocks of consensus sequence large enough to be detected by nucleotide BLAST (BLASTN) comparisons. BLAST programs achieve their remarkable speeds in part by using a search heuristic that begins by looking for exact consecutive letter matches of a specific length (known as "words") between the query and the database [7]. Furthermore, the BLAST algorithm really only examines words from the query sequence that are underrepresented in the database because they are most likely to be extended into statistically significant alignments. Therefore, BLAST will *never* detect homology between query and database sequences that do not share an uncommon stretch of consecutive nucleotide identity as long as the word size. Different implementations of BLAST allow different word size settings so that the user can tune this tradeoff between search speed and the detection of more distant homologies between sequences that may not share longer words but would score as significant if they were aligned.

When searching for RNA homologs with BLASTN, a smaller word size setting is generally preferable. As noted before, highly structured RNAs families tend to share only very short regions of exact identity that are interrupted by insertions or deletions (indels) and paired regions where compensatory mutations are common. Depending on the RNA, longer word sizes may eliminate diverged examples of a motif from being considered as possible matches by the BLAST algorithm. The minimum word size allowed in NCBI-BLAST for nucleotide searches is 7 nt [186]. The WU-BLAST implementation should be a more sensitive tool for RNA homology searching because it allows word sizes as small as 3 nt and a richer set of other scoring options [90].

BLAST remains the method of choice for annotating some large noncoding RNAs with high sequence conservation in genomic sequences. For example, the Mycoplasma

sequencing project at the Broad Institute [2] uses BLASTN comparisons to the European ribosomal RNA database [327] to identify 16S and 23S rRNA sequences in these bacterial genomes. The comparative RNA web (CRW) database [40] contains alignments of rRNAs and large self-splicing introns, and the Rfam database [100] maintains alignments of 5S rRNA, self-splicing introns, RNase P, snoRNAs, miRNAs, and other noncoding RNAs. Many of these RNAs are also commonly annotated in this fashion. For example, the noncoding RNA gene track for the human genome in the UCSC genome browser [124] that includes rRNAs, snoRNAs, and miRNAs was constructed using WU-BLAST searches with various optimized settings specific to each RNA family [66]. As expected, this annotation strategy generally has worse sensitivity and accuracy for smaller noncoding RNAs than it does for rRNAs.

When used iteratively, BLAST can sometimes traverse a surprising spectrum of functional RNA homologs. For example, TPP riboswitches were discovered in fungal genomes using the *E. coli thiC* riboswitch aptamer in an initial NCBI-BLAST query (N. Sudarsan, personal communication). This search identified hits in *Clostridium* species that, when used as queries in a second round of BLASTN searching, located matches in 5' UTR introns of genes related to thiamine biosynthesis in fungi. Aligning these putative TPP riboswitches to bacterial examples showed that they maintained key consensus bases and secondary structures, and the fungal aptamers were subsequently shown to bind thiamine pyrophosphate *in vitro* [268].

One may ask why BLAST word size heuristics should be used for sequence comparisons at all if they may miss valid RNA homologs. Indeed, with today's computers it is feasible to be rigorous and conduct a complete local Smith-Waterman search of genomic databases using the tool SSEARCH from the FASTA3 package [214]. Not surprisingly, this approach was found to be the most sensitive of all primary sequence methods for locating RNA homologs in a comprehensive test of different programs [80].

One practical advantage of NCBI-BLAST over SSEARCH is the ability to search very large sequence databases remotely via a web query [186]. When a local sequence database is available, SSEARCH is a potentially more useful alternative.

2.2.2 Pattern matching

BLAST and Smith-Waterman search procedures detect only primary sequence homology, and methods that integrate what is known about the secondary structure of a functional RNA should have a higher sensitivity and specificity. The simplest of these approaches is to create a template for a functional RNA family that contains information about its structure, including the presence and length of base-paired stems, sites of consensus nucleotides, and constraints on the distances between these elements. Several programs — notably RNAmot [156], RNAbob [62], and PatSearch [101] — implement essentially the same "pattern matching" approach, although they use different search algorithms and vary with respect to how the hybrid consensus/secondary structure query can be specified. The most mature and general pattern matching implementation is the program RNAmotif [175]. RNAmotif allows the user to specify an overall topology of base-paired stems and single-stranded regions that may include pseudoknots, and then write a detailed scoring function that rewards certain feature lengths, combinations of bases at paired positions, or consensus sequences. More interesting options even allow scores based on the compositional complexity of a matched subsequence or the nearest neighbor thermodynamic stability of sequences assigned to a specific stem.

Theoretically, this flexible framework should enable an expert to create a detailed description of functional RNA's structure that can be used to search for homologs with a high specificity and sensitivity. RNAmotif descriptors have been constructively employed in this way to classify RNase P RNAs [167] and locate new SRP RNAs in microbial

genomes [175]. However, pattern matching has more commonly been used to locate candidates for very small RNA structures with questionable results. When RNAmotif was used to identify intrinsic transcription terminators in the *E. coli* genome, it found a strong signal of real terminators between -10 and +60 nt relative to the 3' ends of genes [164]. However, it also had an extremely high false positive rate: 2586 of 6635 total predictions below the gathering score threshold were within protein coding regions. Pattern searching has also been used to locate structures that resemble the hammerhead ribozyme [76] and *in vitro* selected ATP, chloramphenicol, streptomycin, and neomycin B aptamers [157] in genomic sequences. However, none of these predictions has been experimentally validated, and the matches that they report typically contain large insertions between the functional elements or only marginally resemble known sequence constraints. A common mistake is to assume that variable insertions in loops are unconstrained and can be very long. In true functional RNAs these intervening sequences invariably adopt defined RNA structures.

The low specificity of pattern matching approaches can be overcome in the case of riboswitches or other *cis*-regulatory RNA elements because the genomic location of a hit provides independent confirmation that it is a true homolog. Only hits that occur upstream of operons related to the metabolite in question need to be considered and verified by fully aligning them to known motif examples. We created a C++ program called SequenceSniffer (J.E.B., unpublished algorithm) that allows the user to search for motif patterns built from conserved sequence blocks and base-paired stems separated by maximum length constraints. Degenerate hits to the motif (with fewer than a specified number of mismatches) are displayed alongside their genomic contexts for easy evaluation of their regulatory potentials. An example motif pattern for TPP riboswitches is — CUGAGA (200) ACYUGA (5) <<< GNUNNNNC >>> (5) CGNRGGRA — where angle brackets indicate base-paired positions and numbers in parentheses represent the

maximum nt lengths allowed between consensus elements [268]. SequenceSniffer was also useful for identifying additional examples of SAM [326], AdoCbl [201], lysine [269], and purine [178] riboswitches as well as more examples of *B. subtilis* regulatory RNA motifs [19]. Another research group has used the RNApattern program [297] to identify riboswitches and T-boxes during comprehensive comparative efforts to define genes involved in the metabolism of TPP [234], FMN [298], AdoCbl [235, 299], lysine [236], methionine [237], and various amino acids [212, 253].

Despite these successes, pattern matching approaches have two major drawbacks. First, it is time-consuming and difficult to create a pattern descriptor for a complex RNA element that is accurate enough to discriminate homologs from noise. Second, it is easy to unintentionally introduce constraints which rule out valid RNA elements whenever *ad hoc* patterns are manually defined a user in our experience, especially during the early stages of defining a new RNA element when little comparative structural information is available. The covariance model methods described in the next section introduce an automated and principled framework for overcoming these difficulties.

2.2.3 Covariance models

Covariance models (CMs) are generalized probabilistic descriptions of RNA structures based on stochastic context-free grammars (SCFGs) that offer several advantages over primary sequence homology searching methods [67]. CMs can be computationally trained on an input sequence alignment without manual intervention. They provide a more complete probabilistic model of the sequence conservation observed in functional RNA families that incorporates (1) first-order sequence consensus information, (2) second-order sequence covariation information, like base-pairing, where the probability of observing a base in one alignment column depends on the identity of the base in

another column, (3) insert states that allow variable-length nucleotide sequences, and (4) deletions states that allow omission of consensus bases. CM searches are implemented in the Infernal software package, which is distributed with outstanding documentation [310]. Infernal disallows pseudoknots in structural models so that it can use a three-dimensional dynamic programming algorithm for searches. Still, the additional complexity of modeling base pairs comes at a steep computational cost over the simpler profile hidden Markov models (HMMs) commonly used for modeling protein sequences which implement states and transitions for the other three categories of sequence conservation [61]. Until recently, this meant that scanning even a relatively small tRNA covariance model would take years on a sequence the size of the human genome [172].

To overcome this limitation, filtering techniques have recently been devised that speed up CM searches of sequence databases. They use a faster search algorithm to pre-screen input sequences so that only portions of the complete sequence likely to contain homologs are sent to the full CM for evaluation. Rigorous HMM-based filters convert the CM into a simpler HMM envelope model that runs more quickly and guarantee that all sequences matched by the full CM will be recovered [311, 312]. Other heuristic profile HMM filters can accelerate searches even further with negligible decreases in sensitivity for most CMs by relaxing this strict guarantee [313]. The RaveNnA computer program [310] wraps these filtering techniques around the Infernal codebase. We have used RaveNnA to find additional examples of glycine riboswitches [180], to discover divergent homologs of *E. coli* 6S RNA [20], and to define a variety of regulatory RNA motifs in α-proteobacteria [51]. These filtering methods also support local searches of a CM against a sequence database that discover high-scoring partial matches to specific substructures within the entire CM. This procedure can sometimes locate even more-diverged matches than the default global scanning mode.

A covariance model approach has also been adapted to the problem of finding homologs of a functional RNA when a multiple sequence alignment is not available. Given a single sequence and its secondary structure, the program RSEARCH (meant to be the RNA equivalent of SSEARCH) finds and aligns matches in a sequence database by using strand, pair, insertion, and deletion scoring parameters trained on alignments of known RNA structures rather than a multiple sequence alignment of the specific element [148]. Computational filters have also been developed to accelerate RSEARCH searches (Z. Weinberg, personal communication), and we have found them to be useful on occasion for discovering RNA homologs missed by full CM searches.

2.3 Defining the genomic contexts of regulatory RNA elements

We have seen how information about the genomic context of a potential hit discovered by pattern matching is important to validate that it is a true RNA homolog. Knowing genes that are typically regulated by an RNA element can also pinpoint regions that may harbor more diverged variant structures that will only be detected by targeted homology searches. Ideally, we would like to infer transcriptional and translational signals from genome sequences to completely define the collection of transcriptional units (TUs) and genes regulated by a new RNA element. Much of this annotation is created during the standard analysis pipeline applied to newly sequenced genomes [116]. This process uses a variety of computational tools, and it is useful to understand the relative accuracies of different types of predictions and how common errors can be detected.

2.3.1 Predicting transcription start sites

Classical methods for finding σ^{70}-dependent promoters in *E. coli* score potential matches with log-odds weight matrices for the –35 and –10 hexamers [121]. The consensus for these promoter positions is $T_{80}T_{95}G_{45}A_{60}C_{50}A_{96}$ for the –35 region and $T_{82}A_{84}T_{78}A_{65}A_{54}T_{45}$

for the –10 region (also known as the Pribnow box), where subscript numbers are the percentage of promoters with that base occurring at each position [166]. The hexamers are separated by 16–19 bp (optimally 17 bp), and the transcription start site (TSS) is usually a purine located 5–9 bp (optimally 7 bp) downstream of the -10 hexamer.

The general consensus sequence derived for *E. coli* promoters seems to be maintained by housekeeping sigma factors in most bacteria. Long ago, it was reported that RNA polymerase holoenzymes purified from different groups of bacteria all transcribe from a strong T7 phage promoter that normally operates in *E. coli*, supporting the notion that promoter recognition sites in most bacteria are very similar [318]. More recently, alignments of *B subtilis* σ^A promoters have shown that they share a very similar sequence consensus to *E. coli* σ^{70} promoters, at least with respect to the –35 and –10 regions [118]. We have also consistently observed promoters with this consensus upstream of putative regulatory RNA motifs in α-proteobacteria (see Chapter 5). Therefore, it seems reasonable to apply the same promoter identification approaches to define TUs in a wide variety of bacterial species.

However, promoter identification with weight matrices is bedeviled by the small amount of information that is present in a promoter alignment. Even when additional positions flanking the hexamer elements and extra columns near the TSS are also taken into account by these models, a typical result is that adjusting the score threshold to detect 80% of known promoters results in a false positive rate of 1% per position for intergenic sequences [127]. This means that a promoter will be predicted every 50 nt on average (when both strands are scanned). This clearly makes genome-scale predictions problematic and seems to indicate that a majority of promoter-like sequences may require additional positional cues from *cis*-activating factors that bind the DNA template to productively initiate transcription.

More complex promoter models that also learn from the mutual dependence of nucleotide sites and are trained on more of the approximately 60 bp of DNA template associated with RNA polymerase during initiation achieve modest improvements in precision. One example is the neural network promoter prediction (NNPP) program [229, 230], recently updated by adding a new parameter for the distance between –35 and –10 hexamers [37]. A committee support-vector machine (SVM) method trained on sequences from –150 to +50 with respect to the TSS also exists [95], although this algorithm (one might say unfairly) fits *translational* start sequences as well as transcriptional signals because start codons are lined up around ~30 nt downstream of TSSs for many *E. coli* genes. This would obviously be overfitting to a bad assumption for TUs with unusually large UTRs that contain riboswitches. Unusually low DNA stability in promoter regions relative to the rest of the genome can also contribute to improved TSS recognition [141].

Despite the high amount of uncertainty in any individual TSS prediction from these programs, they can still be useful when comparative information is available. When applied to all sequences in an alignment of a proposed regulatory RNA element, TSS prediction programs can give confidence that the relevant portion of the IGR is truly a UTR, or warn that part of the supposed RNA sequence conservation is actually a promoter element. Anecdotally, riboswitches seem to often have near-consensus promoter sequences that are easier to accurately detect. Perhaps it is preferable to begin transcription at these sites constitutively without a requirement for accessory transcription factors because most riboswitches repress gene expression. It is frustrating that none of the newer promoter prediction tools is available as a source code or executable distribution that can be directly integrated into a Unix-style computational pipeline. Of the current tools, the SVM implementation requires Windows, and NNPP is only available as a web-based form.

2.3.2 Predicting transcription terminators

Rho-independent (intrinsic) transcription terminators can be predicted with reasonable accuracy on the basis of sequence information alone. Predictions of transcription terminators have two uses with respect to examining candidates for new *cis*-regulatory RNA elements. First, they define the ends of transcriptional units and operons. Second, they are often specifically associated with regulatory elements that employ transcription attenuation mechanisms.

We have already encountered the program TransTerm that we used to predict regulatory Rho-independent transcription terminators in the BLISS database. TransTerm evaluates the thermodynamic stability of a candidate terminator hairpin with a simplified energy function and adds a U-tail scoring term [72]. It is only applicable to complete genome sequences because it uses annotation of protein reading frames to estimate the confidence of its terminator predictions by assuming that protein-coding regions inside ORFs are devoid of termination signals. In our experience, TransTerm predictions have a very high specificity during genomic scans, but an artificially low sensitivity because the hairpin scoring model is brittle. It misses many valid terminators because it allows at most one gapped position in the base pairing of the hairpin stem.

The program RNall can be used to predict transcription terminators in any arbitrary region of nucleic acid sequence [303]. Rnall predicts terminators based on three criteria: (1) the scaled thermodynamic stability (per nt) of a local RNA structure, (2) a U-tail score, and (3) a weak base pairing potential between the U-tail region and bases upstream of the terminator stem. In our experience, the default parameters for RNall do not allow for long enough stem-loops to predict many true terminators. With suitable adjustments to the three parameters used to discriminate terminators, we have found

that accuracies of ~90% can be achieved while maintaining a specificity of ~90% when predicting riboswitch regulatory terminators (see Section 5.4).

Like promoter sequences, intrinsic transcription terminators appear to function with a wide variety of bacterial polymerases. A detailed analysis of terminators in *B. subtilis* has shown that they differ only slightly from their *E. coli* counterparts [54]. Their hairpins have stems that are slightly longer (by ~2 bp) and more stable (~2 kcal/mol). However, some groups of bacteria do not appear to employ "standard" transcriptional terminators consisting of a stable hairpin followed by a U-tail [72, 290, 306]. Currently, no computational methods exist for predicting other types of transcription terminators (e.g. Rho-*dependent* terminators), and there is no *a priori* reason to assume that these sites could not also be regulated by a riboswitch.

2.3.3 Predicting open reading frames

In order to define what genes a putative RNA motif may be regulating, it is first necessary to predict where the open reading frames are in a microbial genome. This step is often one of the first annotation tasks carried out by a sequencing center after genomic assemblies are available. A typical computational pipeline uses Glimmer [56] to predict the locations of reading frames. Glimmer scores all possible proteins encoded in the six reading frames on a DNA sequence using a 3-periodic Markov model. It then resolves overlapping high-scoring ORFs according to a heuristic decision tree. This method is very sensitive — it misses < 1% of protein coding genes in a typical microbial genome. However, depending on how the model is calibrated and the score cutoffs chosen, this method can be prone to overprediction. Some early genome sequences (e.g. *Aeropyrum pernix* and *Pyrococcus horikoshii*) contain widespread overpredictions of hypothetical ORFs because annotators assumed that practically all of the nucleotide

sequence should be assigned to proteins and that these organisms might have many unique amino acid sequences [193].

The stop codon for a reading frame can be reliably predicted, barring rare sequencing errors, but choosing the correct start codon can be difficult for these algorithms. Practically speaking, perhaps 2-5% of open reading frames near riboswitches have incorrect start codons that greatly overlap the riboswitch and are probably misannotations because alternate downstream start codons exist. Post-processing high-scoring reading frames with the RBSfinder program [272] can sometimes be used to clear up these ambiguities, but this is not always done. These cases can also be easily noticed by conducting a BLAST search of the offending protein reading frame and looking for nonhomologous N-terminal extensions.

2.3.4 Predicting gene functions

Bioinformatic methods for predicting the cellular functions of protein open reading frames are typically based on finding sequence homology to a protein in a model organism that has been experimentally characterized. However, it is not sufficient to take the top-scoring protein hit from a BLASTP similarity search of a database and simply transfer its annotation to the new sequence. This method is prone to a well-known artifact. It will not discriminate between protein that are orthologs (they evolved from the same ancestral sequence and have retained the same function) and proteins that are paralogs (they evolved from an intragenomic gene duplication at some time in the past and are likely to have diverged in function since then). It is common to require best-bidirectional BLAST hits between proteins in two complete organisms to be sure that a protein is not a diverged paralog — where a better match in the genome exists. However, even this procedure can give false orthologs (and consequently annotation propagation) when

dealing with a large protein family where some orthologous proteins are missing from each complete genome.

Several databases have been developed that globally merge similar protein sequences from many organisms into clusters of families in a partially supervised manner. Each resultant cluster corresponds to proteins with similar, if not identical, functions. The COG database contains clusters of protein sequences primarily from microbial genomes created by iteratively merging triangles of BLAST hits and then manually splitting over-clustered groups [276], and the KOG database follows the same procedure with proteins from eukaryotic genomes [274]. The COGNITOR tool reads in BLAST results comparing an amino acid sequence query to proteins already assigned to COGs, totals significant matches to different COGS, and uses a majority rule to assign domains of the query protein to COGs (or KOGs). The Pfam database consists of curated sequence alignments of thousands of protein domains with annotated functions and references [77]. The HMMER program [63, 64] can be used to search a query amino acid sequence against profile HMMs trained on the Pfam alignments to assign functions to protein domains.

The Conserved Domain Database (CDD) subsumes the protein sequence alignments from these three sources (COG, KOG, and Pfam) as well as the SMART database [165]. It clusters these families to reduce redundancy, and adds its own additional domain families [181]. Position-specific scoring matrices (PSSMs) are created from the source alignments so that a query amino acid sequence can be rapidly matched against these profiles using the RPS-BLAST tool available from NCBI [182]. The CDD database is currently the most comprehensive database of protein domains available. However, many of the assignments from even this resource do not provide predictions of specific functional roles for many proteins. A common case is ATP-binding

cassette (ABC) transporters where sequence information cannot currently be used to reliably separate the many related families that are specific for different solutes.

The most rigorous method for clarifying ortholog/paralog relationships to arrive at a specific gene function in these difficult cases is to reconstruct a phylogenetic tree that captures the evolution of the protein family. Usually, it is sufficient to use a simple neighbor-joining tree such as the guide tree that CLUSTALW creates to hierarchically merge sequences into a multiple sequence alignment [280]. More thorough phylogenetic analyses of multiple sequence alignments can be conducted using the various tree-building methods and evolutionary models in the PHYLIP package [74] or other phylogenetic inference tools. There are also cases where genome context, specifically what other genes (whose functions might be more easily predicted from sequence) exist as part of the same operon or a conserved operon structure found in different bacteria, can be used to arrive at a more accurate specification of gene function [130].

2.3.5 Predicting operons

Genes on the same strand separated by fewer than 30 nt are usually part of the same transcriptional unit in *E. coli* [243], and also in many other bacteria [193]. This one-size-fits-all distance approach for predicting operon structure is roughly 80% accurate. A more involved method estimates a genome-specific distance discriminator based on observed differences in the distributions of intergenic distances between adjacent genes on the same strand and on opposite strands. The latter is assumed to represent the non-operon distance distribution, and this method seems to perform better for certain bacterial groups [222]. The same study found that including information about the function of genes (their COGs) and assuming that adjacent genes with related functions are more likely to be part of the same TU (essentially the inverse of predicting gene functions from operon co-occurrence) only slightly improved the accuracy of predictions.

In our experience, it is usually sufficiently accurate to use the 30 nt cutoff when conducting large-scale comparisons where a few missed predictions of TUs continuing across long intergenic distances will not greatly affect the results (e.g. in BLISS). When reporting the downstream genes that may be controlled by a specific instance of a regulatory element, a more inclusive threshold should generally be used. It is customary to note especially long (> 100 nt) intergenic regions when they are suspected to be present between genes in a single operon based on comparative genomics but experimental evidence of this association does not exist, for and example see [235].

2.3.6 Visualizing genomic context

At some point, the output of an RNA homology searching program needs to be combined with genome context data so that each candidate can be easily evaluated. We use an in-house web-based utility and a local MySQL database of CDD predictions to visualize the genomic contexts of results from RNA homology searches. This Perl script uses the Bio::Graphics modules of BioPerl [264] and can output context drawings in bitmap PNG or vector SVG formats. To aid recognition of valid genomic contexts, genes with the most commonly occurring conserved domains are highlighted with the same colors throughout. An example of a low-scoring hit that is a true homolog of the AdoCbl riboswitch based on its genomic context is shown in Figure 2.1.

2.4 Constructing multiple sequence alignments

Once additional homologs of a regulatory RNA with the correct genomic context have been discovered, they should be aligned to existing sequences. This validates their RNA structure and enriches the information about the functional RNA family that is present in the multiple sequence alignment. Although covariance models can be used to align new

Figure 2.1 Genomic context of AdoCbl riboswitch search results

Pictured is a portion of the results (sorted by bit score) for a RaveNnA scan of the RefSeq 14 database for matches to the adenosylcobalamin riboswitch aptamer. Aptamer matches are shown as black arrows on the nucleotide position ruler. Other features (mostly ORFs) are shown as arrows below the ruler. Despite scoring worse than CM matches positioned within known protein reading frames or near unrelated genes, we can be confident that match #561 is a functional AdoCbl riboswitch because it occurs upstream of a coenzyme B_{12} transport operon. Aligning this *Vibrio fisheri* sequence to known cobalamin riboswitches shows that it does indeed preserve the consensus secondary structure. The unusually low score seems to be due to the cumulative effect of omitting two stems that occur in most other AdoCbl riboswitch aptamer sequences and the use of bases that only infrequently occur in other aptamers at several consensus positions.

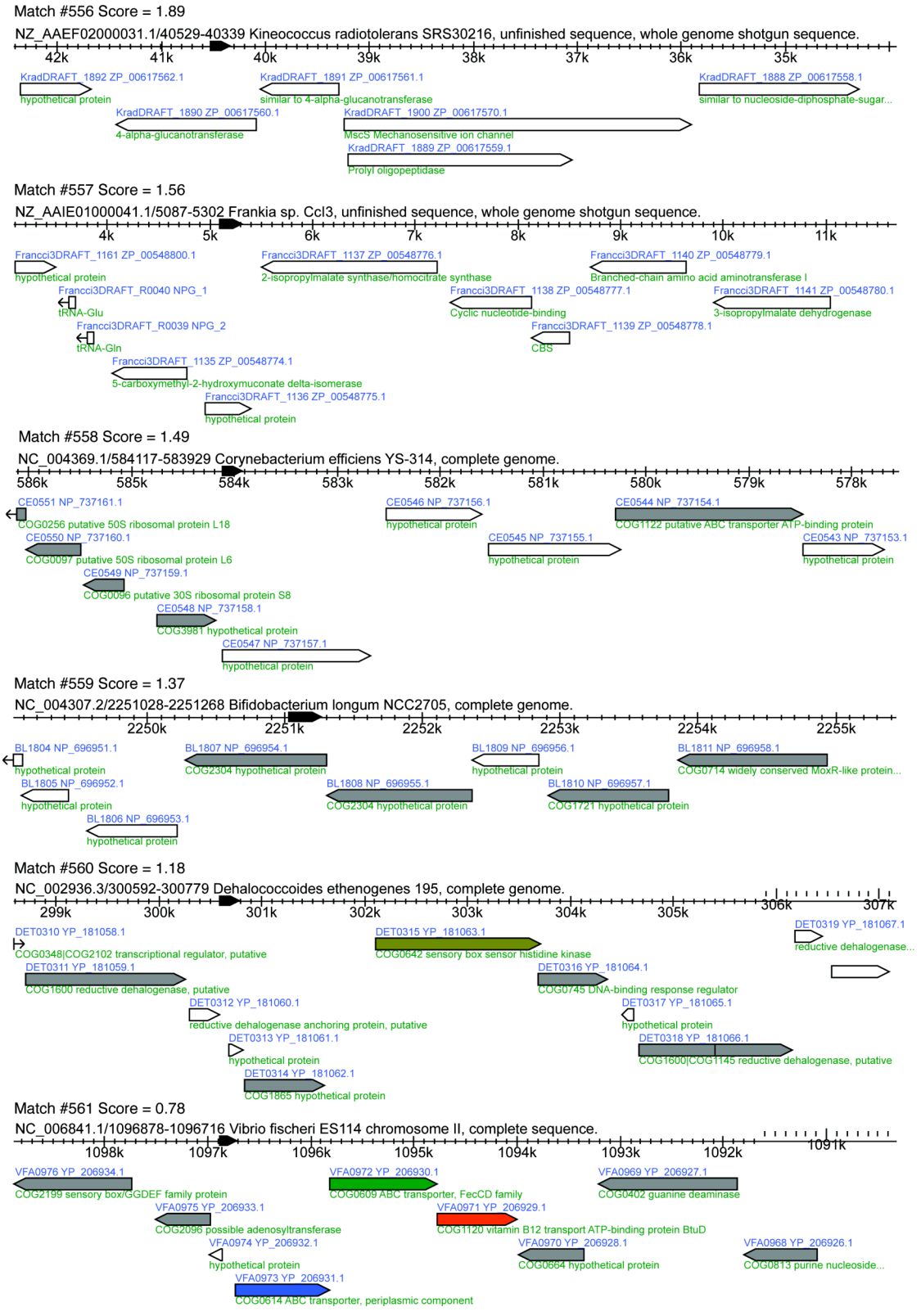


Figure 2.1 Genomic context of AdoCbl riboswitch search results

sequences to an existing alignment, human intervention is necessary to introduce new pairing elements or recognize new RNA structural motifs so that they can be incorporated into the underlying secondary structure model. Various computational tools, experimental methods, and general rules of thumb are useful for constructing high-quality multiple sequence alignments (MSAs) that approach the real RNA structure.

2.4.1 Automated alignments

Covariance models and, to a lesser extent, pattern matching methods provide some information about how a query sequence was aligned to the model as part of specifying a hit. The first step in adding sequences to an existing alignment should be to use these tools. Infernal includes the program cmalign to globally alignment input sequences to an existing CM. The output includes detailed information about how it matches specific nodes in the CM, and the results are usually quite accurate if the model captures the variation present in the sequences to be aligned. The primary limitation of the automated CM alignment procedure is that it cannot incorporate covariance information about pseudoknots. While it is understandable that modeling this information would make database searching prohibitively slow, this need not be the case for carefully alignment of several hundred sequences.

There is no need to include duplicate sequences or sequences from different strains of a given bacteria that may differ by only a few bases in hundreds of nucleotides in an alignment. While the computer will have no problem aligning these additional sequence representatives, they will make manual alignment and adjustments more time-consuming without adding information to the alignment. These representatives can always be recovered later, if desired, by re-scanning the sequence database with an improved model trained on an updated sequence alignment.

2.4.2 Manual alignment editing

Alignments should be stored in the Stockholm file format, which provides for general and readable markup of column, sequence, and file annotation. We primarily use RALEE (RNA ALignment Editor in Emacs) for editing alignments saved as Stockholm files [98]. As the name makes clear, it is an extension to the Emacs text editor that enables alignment editing. Versions of GNU Emacs are available for most computing platforms, although there are sometimes compatibility issues with respect to fonts and some functionality. RALEE features automatic base pair shading and split window viewing for base pairs that are separated by large insertions, as well as the ability to export alignments in a printable PostScript format.

Large blocks of conserved nucleotides should generally be considered anchor points when re-aligning sequences by hand. Smaller stretches of a few consensus residues that always occur adjacent to paired elements are also useful reference points. New sequences may have imperfect matches to these features that reflect compensatory mutations in what are actually paired rather than consensus positions. Large insertions between anchor points in diverged examples of an RNA motif are typically accommodated in new variable stem-loops. Any internal sequence with >10 nt is typically engaged in some pairing interaction that should be apparent in the alignment, although sequences may need to be realigned to make its presence obvious. Aside from these general guidelines for the kinds of variation to be on the lookout for in diverged matches to an RNA motif, it is useful to understand how some aspects of RNA structure impact MSA construction.

2.4.3 Non-Watson-Crick base pairs and common RNA motifs

Base paired stems in functional RNAs tolerate a number of non-Watson-Crick base pairs. The most common non-canonical pairs are G–U wobbles [85], which can be

accommodated within an A-form RNA helix. The next most common are *cis* Watson-Crick A–G base pairs [262], which tend to be found at the ends of helices because of their significantly larger C1'–C1' distances. One shortcoming of many structure and motif prediction programs seems to be the prediction of an unusually high number of G-U pairs. They may also predict pairings between stretches of the same nucleotide 4-or more of the same in a row on one strand that would be likely to slip in reality and lead to heterogeneous ensembles of structures. In our experience, tandem G–U pairs and repeating sequences in conserved helices are rare in riboswitch sequence alignments.

There is a general overrepresentation of single-stranded adenosine residues in 16S ribosomal RNA sequence alignments: 60% of adenosines are unpaired compared to only 30% of other bases [111]. This bias is borne out in other RNAs with complicated tertiary structures [109], and arises from the prevalence of adenine bases in many recurring RNA structural motifs, notably in forming A-minor tertiary interactions that pack in the minor grooves of A-form RNA helices [207]. Common RNA structure motifs and the most relevant references for identifying their conservation patterns are shown in Table 2.1. The identification of these structural motifs in a sequence alignment of a putative RNA element is very strong evidence that it is a functional RNA with higher-order structure.

2.4.4 Thermodynamic structure prediction

The RNAfold program [125] from the Vienna RNA package and Mfold [333] can be used to predict minimum free energy structures according to empirical nearest-neighbor base pair energy parameters and loop energy models. These thermodynamic structure predictions are generally very accurate for short sequences (<30 nt) that adopt independent hairpin structures. In our experience, their performance rapidly degrades in

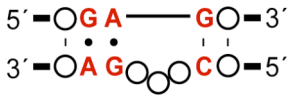
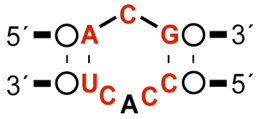
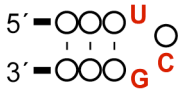
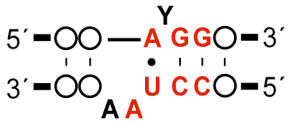
Structural Motif	Example	Consensus	Refs
K-turn	SAM riboswitch		[147, 163]
Bacterial Loop E Motif	Bacterial 5S rRNA		[160, 162]
Sarcin-Ricin Loop Motif	Bacterial 23S rRNA		[160]
T-loop	tRNA T-loop		[152, 200]
C-loop	*E. coli thrS* mRNA		[163]
UNCG Tetraloop	Bacterial 23S RNA		[69]
UNR-type U-turn	tRNA anticodon loop		[108]
GNRA Tetraloop	Group I Intron P4-P6		[140]
Tetraloop Receptor	Group I Intron P4-P6		[42]

Table 2.1 Common RNA Structural Motifs

Consensus structures of recurring RNA structure motifs that commonly form tertiary interactions are depicted with highly conserved nucleotides in red and commonly occurring nucleotides in black. Unfilled circles represent positions without a consensus base. Thick lines represent nucleotide sequences of variable length, and thin lines are used to show strand connections. Watson-Crick base pairs are shown as dashes and dots represent non-canonical base pairs. The references provided describe the conserved base pairing and stacking interactions in each motif and sometimes feature less common sequence variants that adopt the same structure. The IUPAC degenerate nucleotide abbreviations used are R = A, G; Y = U, C; K = G, U.

longer sequences (>50 nt) as the number of alternate structures that is possible multiplies and tertiary interactions including pseudoknots, which are not considered by these algorithms, become increasingly important for determining an RNA sequence's fold. We generally employ these programs to predict optimal base pairing in variable loop regions where it is clear that a structure that does not interact with the core aptamer fold forms. They can also be useful for brainstorming possible structures that a single sequence can adopt in the early stages of creating a secondary structure model. Compensatory mutations are generally stronger evidence of base pairing. If they also support any of the hypothetical pairings, then these helices should be incorporated into the phylogenetic model.

Finally, duplex predictions can be useful for predicting variable P1 stems or alternate structures involved in gene expression that are possible between pieces of an aptamer and downstream sequences. Here folding predictions can be simplified by extracting two nt regions from the alignment and looking for stable pairing between them without the other sequences present. This strategy will work even when the aptamer structure is not accurately predicted by a full thermodynamic treatment or when long-distance expression platform pairs (to a RBS, for example) are interrupted by long stretches of intervening and irrelevant bases. The RNAduplex program from the Vienna RNA package is well suited for this analysis.

2.4.5 In-line probing

We experimentally corroborate secondary structure models for conserved RNA elements using in-line probing [260]. In this assay, 5'-radiolabeled RNA is incubated for 1-2 days at 25°C in a slightly basic buffer (pH = 8.3) containing somewhat elevated Mg^{2+} concentrations (10-20 mM). The extent of spontaneous cleavage during this time at each internucleotide linkage in an RNA molecule is determined by separating degradation

products on a polyacrylamide gel with single nucleotide resolution. RNA cleavage occurs most rapidly at sites where nucleophilic attack by the 2' oxygen of a ribose approaches an "in-line" geometry with respect to the phosphorus atom and adjoining 5' oxygen leaving group. Typically, linkages next to base paired nucleotides in a structured RNA are rigidly held in a conformation that is far from an optimal in-line geometry. Therefore these sites cleave slowly. In contrast, internucleotide linkages that are in flexible regions of an RNA molecule occasionally sample an in-line geometry and are cleaved more rapidly. Therefore, regions with relatively low levels of degradation product in an in-line probing gel typically correspond to base paired or other structured regions of an RNA.

Mapping in-line probing information onto a secondary structure model can rule out certain pairings, but it may not unambiguously reveal the actual structure. There are many other structural probing methods that can be used on RNA, including probing with metals, ribonucleases, and chemical reagents, to gain other types of information [36]. We generally prefer in-line probing to these other methods because it does not rely on the addition of non-physiological metals or reagents that could potentially affect the native RNA conformation. This consideration is especially important when the examining structural changes in riboswitch aptamers that occur upon binding to small molecules.

2.5 Sharing the information in a sequence alignment

The process of iteratively searching for new examples of a *cis*-regulatory RNA element in a sequence database, adding new homologs with the correct genomic contexts to a multiple sequence alignment, and improving the secondary structure model eventually converges. The resultant multiple sequence alignment and table of regulated operons is a rich source of information about an RNA family that can be used to draft a detailed multi-level consensus structure figure. It can also be used to predict additional non-canonical base interactions using mutual information analyses and analyze the preferred

regulatory mechanisms of the RNA element (see Chapter 5). Finally, the MSA should be made publicly available by submitting it to the Rfam database and the accompanying insight gained about ORF positions and functions should be used to improve genomic annotations when possible.

2.5.1 Drafting consensus structures

There is currently only one strain of *Bacillus subtilis* sequenced [155], but the genomic sequences of 12 strains of *Bacillus anthracis* are available due to a targeted effort by TIGR to learn about this pathogen [70, 226]. This is an extreme case of the sequence duplication that will carry over through any search results and automated alignments of a new RNA element due to the inclusion of sequences from related organisms. Therefore, prior to constructing a consensus sequence and structure (or any other quantitative analysis of an alignment) it is necessary to down-weight the contribution of similar sequences in the MSA. Infernal uses the Gerstein-Sonnhammer-Chothia (GSC) algorithm [87] to weight the contribution of individual sequences in a multiple sequence alignment to CM emission and transition probabilities before training. The GSC approach weights sequences based on their distance (in nodes) from the root of a neighbor-joining phylogenetic tree constructed from the MSA. The WEIGHT program (available as part of the SQUID library distributed with Infernal) can be used to perform the same calculations and directly add sequence weights to a Stockholm file

A multiple sequence alignment is a rich source of information about the conservation and architecture of a functional RNA, but it is not a compact representation. When communicating this information to a general audience in a presentation or paper, the goal is often to create a consensus structure figure that has a higher information density and "rewards further study" [288]. RNA structure logos [96] and function logos [81] display the information content of columns in an RNA sequence alignment, but fail

to convey its structure. Various utilities distributed with the Vienna RNA package associated with the RNAalifold program [126] can automatically draw more interesting mountain plots that show base pairs as colored bands that reflect the combinations of pairs observed between those positions or create secondary structure figures with consensus information. However, we generally use Adobe Illustrator to draft RNA structure figures because of the ultimate flexibility that it allows.

For data preparation, we use a Perl script that adds per-column annotation lines representing a multi-level consensus to a Stockholm file. With typical settings this shows nucleotides conserved in 50%, 75%, 90% and 97% of the (weighted) sequences in the MSA — an almost logarithmic progression of added uncertainty. We usually depict nucleotide identities when a single nucleotide is ≥75% conserved and the presence of a nucleotide when one exists at a given position in ≥50% of representatives. Otherwise we draw thick lines to represent variable insertions. Stem-loops that are sometimes present are drawn in gray and shown as insets. When there are clearly two distinct subsets of structures in a particular region it is best to create separate consensuses for each and illustrate the possibilities with boxed insets (e.g. 6S RNA terminal loops, in Section 6.4). Examples of these graphics can be found in Chapters 3–6, with the most information-rich versions in Chapter 5.

2.5.2 Submitting alignments to the Rfam database

Once an alignment of a riboswitch or putative RNA element has been completed, it should be submitted to the Rfam Database [100], currently maintained at the Wellcome Trust Sanger Center (http://www.sanger.ac.uk) and mirrored with a different interface at Washington University in St. Louis (http://rfam.wustl.edu). The Rfam database collects multiple sequence alignments of functional RNAs from all domains of life, and uses covariance models to periodically search for additional examples of these elements in

updates to the EMBL sequence database. The eukaryotic genome annotation project ENSEMBL has recently incorporated Rfam into its annotation pipeline for eukaryotic genomes [27]. Rfam predictions have also been included in new interactive environments for annotating microbial genomes [293] and are appearing in new microbial genomes submitted to GenBank (e.g. Burkholderia cenocepacia AU 1054 and other recent DOE Joint Genome Institute projects). Thus, submitting an alignment to Rfam not only makes it available from a public repository, it also ensures that an alignment will remain relevant in perpetuity. As for other biological databases such as the Protein Data Bank [59], an Rfam accession number should be requested for a new RNA family and included in publications describing that element.

Figure 2.2 shows the main page of the Rfam version 7.0 (March 2005) entry for the lysine riboswitch (Rfam:RF00168). It includes a secondary structure drawing, brief description and classification of the RNA family, literature references, acknowledgement of the author and source of the alignment, and technical information describing how the covariance model was created and scanned. The manual SEED alignments used to create the covariance model and FULL automated alignments constructed with Infernal from all matches in the sequence database may be retrieved in many common MSA formats. There is also a link to view the matches mapped onto a species tree constructed from the NCBI taxonomy hierarchy. Note that the numbers next to each species name in this tree do not necessarily represent the multiplicity of a functional RNA in a single genome. It lumps together occurrences in all sequence records in the database that have been classified as derived from that species. For example, the same lysine riboswitch is found in each of the complete genomes of several different strains of *E. coli* (K12, O157:H7, CFT073, etc.), as well as earlier independent submissions of several chromosome regions (e.g. *E. coli* lysC gene promoter region).

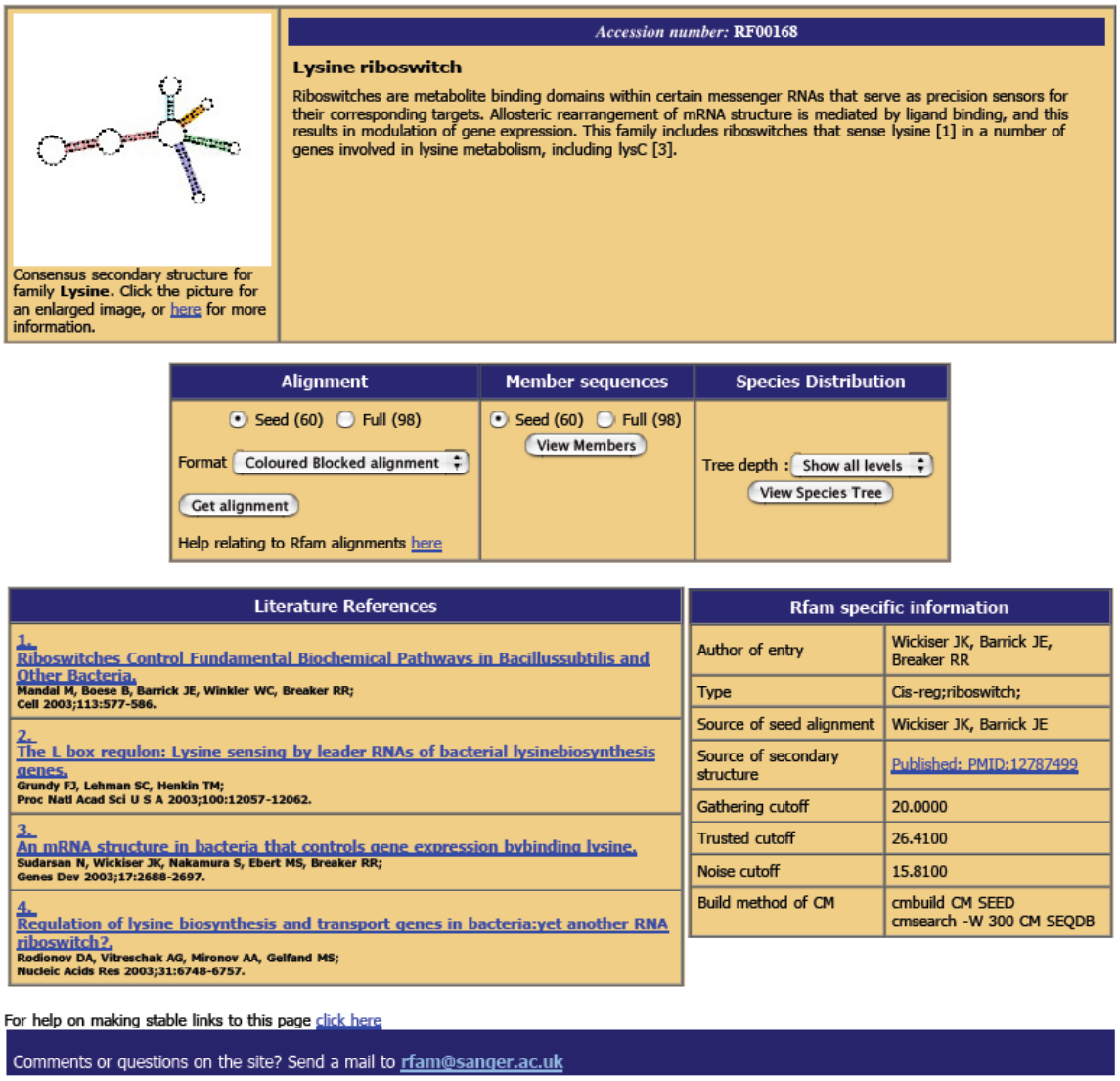


Figure 2.2 Lysine riboswitch entry from the Rfam database

Various aspects of this example Rfam database entry are described in the text.

When an Rfam covariance model is created, a score cutoff that separates real from false hits must be empirically determined (see Section 2.2.3). The *trusted cutoff* represents the lowest score of any sequence match believed to belong to an RNA family. The *noise cutoff* is the highest scoring sequence hit that is believed to *not* be a true example of the functional RNA. Rfam maintains a zero false positive policy. Therefore, a *gathering cutoff*, used to determine what matches from a database search should be included in the automated full alignment, is chosen between the trusted and noise cutoffs, even if this means that known RNA representatives with lower scores must be left out. In practice, incompleteness is only an issue for small motifs where there is not much information content in the sequence alignment or for rare divergent variants of a functional RNA that are not well-represented in the SEED alignment.

It is important to understand this scoring system to be able to intelligently interpret Rfam results and notify the curators of rare errors. One known problem with the way automated updates are currently conducted is that spurious matches that score above the *gathering cutoff*, but below *trusted cutoff*, may be encountered as more sequences are added to the database. Also, despite the best efforts of the curators, they are not experts on the >500 RNA families in the database, and occasionally the gathering cutoff they choose is slightly incorrect. Thus, despite having by far the lowest score of 26.32 compared to the next lowest score of 62.5, one spurious lysine riboswitch (of 98 total sequences) consisting mostly of AU-repeats from a sequence record described as "Human DNA sequence from clone DASS-81K16 on chromosome 6" is included in the Lysine riboswitch alignment. There are also caveats for interpreting some of the sequence data in the underlying database. For example, some clones from ongoing eukaryotic sequencing projects deposited in the EMBL and GenBank databases that contain riboswitch aptamer sequences are clearly contaminating bacterial DNA.

In this vein, it is worth mentioning that Rfam does not actually run the covariance model searches on the entire sequence database because this would require prohibitively large amounts of computing time. Instead, it uses faster sequence comparison methods to eliminate sequences that are unlikely to contain homologs, then runs the CM search only in the remaining "filtered" subset. Since this filtering step might eliminate more than 99% of the original sequences from consideration, it can greatly accelerate searches. However, there is also the possibility that it removes some *bona fide* RNA matches from consideration. Version 7.0 of the Rfam database uses BLAST comparisons with a relaxed E-value threshold for filtering (the "rfam_scan.pl" Perl script provided on the web site automates this analysis). However, BLAST has a relatively poor specificity, and the Rfam maintainers plan to transition to the more sensitive HMM-based heuristic filters described in Section 2.2.3 in future releases.

2.5.3 Improving genome annotation

One of the most frustrating aspects of comparative genomics currently is that there is no mechanism to correct annotation errors in many databases. Many of the analyses that we have described for characterizing riboswitches discover small errors in microbial genomic annotations in the RefSeq database and suggest unambiguous corrections (Figure 2.3). Hypothetical ORFs that overlap riboswitch predictions and do not share sequence homology with valid protein reading frames in other organisms are overpredictions that should be eliminated. Misannotated start codons that cause genes to overlap riboswitches with unconserved N-terminal amino acid extensions should be corrected when compatible downstream start codons with better ribosome-binding sites exist.

Figure 2.3 Correcting genome annotation

The genomic contexts of several high-scoring matches from a RaveNnA scan of the RefSeq 17 database for AdoCbl riboswitch aptamers are shown. In each case the riboswitch prediction contradicts or clarifies existing annotation. ***(A)*** Protein reading frame overpredictions. Hypothetical proteins without conservation to protein products from other genomes overlap each of these riboswitch aptamers. These ORFs are unlikely to be biologically significant. ***(B)*** Start codon misannotations. Genes on the same and opposite strands have start codons that substantially overlap these riboswitch examples. In both cases the true start codons are probably downstream in the same reading frame. ***(C)*** Ambiguous protein functions. The putative Fe^{3+} siderophore ABC transporter operon is likely to really be specific for coenzyme B_{12} based on the upstream AdoCbl riboswitch. The putative hupN gene product (a nickel cheletase) is likely to be a paralog that is truly involved in Co^{2+} insertion into AdoCbl.

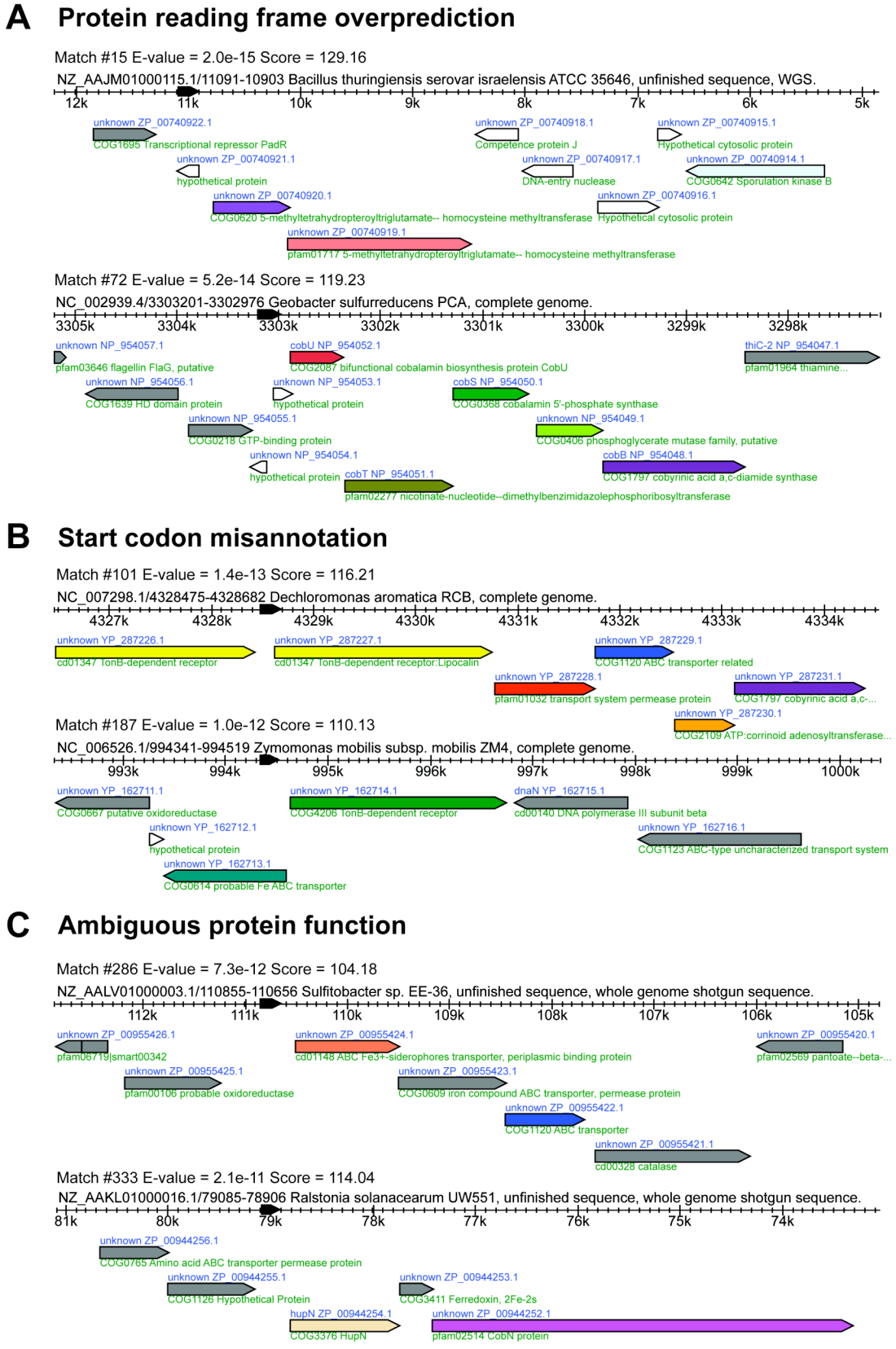


Figure 2.3 Correcting genome annotation

Riboswitch predictions also offer a golden opportunity to disambiguate the functions of downstream genes. This additional cue is especially important for properly identifying the exact substrates of transporter proteins where amino acid sequence homology fails to accurately delineate between different families. It can also discriminate between members of duplicated enzyme families that catalyze related reactions in completely different metabolic pathways. This value-added component is not unique to riboswitch predictions; it is relevant for other regulatory sequences such as Fur repressor binding sites regulating iron-related genes [15]. However, the large sizes and highly conserved signatures of riboswitch aptamers make them easier to accurately predict from sequence data alone than other regulatory sequences like the short inverted repeats recognized by most DNA-binding transcription factor proteins. Careful analyses of the TPP [234], AdoCbl [235], FMN [298], and lysine [236] regulons have provided a wealth of new information about genes involved in metabolism of these coenzymes, including predictions of new gene families related to their transport and alternate biosynthetic pathways.

Genome annotation teams have developed several different organizational models to include expert corrections that improve on the groundwork laid by automated gene prediction tools [265]. For microbial genomes, many different annotation environments have been created that seek to integrate and display the outputs of different computational analyses and allow manual intervention. They implement different collaborative models to enable coordinated updates by teams of annotators. The tool Manatee [1] allows any user to submit annotation without peer review, PeerGAD [53] registers approval of annotation changes by peers, and the ASAP environment allows any user to create annotation but requires a project leader to accept changes [91]. Although these tools are used for continuing annotation updates for a limited number of organisms (e.g. ASAP is used for several enterobacteria species),

curated annotation ceases for a vast majority of microbial genomes after publication and final acceptance into GenBank. Thus, while there are continuing efforts to maintain updated and evolving information linked to the scientific literature for the model microbial organisms *E. coli* [233] and *B. subtilis* [194], there does not exist a central mechanism to update venerable genomes such as *Haemophilus influenzae*, *Thermatoga maritima*, and *Aquifex aeolicus*. Consequently, their RefSeq entries have never been revised to a new version number. Although these records appear to be updated occasionally with new computational predictions of gene functions this process is not transparent and much of their annotation is currently at least 5 years stale.

Currently, the avenues for incorporating the corrections that a comparative genomics analysis creates into genome databases are limited. The Fellowship for Interpretation of Genomes (FIG) recently launched a Project to Annotate 1000 Genomes that focuses on using expert comparative analyses to annotate the protein components of biological "subsystems" [211]. The TPP, FMN, and AdoCbl biosynthetic pathways in FIG have been annotated based on work that integrated riboswitch predictions to predict genes involved in these pathways. Eukaryotic genome browsers such as ENSEMBL [27] and the UC Santa Cruz Genome Browser [124] allow anyone to contribute "tracks" that overlay the standard annotation. However, in all of these cases this additional annotation is never incorporated into the original GenBank flat files or core eukaryotic databases. It remains a supplement that is invisible to casual users and must typically be updated in parallel when the underlying sequence changes.

New database models are needed that allow continuing community annotation to improve and correct all genomic features. One model might follow the lead of the Wikipedia online encyclopedia (http://www.wikipedia.org). Wikipedia allows any visitor to change most entries, but maintains a record of all page edits and employs volunteers to ensure the quality of updates and revert vandalism. If a casual visitor notices a

misspelled word in the entry for "riboswitch" (http://en.wikipedia.org/wiki/Riboswitch) they can easily correct it. We need to make this level of interaction possible for data in the central repositories of genomic sequences. Many of the annotation errors that could be corrected in the process of a comparative analysis of riboswitches and their regulons are relatively minor individually (akin to spelling errors in Wikipedia), but they become substantial in aggregate. They are repeated annoyances to programs that seek to automatically extract information from these sources. In a sense, this is a failure of the scientific publication model. We need to ensure that primary sources are immediately corrected whenever new experimental observations are reported, perhaps by creating batch update tools and encouraging updates to be released and cited with publications.

3 New regulatory RNA motifs in *Bacillus subtilis*

3.1 Introduction

The first two chapters described general approaches for discovering and defining new *cis*-regulatory RNA motifs in bacteria. In the next four chapters we turn our attention to the RNA elements that we characterized with these methods. This chapter begins by describing eight regulatory RNA elements that we identified in the genome of *B. subtilis* using the first version of the BLISS database [19]. The consensus secondary structure models that we initially predicted for these motifs are shown in Figure 3.1, and some of the relevant properties of each motif are summarized in Table 3.1. Three of these RNA motifs have subsequently been proven to function as metabolite-sensing riboswitches. In these cases, I relate how their cognate metabolites were discovered and briefly discuss their unique properties. Updated secondary structure models and taxonomic distribution information for these proven riboswitches are provided in Chapter 5. Several of the five remaining *B. subtilis* RNA motifs have many of the hallmarks of known riboswitches, but the biological functions of these "orphan riboswitches" are still unknown. I present our current understanding of the genetic regulons and secondary structures of these putative regulatory RNA elements.

3.2 The *glmS* element is a metabolite-dependent ribozyme that senses glucosamine-6-phosphate

The *glmS* element was originally discovered upstream of the monocistronic *glmS* gene in 18 Gram-positive organisms (Figure 3.2 — this and all following figures are located at the end of the chapter). The *glmS* gene encodes glucosamine/fructose-6-phosphate aminotransferase, the enzyme that produces glucosamine-6-phosphate (GlcN6P) from

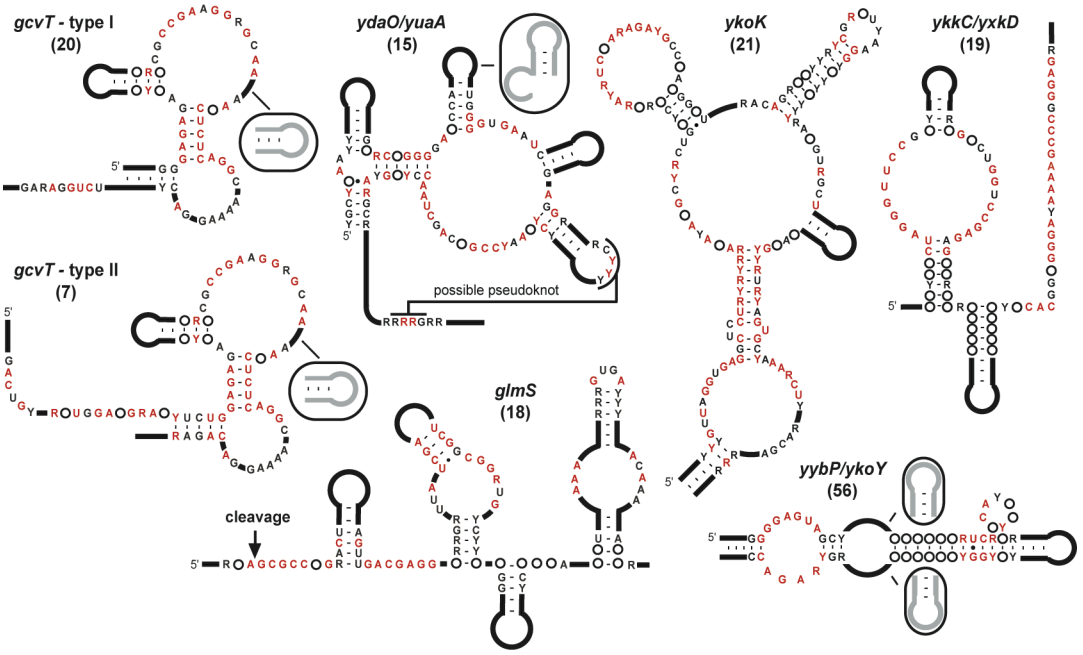


Figure 3.1 Secondary structure models for *Bacillus subtilis* regulatory RNA motifs

Red and black letters identify nucleotides whose sequence identities are conserved in greater than 80% and 95% of the representatives, respectively. Purine (R) or pyrimidine (Y) designations are given when a single nucleotide does not reach this conservation threshold. Unfilled circles represent nucleotides whose presence but not identity is conserved. Dark lines symbolize inserted stretches of nucleotides whose length is variable. Numbers in parentheses indicate how many sequence representatives are in the final alignment for each RNA motif. Two variants of the *gcvT* element (type I and type II) are depicted. The *ykvJ* and *ylbH* element alignments (not pictured) contain too few members to construct an accurate consensus model.

RNA Element	#	Distribution	Term	Reg	Gene Functions
glmS	18	B/C, Fus	NA	–	GlcN6P synthetase
gcvT	27	B/C, α, β, γ, Act	+	–	Glycine cleavage system, Na^+/alanine symporter
ydaO/*yuaA*	15	B/C, Act	+	–	K^+ transporters, membrane metalloproteases
ykkC/*yxkD*	19	B/C, Cya, α, β, γ, ε	+	+	Nitrate/sulfonate/bicarbonate and multidrug resistance transporter
ykoK	21	B/C, β, γ, Act	+	–	Divalent metal transporter
yybP/*ykoY*	56	B/C, Cya, α, β, γ, Act	+	+	Cation transport ATPase
ykvJ	9	B/C	+	?	PP-loop ATPase, tetrahydrobiopterin synthase, GTP cyclohydrolase related
ylbH	6	B/C	–	?	N6-adenine methylase, phosphopantetheine adenylyltransferase

Table 3.1 Characteristics of *Bacillus subtilis* regulatory RNA motifs

Motifs are named for the initial gene of each downstream operon in *B. subtilis*. The number of sequence representatives (#), evolutionary distribution (Distribution), presence of an intrinsic terminator before downstream start codons in Gram-positive bacteria (Term), and predicted effect of metabolite binding to the RNA element on gene expression (Reg) are shown for each element. Positive regulation was predicted when the conserved RNA element overlaps downstream regulatory transcription terminator stems such that the two structures are likely to be mutually exclusive. Negative regulation, presumably mediated by an antiterminator stem overlapping the conserved element, was predicted otherwise. ? = no prediction of regulation because the conserved element was poorly defined. Bacterial classification abbreviations: *Bacillus*/*Clostridium* (B/C), α-proteobacteria (α), β-proteobacteria (β), γ-proteobacteria (γ), ε-proteobacteria (ε), cyanobacteria (Cya), fusobacteria (Fus), actinobacteria (Act).

fructose-6-phosphate using glutamine as the amine donor. This reaction is the rate-limiting first step of the biosynthesis of the cell-wall precursor UDP-N-acetyl glucosamine. During our original in-line probing experiments to characterize the structure of this element, our RNA construct exhibited an extraordinarily high level of cleavage at a specific internucleotide linkage that exceeded the maximal theoretical rate of uncatalyzed backbone breakdown [68, 260].

Further studies showed that the presence of GlcN6P, but not closely related metabolites such as glucose-6-phosphate, further stimulated the cleavage rate of this 5' UTR element by as much as 1,000 fold [325]. The *glmS* element was a new type of metabolite-responsive ribozyme. A minimal ribozyme consisting of only the first two (of four) paired domains (Figure 3.3) is still capable of self-cleavage at a lower rate, and only a single nucleotide 5' of the cleavage site is necessary for activity. Deactivating mutations in the ribozyme domain that compromise its cleavage rate cause proportional derepression of a reporter gene fused to the *glmS* mRNA leader. Therefore, the *glmS* ribozyme serves as a ribozyme-riboswitch that turns off expression of the enzyme responsible for GlcN6P biosynthesis when sufficient levels of GlcN6P are present. The precise mechanism by which ribozyme cleavage leads to reduced gene expression remains to be elucidated. It is possible that it destabilizes the mRNA transcript and makes it prone to further degradation by cellular nucleases.

Subsequent work on the *glmS* ribozyme [Rfam:RF00234] has been aimed at determining its molecular resolution structure and the chemical mechanism of metabolite-triggered self-cleavage. Our laboratory has presented evidence that Mg^{2+} ions play a structural role in ribozyme folding rather than a catalytic role in *glmS* cleavage [239], unlike other natural ribozymes that catalyze phosphodiester bond transfer using divalent metal ion mechanisms such as the Group I intron [198, 266]. Work by others has further mapped the molecular discrimination characteristics of the

glmS ribozyme in order to understand whether GlcN6P acts as a catalytic cofactor (that takes part in the chemical mechanism of cleavage) or an allosteric effector (that induces a structural change necessary for cleavage). One of the most interesting findings was that tris(hydroxymethyl)aminomethane (TRIS) stimulates ribozyme cleavage at the typical millimolar concentrations it is present at when used as a buffering agent [184]. In 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) buffer, GlcN6P addition activates cleavage above the background rate by a staggering factor of >100,000-fold. The realization that TRIS can substitute for GlcN6P and pH profiles of cleavage in the presence of different ligand analogs support a model where an amine group within the ligand is required for catalysis. Further observations from nucleotide analog interference mapping (NAIM) experiments indicate that suppression of interference (i.e. NAIS) occurs within the catalytic core when glucosamine versus glucosamine-6-phosphate is used as ligand to trigger cleavage [136]. These data provide further support for the cofactor model because they imply that GlcN6P binds very near the cleavage site.

Shortly after our initial findings, multiple research groups including our own noted that a pseudoknot supported by compensatory mutations forms outside of the minimal active construct between sequences 3' of the reported consensus and the loop of the third paired domain [287, 319]. A second pseudoknot within the ribozyme core between the 5' single-stranded tail, including the cleavage site, and nucleotides within the second paired domain has been proposed from indirect mutational evidence [259]. Recently the x-ray crystal structures of precursor and product forms of the *glmS* ribozyme from *Thermoanaerobacter tengcongensis* have been determined (D.J. Kline and A.R. Ferre-D'Amare, personal communication). They show a doubly-pseudoknotted core that contains pairing similar to the first proposed 5' pseudoknot but shifted by several base pairs. Neither of these structures contains electron density for bound GlcN6P, and therefore they cannot conclusively prove whether it acts as a cofactor. Still, the authors

argue against a role for GlcN6P as an allosteric effector based on the small changes that they observe in the ribozyme's structure before and after cleavage. They also found that some mutations in pockets near the active site differentially affect TRIS and GlcN6P cleavage activation, possibly because they affect accessibility or docking of these two effector molecules differently. Other studies that crosslink the ribozyme in an active form or add Mg^{2+} for folding before glucosamine-6-phosphate to trigger cleavage are also consistent with a cofactor role [114].

3.3 The *gcvT* element is a cooperative riboswitch that binds glycine

The conserved element associated with the *B. subtilis gcvT* gene was originally classified into two similar structural types (I and II) with distinctive 5' and 3' ends that flanked a common central core (Figure 3.1 and Figure 3.5). A 195-nt RNA construct from *B. subtilis* encompassing a type I element exhibited evidence of extensive structure formation when it was subjected to in-line probing (data not shown).

The *gcvT* element occurs most often upstream of the *gcvTHP* operon encoding the glycine cleavage system, a multi-subunit complex that produces 5-10-methylene-tetrahydrofolate, ammonia, and carbon dioxide by breaking down glycine (Figure 3.4). This methylene-carrying form of tetrahydrofolate (THF) can be used to directly convert another molecule of glycine to serine or to convert 2'-deoxyuridyl-5'-monophosphate (dUMP) into 2'-deoxythymidyl-5'-monophosphate (dTMP). It may also be converted into other alkylated THF derivates involved in a variety of metabolic pathways including purine and methionine biosynthesis. The *gcvT* element often occurs adjacent to putative Na^+/alanine symporters, and in one instance it is located upstream of γ-aminobutyrate permease. Interestingly, the element also occurs upstream of a redundant serine hydroxymethyltransferase [45] and L-serine deaminase in *Mycobacterium tuberculosis*,

and a glycine/D-amino acid oxidase in *Mesorhizobium loti*. Both of these latter systems are used to convert glycine into pyruvate.

After some difficulty, we were able to show that both type I and type II variants of *gcvT* bind glycine with high selectivity [180]. The breakthrough discovery was that most of the time two *gcvT* motifs occurred in tandem. The type I and type II conserved upstream and downstream sequences are really portions of the opposite aptamer copy. Dissecting these two tandem copies into single aptamers often resulted in inactive RNA constructs. Furthermore, the structure that we initially proposed for each aptamer was flawed. An alternate interpretation of the in-line probing data that also agreed with the sequence alignment was possible. Searching for more aptamers with this reorganized secondary structure model (Figure 3.6*A*) found many more examples [Rfam:RF00504].

Of even greater interest, constructs that include two tandem aptamers bind two molecules of glycine cooperatively, i.e. binding at one site increases the affinity for subsequent binding at the other site. The overall effect is that the aptamers go from unbound to fully saturated over a narrower glycine concentration range than is possible with single binding site. Amazingly, the Hill coefficient for glycine binding is 1.6 (or 0.8 per aptamer binding site), which is comparable to the 2.8 value measured for hemoglobin's four subunits (or 0.7 per binding site). Using *in vitro* transcription assays to measure termination efficiency, it was possible to show that this "more-digital" binding response is translated into a sharper change in gene control as glycine levels rise. We have speculated that this unique mechanism might be required by a bacterium's need to reserve a certain level of glycine for protein synthesis but switch quickly to a regime where glycine is salvaged for energy if its concentration becomes higher.

It is also unusual that the *B. subtilis gcvTHP* riboswitch is an ON switch. It activates gene expression when glycine levels are high by disrupting an intrinsic transcription terminator hairpin that overlaps P1 of the second aptamer copy. Most other

riboswitches that have been characterized are OFF switches. Adenine riboswitches employ the only other confirmed natural aptamers known to activate gene expression [179]. Other examples of glycine riboswitches that regulate putative sodium-alanine symporters (that probably transport glycine) operate as OFF switches, making this the only riboswitch aptamer known to harness both ON and OFF genetic logic.

3.4 The *ykvJ* element is a miniature riboswitch that binds preQ$_1$

The *ykvJ* element was initially relegated to the supplement of our publication on *B. subtilis* motifs because we found only nine sequence representatives and were unsure of its structure. (This is why it is missing from Figure 3.1.) Genes in the downstream operons were annotated as PP-loop ATPases, tetrahydrobiopterin synthases, and GTP cyclohydrolase-related enzymes, but pterin compounds did not bind to our RNA construct. Soon thereafter, genes in the downstream *ykvJKLM* operon in *B. subtilis* were implicated in the biosynthesis of the hypermodified nucleotide queuosine by biochemical experiments after they were identified by a comparative candidate gene approach [227]. These genes have been renamed *queCDEF* although a precise role in queuosine biosynthesis is only certain for the QueF enzyme, which reduces a precursor molecule known as preQ$_0$ to preQ$_1$, the next intermediate in the pathway [294].

With this key functional clue, we probed this element for binding to intermediates in the queuosine biosynthetic pathway (A. Roth, E. Regulski, J.E. Barrick, and R.R. Breaker, unpublished data). We discovered that the molecule preQ$_1$ (7-aminomethyl-7-deazaguanine) specifically modulates the structure of the conserved queC element as judged by in-line probing (Figure 3.8). This aptamer binds the earlier precursor preQ$_0$ (7-cyano-7-deazaguanine) at five times the concentration that it senses preQ$_1$, which may indicate that it also senses this molecule under physiological conditions.

By finding additional examples of the $preQ_1$ riboswitch upstream of orthologs of the *queCDEF* biosynthetic genes in other bacterial species we improved our structural model [Rfam:RF00522]. Re-searching this improved model predicts new candidates for queuosine transporters and salvage genes (Figure 3.7). The $preQ_1$ riboswitch is remarkable from the standpoint of its small size relative to other known riboswitches. It consists of a single stem loop followed by an A-rich strand that probably makes A-minor contacts [207] to the P1 stem and interacts with conserved sequences in the main loop. A minimal 34-nt *B. subtilis* construct binds $preQ_1$ with a dissociation constant of 50 nM.

3.5 The *ydaO*/*yuaA* element

The *ydaO*/*yuaA* element [Rfam:RF00379] occurs upstream of these two genes in *B. subtilis* (Figure 3.9). Regulatory terminator structures are positioned such that we predict this motif functions as a genetic OFF switch. Structural probing of the corresponding RNA transcript supports the formation of a pseudoknot between a conserved loop and a complementary sequence located downstream, suggesting that this RNA has a complex tertiary structure (Figure 3.10). The *ydaO* gene product is a predicted amino acid transporter, whereas the protein products of the *yuaA*-*yubG* operon constitute a K^+ transporter and have been recently renamed KtrA and KtrB (29). The remaining genes appear to be membrane metalloendopeptidases, cell wall-associated hydrolases, and oligopeptide transporters involved in remodeling the cell wall. *B. subtilis* strains defective in KtrAB are sensitive to sudden osmotic shock. Therefore, the *ydaO*/*yuaA* element could possibly respond to a specific osmoprotectant molecule, or it might bind to a compound whose concentration changes during cell wall remodeling in response to damage.

3.6 The *ykkC*/*yxkD* element

The *ykkC*/*yxkD* element [Rfam:RF00442] is composed of a two-stem junction and a long 3' conserved region that lacks obvious local secondary structure (Figure 3.11, Figure 3.12, and Figure 3.13). However, the 3' conserved region often overlaps the GC-rich stem of a transcriptional terminator in Gram-positive organisms. We probed the full *ykkC* and *yxkD* leaders (data not shown) as well as a *ykkC* construct with the right hand shoulder of the terminator stem deleted (Figure 3.13). In all three constructs, the 3'-conserved stretch of nucleotides displays reduced spontaneous cleavage. It is possible that this region is sequestered in the terminator stem when present and takes part in an alternate structure when a portion of the stem is deleted.

The *ykkCD* operon produces a multidrug resistance efflux pump with a broad specificity [135], while the *yxkD* gene encodes a conserved protein of unknown function. In proteobacteria, the *ykkC*/*yxkD* RNA element usually occurs upstream of an operon that encodes all three subunits of an ABC-type transporter related to nitrate/sulfate/bicarbonate transport systems followed by two copies of a different uncharacterized gene. The *Synechocystis speB* gene affiliated with this element does not have the expected agmatinase or arginase function and appears to be involved in an uncharacterized reaction of secondary metabolism [252]. One possible interpretation of these mixed gene functions is that the *ykkC/yxkD* element switches ON efflux pumps and detoxification systems in response to harmful environmental molecules.

3.7 The *ykoK* element

The *ykoK* motif [Rfam:RF00380] is the most elaborate of the riboswitch candidates (Figure 3.14, Figure 3.15, and Figure 3.16). A 221-nucleotide transcript of this region folds into a highly structured RNA. Genes downstream of *ykoK* elements are similar to a variety of divalent metal transporters including those specific for Mg^{2+}, Mn^{2+}, Co^{2+}, Ni^{2+},

and Fe^{2+}. Since it is difficult to assign precise functions for transporters based on protein sequence similarity alone, it is possible that all of these regulated genes could primarily transport the same metal cation. It seems unlikely that such a large RNA would be necessary to sense an easily bound divalent metal, and so a more complex metabolite target might be involved. For example, the coenzyme B_{12} riboswitch, which has the most complex natural aptamer of all known riboswitches, is involved in regulating cobalt transporters in some bacteria [201, 235]. This arrangement ensures that cobalt transport matches the demand of coenzyme B_{12} for its obligatory cobalt ligand. Perhaps a similar genetic logic applies to the *ykoK* element. It is positioned with respect to terminators in low G+C Gram-positive bacteria such that it could serve as an OFF switch in response to a cellular compound that requires a metal ligand.

3.8 The *yybP*/*ykoY* element

The *yybP*/*ykoY* element [Rfam:RF00080] is the most common and widely distributed of the new *B. subtilis* riboswitch candidates and the only one that is also known to occur in *E. coli* (Table 3.1 and Figure 3.17). In fact, one of the *E. coli* examples of this element (upstream of *ygjT*) had been previously discovered by a computational screen designed to identify small noncoding RNAs in *E. coli* using promoter and terminator predictions and experimentally observed as a 189-nt RNA fragment by Northern blotting [11]. This small fragment probably corresponds to a prematurely terminated form of this *yybP*/*ykoY* regulatory element or a degradation intermediate of the mRNA leader. Similar 5' UTR fragments of TPP and FMN riboswitches have been recovered by shotgun cloning of small RNAs in *E. coli* [302]. The presence of two copies of this element in both *B. subtilis* and *E. coli*, as well as its constant occurrence upstream of related genes in diverse bacteria, leaves little doubt that it functions as a *cis*-regulatory element.

In-line probing confirms the existence of a complex RNA structure that is consistent with the secondary structure model (Figure 3.18 and Figure 3.19). This compact structure has a highly conserved internal bulge that is a fixed distance of ten base pairs (one complete helical turn) from a second asymmetric bulge with a specific consensus sequence that appears to be opposite continuous base pairing on the other strand. Presumably, these two regions make a tertiary contact, and they might enclose the molecular effector of this presumed orphan riboswitch. The main conserved bulge of the *yybP*/*ykoY* element may contain a structural motif that mediates this interaction. Portions of its consensus sequence resemble a sarcin/ricin type loop E motif [160] or K-turn [147, 163]. Interestingly, the *yybP*/*ykoY* element appears to be a genetic ON switch in *B. subtilis*. A transcription terminator hairpin is mutually exclusive with the conserved structure because it overlaps the 3' side of the P1 helix.

The *yybP*/*ykoY* element resides upstream of two separate monocistronic transcripts in *B. subtilis* and *E. coli*, and is found mainly upstream of genes classified into four groups. The first cluster includes *E. coli ygjT* and *B. subtilis ykoY*, which are similar but not orthologous to an *E. coli* gene (*terC*) that encodes a membrane protein with a poorly defined function related to tellurium resistance. The second group encodes a cation transport ATPase, while the final two clusters are predicted simply to be families of related membrane proteins (one includes *E. coli yebN*). No function has been assigned to the annotated *B. subtilis yybP* reading frame from sequence similarity. One associated gene from *Corynebacterium glutamicum* is suggestively similar to an arsenite efflux pump. The diverse and nonspecific functions of the genes in its regulon do not readily suggest a possible target metabolite for the *yybP*/*ykoY* element, although it seems to be broadly related to cation homeostasis.

3.9 The *ylbH* element

The *ylbH* motif was the least-promising riboswitch candidate that we characterized. We found only six examples, all in low G+C Gram-positive bacteria (Figure 3.20). This was not enough comparative information to posit a detailed consensus structural model, but in-line probing did generally agree with stems we proposed (Figure 3.21). This putative regulatory RNA element always occurs upstream of an N6-adenine-specific methylase and phosphopantetheine adenylyltransferase. The latter enzyme catalyzes the penultimate step in coenzyme A (CoA) biosynthesis, but we were unable to detect any binding of CoA or related molecules to our RNA construct by in-line probing.

Subsequently, we recognized that the *ylbH* sequence conservation is more likely to be an RNA element on the opposite strand. Specifically, note how the first predicted hairpin (last for the reverse complement sequence) is a canonical transcription terminator on the other strand in all sequences. The reverse complement of the *ylbH* RNA motif always occurs upstream of a conserved hypothetical gene of unknown function and could play a role in its regulation.

3.10 Methods

Bioinformatics strategies

An archived version of the BLISS database (v1) is available on the web (http://bliss.biology.yale.edu). Briefly, intergenic regions with a minimum length of 30 nucleotides from 91 complete genomes were analyzed. Conservation between each *B. subtilis* IGR and other intergenic sequences was identified by BLASTN searches [7], and IGR matches were pairwise aligned using the FASTA3 package [214] to highlight additional conservation. For each genome, gene functions were assigned uniformly using the COG database [276], and intrinsic transcription terminators were predicted by

a modified version of the program TransTerm [72]. A web interface allows IGR sequence alignments and associated evidence of riboswitch function to be interactively viewed and annotated. Refer to Chapter 1 for a more detailed description of the BLISS database. Promising secondary structure models were iteratively refined and extended by motif searching with the program SequenceSniffer (unpublished algorithm) and additional BLAST searches.

Ribozyme assays and in-line probing

Templates for transcription were PCR amplified from chromosomal DNA extracted from *B. subtilis* strains (Bacillus Genetic Stock Center, Columbus, OH) 1A40 (utilized for preparation of *yybP*, *ydaO*, *gcvT*, *glmS*), 1A210 (*yuaA*), and 1A234 (*ykkC*, *yxkD*, *ykoK*, *ykvJ*). RNA molecules were prepared by *in vitro* transcription using RiboMAX transcription kits (Promega). Bimolecular ribozyme assays were conducted using ~5 nM 5´ 32P-labeled RNA substrate that was incubated for 5 min at 23ºC in the presence of 50 mM Tris-HCl (pH 7.5 at 23°C), 200 mM KCl, 10 mM $MgCl_2$, 100 nM ribozyme, and in the absence or presence of 100 μM effector as indicated for each experiment. Reactions were terminated with an equal volume of 2X gel loading buffer (90 mM Tris base, 90 mM borate, 8 M urea, 20% sucrose (w/v), 1 mM EDTA, 0.1% SDS, 0.05% xylene cyanol FF, 0.05% bromophenol blue), which was supplemented with EDTA to a final concentration of 100 mM. The products were separated using denaturing 20% PAGE and analyzed by using a PhosphorImager (Molecular Dynamics). In-line probing assays were carried out using methods adapted from those described elsewhere [260].

Figure 3.2 Distribution and multiple sequence alignment of the glmS ribozyme

The distribution table (above) displays the genomic context of each *glmS* riboswitch in the multiple sequence alignment (below). Accession numbers refer to GenBank nucleotide records and positions are for the starred columns in the corresponding alignment. For each element instance, locus tags (e.g. BH2815) and/or functional assignments (e.g. COG0403) are provided for genes in the putative downstream operon. The genomic context table is followed by a key that describes the functions of commonly occurring COGs. The sequence alignment includes the same list of elements, indexed by their organism abbreviations and sometimes also by the first gene of the downstream operon. The structure line (SS) shows conserved base pairing. The consensus line (Cons) shows positions with ≥95% (uppercase) and ≥80% (lowercase) sequence conservation (R = A, G; Y = C, U). Motif representatives that share ≥90% sequence identity were eliminated for this calculation. Single nucleotides highlighted in red differ between the database sequence and that determined by sequencing clones prepared in our laboratory. Other colored backgrounds indicate base pairing in individual aligned sequences. Periods and dashes represent gapped positions in aligned sequences. Putative intrinsic transcription terminator hairpins are underlined in other sequences.

Abb	Organism	Accession		Position		Genes
Bacillus/Clostridia						
Ban	*Bacillus anthracis*	NC_003997.3	+	157538	157672	COG0449
Bce	*Bacillus cereus*	NC_004722.1	+	163187	163321	COG0449
Bha	*Bacillus halodurans*	NC_002570.1	+	288777	288923	COG0449
Bsu	*Bacillus subtilis*	NC_000964.1	+	200017	200163	COG0449
Cac	*Clostridium acetobutylicum*	NC_003030.1	+	179926	180064	COG0449
Cpe	*Clostridium perfringens*	NC_003366.1	-	2678977	2679173	COG0449
Cte	*Clostridium tetani*	NC_004557.1	-	2702212	2702464	COG0449
Dha	*Desulfitobacterium hafniense*	NZ_AAAW00000000.2	-	3522	3675	COG0449
Efa	*Enterococcus faecalis*	NC_004668.1	-	2055323	2055449	COG0449
Gst	Geobacillus stearothermophilus		+	152	302	COG0449
	gnl|OUACGT_1422|bstearo.fasta.screen.Contig138					
Lpl	*Lactobacillus plantarum*	NC_004567.1	+	761460	761590	COG0449
Lin	*Listeria innocua*	NC_003212.1	+	755745	755920	COG0449
Lmo	*Listeria monocytogenes*	NC_003210.1	+	756469	756643	COG0449
Oih	*Oceanobacillus iheyensis*	NC_004193.1	-	258887	259034	COG0449
Sau	*Staphylococcus aureus*	NC_002745.1	-	2212077	2212272	COG0449
Sep	*Staphylococcus epidermidis*	NC_004461.1	-	1803993	1804189	COG0449
Tte	*Thermoanaerobacter tengcongensis*	NC_003869.1	+	2101458	2101594	COG0449
Fusobacteria						
Fnu	*Fusobacterium nucleatum*	NC_003454.1	+	1095667	1095791	COG0449

COG	Gene	Description
COG0449	glmS	Glucosamine 6-phosphate synthetase

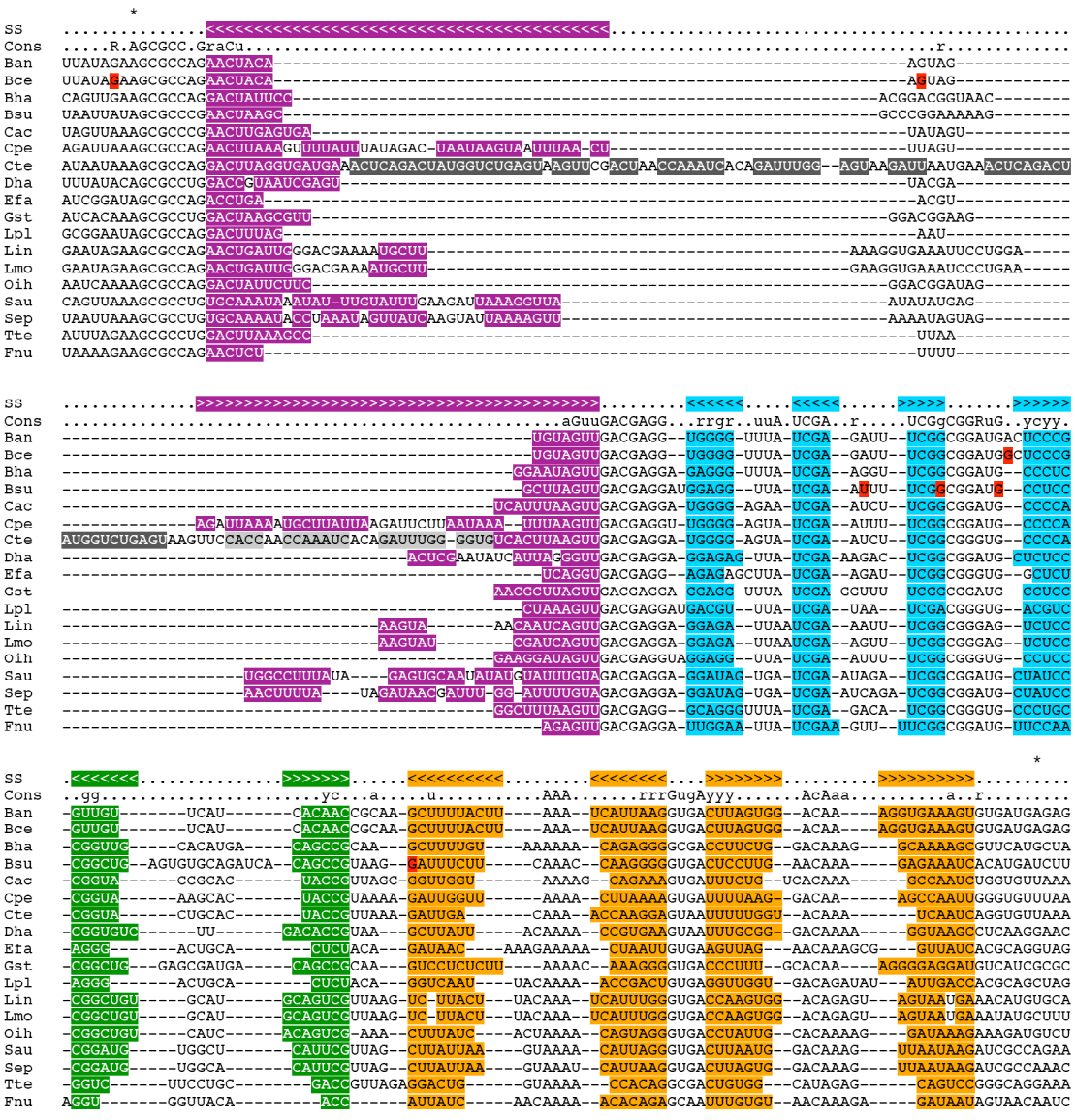


Figure 3.2 Distribution and multiple sequence alignment of the glmS ribozyme

Figure 3.3 Consensus structure and function of the *glmS* ribozyme

(A) Consensus secondary structure. Circles replace unconserved nucleotides that are always present at a position, and lines replace sequences of variable length. Red and black nucleotides are conserved in at least 80% and 95% of representatives, respectively. **(*B*)** Sequence and predicted secondary structure of the *glmS* ribozyme that has been engineered to function as a bimolecular construct. **(*C*)** Metabolite-dependent ribozyme function of the conserved *glmS* element. Reactions were conducted in the absence (–) or presence (+) of ribozyme and 100 μM effector as indicated for each lane. Sub and Clv identify the substrate and 5' cleavage product, respectively.

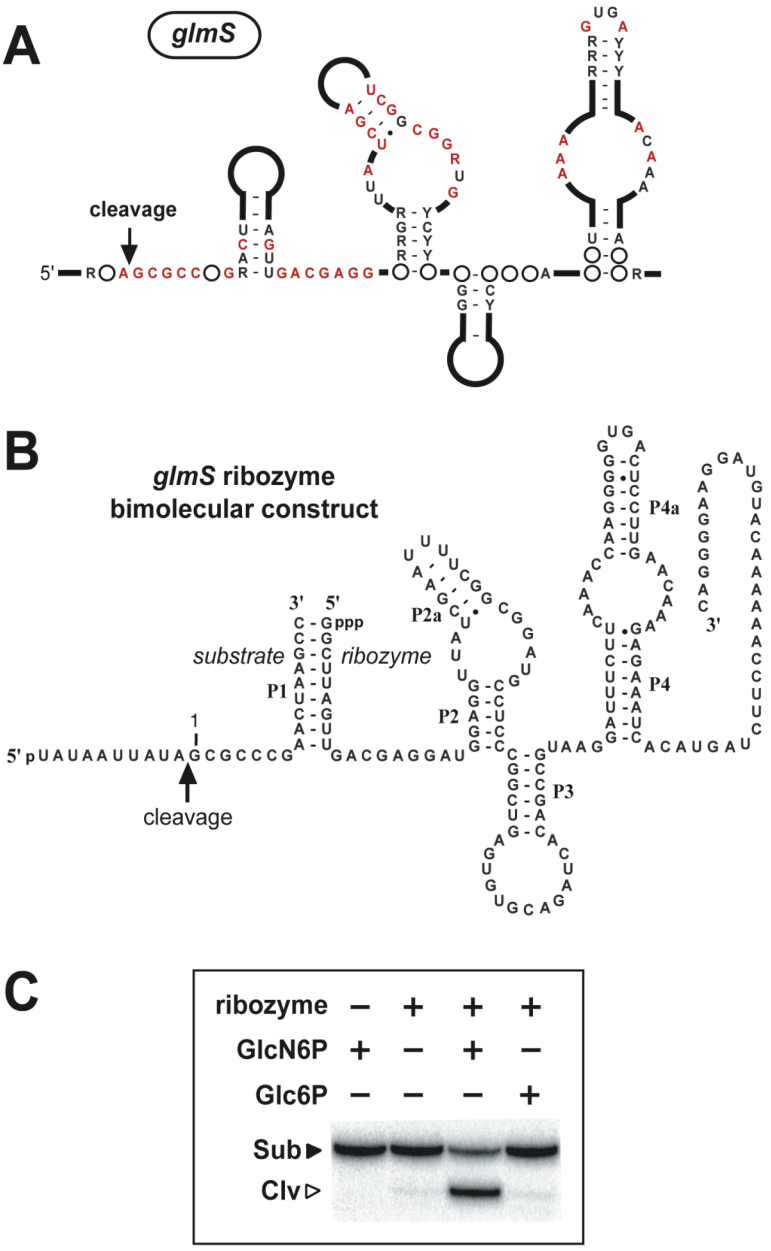


Figure 3.3 Consensus structure and function of the *glmS* ribozyme

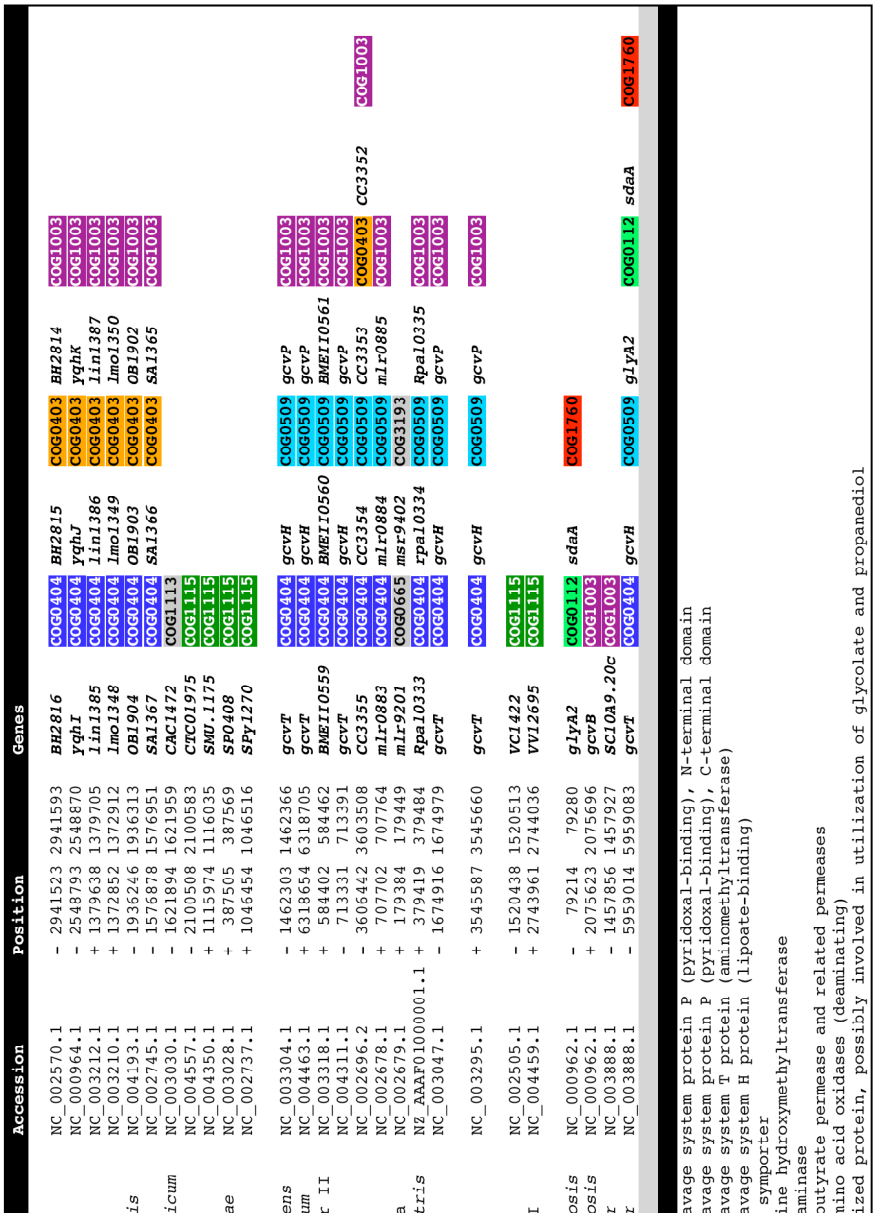


Abb	Organism	Accession	Position	Genes							
Bacillus/Clostridium											
# Bha	*Bacillus halodurans*	NC_002570.1	- 2941523 2941593	***BH2816***	COG0404	***BH2815***	COG0403	***BH2814***	COG1003		
# Bsu	*Bacillus subtilis*	NC_000964.1	- 2548793 2548870	***yqhI***	COG0404	***yqhJ***	COG0403	***yqhK***	COG1003		
# Lin	*Listeria innocua*	NC_003212.1	+ 1379638 1379705	***lin1385***	COG0404	***lin1386***	COG0403	***lin1387***	COG1003		
# Lmo	*Listeria monocytogenes*	NC_003210.1	+ 1372852 1372912	***lmo1348***	COG0404	***lmo1349***	COG0403	***lmo1350***	COG1003		
# Oih	*Oceanobacillus iheyensis*	NC_004193.1	- 1936246 1936313	***OB1904***	COG0404	***OB1903***	COG0403	***OB1902***	COG1003		
# Sau	*Staphylococcus aureus*	NC_002745.1	- 1576878 1576951	***SA1367***	COG0404	***SA1366***	COG0403	***SA1365***	COG1003		
* Cac	*Clostridium acetobutylicum*	NC_003030.1	- 1621894 1621959	***CAC1472***	COG1113						
* Cte	*Clostridium tetani*	NC_004557.1	- 2100508 2100583	***CTC01975***	COG1115						
* Smu	*Streptococcus mutans*	NC_004350.1	+ 1115974 1116035	***SMU.1175***	COG1115						
* Spn	*Streptococcus pneumoniae*	NC_003028.1	+ 387505 387569	***SP0408***	COG1115						
* Spy	*Streptococcus pyogenes*	NC_002737.1	+ 1046454 1046516	***SPy1270***	COG1115						
Alpha Proteobacteria											
# Atu	*Agrobacterium tumefaciens*	NC_003304.1	- 1462303 1462366	***gcvT***	COG0404	***gcvH***	COG0509	***gcvP***	COG1003		
# Bja	*Bradyrhizobium japonicum*	NC_004463.1	+ 6318654 6318705	***gcvT***	COG0404	***gcvH***	COG0509	***gcvP***	COG1003		
# Bme	*Brucella melitensis chr II*	NC_003318.1	+ 584402 584462	***BMEII0559***	COG0404	***BMEII0560***	COG0509	***BMEII0561***	COG1003		
# Brs	*Brucella suis chr II*	NC_004311.1	- 713331 713391	***gcvT***	COG0404	***gcvH***	COG0509	***gcvP***	COG1003		
# Ccr	*Caulobacter crescentus*	NC_002696.2	- 3606442 3603508	***CC3355***	COG0404	***CC3354***	COG0509	***CC3353***	COG0403	***CC3352***	COG1003
# Mlo	*Mesorhizobium loti*	NC_002678.1	+ 707702 707764	***mlr0883***	COG0404	***mlr0884***	COG0509	***mlr0885***	COG1003		
# Mlo	*Mesorhizobium loti pMLa*	NC_002679.1	+ 179384 179449	***mlr9201***	COG0665	***msr9402***	COG3193				
# Rpa	*Rhodopseudomonas palustris*	NZ_AAAF01000001.1	+ 379419 379484	***Rpal0333***	COG0404	***rpal0334***	COG0509	***Rpal0335***	COG1003		
# Sme	*Sinorhizobium meliloti*	NC_003047.1	- 1674916 1674979	***gcvT***	COG0404	***gcvH***	COG0509	***gcvP***	COG1003		
Beta Proteobacteria											
# Rso	*Ralstonia solanacearum*	NC_003295.1	+ 3545587 3545660	***gcvT***	COG0404	***gcvH***	COG0509	***gcvP***	COG1003		
Gamma Proteobacteria											
* Vch	*Vibrio cholerae chr I*	NC_002505.1	- 1520438 1520513	***VC1422***	COG1115						
* Vvu	*Vibrio vulnificus chr I*	NC_004459.1	+ 2743961 2744036	***VV12695***	COG1115						
Actinobacteria											
# Mtu	*Mycobacterium tuberculosis*	NC_000962.1	- 79214 79280	***glyA2***	COG0112	***sdaA***	COG1760				
# Mtu	*Mycobacterium tuberculosis*	NC_000962.1	+ 2075623 2075696	***gcvB***	COG1003						
# Sco	*Streptomyces coelicolor*	NC_003888.1	- 1457856 1457927	***SC10A9.20c***	COG1003						
# Sco	*Streptomyces coelicolor*	NC_003888.1	- 5959014 5959083	***gcvT***	COG0404	***gcvH***	COG0509	***glyA2***	COG0112	***sdaA***	COG1760

Type I = #, Type II = *

COG	Gene	Description
COG0403	***gcvP_1***	Glycine cleavage system protein P (pyridoxal-binding), N-terminal domain
COG1003	***gcvP_2***	Glycine cleavage system protein P (pyridoxal-binding), C-terminal domain
COG0404	***gcvT***	Glycine cleavage system T protein (aminomethyltransferase)
COG0509	***gcvH***	Glycine cleavage system H protein (lipoate-binding)
COG1115	***yaaJ***	Na+/alanine symporter
COG0112	***glyA***	Glycine/serine hydroxymethyltransferase
COG1760	***sdaA***	L-serine deaminase
COG1113	***ansP***	Gamma-aminobutyrate permease and related permeases
COG0665	***dadA***	Glycine/D-amino acid oxidases (deaminating)
COG3193	***glcG***	Uncharacterized protein, possibly involved in utilization of glycolate and propanediol

Figure 3.4 Distribution of the *gcvT* element

Refer to the legend of Figure 3.2 for details.

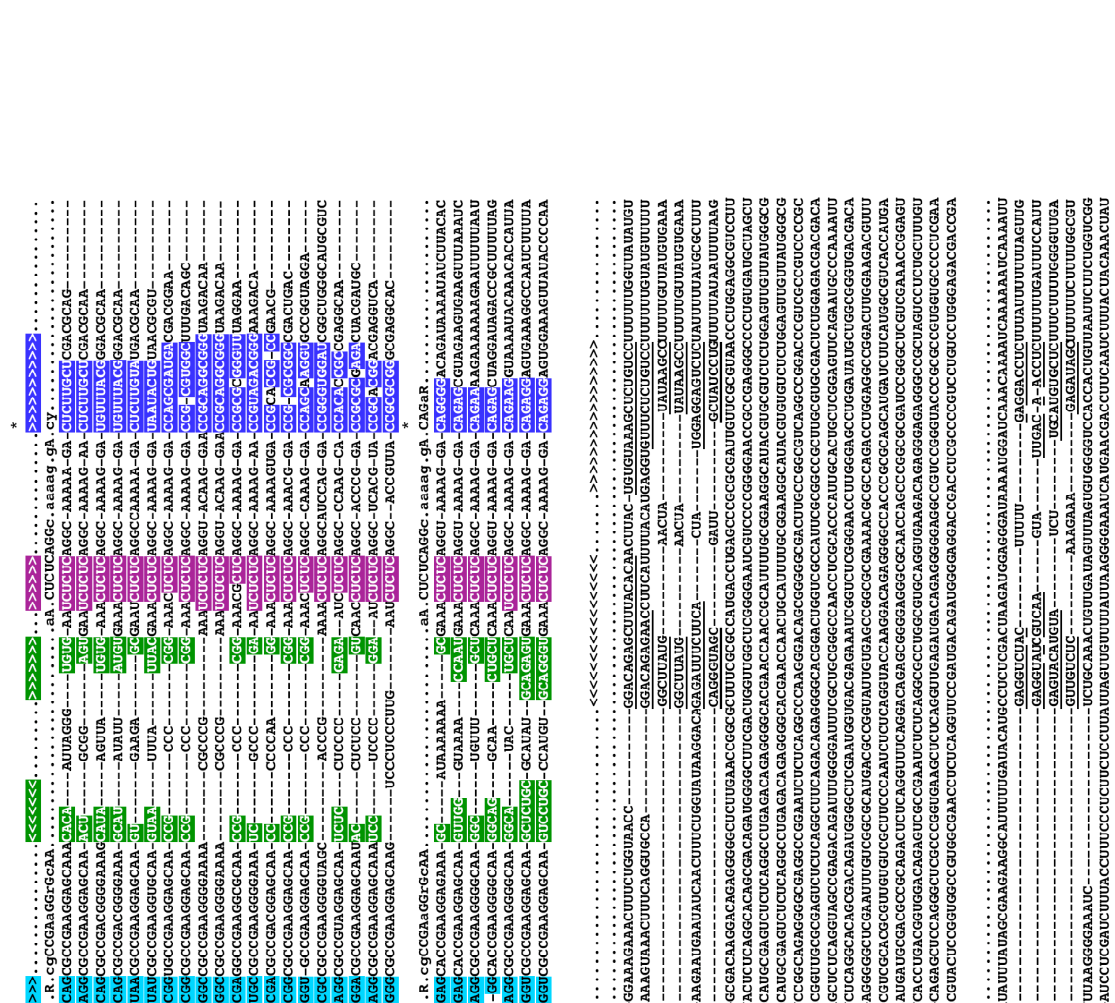


Figure 3.5 Sequence alignment of the *gcvT* element

Refer to the legend of Figure 3.2 for details.

Figure 3.6 Consensus structure and in-line probing of the glycine riboswitch

(A) Consensus structure. Red and black positions indicate ≥95% and ≥80% conservation of a particular nucleotide, respectively. Purine (R) or pyrimidine (Y) designations are used when a single nucleotide is not 80% conserved. Solid black lines indicate variable length insertions of unconserved nucleotides, and insets with solid grey lines are optional sequence insertions that are not present in all examples of an element. Unfilled circles represent single nucleotides whose presence (but not sequence) is conserved. P1 through P4 identify common base-paired elements. ORF, open reading frame. ***(B)*** Secondary structure model for the *Vibrio cholerae* glycine riboswitch. Changes in the spontaneous cleavage pattern that occur in the pictured VC I-II RNA construct as glycine is added are depicted. Numbers adjacent to sites of changing spontaneous cleavage correspond to gel bands denoted with asterisks in (*C*). 5' guanosyl residues (g) not present in the genomic sequence were added to improve *in vitro* transcription yields with T7 RNA polymerase. ***(C)*** In-line probing gel. Spontaneous cleavage products of VC I-II were separated by polyacrylamide gel electrophoresis (PAGE). NR, T1, $^{-}$OH, and – represent no reaction, partial digest with RNase T1, partial digest with alkali, and in-line probing in the absence of added ligand, respectively. Precursor RNA (Pre) and some fragment bands corresponding to T1 digestion (cleaves after G residues) are labeled. Numbered asterisks identify locations of major structural modulation in response to increasing glycine concentrations. Reactions in the two rightmost lanes contained 1 mM of the amino acids noted. Brackets labeled I and II identify RNA fragments that correspond to cleavage events in the type I and type II aptamers.

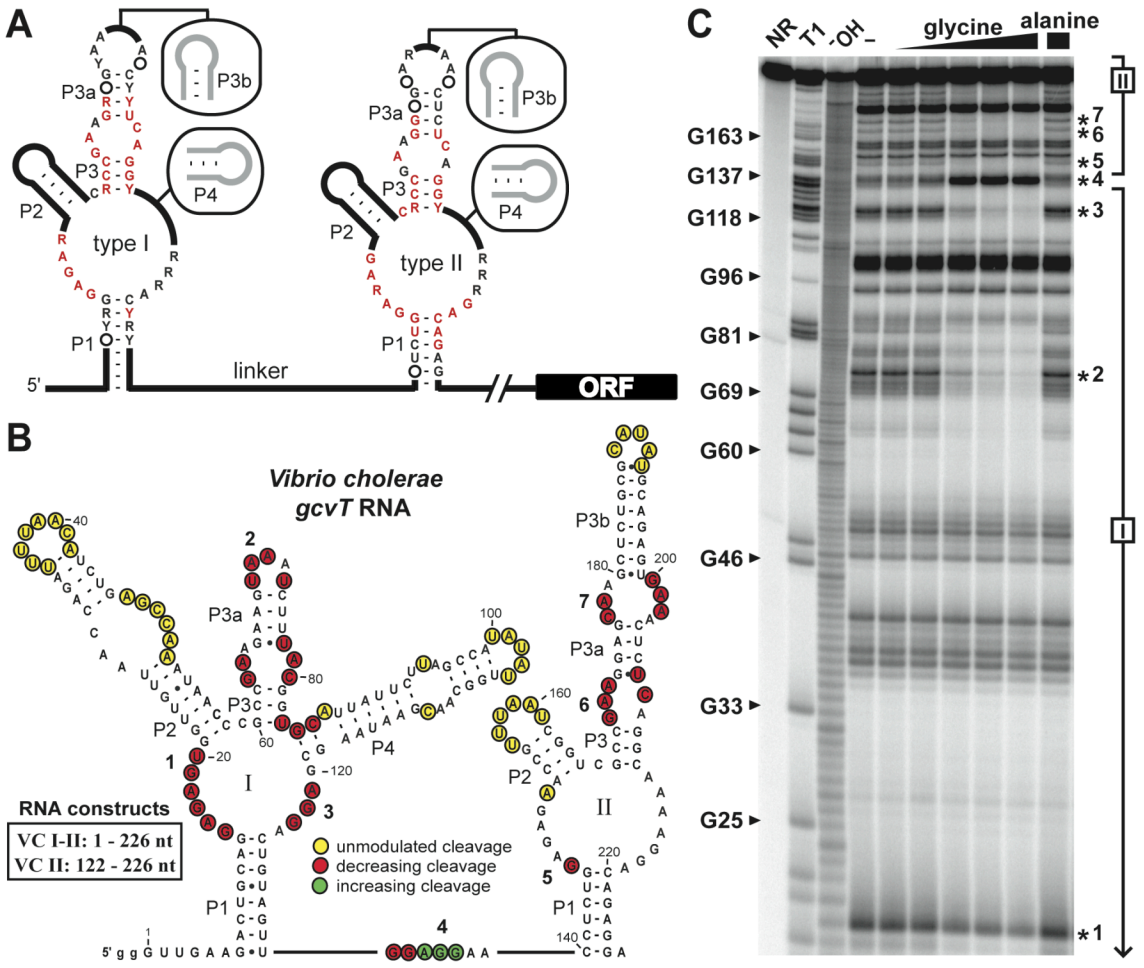


Figure 3.6 Consensus structure and in-line probing of the glycine riboswitch

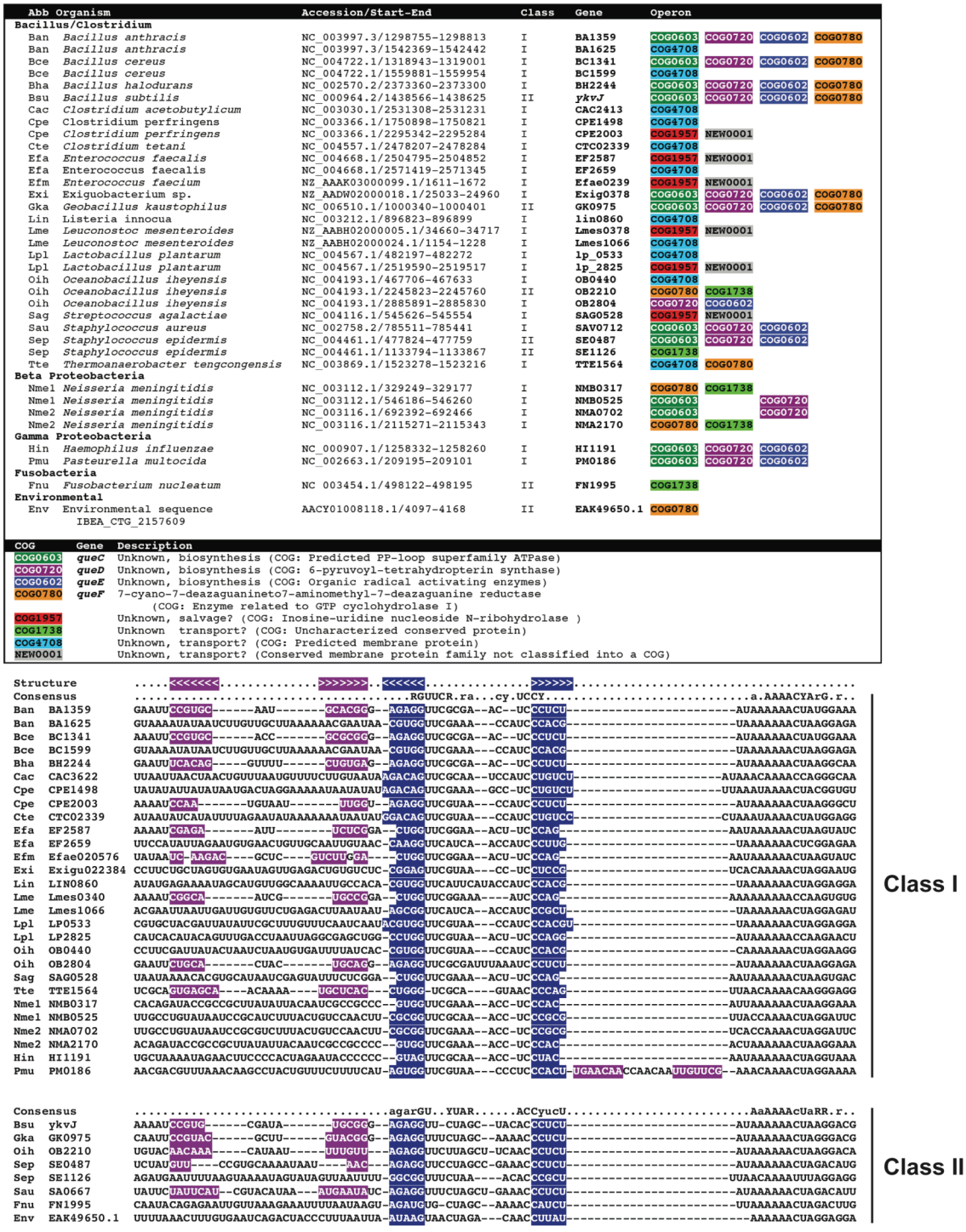


Figure 3.7 Distribution and multiple sequence alignment of the preQ1 riboswitch

Refer to the legend of Figure 3.2 for details. Only the locus tag designation for the first gene in the operon is shown in the context table.

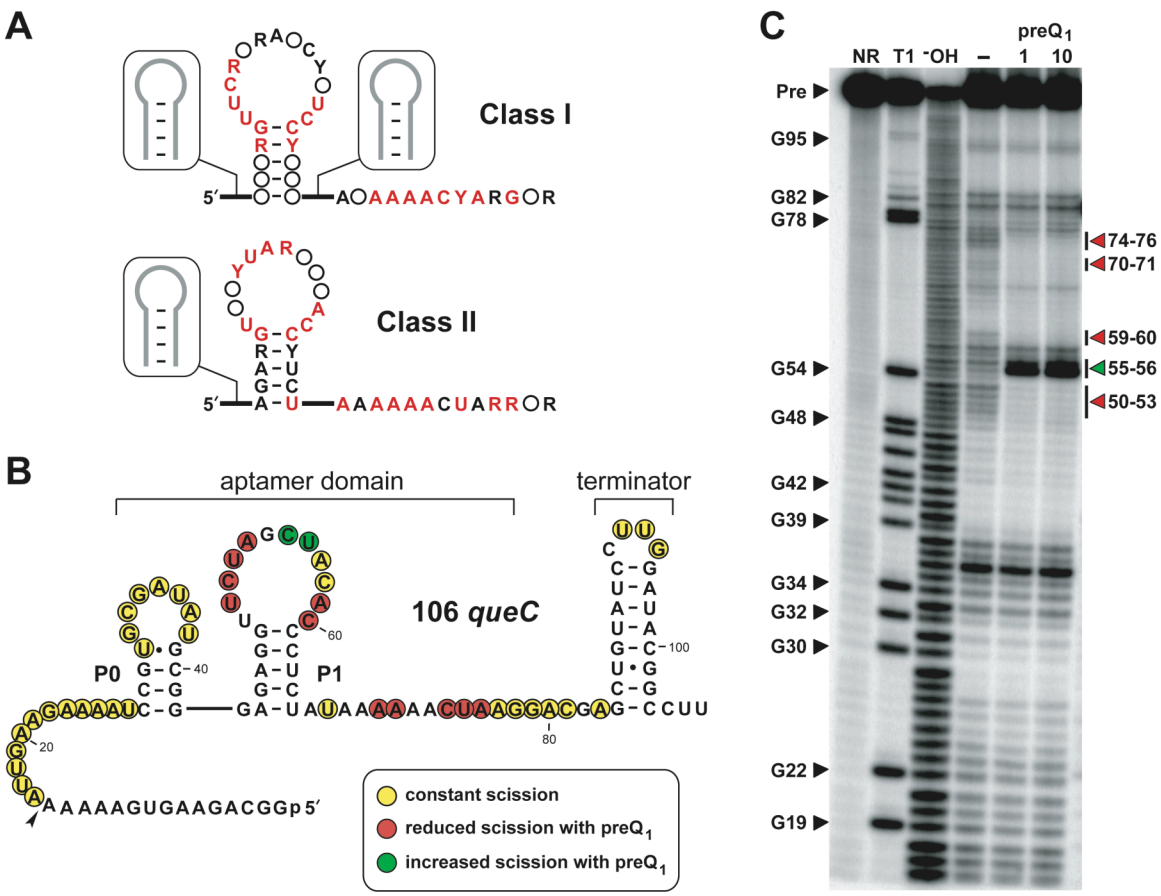


Figure 3.8 Consensus structure and in-line probing of the preQ$_1$ riboswitch

Refer to the legend of Figure 3.6 for details. In (*C*) the concentrations of added preQ$_1$ are 1 and 10 μM, and sites that modulate in the presence of preQ$_1$ are labeled with arrowheads and position numbers on the right side of the gel.

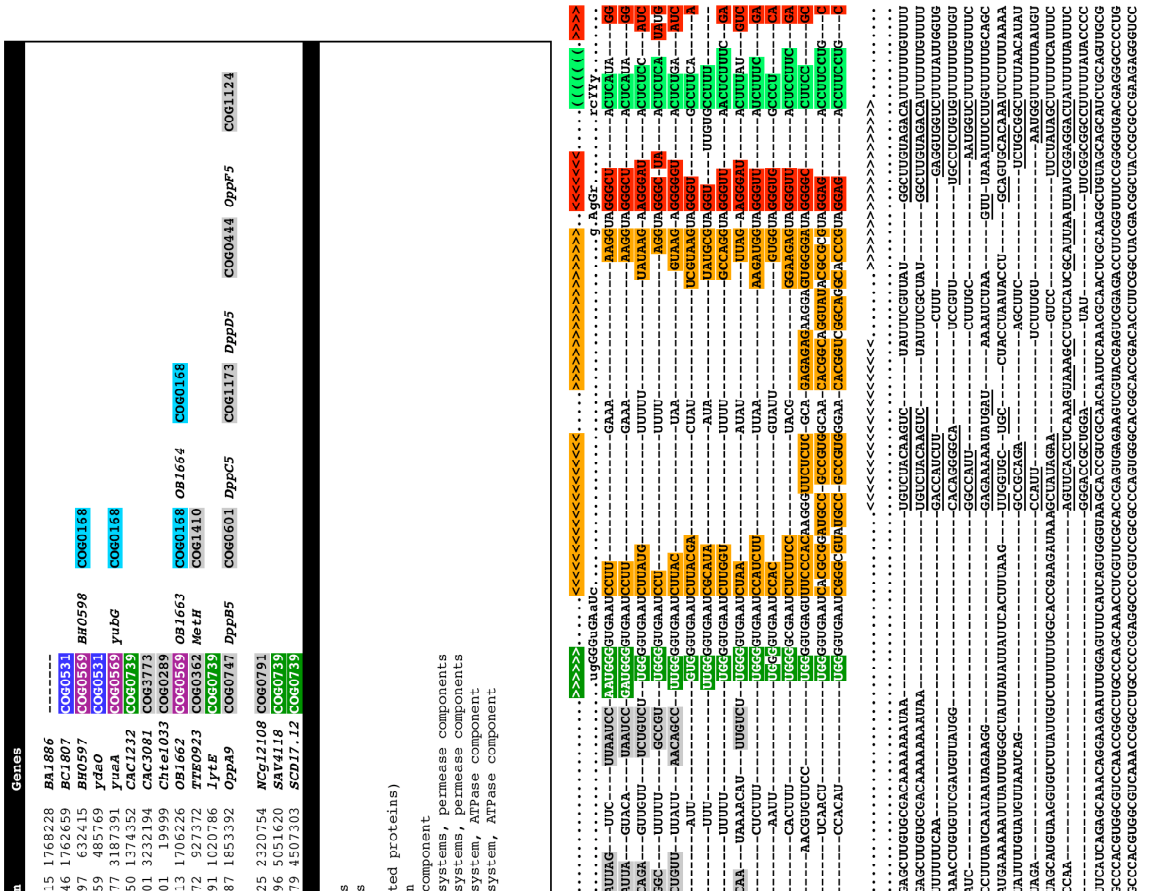


Figure 3.9 Distribution and multiple sequence alignment of the *ydaO*/*yuaA* element

Refer to the legend of Figure 3.2 for details.

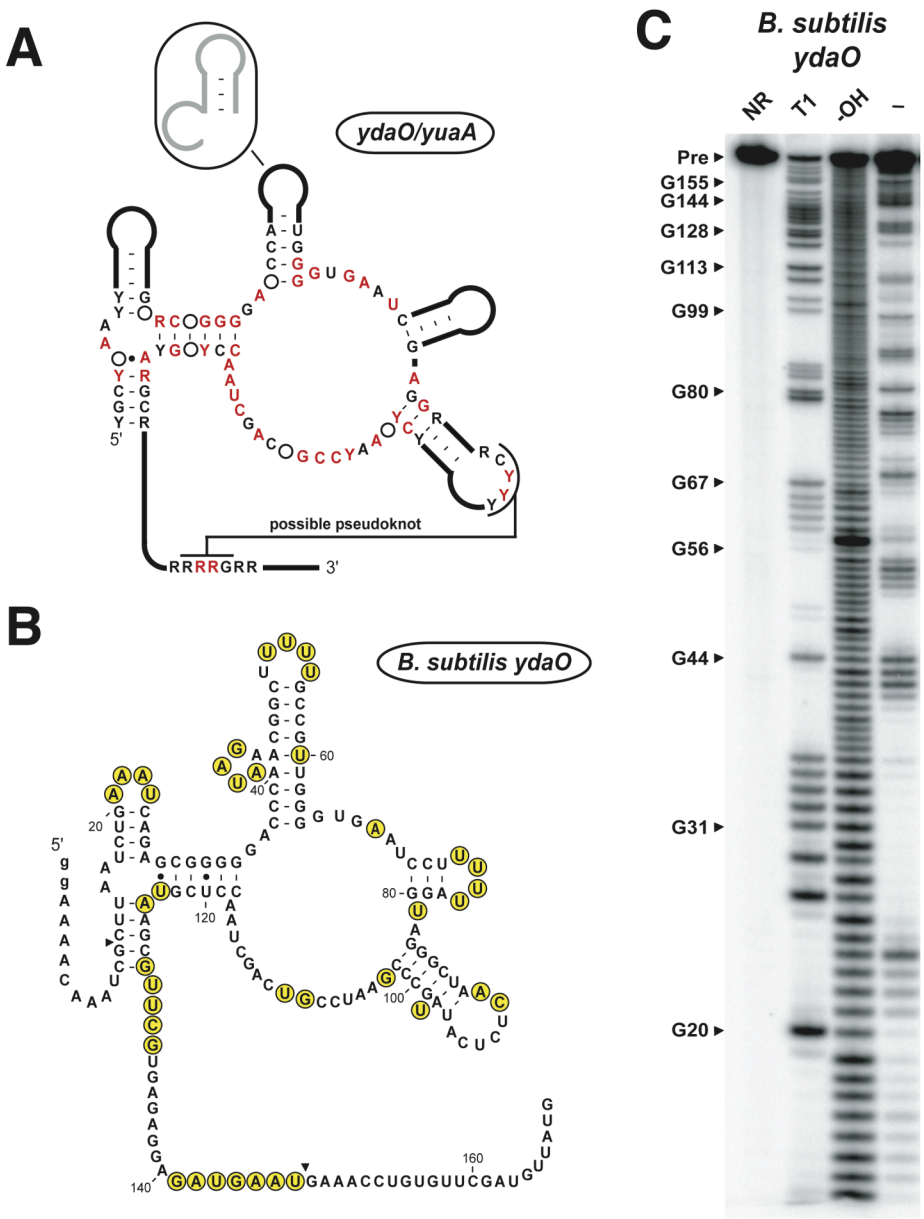


Figure 3.10 Consensus structure and in-line probing of the *ydaO/yuaA* element

Refer to the legend of Figure 3.6 for details.

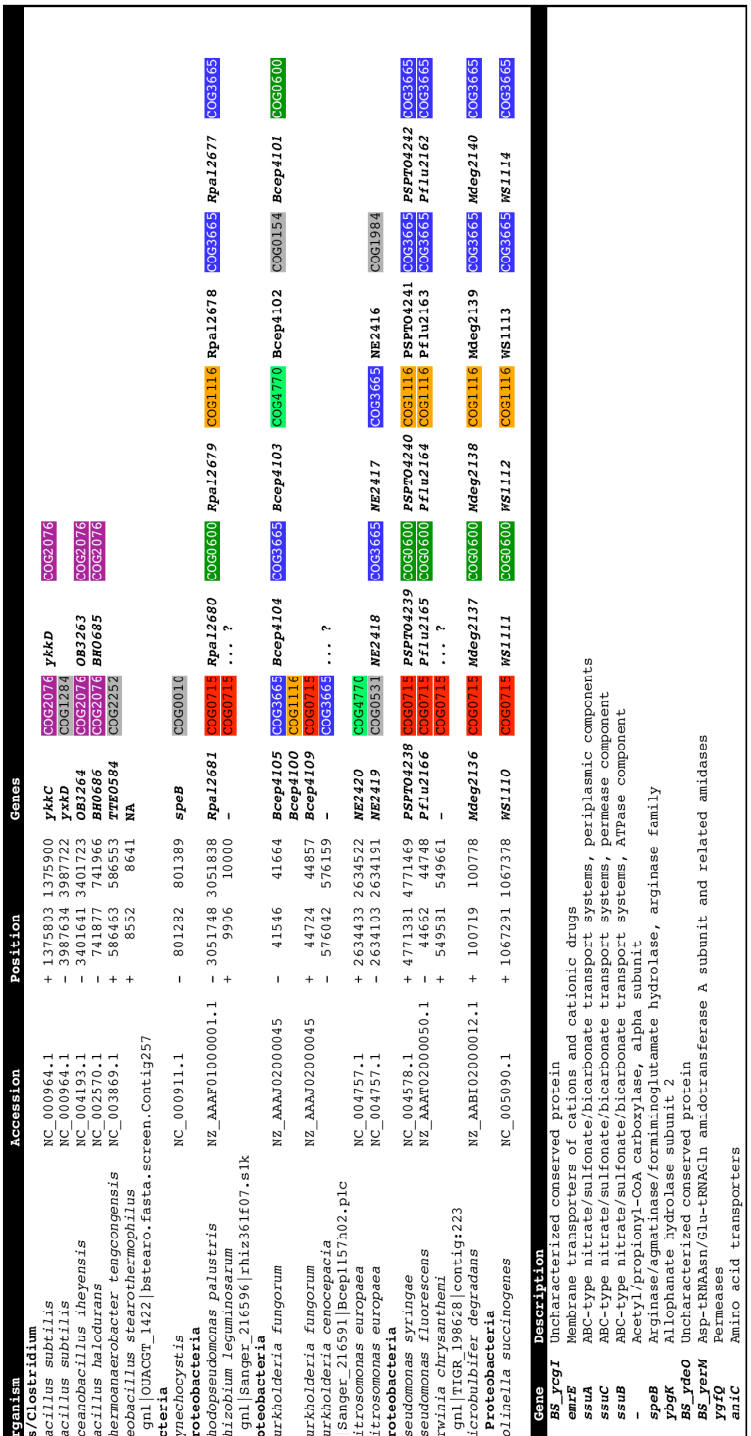


Abb	Organism	Accession	Position		Genes									
Bacillus/Clostridium														
Bsu	*Bacillus subtilis*	NC_000964.1	+	1375803	1375900	***ykkC***	COG2076	***ykkD***	COG2076					
Bsu	*Bacillus subtilis*	NC_000964.1	-	3987634	3987722	***yxkD***	COG1284							
Oih	*Oceanobacillus iheyensis*	NC_004193.1	-	3401641	3401723	***OB3264***	COG2076	***OB3263***	COG2076					
Bha	*Bacillus halcdurans*	NC_002570.1	-	741877	741966	***BH0685***	COG2076	***BH0685***	COG2076					
Tte	*Thermoanaerobacter tengcongensis*	NC_003869.1	+	586453	586553	***TTE0584***	COG2252							
Gst	*Geobacillus stearothermophilus*		+	8552	8641	**NA**								
	gnl|OJACGT_1422|bstearo.fasta.screen.Contig257													
Cyanobacteria														
Syn	*Synechocystis*	NC_000911.1	-	801232	801389	***speB***	COG0010							
Alpha Proteobacteria														
Rpa	*Rhodopseudomonas palusttris*	NZ_AAAF01000001.1	-	3051748	3051838	***Rpal2681***	COG0715	***Rpal2680***	COG0600	***Rpal2679***	COG1116	Rpal2678	COG3665	***Rpal2677***	COG3665
Rle	*Rhizobium leguminosarum*		+	9936	10000	-	COG0715	... ?							
	gnl|Sanger_216596|rhiz361f07.s1k														
Beta Proteobacteria															
Bfu	*Burkholderia fungorum*	NZ_AAAJ02000045	-	41546	41664	***Bcep4105***	COG3665	***Bcep4104***	COG3665	***Bcep4103***	COG4770	Bcep4102	COG0154	***Bcep4101***	COG0600
						Bcep4100	COG1116								
Bfu	*Burkholderia fungorum*	NZ_AAAJ02000045	+	44724	44857	***Bcep4109***	COG0715								
Bce	*Burkholderia cenocepacia*		-	576042	576159	-	COG3665	... ?							
	gnl Sanger_216591|Bcep1157n02.p1c														
Neu	*Nitrosomonas europaea*	NC_004757.1	+	2634433	2634522	***NE2420***	COG4770								
Neu	*Nitrosomonas europaea*	NC_004757.1	-	2634103	2634191	***NE2419***	COG0531	***NE2418***	COG3665	***NE2417***	COG3665	NE2416	COG1984		
Gamma Proteobacteria															
Psy	*Pseudomonas syringae*	NC_004578.1	+	4771331	4771469	***PSPTO4238***	COG0715	***PSPTO4239***	COG0600	***PSPTO4240***	COG1116	PSPTO4241	COG3665	***PSPTO4242***	COG3665
Pfl	*Pseudomonas fluorescens*	NZ_AAAN02000050.1	-	44652	44748	***Pf1u2166***	COG0715	***Pf1u2165***	COG0600	***Pf1u2164***	COG1116	Pf1u2163	COG3665	***Pf1u2162***	COG3665
Ech	*Erwinia chrysantheni*		+	549531	549661	-	COG0715	... ?							
	gnl|TIGR_198628|contig:223														
Mde	*Microbulbifer degradans*	NZ_AABI02000012.1	+	100719	100778	***Mdeg2136***	COG0715	***Mdeg2137***	COG0600	***Mdeg2138***	COG1116	Mdeg2139	COG3665	***Mdeg2140***	COG3665
Epsilon Proteobacteria															
Wsu	*Wolinella succinogenes*	NC_005090.1	+	1067291	1067378	***WS1110***	COG0715	***WS1111***	COG0600	***WS1112***	COG1116	WS1113	COG3665	***WS1114***	COG3665

COG	Gene	Description
COG3665	***BS_ycgI***	Uncharacterized conserved protein
COG2076	***emrE***	Membrane transporters of cations and cationic drugs
COG0715	***ssuA***	ABC-type nitrate/sulfonate/bicarbonate transport systems, periplasmic components
COG0600	***ssuC***	ABC-type nitrate/sulfonate/bicarbonate transport systems, permease component
COG1116	***ssuB***	ABC-type nitrate/sulfonate/bicarbonate transport systems, ATPase component
COG4770	-	Acetyl/propionyl-CoA carboxylase, alpha subunit
COG0010	***speB***	Arginase/agmatinase/formimoglutamate hydrolase, arginase family
COG1984	***ybgK***	Allophanate hydrolase subunit 2
COG1284	***BS_ydeO***	Uncharacterized conserved protein
COG0154	***BS_yerM***	Asp-tRNAAsn/Glu-tRNAGln amidotransferase A subunit and related amidases
COG2252	***ygfQ***	Permeases
COG0531	***anlC***	Amino acid transporters

Figure 3.11 Distribution of the *ykkC*/*yxkD* element

Refer to the legend of Figure 3.2 for details.

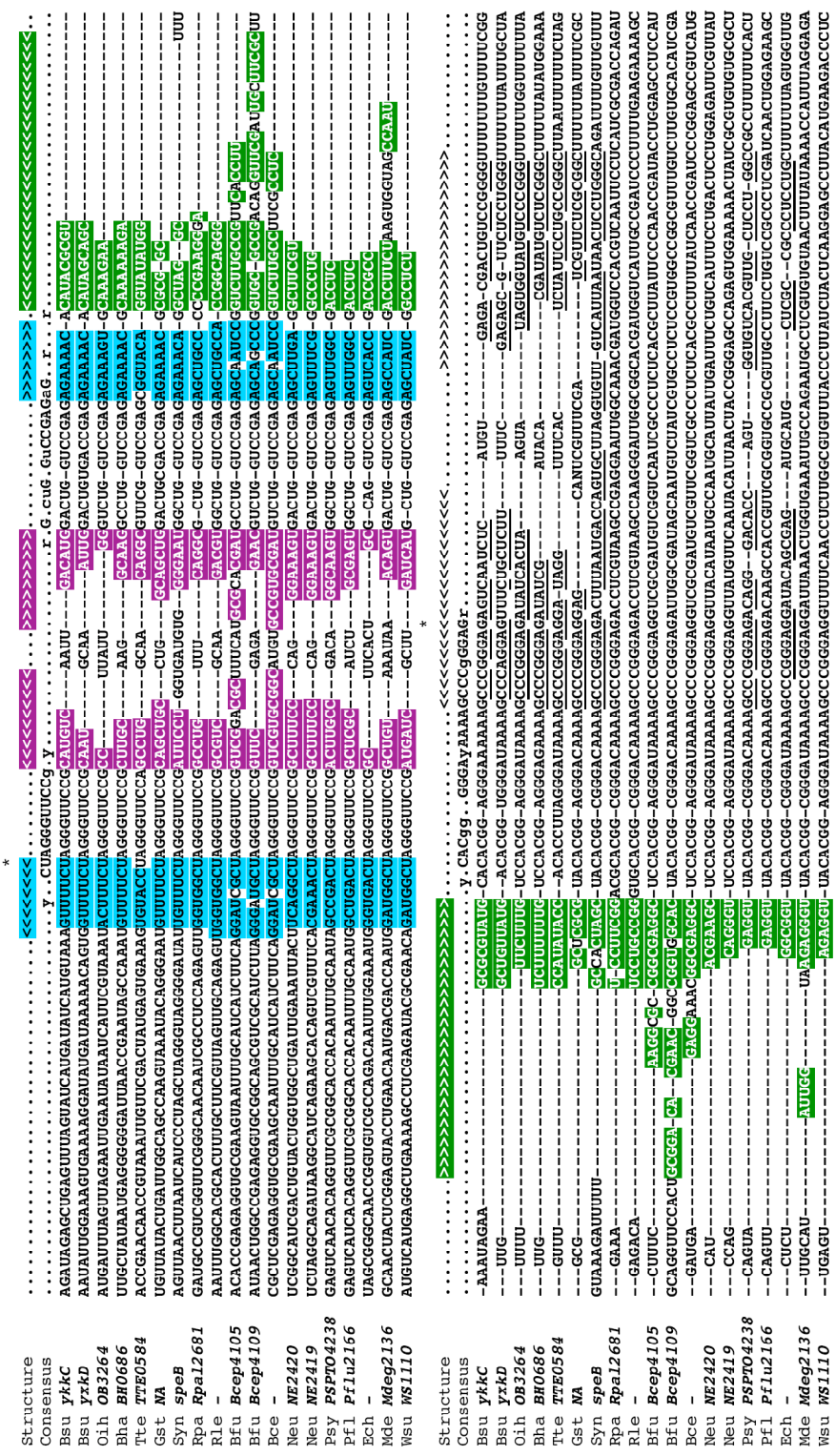


Figure 3.12 Sequence alignment of the *ykkC*/*yxkD* element

Refer to the legend of Figure 3.2 for details.

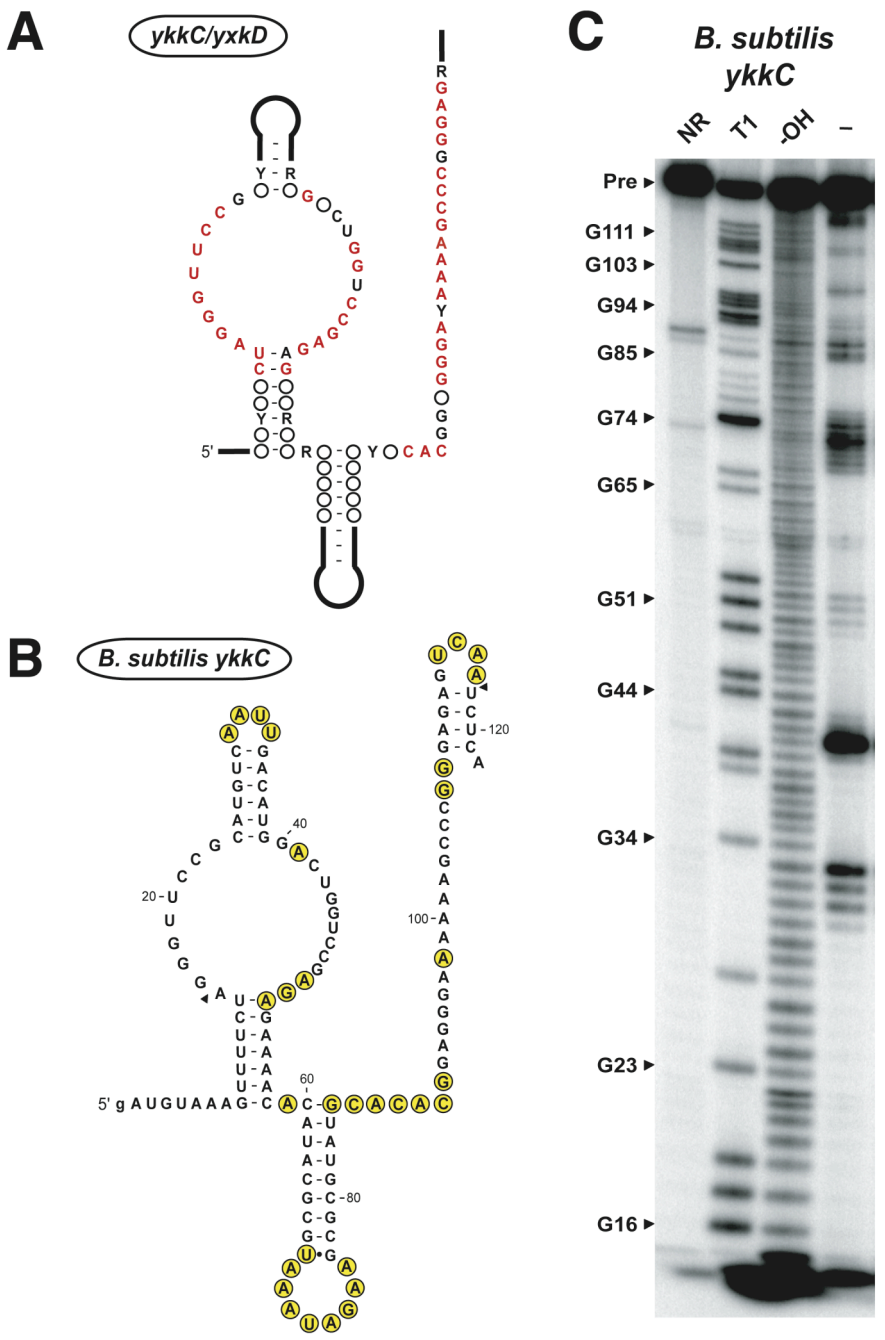


Figure 3.13 Consensus structure and in-line probing of the *ykkC/yxkD* element

Refer to the legend of Figure 3.6 for details.

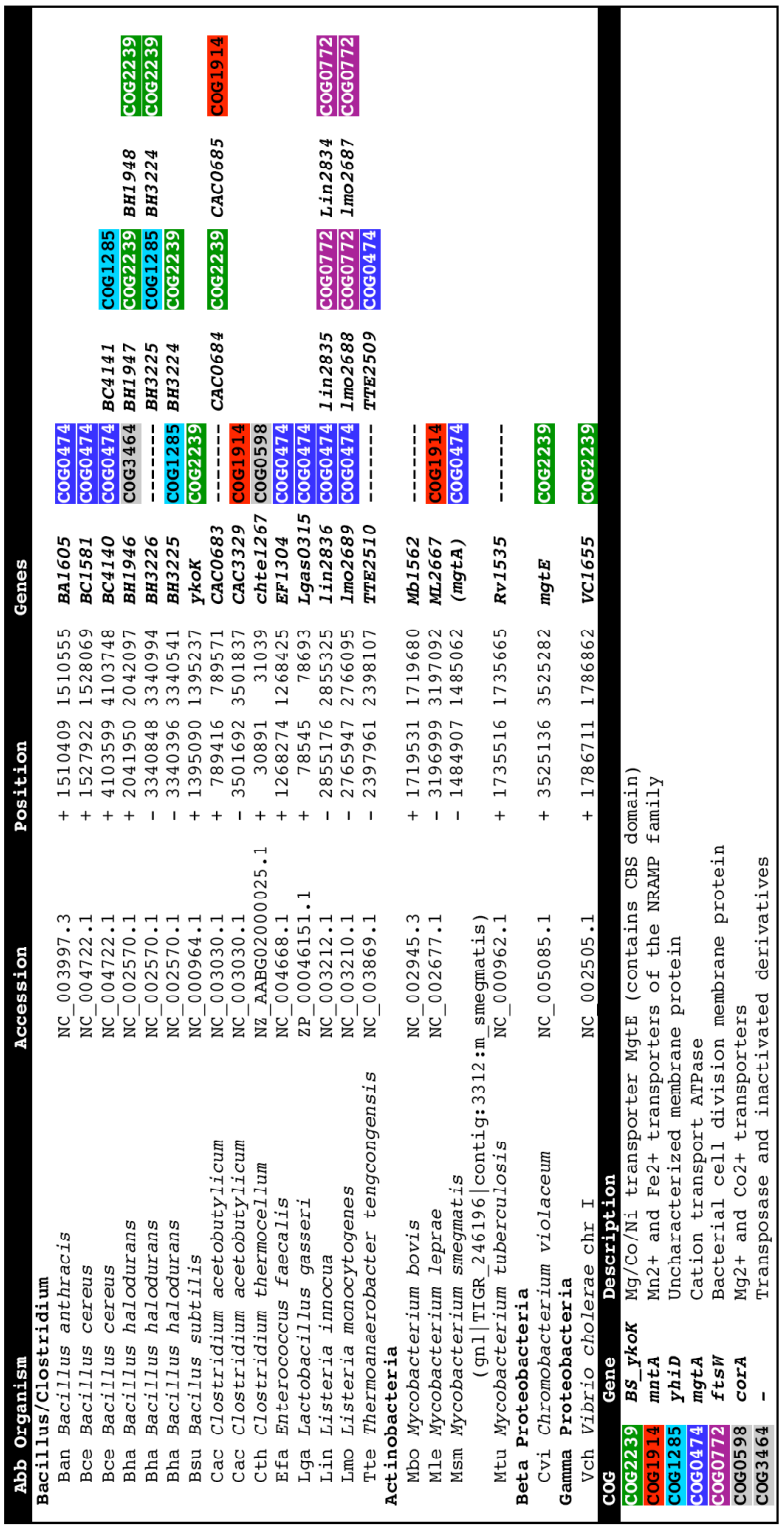


Figure 3.14 Distribution of the *ykoK* element

Refer to the legend of Figure 3.2 for details.

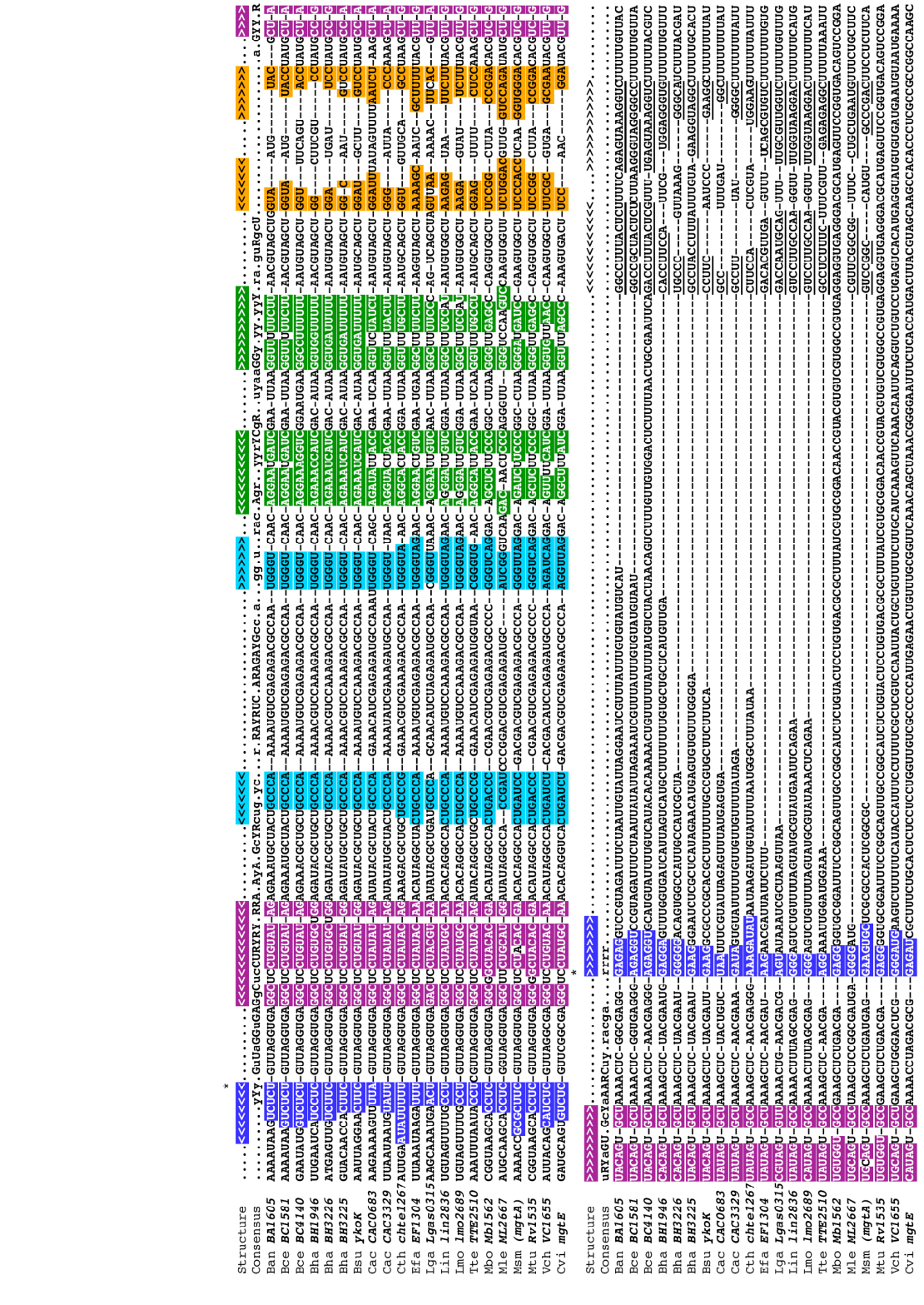


Figure 3.15 Sequence alignment of the *ykoK* element

Refer to the legend of Figure 3.2 for details.

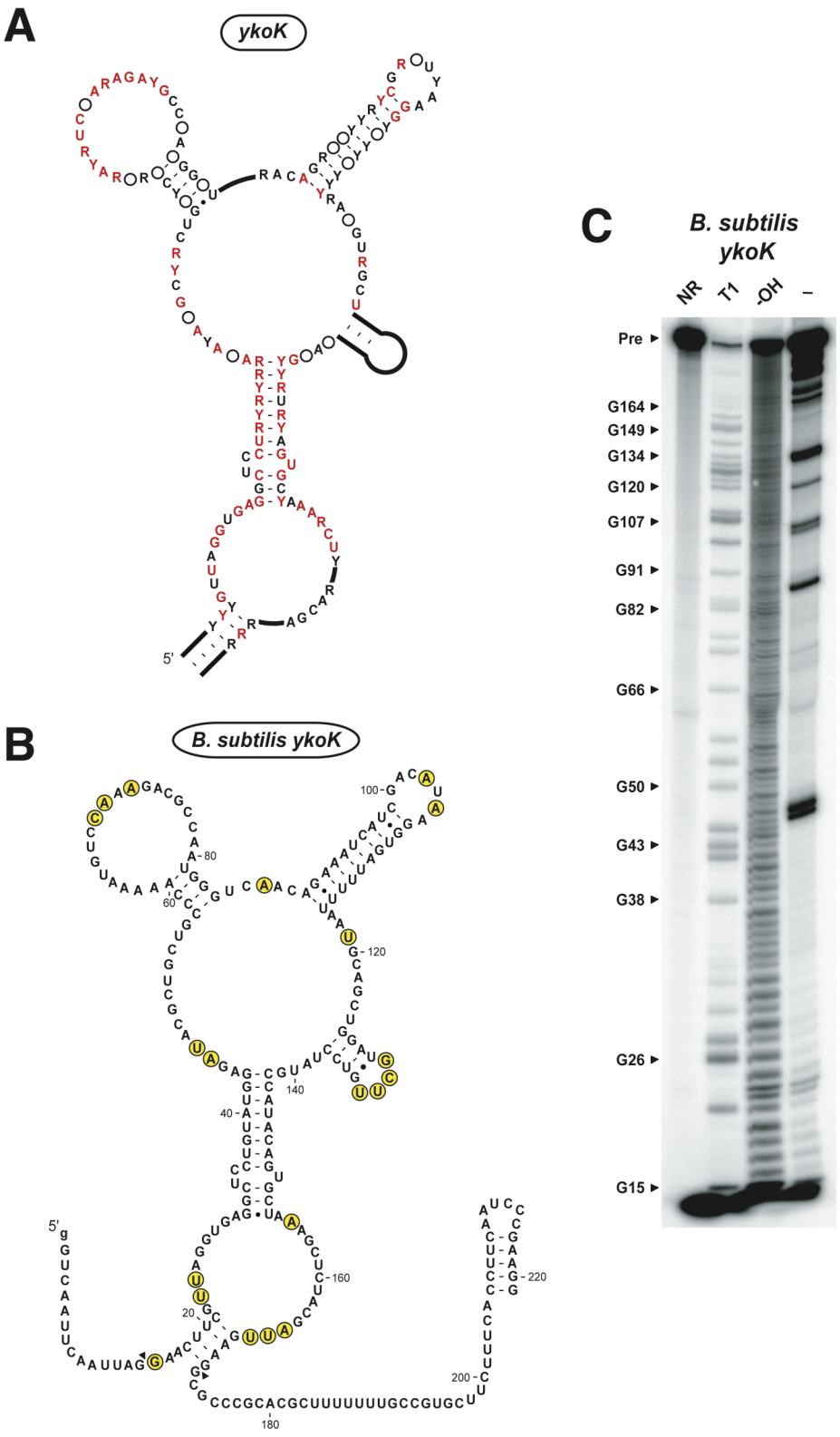


Figure 3.16 Consensus structure and in-line probing of the *ykoK* element

Refer to the legend of Figure 3.6 for details.

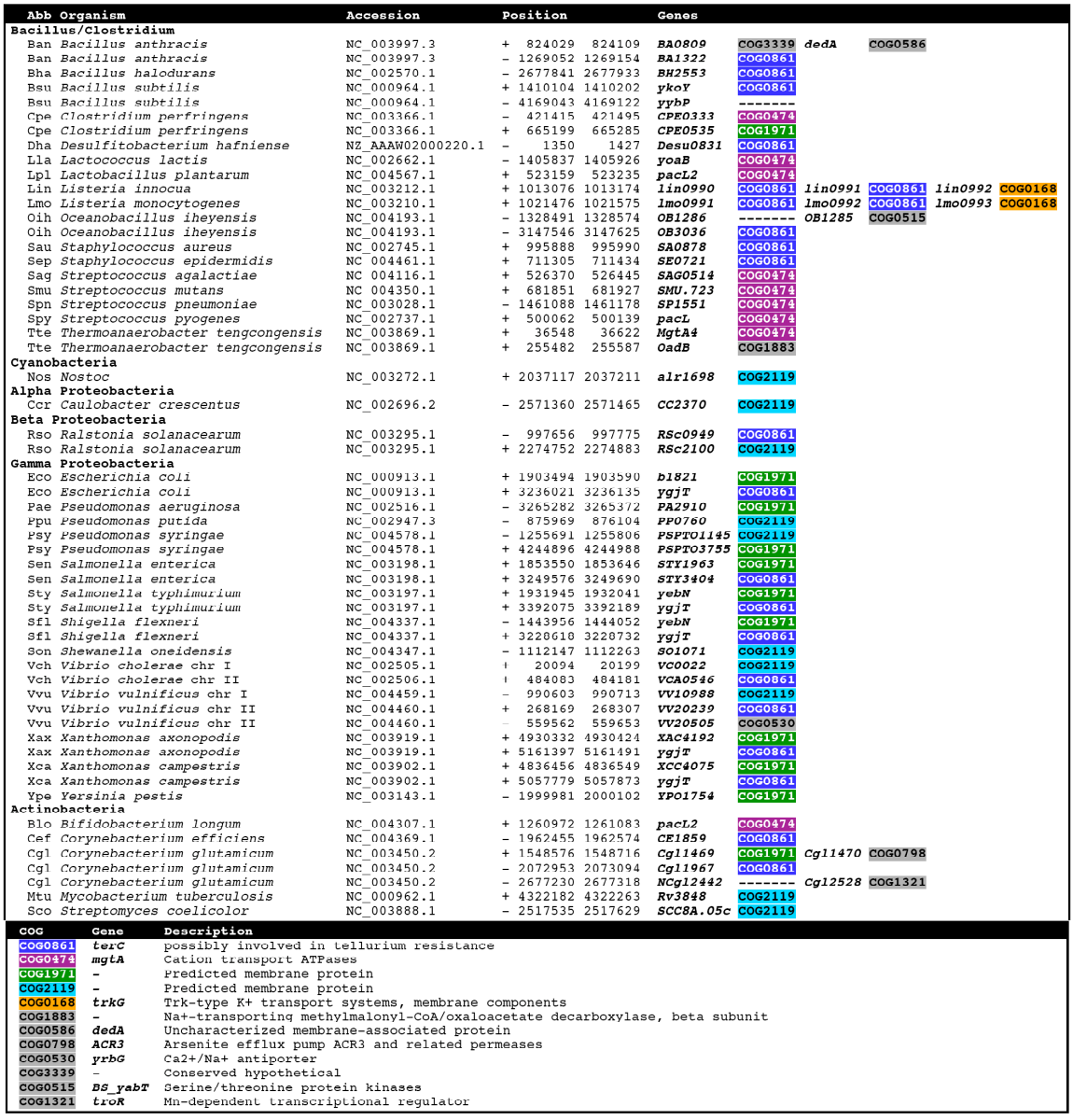


Abb	Organism	Accession	Position	Genes					
Bacillus/Clostridium									
Ban	*Bacillus anthracis*	NC_003997.3	+ 824029 824109	*BA0809*	COG3339	*dedA*	COG0586		
Ban	*Bacillus anthracis*	NC_003997.3	- 1269052 1269154	*BA1322*	COG0861				
Bha	*Bacillus halodurans*	NC_002570.1	- 2677841 2677933	*BH2553*	COG0861				
Bsu	*Bacillus subtilis*	NC_000964.1	+ 1410104 1410202	*ykoY*	COG0861				
Bsu	*Bacillus subtilis*	NC_000964.1	- 4169043 4169122	*yybP*	-------				
Cpe	*Clostridium perfringens*	NC_003366.1	- 421415 421495	*CPE0333*	COG0474				
Cpe	*Clostridium perfringens*	NC_003366.1	+ 665199 665285	*CPE0535*	COG1971				
Dha	*Desulfitobacterium hafniense*	NZ_AAAW02000220.1	- 1350 1427	*Desu0831*	COG0861				
Lla	*Lactococcus lactis*	NC_002662.1	- 1405837 1405926	*yoaB*	COG0474				
Lpl	*Lactobacillus plantarum*	NC_004567.1	+ 523159 523235	*pacL2*	COG0474				
Lin	*Listeria innocua*	NC_003212.1	+ 1013076 1013174	*lin0990*	COG0861	*lin0991*	COG0861	*lin0992*	COG0168
Lmo	*Listeria monocytogenes*	NC_003210.1	+ 1021476 1021575	*lmo0991*	COG0861	*lmo0992*	COG0861	*lmo0993*	COG0168
Oih	*Oceanobacillus iheyensis*	NC_004193.1	- 1328491 1328574	*OB1286*	-------	*OB1285*	COG0515		
Oih	*Oceanobacillus iheyensis*	NC_004193.1	- 3147546 3147625	*OB3036*	COG0861				
Sau	*Staphylococcus aureus*	NC_002745.1	+ 995888 995990	*SA0878*	COG0861				
Sep	*Staphylococcus epidermidis*	NC_004461.1	+ 711305 711434	*SE0721*	COG0861				
Sag	*Streptococcus agalactiae*	NC_004116.1	+ 526370 526445	*SAG0514*	COG0474				
Smu	*Streptococcus mutans*	NC_004350.1	+ 681851 681927	*SMU.723*	COG0474				
Spn	*Streptococcus pneumoniae*	NC_003028.1	- 1461088 1461178	*SP1551*	COG0474				
Spy	*Streptococcus pyogenes*	NC_002737.1	+ 500062 500139	*pacL*	COG0474				
Tte	*Thermoanaerobacter tengcongensis*	NC_003869.1	+ 36548 36622	*MgtA4*	COG0474				
Tte	*Thermoanaerobacter tengcongensis*	NC_003869.1	+ 255482 255587	*OadB*	COG1883				
Cyanobacteria									
Nos	*Nostoc*	NC_003272.1	+ 2037117 2037211	*alr1698*	COG2119				
Alpha Proteobacteria									
Ccr	*Caulobacter crescentus*	NC_002696.2	- 2571360 2571465	*CC2370*	COG2119				
Beta Proteobacteria									
Rso	*Ralstonia solanacearum*	NC_003295.1	- 997656 997775	*RSc0949*	COG0861				
Rso	*Ralstonia solanacearum*	NC_003295.1	+ 2274752 2274883	*RSc2100*	COG2119				
Gamma Proteobacteria									
Eco	*Escherichia coli*	NC_000913.1	+ 1903494 1903590	*b1821*	COG1971				
Eco	*Escherichia coli*	NC_000913.1	+ 3236021 3236135	*ygjT*	COG0861				
Pae	*Pseudomonas aeruginosa*	NC_002516.1	- 3265282 3265372	*PA2910*	COG1971				
Ppu	*Pseudomonas putida*	NC_002947.3	- 875969 876104	*PP0760*	COG2119				
Psy	*Pseudomonas syringae*	NC_004578.1	- 1255691 1255806	*PSPTO1145*	COG2119				
Psy	*Pseudomonas syringae*	NC_004578.1	+ 4244896 4244988	*PSPTO3755*	COG1971				
Sen	*Salmonella enterica*	NC_003198.1	+ 1853550 1853646	*STY1963*	COG1971				
Sen	*Salmonella enterica*	NC_003198.1	+ 3249576 3249690	*STY3404*	COG0861				
Sty	*Salmonella typhimurium*	NC_003197.1	+ 1931945 1932041	*yebN*	COG1971				
Sty	*Salmonella typhimurium*	NC_003197.1	+ 3392075 3392189	*ygjT*	COG0861				
Sfl	*Shigella flexneri*	NC_004337.1	- 1443956 1444052	*yebN*	COG1971				
Sfl	*Shigella flexneri*	NC_004337.1	+ 3220610 3220732	*ygjT*	COG0861				
Son	*Shewanella oneidensis*	NC_004347.1	- 1112147 1112263	*SO1071*	COG2119				
Vch	*Vibrio cholerae* chr I	NC_002505.1	+ 20094 20199	*VC0022*	COG2119				
Vch	*Vibrio cholerae* chr II	NC_002506.1	+ 484083 484181	*VCA0546*	COG0861				
Vvu	*Vibrio vulnificus* chr I	NC_004459.1	- 990603 990713	*VV10988*	COG2119				
Vvu	*Vibrio vulnificus* chr II	NC_004460.1	+ 268169 268307	*VV20239*	COG0861				
Vvu	*Vibrio vulnificus* chr II	NC_004460.1	- 559562 559653	*VV20505*	COG0530				
Xax	*Xanthomonas axonopodis*	NC_003919.1	+ 4930332 4930424	*XAC4192*	COG1971				
Xax	*Xanthomonas axonopodis*	NC_003919.1	+ 5161397 5161491	*ygjT*	COG0861				
Xca	*Xanthomonas campestris*	NC_003902.1	+ 4836456 4836549	*XCC4075*	COG1971				
Xca	*Xanthomonas campestris*	NC_003902.1	+ 5057779 5057873	*ygjT*	COG0861				
Ype	*Yersinia pestis*	NC_003143.1	- 1999981 2000102	*YPO1754*	COG1971				
Actinobacteria									
Blo	*Bifidobacterium longum*	NC_004307.1	+ 1260972 1261083	*pacL2*	COG0474				
Cef	*Corynebacterium efficiens*	NC_004369.1	- 1962455 1962574	*CE1859*	COG0861				
Cgl	*Corynebacterium glutamicum*	NC_003450.2	+ 1548576 1548716	*Cg11469*	COG1971	*Cg11470*	COG0798		
Cgl	*Corynebacterium glutamicum*	NC_003450.2	- 2072953 2073094	*Cg11967*	COG0861				
Cgl	*Corynebacterium glutamicum*	NC_003450.2	- 2677230 2677318	*NCg12442*	-------	*Cg12528*	COG1321		
Mtu	*Mycobacterium tuberculosis*	NC_000962.1	+ 4322182 4322263	*Rv3848*	COG2119				
Sco	*Streptomyces coelicolor*	NC_003888.1	- 2517535 2517629	*SCC8A.05c*	COG2119				

COG	Gene	Description
COG0861	*terC*	possibly involved in tellurium resistance
COG0474	*mgtA*	Cation transport ATPases
COG1971	-	Predicted membrane protein
COG2119	-	Predicted membrane protein
COG0168	*trkG*	Trk-type K+ transport systems, membrane components
COG1883	-	Na+-transporting methylmalonyl-CoA/oxaloacetate decarboxylase, beta subunit
COG0586	*dedA*	Uncharacterized membrane-associated protein
COG0798	*ACR3*	Arsenite efflux pump ACR3 and related permeases
COG0530	*yrbG*	Ca2+/Na+ antiporter
COG3339	-	Conserved hypothetical
COG0515	*BS_yabT*	Serine/threonine protein kinases
COG1321	*troR*	Mn-dependent transcriptional regulator

Figure 3.17 Distribution of the *yybP*/*ykoY* elements

Refer to the legend of Figure 3.2 for details.

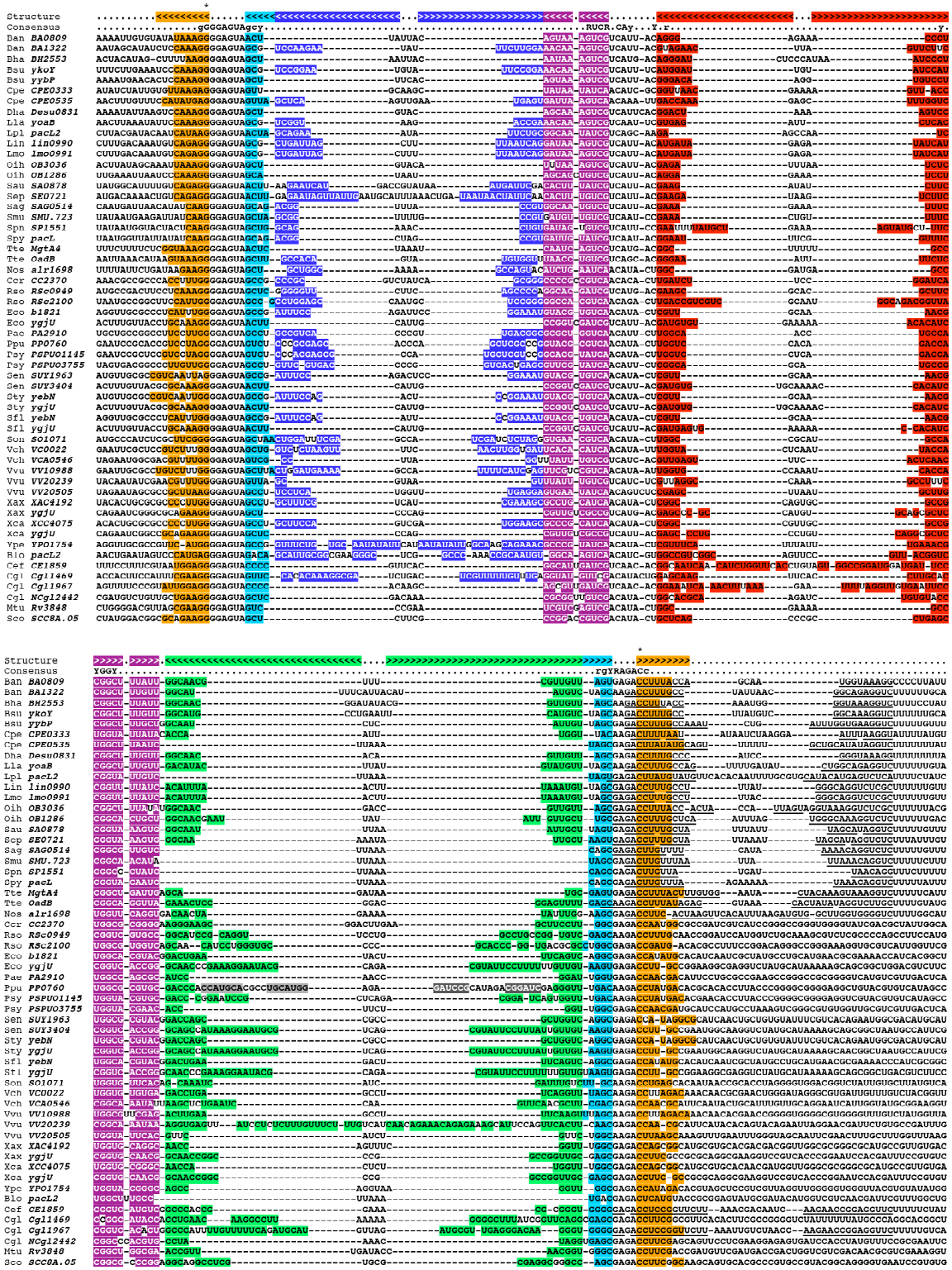


Figure 3.18 Sequence alignment of the *yybP*/*ykoY* element

Refer to the legend of Figure 3.2 for details.

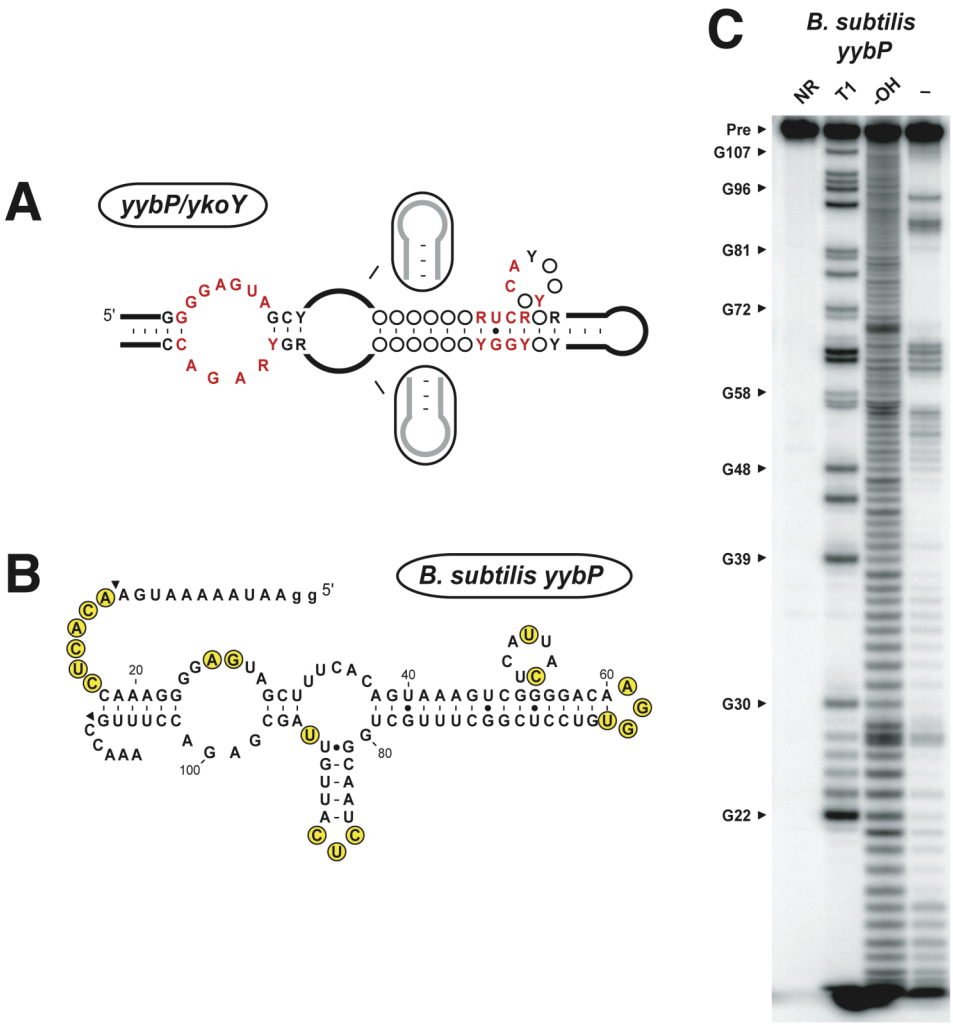


Figure 3.19 Consensus structure and in-line probing of the *yybP/ykoY* element

Refer to the legend of Figure 3.6 for details.

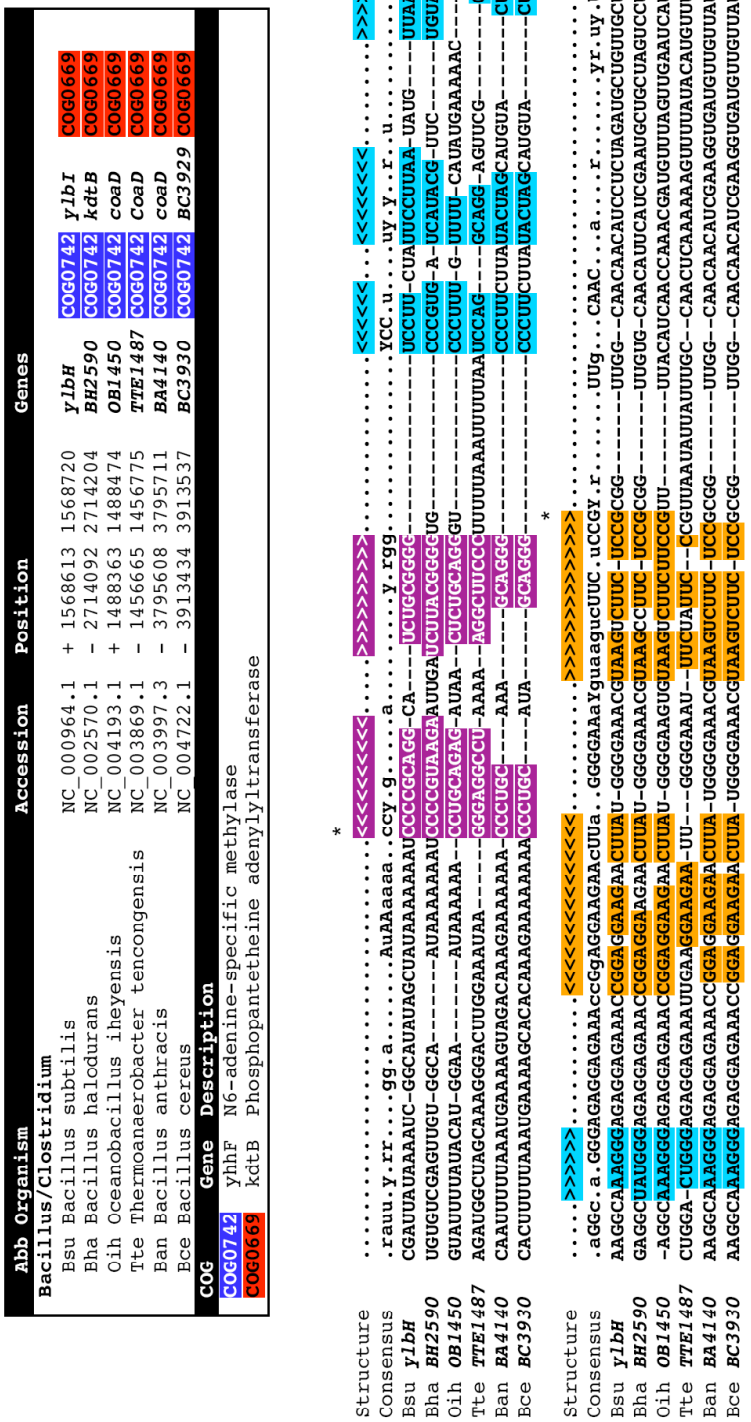


Figure 3.20 Distribution and sequence alignment of the *ylbH* element

Refer to the legend of Figure 3.2 for details.

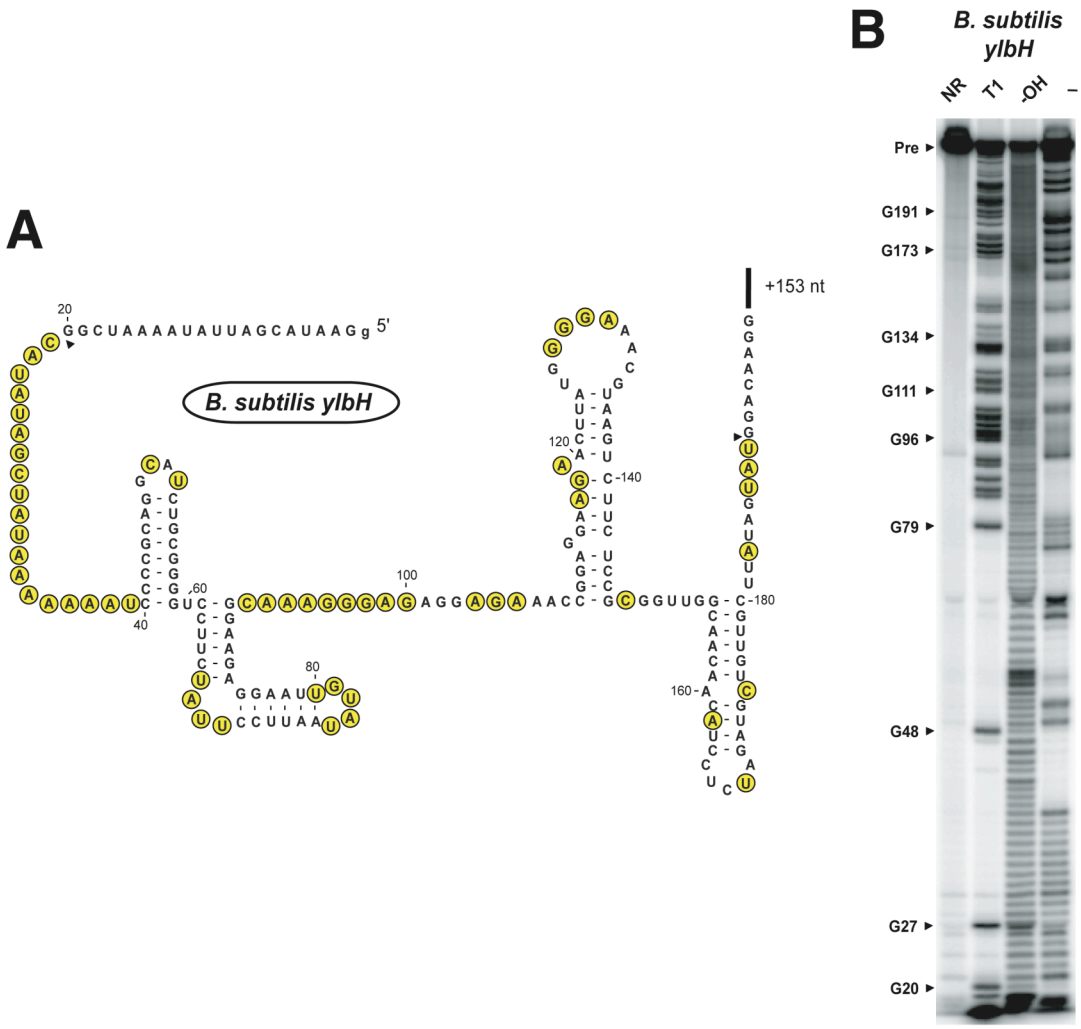


Figure 3.21 In-line probing of the *ylbH* element

Refer to the legend of Figure 3.6 for details.

4 New regulatory RNA motifs in *Agrobacterium tumefaciens*

4.1 Introduction

Most riboswitches described in the previous chapters are found predominantly in low G+C Gram-positive bacteria, and representatives of all these riboswitch classes are present in the genome of *Bacillus subtilis*. We speculated that other bacterial groups might harbor different *cis*-regulatory RNA elements, some of which could be novel riboswitches. In order to discover lineage-specific RNA motifs we concentrated on examining intergenic regions in the genome of the α-proteobacterium *Agrobacterium tumefaciens* using the second incarnation of the BLISS database. We were able to identify five motifs specific to α-proteobacteria that are likely to be RNA elements (Figure 4.1 and Figure 4.2).

4.2 The *metA* element is a second class of SAM riboswitch

The *metA* RNA element [Rfam:RF00521] is found in a variety of α-proteobacteria, and there are even a few occurrences in other proteobacteria and bacteroidetes. This RNA motif was originally identified upstream of the *metA* gene in *A. tumefaciens*, but was subsequently found preceding other genes related to methionine and *S*-adenosylmethionine (SAM) biosynthesis. It has a compact structure with a single stem (P1) and pseudoknot (P2) that are both supported by compensatory mutations in a collection of more than 70 sequence representatives. Usually a putative transcription start site with near-consensus -35 and -10 promoter elements is located a few base pairs upstream of the DNA sequence corresponding to the first nucleotide of P1.

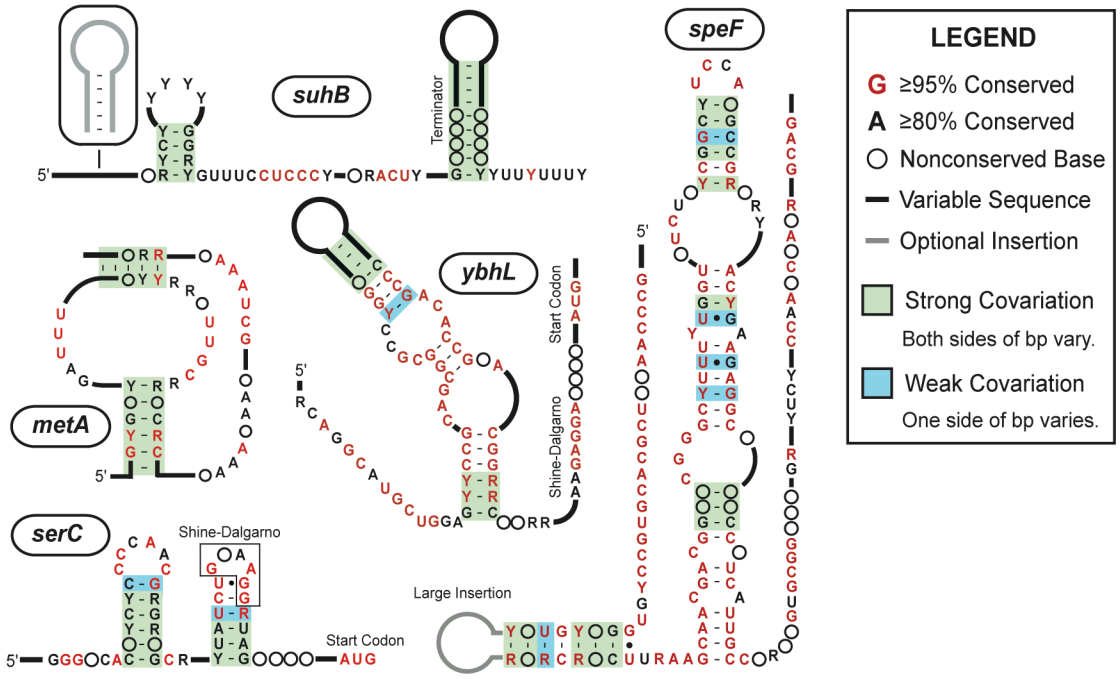


Figure 4.1 Secondary structure models of *A. tumefaciens* regulatory RNA motifs

Red and black positions for each RNA element indicate ≥95% and ≥80% conservation of a particular nucleotide, respectively. Purine (R) or pyrimidine (Y) designations are used when a single nucleotide is not 80% conserved. Solid black lines indicate variable length insertions of unconserved nucleotides, and insets with solid grey lines are optional sequence insertions that are not present in all examples of an element. Unfilled circles represent single nucleotides whose presence (but not sequence) is conserved. Base pairs supported by strong (both bases in the pair vary) and weak (only one base in the pair varies) sequence covariation in a motif alignment have green and blue shaded backgrounds, respectively.

Classification and Organism	serC	speF	suhB	ybhL	metA
α-proteobacteria					
Rhizobiales					
Agrobacterium tumefaciens	1	1	2	1	2
Bartonella henselae	-	1	-	-	-
Bartonella quintana	-	1	-	-	-
Bradyrhizobium japonicum	1	1	5	-	3
Brucella melitensis	1	1	-	1	1
Brucella suis	1	1	-	1	1
Mesorhizobium sp.	1	1	1	1	2
Mesorhizobium loti	1	1	1	1	2
Rhodopseudomonas palustris	1	1	2	-	2
Sinorhizobium meliloti	1	1	1	1	2
Caulobacterales					
Caulobacter crescentus	1	-	4	-	-
Rhodobacterales					
Rhodobacter sphaeroides	1	-	-	-	1
Silicibacter	1	-	-	-	-
Rhodospirillales					
Magnetospirillum magnetotacticum	1	1	3	-	2
Rhodospirillum rubrum	-	-	2	-	1
Sphingomonadales					
Novosphingobium aromaticivorans	1	-	3	-	-
β-proteobacteria					
Bordetella bronchiseptica	-	-	-	-	1
Bordetella parapertussis	-	-	-	-	1
γ-proteobacteria					
Coxiella burnetii	-	-	-	-	1
Bacteroidetes					
Bacteroides thetaiotaomicron	-	-	-	-	1
Porphyromonas gingivalis	-	-	-	-	1
Environmental Sequences	6	-	3	-	52

Figure 4.2 Phylogenetic distributions of *A. tumefaciens* regulatory RNA motifs

Element names correspond to genes adjacent to *A. tumefaciens* representatives.

Many *metA* representatives also contain GC-rich stem-loops followed by U-tails that may function as intrinsic transcription terminators between P2 and the downstream ORF. This arrangement is characteristic of known riboswitches, and suggested that the *metA* RNA was a regulatory element that functioned as a genetic OFF switch [19]. Gram-positive bacteria are known to make extensive use of SAM-sensing riboswitches (Figure 4.3*B*) to repress a similar collection of methionine biosynthesis genes when SAM becomes abundant in the cell (Figure 4.3*C*), often with expression platforms that use transcription terminators [71, 185, 237, 326]. Taking into consideration these factors, we tested whether the simpler *metA* motif also functions as a natural SAM aptamer.

RNA constructs corresponding to nucleotides -230 to -75 relative to the translation start site of the *A. tumefaciens metA* gene [GenBank:NC_003304.1/2703291-2703446] were prepared by *in vitro* transcription. The resulting 156 nt RNA (156 *metA*) contains the majority of the intergenic region but excludes the proposed terminator stem. In-line probing assays revealed that the 156 *metA* structure is greatly modulated in response to SAM concentrations ranging from 1 nM to 6 mM (Figure 4.4*A*). Mapping spontaneous cleavage patterns onto the secondary structure model for 156 *metA* (Figure 4.4*B*) reveals that all SAM-induced changes occur within the conserved *metA* sequence element. There are incidents of both increased and decreased rates of spontaneous RNA cleavage, indicating that SAM does not facilitate general RNA degradation. Rather, SAM associates with 156 *metA* to induce a precise structure that stabilizes certain RNA regions and destabilizes others as has been seen for all riboswitches characterized previously. An apparent K_d value of ~1 μM (Figure 4.4*C*) for the RNA-SAM complex was determined by plotting the normalized fraction of RNA cleaved in several regions against the logarithm of the SAM concentration.

Figure 4.3 The *metA* RNA element

(*A*) Sequence alignment of representative *metA* RNAs. Shaded nucleotides represent conserved base pairing regions as indicated by angle brackets in the secondary structure line (SS). Lowercase and uppercase letters in the consensus line (Cons) indicate 80% and 95% sequence conservation, respectively. Organism abbreviations: Atu, *Agrobacterium tumefaciens*; Bja, *Bradyrhizobium japonicum*; Bme, *Brucella melitensis*; Mma, *Magnetospirillum magnetotacticum*; Mlo, Mesorhizobium loti; Rsp, *Rhodobacter sphaeroides*; Rpa, *Rhodopseudomonas palustris*; Sme, *Sinorhizobium meliloti*; Cbu, *Coxiella burnetii*; Bth, *Bacteroides thetaiotaomicron*; Bbr, *Bordetella bronchiseptica*. **(*B*)** Consensus sequence and structure of the SAM-I riboswitch aptamer found in Gram-positive bacteria. The consensus is updated from [326] and depicted using the same conventions as Figure 4.1*A*. The SAM-II aptamer structure is shown again for comparison. **(*C*)** Genes in the methionine and SAM biosynthetic pathways found downstream of SAM-I and SAM-II riboswitches.

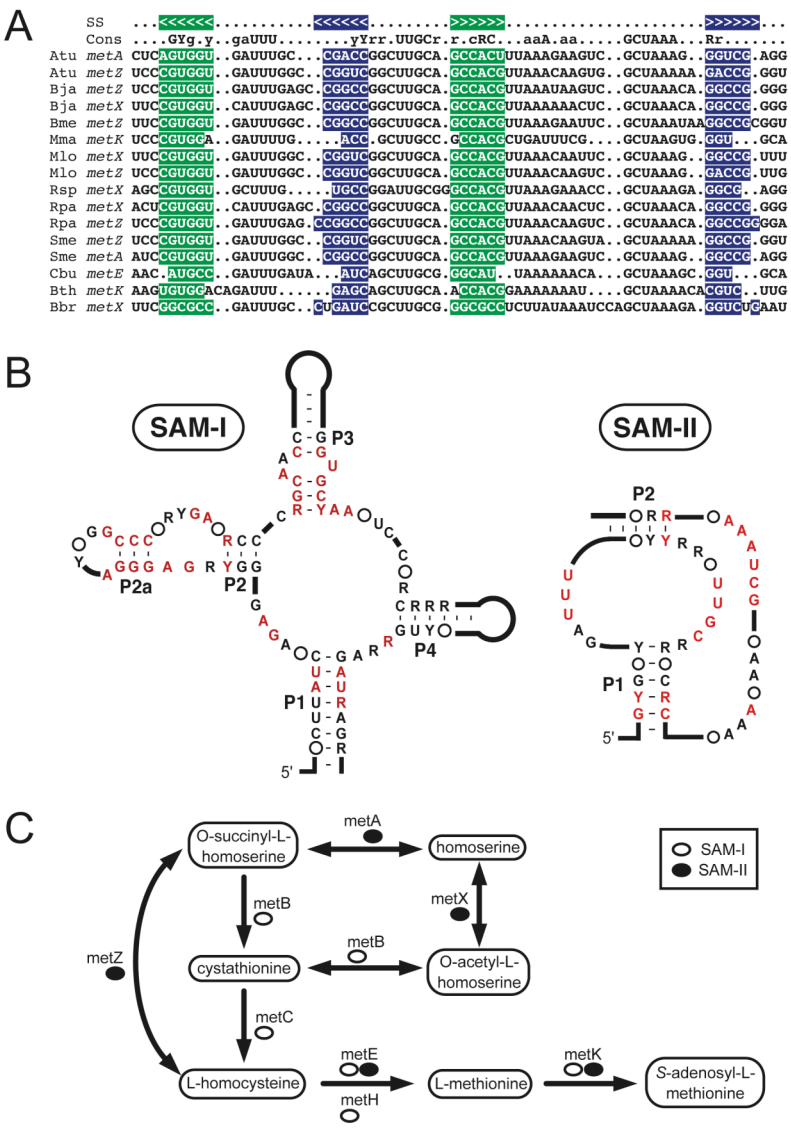


Figure 4.3 The metA RNA element

Figure 4.4 The *metA* element binds SAM

The *metA* element binds SAM. **(*A*)** In-line probing of 156 *metA* RNA from *A. tumefaciens*. ^{32}P-labeled RNA (NR, no reaction) and products resulting from partial digestion with nuclease T1 (T1), partial digestion with alkali ($^{-}$OH), and spontaneous cleavage during a 40 h incubation in the presence of varying of SAM concentrations (1 μM to 6 mM) were separated by polyacrylamide gel electrophoresis. Marker bands generated by T1 digestion corresponding to the positions of certain G residues and full length 156 *metA* RNA (Pre) are labeled. **(*B*)** Sequence and secondary structure model for *A. tumefaciens metA* RNA. Sites of structural modulation for 156 *metA* derived from in-line probing are circled with red, green and yellow representing reduced, increased, and constant scission in the presence of SAM, respectively. **(*C*)** Dependence of spontaneous cleavage in various regions of 156 *metA* on the concentration of SAM. Band intensities for the five regions (labeled 1-5) on the in-line probing gel in (*A*) were quantitated and normalized to the maximum modulation observed. Data from each of these sites corresponds to an apparent K_d of ~1 μM (producing half maximal modulation of cleavage) when plotted against the logarithm of the SAM concentration. Theoretical curves for single ligand binding at sites where cleavage increases (black) and decreases (gray) with a K_d of 1 μM are shown for comparison.

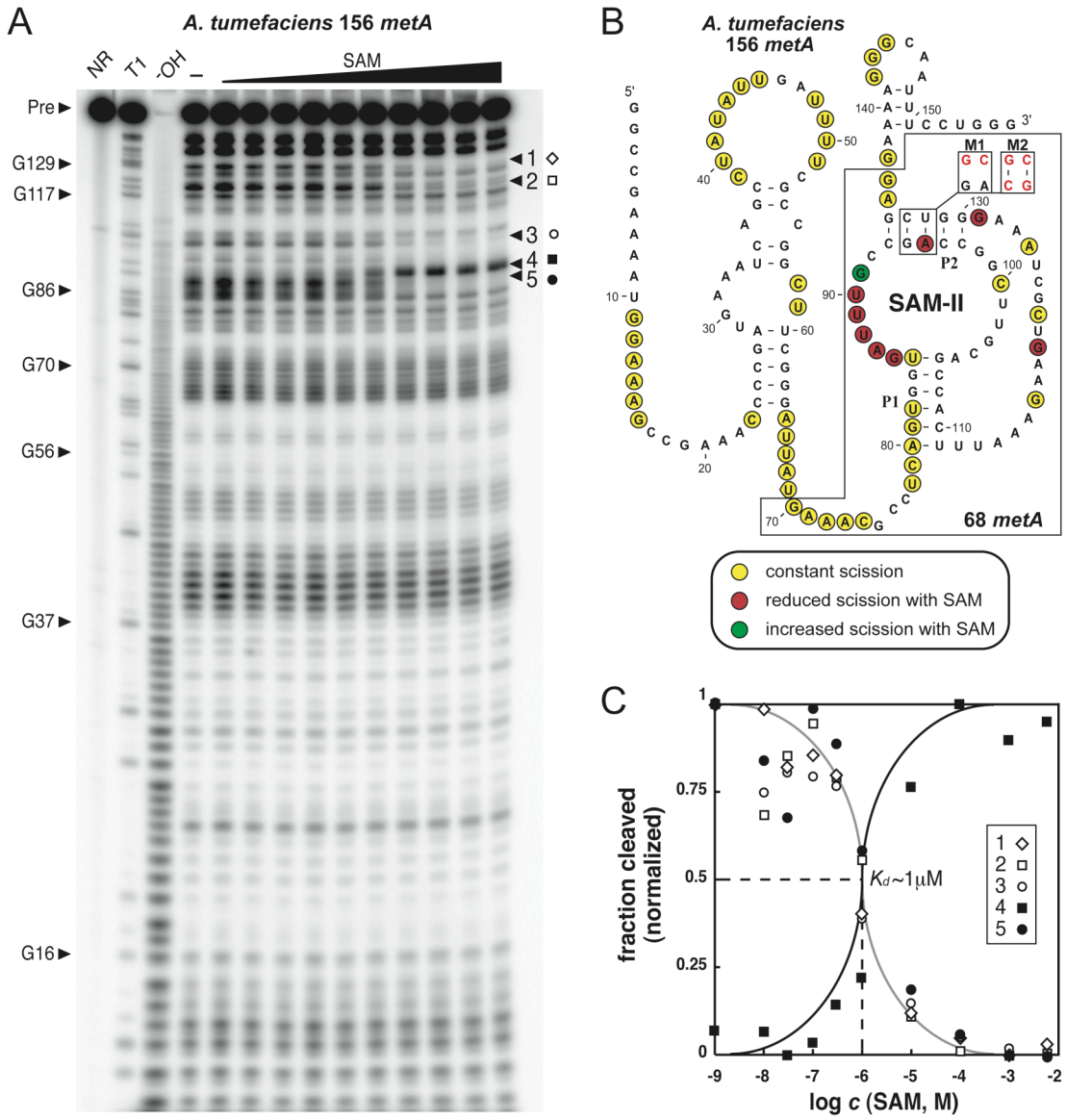


Figure 4.4 The *metA* element binds SAM

These results suggested that only the conserved core of this RNA is necessary for SAM recognition. Indeed, a smaller 68 nt *metA* RNA (68 *metA*) encompassing only nucleotides -161 to -94 [GenBank:NC_003304.1/2703360-2703426] binds with an affinity of ~10 μM and displays a similar change in its spontaneous cleavage pattern (data not shown). Using 68 *metA*, we examined the importance of the formation of the pseudoknot stem (P2) for SAM binding by making two variants (Figure 4.4*B*). One variant carries disruptive mutations (M1: U132→C, C133→G), and the other carries these mutations and the corresponding compensatory mutations (M2: M1, G94→C, A95→G). These RNAs were subjected to in-line probing in the presence of 1 mM SAM (data not shown). Under these conditions, the spontaneous cleavage pattern of M1 did not change in response to SAM. In contrast, M2 exhibited wild-type levels of structural modulation. These results are consistent with covariation in the *metA* sequence alignment that suggests P2 stem formation is required for SAM binding.

We obtained further proof of direct binding between SAM and the *A. tumefaciens* *metA* RNA by equilibrium dialysis. Adding 10 μM of 156 *metA* to one side of an equilibrium dialysis chamber containing 100 nM *S*-adenosyl-L-methionine-(methyl-^{3}H) ([^{3}H]SAM) shifted the distribution of [^{3}H]SAM to favor the RNA side of the membrane by 2.6-fold. A greater shift was not observed because our [^{3}H]SAM sample contained an appreciable amount of radiolabeled breakdown products (see Methods). If 125 μM of unlabeled SAM or the related compound S-adenosyl-L-homocysteine (SAH) are subsequently added to similarly-prepared dialysis chambers, only SAM is able to compete with [^{3}H]SAM and shift the ratio of tritium back to 1. This result demonstrates that 156 *metA* strongly discriminates against the demethylated form of SAM.

The genomic distribution of the *metA* element and its function as a receptor for SAM are consistent with its proposed function as a SAM riboswitch. SAM-II riboswitches

found in α-proteobacteria have a consensus sequence and secondary structure that are distinct from SAM-I riboswitches found in Gram-positive bacteria. A SAM-I riboswitch (the 124 *yitJ* aptamer from *B. subtilis*) has been shown to have a K_d for SAM of ~4 nM [326]. In contrast, the minimized aptamer from the *A. tumefaciens* SAM-II riboswitch upstream of *metA* has a much poorer affinity for SAM (68 *metA*, K_d ~10 μM). It has been shown that *in vitro* selected RNA aptamers that have greater information content generally exhibit greater ligand affinity [41]. The SAM-I and SAM-II aptamers follow this general trend, as low-affinity SAM-II aptamers require only two paired elements and 24 nucleotides to be >80% conserved (Figure 4.3*B*). In comparison, SAM-I aptamers incorporate at least four paired stems and 54 conserved nucleotides.

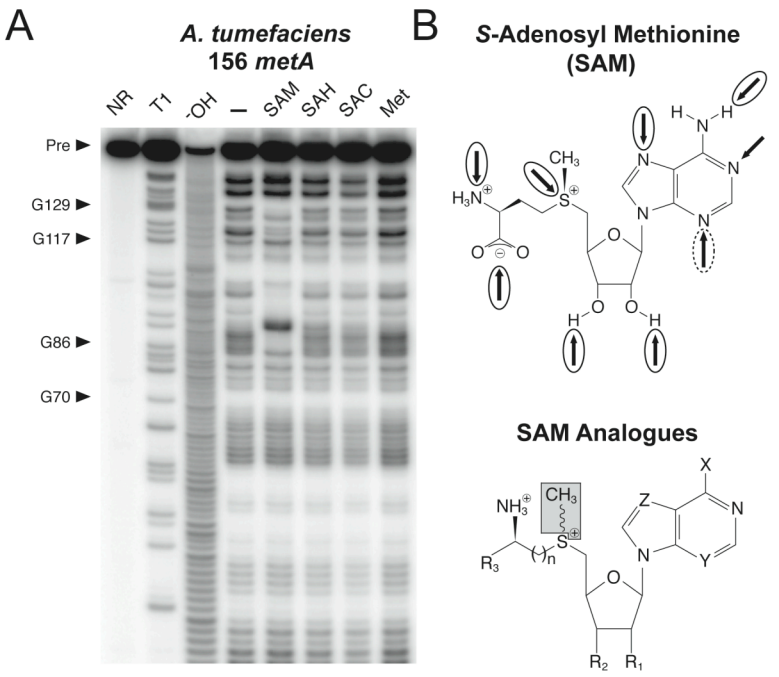
The poorer affinity of the SAM-II aptamer does not necessarily mean that it would exhibit inferior *in vivo* genetic control as a riboswitch. The physiological environments for these riboswitches may be quite different since they operate in divergent groups of bacteria. Furthermore, the kinetics of transcription and ligand binding appear to be more important than equilibrium binding constants for determining whether a flavin mononucleotide (FMN) riboswitch triggers transcription termination [317]. The K_d for the truncated SAM-II aptamer examined in this study is roughly equal to the SAM concentrations needed to trigger transcription termination by SAM-I riboswitches *in vitro* [185, 326]. Furthermore, the affinity of the SAM-II RNA is probably more than sufficient to sense SAM at biologically relevant concentrations. Endogenous SAM levels have been estimated to range from roughly 30 μM to 200 μM in *E. coli* cells grown in rich media [220]. Regardless, the ability of the SAM-II motif to function as an efficient riboswitch might be compromised if it were less capable of discriminating against metabolites with structures similar to SAM than the SAM-I aptamer. Therefore, we investigated the molecular specificity of the SAM-II riboswitch in more detail.

We performed in-line probing assays with 156 *metA* in the presence of various SAM analogues to measure the discrimination of the SAM-II aptamer against related metabolites (Figure 4.5). No RNA structure modulation was seen in the presence of 1 mM SAH, *S*-adenosyl-L-cysteine (SAC), or methionine (Figure 4.5*A*). A more detailed molecular recognition study (Figure 4.5*C*) was conducted using a variety of chemically synthesized SAM derivatives (see Methods) containing systematic single substitutions of functional groups that could potentially be recognized by the SAM-II aptamer (compounds *a*-*f*). It is important to note that the biologically active form of SAM used in our initial tests has the (-) sulfonium configuration [112], while the chemically synthesized compounds are racemic (±). Only two of these compounds modulated the riboswitch structure at a concentration of 1 mM. Full titrations indicated that racemic SAM (compound *a*) had a roughly two-fold higher K_d than (-) SAM and that the 3-deaza SAM analogue (compound *e*) bound with a 50-fold higher K_d.

These analogue binding studies indicate that the SAM-II aptamer creates a binding compartment that recognizes functional groups on the entire surface of SAM. SAM-II discriminates more than 1000 fold against binding SAM analogues lacking the ribose 2'- or 3'-hydroxyl groups and SAM analogues with single substitutions of the adenine 3-aza, 6-amino, or 7-aza groups. A majority of this affinity loss probably comes from disrupting hydrogen bonds or electrostatic interactions between the aptamer and metabolite. Secondary consequences of the chemical changes, such as altering the preferred ribose sugar pucker or purine ring electronic characteristics, may also contribute to weaker binding. Removal of either the carboxyl or amino group from the methionyl moiety is similarly detrimental and could disrupt hydrogen bonds or electrostatic interactions that the aptamer may form with the amino acid zwitterion. Not surprisingly, the SAM-II aptamer also discriminates against the removal of the *S*-methyl

Figure 4.5 Molecular recognition characteristics of SAM-II aptamers

(*A*) In-line probing of *A. tumefaciens* 156 *metA* RNA in the presence of 1 mM SAM, SAH, SAC, and Met. See the Figure 4.4 legend for an explanation of the labels. **(*B*)** Chemical structures of SAM and a generalized SAM analogue. Arrows represent possible hydrogen bonds and electrostatic interactions that could serve as points of recognition by the aptamer. Circled interactions were determined to have strong (solid) or weak (dashed) contributions to binding affinity in singly substituted chemical analogues. Recognition of the N1 position of SAM was not tested. **(*C*)** Apparent K_d values of SAM analogues for binding to 156 *metA.* Columns (n, X, Y, Z, R_1, R_2, R_3) correspond to groups on the core structure in (*B*). The *S*-methyl group (gray box) is not present for SAH and SAC.



C

Compound	n	X	Y	Z	R_1	R_2	R_3	S-methyl	K_D (μM)
Methionine	--	--	--	--	--	--	--	--	> 1,000
SAH	2	NH_2	N	N	OH	OH	CO_2H	no	> 1,000
SAC	1	NH_2	N	N	OH	OH	CO_2H	no	> 1,000
(-)-SAM	2	NH_2	N	N	OH	OH	CO_2H	(-)-CH_3	1
a	2	NH_2	N	N	OH	OH	CO_2H	(±)-CH_3	2
b	2	OH	N	N	OH	OH	CO_2H	(±)-CH_3	> 1,000
c	2	H	N	N	OH	OH	CO_2H	(±)-CH_3	> 1,000
d	2	NH_2	N	CH	OH	OH	CO_2H	(±)-CH_3	> 1,000
e	2	NH_2	CH	N	OH	OH	CO_2H	(±)-CH_3	100
f	2	NH_2	N	N	H	OH	CO_2H	(±)-CH_3	> 1,000
g	2	NH_2	N	N	OH	H	CO_2H	(±)-CH_3	> 1,000
h	2	NH_2	N	N	OH	OH	H	(±)-CH_3	> 1,000
i	1	NH_2	N	N	OH	OH	CO_2H	(±)-CH_3	> 1,000

Figure 4.5 Molecular recognition characteristics of SAM-II aptamers

group that is critical for the function of SAM as a coenzyme, probably because it recognizes positive charge on the sulfonium center. Finally, shortening the methionine side chain by one methylene group prevents SAM binding, most likely by creating a distance constraint that prevents the simultaneous recognition of the methionyl and adenosyl moieties.

We have not investigated whether the 1-aza group of adenine is required for binding, but it is possible that the Watson-Crick face of the adenine base is recognized by a canonical base pair to an aptamer uracil, like that found in the adenine riboswitch [179, 208, 255]. There are six absolutely-conserved uridine residues at putatively single-stranded positions in the SAM-II riboswitch structure that could fulfill this role (Figure 4.3*B*). The molecular recognition determinants for ligand binding by the SAM-II aptamer are summarized in Figure 4.5*B*.

The SAM-I riboswitch binds SAH and SAC ~100- and ~10,000-fold poorer than SAM, respectively [326]. The SAM-II aptamer discriminates greater than 1000-fold against both these compounds, and therefore SAM-II appears to be at least as sensitive to the presence of the *S*-methyl group as SAM-I. Further binding studies with a panel of SAM analogues modified at the sulfonium center indicate that SAM-I tolerates these changes much better than SAM-II [169]. We are unable to quantitate discriminations of greater than 1000-fold against analogues for SAM-II due to its poorer overall K_d. However, our findings indicate that the smaller size of the SAM-II aptamer does not prevent it from attaining the same exquisite discrimination required for efficient genetic control that is exhibited by the aptamers of SAM-I riboswitches.

4.3 The *serC* element

The short *serC* RNA element [Rfam:RF00517] consists of two conserved, base-paired stems (Figure 4.6 and Figure 4.7 — these and all following figures are located at the end

of the chapter). Putative transcription start sites associated with near-consensus upstream promoter elements directly precede all examples of this motif, and the start codon for the *serC* gene is at most 11 nucleotides downstream of the final hairpin. This arrangement suggests that formation of the final hairpin would repress translation by sequestering the ribosome binding site within the 3' side of its base-paired stem and GNRA tetraloop. In-line probing of an RNA corresponding to nucleotides -46 to +11 relative to the *serC* start codon in *A. tumefaciens* [GenBank:NC_003305.1/788249-788193] supports this structure.

The *serC* motif is located upstream of an operon encoding serine transaminase (SerC) and phosphoglycerate dehydrogenase (SerA) enzymes in many α-proteobacteria. Together, these proteins convert 3-phosphoglycerate into 3-phosphoserine during the first two steps of serine biosynthesis. SerC can also catalyze a related step in pyridoxal 5'-phosphate (PLP) biosynthesis involving a similar substrate. We have tested whether L-serine, L-threonine, PLP, pyridoxal, pyridoxine, pyridoxamine, or 4-pyridoxic acid are capable of directly binding to the *A. tumefaciens* RNA. None of these compounds has any effect on RNA structure as judged by in-line probing (data not shown). It is possible that an RNA-binding protein could be responsible for sensing a relevant metabolite, binding to the relatively small *serC* element, and derepressing translation.

4.4 The *speF* element

The extended *speF* element [Rfam:RF00518] is found upstream of proteins classified into COG0019 in several α-proteobacteria (Figure 4.8 and Figure 4.9). Primary sequence conservation begins at the 5' end near a putative transcription start site and continues into a base-paired stem that is topped with a large insertion that can form a four-stem junction in some representatives. Following this stem a stretch of ~80

conserved nucleotides appears to fold into a long bulged stem-loop. This model is tentatively supported by compensatory mutations at a few positions in the alignment, except in the outermost paired helix where the sequence is absolutely conserved. The model is also supported by in-line probing patterns for the RNA corresponding to nucleotides -400 to +3 relative to the *speF* translation start site in *A. tumefaciens* [GenBank:NC_003305.1/205774-205372]. There appear to be further conserved blocks of sequence within the more than 150 nt remaining before the *speF* start codon, but we were unable to confidently assign secondary structures in this region from comparative sequence data or in-line probing results.

Although COG0019 encodes diaminopimelate decarboxylases (*lysA*) in other groups of bacteria, a phylogenetic tree of protein sequences indicates that the genes downstream of this motif are orthologs of *B. subtilis speF*, an ornithine decarboxylase enzyme that catalyzes one of the first steps in polyamine biosynthesis. We have tested whether metabolites related to this pathway bind directly to the *A. tumefaciens* intergenic region and cause structural changes detectable by in-line probing. There is no measurable binding of L-ornithine, L-lysine, *meso*-diaminopimelate, putrescine, cadaverine, or spermidine to the *speF* RNA construct used in this study (data not shown).

4.5 The *suhB* element

The *suhB* element [Rfam:RF00519] was originally recognized upstream of one of nine *A. tumefaciens* (Figure 4.10 and Figure 4.11) ORFs encoding proteins with similarity to archeal fructose-1,6-bisphosphatases (COG0483). After more matches were found, it became clear that this motif was most likely not a *cis*-acting regulatory element for the *suhB* gene but was more likely to be a small noncoding RNA that is transcribed from the opposite strand relative to the *suhB* gene. In this orientation, each representative carries

a putative promoter and intrinsic transcription terminator flanking the conserved sequence domain. Further searches for this motif revealed that multiple copies are present in many α-proteobacterial genomes (e.g. 5 in *Bradyrhizobium japonicum* and 4 in *Caulobacter crescentus*) and that it is not associated with specific neighboring genes. The only evolutionarily conserved secondary structure in the *suhB* noncoding RNA, aside from the terminator stem, appears to be a short helix near its 5' end. In-line probing of an RNA corresponding to a portion of one *A. tumefaciens* intergenic region containing this motif [GenBank:NC_003305.1/979721-979594] also indicates that its characteristic conserved sequences reside in unstructured regions, suggesting that this family could be involved in some form of antisense gene regulation or other small noncoding RNA function [307].

4.6 The *ybhL* element

The *ybhL* RNA motif [Rfam:RF00520] appears to be restricted to bacteria from the *Rhizobiales* order (Figure 4.12 and Figure 4.13). In-line probing data from an RNA corresponding to nucleotides -139 to +21 relative to the translation start site of the *ybhL* gene in *A. tumefaciens* [GenBank:NC_003304.1/2665399-2665558] indicate that this element folds into a doubly-bulged hairpin of ~60 nt. Sequence covariation substantiates the formation of the outermost and innermost paired stems. A putative transcription start site is located close to the beginning of the hairpin within a region that appears highly conserved in our limited number of sequence examples. This RNA motif always occurs upstream of genes related to the *E. coli ybhL* gene (COG0670), a putative integral membrane protein. Because the function of *ybhL* is not known, we were unable to formulate any hypotheses for the role of this RNA element.

4.7 Conclusions

We characterized five novel structured RNA elements by focusing our comparative sequence analysis of IGRs on α-proteobacterial genomes. One of the five newly identified motifs from *A. tumefaciens* proved to function as a new class of riboswitch that senses *S*-adenosylmethionine (SAM). This SAM-II aptamer, found primarily in α-proteobacteria, has a much smaller conserved structure than the aptamer of the SAM-I riboswitch from low G+C Gram-positive bacteria. Despite having an overall lower affinity for SAM, the SAM-II aptamer appears to be adapted for precise genetic control and discriminates against closely related compounds at least well as the SAM-I aptamer. Although multiple RNA solutions to small-molecule binding challenges are often found by *in vitro* selection (e.g. ATP aptamers; [129, 246, 247]), this was the first report that nature also exploits the structural diversity of RNA and employs multiple, unique mRNA motifs to sense a single metabolite.

After this study, a third SAM riboswitch (the S_{MK} box) was reported to regulate the production of SAM synthase (the *metK* gene product) in lactic acid bacteria [83]. This bacterial order is classified within the Bacillus/Clostridium group, but it was known that some Lactobacilliales genera lacked the SAM-I riboswitches that usually regulate *metK*. This SAM-III riboswitch has a simpler consensus secondary structure than SAM-I. It consists of two paired regions bracketing a complex conserved asymmetric bulge, and the terminal stem-loop accommodates very long (>100 nt) insertions. It has been reported to regulate translation initiation by directly sequestering the downstream ribosome binding site within a conserved pseudoknotted pairing to bases in its internal bulge upon binding SAM. It is remarkable that bacteria have developed three separate SAM-sensing riboswitches. *S*-adenosylmethionine must be a particularly important

cofactor for cells to monitor, and many RNA structures must be especially suited to recognize this coenzyme molecule.

4.8 Methods

Bioinformatics

An updated version of the BLISS database (v2), containing the results of an all-versus-all BLAST comparison of IGRs from 116 microbial genomes, was used to manually examine several α-proteobacterial genomes for conserved RNA elements. BLISS web pages display alignments of homology between bacterial IGRs along with compilations of sequence statistics, species distributions, and neighboring gene function assignments from the COG database [274] in a collaborative annotation environment. Further matches to the five motifs were found by iterative BLAST and filtered covariance model searches [311, 312] of unfinished bacterial genomes and environmental sequences [295]. Phylogenetic trees were constructed with CLUSTALW [280] to clarify the specific functions of some genes assigned to ambiguous COGs.

In-line probing assays

RNA preparation, radiolabeling, and in-line probing assays were performed essentially as described previously [260]. DNA templates for *in vitro* transcription with T7 RNA polymerase promoters were prepared by whole-cell PCR from *A. tumefaciens* strain GV2260, except for 68 *metA* RNA mutants M1 and M2 where overlapping synthetic oligonucleotides were extended with reverse transcriptase. For each in-line probing reaction, ~1 nM 5' ^{32}P-radiolabeled RNA was incubated for 40-48 h in a mixture of 50 mM Tris-HCl (pH 8.3 at 25°C), 20 mM $MgCl_2$, 100 mM KCl, and various compounds as indicated. All compounds used for in-line probing or chemical synthesis were purchased

from Sigma-Aldrich. SAM analogues were prepared as diastereomeric mixtures by the reaction of *S*-adenosylhomocysteine derivatives [30, 31] and excess methyl iodide [32].

Equilibrium dialysis

Assays were performed by adding 100 nM *S*-adenosyl-L-methionine-(methyl-^{3}H) to side 'a' and 10 μM *metA* RNA to side 'b' of a DispoEquilibrium Biodialyser with a 5 kDa MWCO (The Nest Group, Inc., Southboro, MA) in 40 mM $MgCl_2$, 200 mM KCl, 200 mM Tris-HCl (pH 8.5 at 23°C). The sample remaining on side 'a' of the chamber after 10 h of incubation at 23°C was replaced with fresh buffer to increase the final binding signal by preferentially removing non-interacting, radiolabeled metabolite breakdown products [202]. After a second 10 h incubation, the counts in each chamber were recorded. Unlabeled SAM or SAH was added to a concentration of 125 μM in side 'a' and the counts were measured again after a final 10 h incubation.

Abb	Organism	Accession/Start-End	Genes
Alpha Proteobacteria			
Atu	*Agrobacterium tumefaciens* C58 chr linear	NC_003305.1/788250-788208	*serC* *serA*
Bja	*Bradyrhizobium japonicum* USDA 110	NC_004463.1/8141404-8141361	*serC* *serA*
Bme	*Brucella melitensis* 16M chr I	NC_003317.1/358102-358145	*serC* *serA*
Brs	*Brucella suis* 1330 chr I	NC_004310.1/1631166-1631123	*serC* *serA*
Ccr	*Caulobacter crescentus* CB15	NC_002696.2/3479617-3479574	*serC* *serA*
Mlo	*Mesorhizobium loti* MAFF303099	NC_002678.1/3114177-3114135	*serC* *serA*
Mes	*Mesorhizobium* BNC1	NZ_AAED01000001.1/117537-117580	*serC* *serA*
Mma	*Magnetospirillum magnetotacticum* MS-1	NZ_AAAP01003860.1/22623-22667	*serC* *serA*
Nar	*Novosphingobium aromaticivorans* DSM 12444	NZ_AAAV02000003.1/53466-53509	*serC* *serA*
Rsp	*Rhodobacter sphaeroides* 2.4.1	NZ_AAAE01000117.1/33042-33001	*serC* *serA*
Rpa	*Rhodopseudomonas palustris* CGA009	NC_005296.1/4861073-4861031	*serC* *serA*
Sil	*Silicibacter* TM1040	NZ_AAFG01000007.1/111863-111905	*serC* *serA*
Sme	*Sinorhizobium meliloti* 1021	NC_003047.1/2938127-2938085	*serC* *serA*
Environmental			
Env1	Environmental sequence IBEA_CTG_2124392	AACY01024091.1/158-199	*serC*
Env2	Environmental sequence IBEA_CTG_2159813	AACY01068735.1/1559-1517	*serC* *serA*
Env3	Environmental sequence IBEA_CTG_2076514	AACY01075427.1/1481-1439	*serC*
Env4	Environmental sequence IBEA_CTG_2157737	AACY01079269.1/306-348	*serC* *serA*
Env5	Environmental sequence IBEA_CTG_2082756	AACY01092568.1/320-362	*serC* *serA*
Env6	Environmental sequence IBEA_CTG_UBAK909TF	AACY01564294.1/483-441	*serC*

Gene	Description
serC	Phosphoserine transaminase
serA	Phosphoglycerate dehydrogenase

```
                                              *                                            *
SS      .-35.................-10.................<<<<<<......>>>>>>....<<<<<<<....>>>>>>>.......................
Cons    UUG..........y..................r...gYgGG.cAc.ycycCCcAaCGrgr.gCr...yuaUCUG.aAGGRuag............AUGry.R.....
Atu     uugacgGAGGCCGGCGACAGGA.auagca.ACGGCGGCAGGACACCUCUCCCUAACGAGAGGCGC..CUAUCUGAAAGGGUAG.UCAA.......augACAGAUAUA
Bja     uugacgACUCUUCCCGUUUCCCGuaucag..CCGCGGCGGGACACCUCCCCCCAACGGGAGGCUUA.CUAUCUGGAAGGGUAG.AAU........augACUGUAGCG
Bme     uugaggAAAAGCGCGCUUUCGGCuauccc..GGUAAGUGGGCCACGUCUCCCCAACGAGACGCGUG.CUAUCUGAGAGGAUAG.AAACAAC....augACGACGAUU
Brs     uugaggAAAAGCGCGCUUUCGGCuauccc..GGUAAGUGGGCCACGUCUCCCCAACGAGACGCGUG.CUAUCUGAGAGGAUAG.AAACAAC....augACGACGAUU
Ccr     uugcaaUUCCCCUCCAGACGCGGuuucuc...GCCGGCGGGCCACGCCCCCCCAACGGGGCGCAGU.CCAUCUGUAAGGAUGG.GAGACAUCGCUaugACCACGACC
Mlo     uugacgCCCGAGGCCUGAACGC.auaucg.CCGUCCGUGGGCCACCCCUCCCCAACGAGGGGCCA..CUAUCUGGAAGGAUAG.ACCACG.....augACGACGCAU
Mes     uugcgcUUUGUCUCCGCUUCGGGuaaugc.CCG.CAGCGGGCUACCCCUCCCCACCGAGGGGCGAG.CUAUCUGGAAGGAUAG.GAACAAC....augACCAUGCUC
Mma     uugcgaAUCGCUCCUGGCCCUC.auacac..UCCGCGCGGGCCACGGUCCCCCAACGGCCCGCGU.CCCUUCUGCGAGGGAGGGUUUUUUAU...augAGCGAUGUU
Nar     uugacaUAUCCCGCCGCGCCAC.uagagc..GACCGGCGGGCCACGCGGUCCCAACGCCGCGCGUG.CUCUCUGCAAGGAGAG.ACUAC......augUCUGCUACG
Rsp     uugcgcCAGAGCCUGCCGGGGCGuaaugg.CGA.CAGCGGGACACCUCUCCCCAACGAGAGGCAUC..UAUCUGGAAGGAUA..CC.........augUCUGAUCUG
Rpa     uugacuUUCGAGGCCAAAGGAC.uauccc..CAGCAGCGGGAAACCUCUCCCACCGAGAGGCAG..CUAUCUGGAAGGAUAG.ACU........augACUGUAGCG
Sil     uugcgcGGAAUCAUCCUCCCGC.auaucg.CUGUCGGUGGGACACCCCUCCCCAACGAGGGGCGC..UUAUCUGGAAGGGUAG.CAUUU......augGCUAUUGCA
Sme     uugacuUCCUCUCCGAGCCGC..aauauuCCCGGCGGCGGGACACCCCUCCCCAACGAGGGGCUU..CUAUCUGAGAGGAUAG.CAU........augACGAAGCCU
Env1    uuguuaAAGUCUUUCGAAAUC..uauaucCGUUUUUGUGGGACACCCCUCCCCAACGAGGGGCAUA..UAUCUGUAAGGGUA..AAAUAA.....augAAUAUAGAA
Env2    uuguuaAAGGCAGGAUUAUCAA.uuauuu.CGAUACAUGGGACACCUCUCCCAACGAGAGGCAA..UUAUCUGUAAGGGUAA.CAGAC......augGAUACUAAA
Env3    uugauaAAUGUAAAGUUUUGCAGuauuaa...AGAAAUGGGACAGCUCUCCCCAACGAGAGUCAA..CUAUCUGGAAGGGUAG.AGAAC......augACAAAGAAA
Env4    uugacuAAUCUUAUUAUAAGU..uauauc.UGACUCUUGGGACACUCCUCCCCAACGAGGAGCAA..CUAUCUGUAAGGGUAG.AUAA.......augAUAGAAAAA
Env5    uugacuAAUCUUACUUUAAGU..uauauu.CCACUCUUGGGACACUCCUCCCCAACGAGGAGCAA..CUAUCUGUAAGGGUAG.AGUU.......augAUAGAAAAA
Env6    uugacuAAUCUUAUUAUAAGU..uauauc.UCACUCUUGGGACACUCCUCCCCAACGAGGAGCAA..CUAUCUGUAAGGGUAG.AUAA.......augAUAGAAAAA
```

Figure 4.6 Distribution and multiple sequence alignment of the *serC* element

The distribution table (above) displays the genomic context of each *serC* element in the multiple sequence alignment (below). Accession numbers refer to GenBank nucleotide records. Functional assignments are provided for the downstream genes that are likely to be regulated if the element is located in the 5' UTR of an mRNA. In the sequence alignment, lowercase and uppercase letters in the consensus line indicate 80% and 95% sequence conservation, respectively. Putative elements related to transcription and translation initiation are shown in lowercase letters with shaded backgrounds: orange, promoter -35 and -10 boxes; green, start codons; red, ribosome-binding sites. Other shaded nucleotides represent conserved base pairing regions. Periods represent gaps, and dashes indicate the end of an incomplete sequence record.

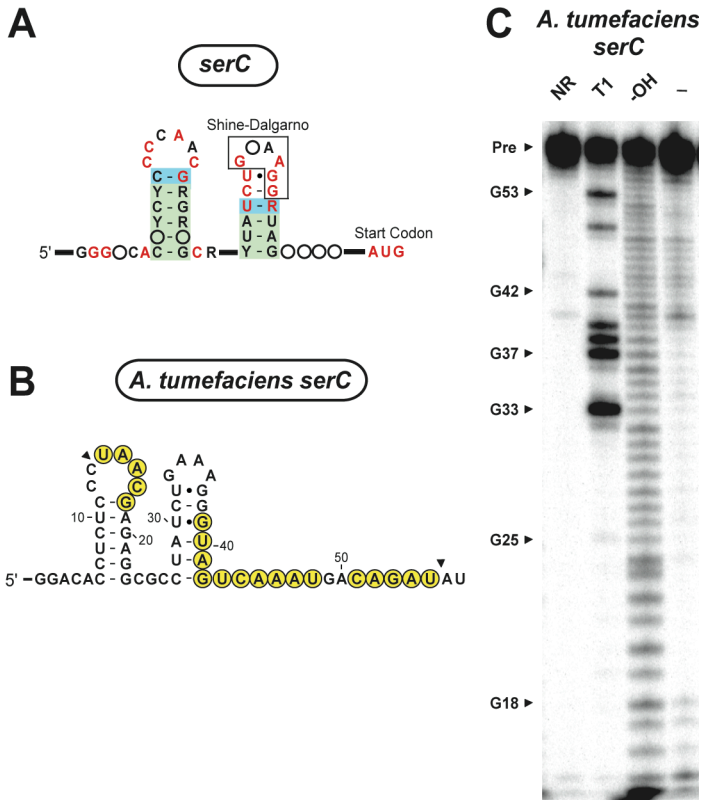


Figure 4.7 Consensus structure and in-line probing of the *serC* element

(*A*) Consensus structure. Refer to the legend of Figure 4.1 for details. **(*B*)** Secondary structure model for the *A. tumefaciens serC* RNA construct. Spontaneous cleavage products in the in-line probing gel are mapped onto nucleotides in the experimental RNA construct probed from *A. tumefaciens* with shaded circles. The boundaries of the region where this mapping was possible are demarked by arrowheads. **(*C*)** In-line probing gel. The lanes are no reaction (NR), partial RNase T1 digestion (T1), partial alkali digestion ($^-$OH), and spontaneous cleavage during a 40 hr incubation (–). The band labeled Pre is the full-length precursor RNA. Some G-specific RNase T1 cleavage products (G18, G20, G22, G23) expected in the T1 lane are missing for *serC*, presumably due to RNA structures that preclude enzyme action. For other constructs, 5' guanosyl residues (g) were sometimes added to improve *in vitro* transcription yields with T7 RNA polymerase.

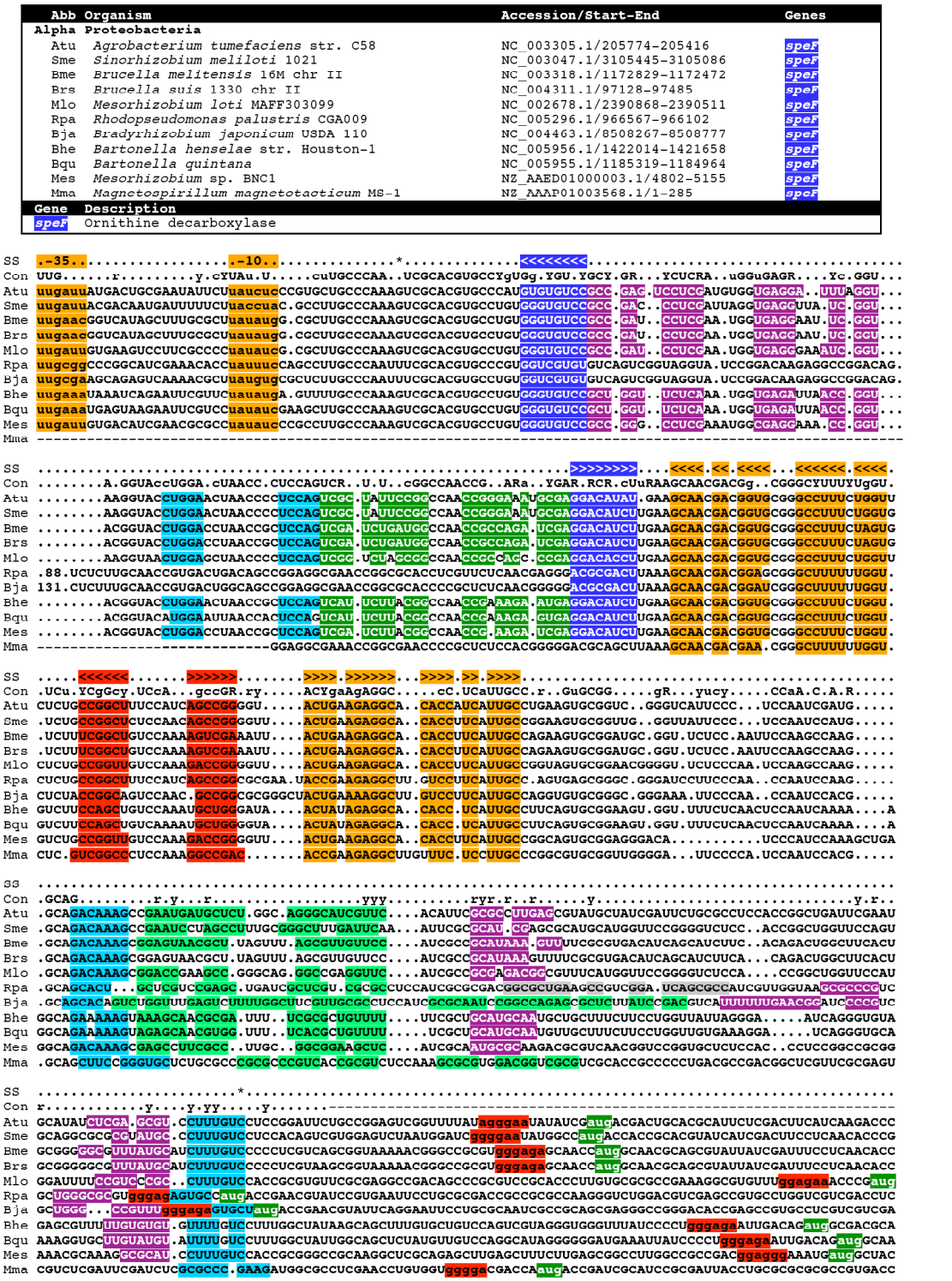


Figure 4.8 Distribution and multiple sequence alignment of the *speF* element

Refer to the legend of Figure 4.6 for details.

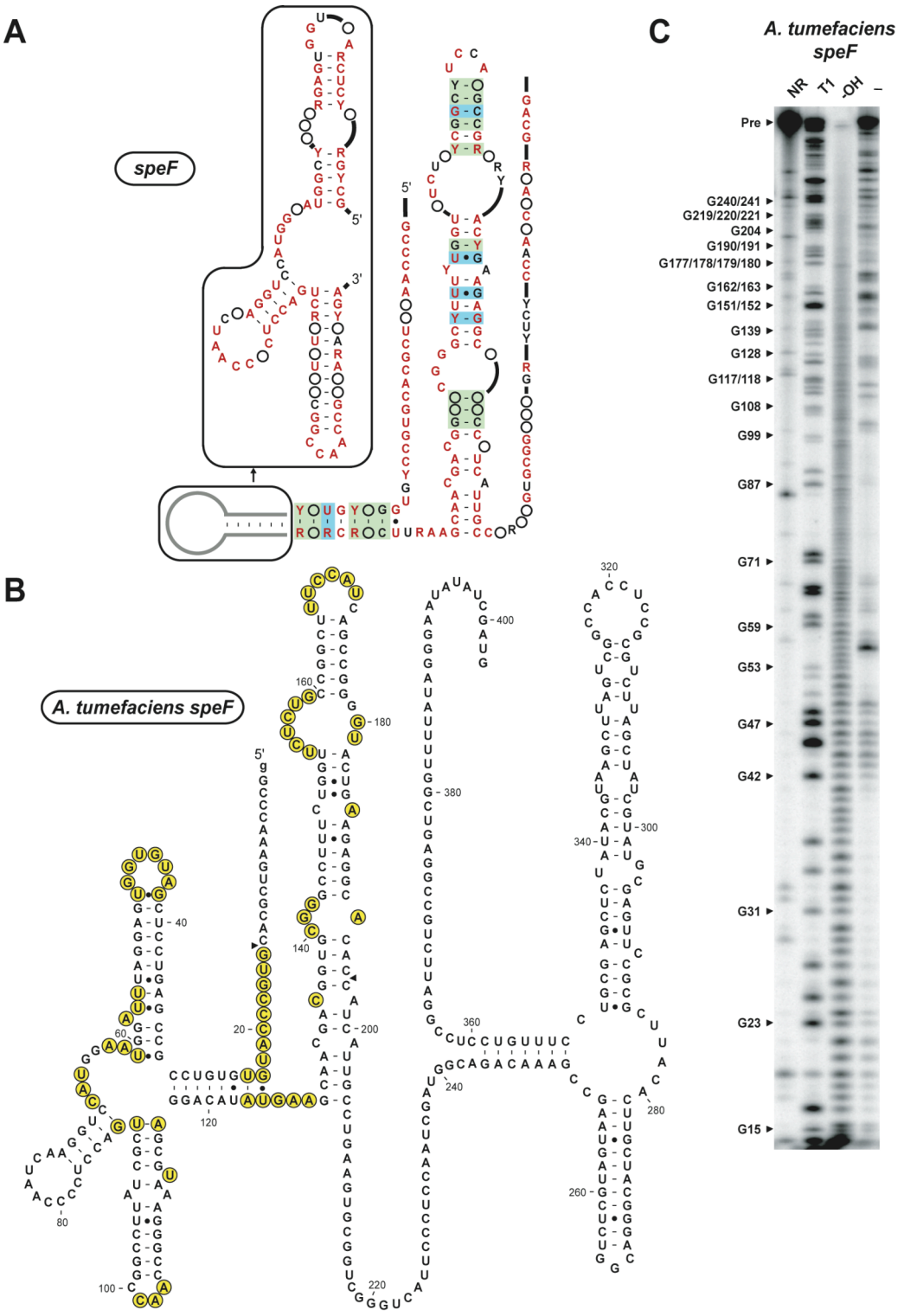


Figure 4.9 Consensus structure and in-line probing of the *speF* element

Refer to the legend of Figure 4.7 for details.

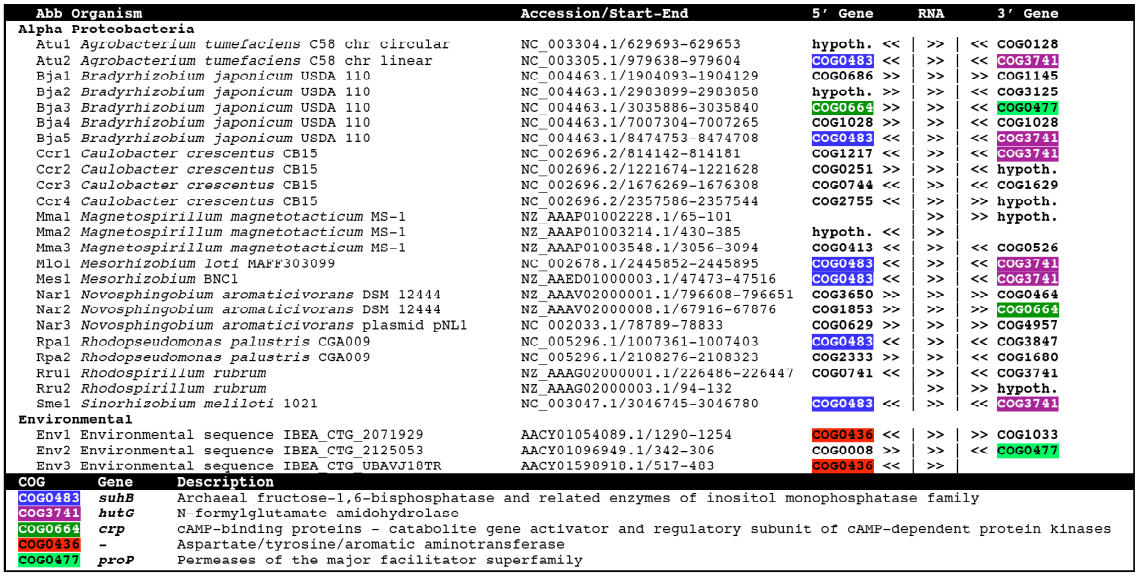


Figure 4.10 Distribution and multiple sequence alignment of the *suhB* element

Refer to the legend of Figure 4.6 for details.

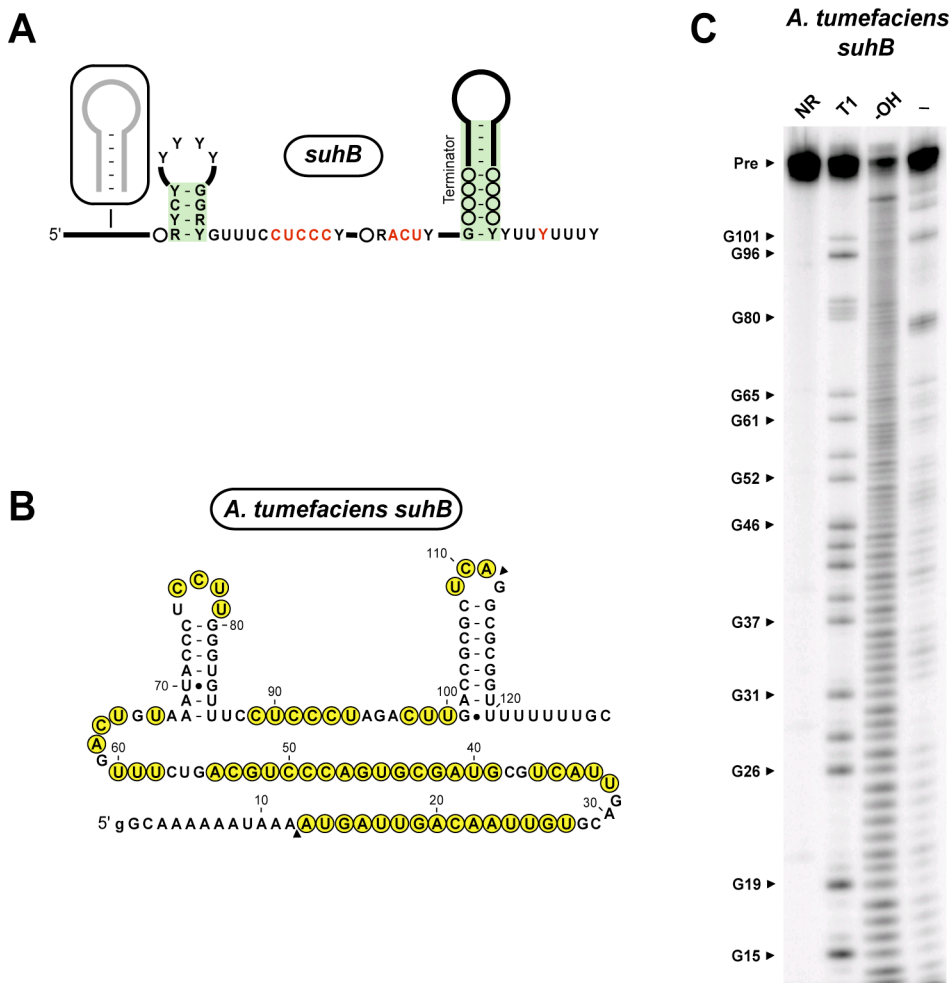


Figure 4.11 Consensus structure and in-line probing of the *suhB* element

Refer to the legend of Figure 4.7 for details. Note that the initial stem appears to be extended by several closing base pairs compared to the consensus structure in this particular sequence.

Abb	Organism	Accession/Start-End	Genes
Alpha Proteobacteria			
Atu	*Agrobacterium tumefaciens* str. C58 chr circular	NC_003304.1/2665461-2665537	COG0670
Bme	*Brucella melitensis* 16M chr I	NC_003317.1/1909433-1909342	COG0670
Brs	*Brucella suis* 1330 chr I	NC_004310.1/93635-93726	COG0670
Mlo	*Mesorhizobium loti* MAFF303099	NC_002678.1/3454185-3454105	COG0670
Mes	*Mesorhizobium* BNC1	NZ_AAED01000002.1/530436-530361	COG0670
Sme	*Sinorhizobium meliloti* 1021	NC_003047.1/3506027-3506125	COG0670

COG	Gene	Description
COG0670	***ybhL***	Integral membrane protein, interacts with FtsH

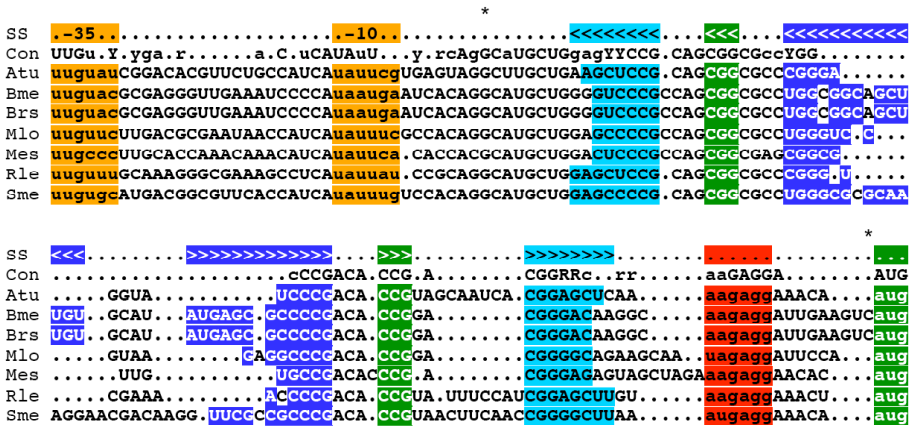


Figure 4.12 Distribution and multiple sequence alignment of the *ybhL* element

Refer to the legend of Figure 4.6 for details.

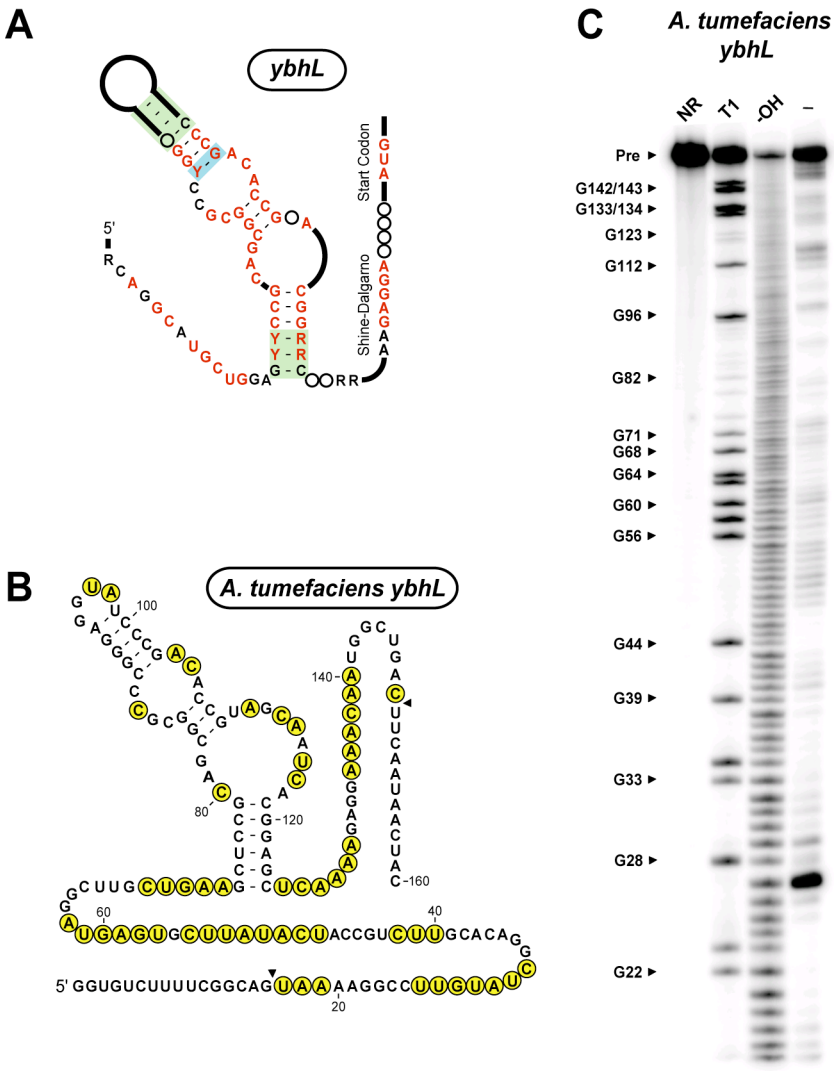


Figure 4.13 Consensus structure and in-line probing of the *ybhL* element

Refer to the legend of Figure 4.7 for details.

5 The distributions, mechanisms, and structures of metabolite-binding riboswitches

5.1 Introduction

Due to the very recent discovery of riboswitches, many aspects of this new regulatory paradigm have not yet been critically and quantitatively surveyed. We present here a large-scale comparative analysis based on updated searches of microbial genomes, environmental sequences, and selected eukaryotic organisms for ten classes of metabolite-binding riboswitch aptamers. With the results we define the overall taxonomic distribution of each riboswitch class and its preferred gene control mechanisms. This expanded data set has also allowed us to systematically reevaluate and refine the secondary structure models of many riboswitch aptamer domains. Notably, we are able to predict new base pairing interactions in several riboswitches with a procedure that estimates the statistical significance of mutual information scores between alignment columns.

5.2 Riboswitch identification

Metabolite-binding riboswitch aptamers are typical of complex RNA sequences that must adopt precise three-dimensional shapes to perform their molecular functions. A conserved scaffold of base-paired helices organizes the overall fold of the riboswitch. The identities of bases within most of these helices vary during evolution, but they change in a correlated manner that preserves base pairing to maintain the same overall secondary structure. On the other hand, the identities of nucleotide bases that directly contact a ligand molecule or stabilize intricate tertiary interactions necessary to assemble a binding pocket are highly conserved during evolution. Additionally,

riboswitches generally tolerate long insertions with variable sequences at certain characteristic sites within their structures. These variable insertions typically form stable RNA stem-loops or hairpins so that they do not interfere with riboswitch folding.

Our goal is to identify all occurrences of a conserved riboswitch aptamer in a sequence database. To do this, we must extrapolate from known sequences to more diverged examples. A variety of search techniques appropriate for finding short regions of base conservation and base-paired stems separated by variable insertions have proven useful for this task when they also consider the genomic context of candidate riboswitch matches. All of the riboswitch classes that have been discovered to date are *cis*-regulatory elements. They are found almost exclusively near protein-coding genes related to the metabolism of their target molecules. Therefore, diverged examples of these riboswitches can be recognized as true positives based on the independent observation that they are located near genes with expected functions even when the search method finds many higher-scoring false positive hits. By incorporating these low-scoring hits into a new structural model and re-searching the sequence database, we can iteratively refine our description of a riboswitch aptamer.

Many riboswitches were first identified as widespread RNA elements on the basis of a highly conserved "box" sequence within their structures. BLAST searches for the B12-box, THI-box, and S-box sequences are effective for discovering many examples of the AdoCbl, TPP, and SAM-I riboswitches, respectively. More detailed search techniques score how well a sequence matches a template of conserved bases and base pairing constraints manually constructed from known examples of the riboswitch aptamer. This sort of generalized pattern matching is implemented by the RNAmotif program [175]. A different motif-based strategy computationally identifies ungapped blocks of conservation, modeled with weight matrices, separated by regions of variable sequence that are characteristic of a given riboswitch [3]. While these methods can be

effective, they generally do not fully exploit the information contained in a multiple sequence alignment of a functional RNA or efficiently identify diverged RNA structures.

Covariance models (CMs) are generalized probabilistic descriptions of RNA structures based on stochastic context-free grammars that offer several advantages over other RNA homology search methods [67]. CMs can be computationally trained on an input sequence alignment without time-consuming manual intervention. They also provide a more complete model of the sequence and structural conservation observed in functional RNA families that incorporates (1) first-order sequence consensus information, (2) second-order covariation, like base-pairing, where the probability of observing a base in one alignment column depends on the identity of the base in another column, (3) insert states that allow variable-length insertions, and (4) deletion states that allow omission of consensus nucleotides. This additional complexity comes at a computational cost, but several HMM-based filtering techniques have recently been developed that make CM searches of large databases practical [311, 312, 313]. We have previously used CMs to find divergent homologs of *E. coli* 6S RNA [20] and define a variety of regulatory RNA motifs in α-proteobacteria [51]. The Rfam database [100] maintains hundreds of covariance models for identifying a wide variety of functional RNAs, including riboswitches.

In the present study, we used covariance models to systematically search for ten classes of metabolite-binding riboswitches in microbial genomes, environmental sequences, and selected eukaryotic organisms. The riboswitch sequence alignments that we used to train CMs were derived from a variety of published and unpublished sources (Table 5.1). This list includes a previously unreported riboswitch that binds the metabolite $preQ_1$ (A. Roth, E. Regulski, J.E. Barrick, and R.R. Breaker, unpublished data) to inhibit expression of an operon encoding genes necessary for queuosine

Riboswitch Aptamer	Rfam Accession	Seed Alignment	Other Alignments
Thiamine Pyrophosphate (TPP)	RF00059	[268, •]	[234]
Adenosylcobalamin (AdoCbl)	RF00174	[201]	[299]
Lysine	RF00168	[269]	[236]
Glycine	RF00504	[180]	
S-Adenosylmethionine Class 1 (SAM-I)	RF00162	[326]	[103, 237]
Flavin Mononucleotide (FMN)	RF00050	[298]	
Guanine and Adenine (Purine)	RF00167	[178]	
Glucosamine-6-Phosphate (GlcN6P)	RF00234	[19]	
7-Aminoethyl 7-Deazaguanine ($preQ_1$)	RF00522	[•]	
S-Adenosylmethionine Class 2 (SAM-II)	RF00521	[51]	

Table 5.1 Sources of riboswitch aptamer sequence alignments

Riboswitch aptamers are named by the metabolite that they sense with standard abbreviations in parentheses. Rfam database numbers are provided for each riboswitch along with references for the "seed" alignments that we used to train covariance models for database searches in this study, and other published alignments. [•] = J.E. Barrick, unpublished data.

biosynthesis in *B. subtilis* [227]. The $preQ_1$ riboswitch aptamer was first described as the *ykvJ* RNA motif in recent a survey of structured regulatory RNA sequences of unknown function [19].

5.3 Riboswitch distributions

The phylogenetic distributions of ten riboswitch classes are depicted in Figure 5.1. The TPP riboswitch is the only metabolite-binding RNA known to occur outside of bacteria. It is found in euryarchaeal, fungal, and plant genomes. Several fungal genomes have multiple copies of the TPP riboswitch (as many as three) regulating different genes. AdoCbl is the most widespread riboswitch in bacteria, but TPP, FMN, and SAM-I are also common in many groups. Riboswitches that sense glycine and lysine have more fragmented distributions: they are widespread in certain groups, but missing from substantial ranges of the bacterial spectrum. Finally, the GlcN6P, purine, $preQ_1$, and SAM-II riboswitches appear to be present in only a few groups of bacteria. Interestingly, the SAM-I and SAM-II aptamer distributions overlap slightly. We have found examples of both SAM-sensing riboswitch classes in α-Proteobacteria, γ-Proteobacteria, and Bacteroidetes, but no bacterium for which a complete genome sequence is available seems to employ both riboswitches.

It seems likely that many of the relatively isolated examples where riboswitches occur only sparsely in certain clades (e.g. SAM-I, SAM-II, purine, and $preQ_1$ in γ-Proteobacteria) may be examples of horizontal DNA transfer. There is some evidence that this process has been important for the dispersal of riboswitches into new bacterial genomes. Entire transcriptional units containing AdoCbl riboswitches and their associated biosynthetic operons have apparently been copied recently from *Bacillus*/*Clostridium* species to enterobacteria [299]. In phylogenetic trees inferred for

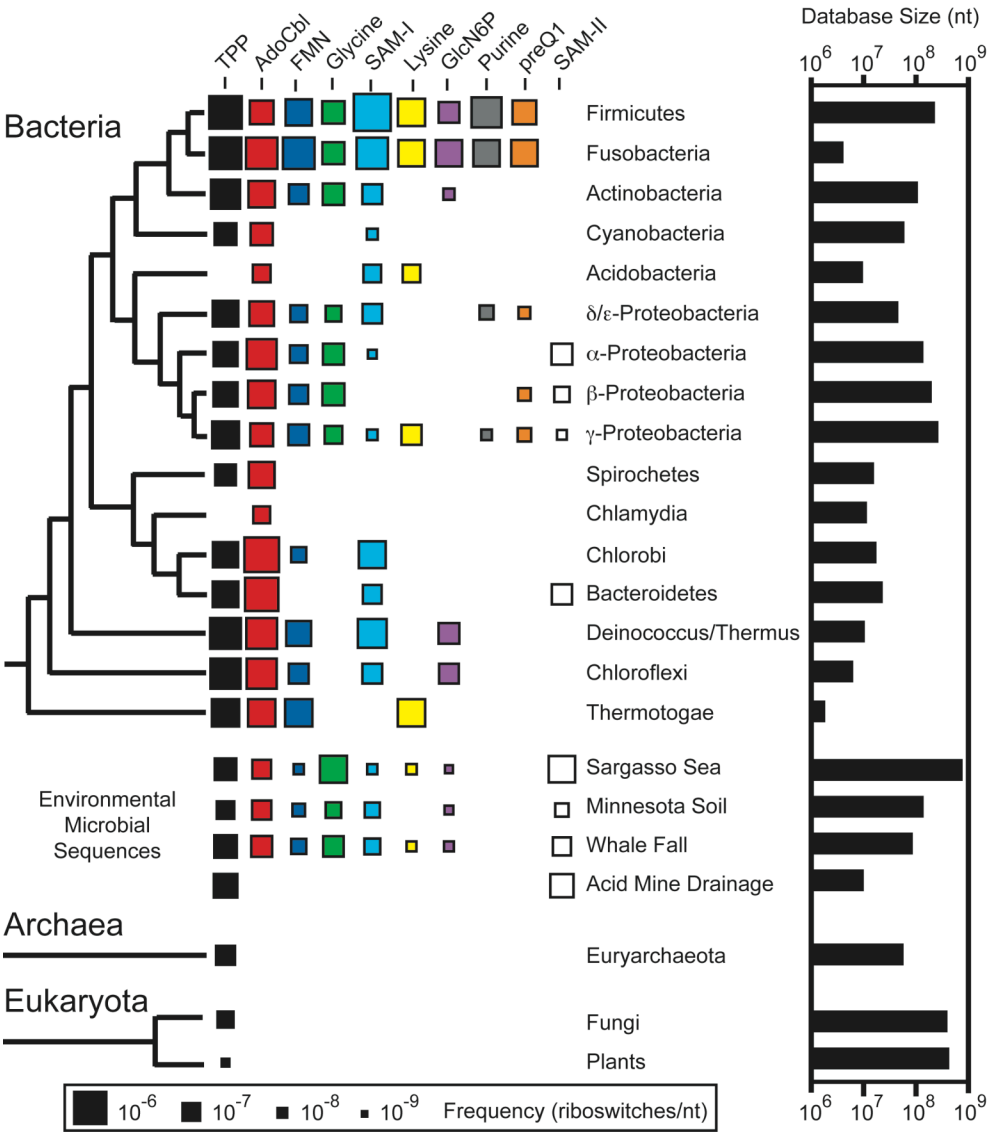


Figure 5.1 Riboswitch phylogenetic distributions

The dimensions of each square are proportional to the frequency with which a given riboswitch occurs in the corresponding taxonomic group. A phylogenetic tree with the standard accepted branching order of each group of organisms is shown on the left. For bacteria, this tree is adapted from [176] with the addition of Fusobacteria [189]. On the right is a graph depicting the total number of nucleotides from each taxonomic division in the sequence databases that were searched for riboswitches.

both AdoCbl riboswitch aptamers and the downstream proteins, these specific enterobacterial sequences are placed within branches dominated by Firmicute sequences. This disagreement with the accepted evolutionary history of these bacterial species, in contrast to the expected placement of other AdoCbl riboswitches and biosynthetic proteins from enterobacteria with γ-proteobacterial sequences, indicates that these transcriptional units were recently horizontally acquired. In contrast, this evidence of recent and selective horizontal transfer was not observed in phylogenetic trees of lysine riboswitch aptamers, despite their disjoint distribution across different taxonomic groups [236]. It may be that the genes that the glycine and lysine riboswitches regulate are not passed as readily between different bacterial groups because they are involved in central metabolism rather than dispensable cofactor biosynthesis.

Gram-positive low G+C bacteria (also known as *Bacillus*/*Clostridium* species or Firmicutes) make the most use of riboswitch regulation. Every riboswitch except SAM-II is widespread in this clade, and most aptamer classes occur multiple times per genome. A detailed accounting of genes controlled by riboswitches in *B. subtilis* has previously been worked out [178]. Now we know that 29 riboswitches (5 TPP, 1 AdoCbl, 2 FMN, 1 glycine, 11 SAM-I, 2 lysine, 1 GlcN6P, 4 Guanine, 1 Adenine, and 1 $preQ_1$) control some 73 genes in this soil bacterium.

As a whole, γ-proteobacteria employ a variety of riboswitch classes that is comparable to the diversity found in low G+C Gram-positive species. However, individual proteobacterial genomes invariably have fewer riboswitch classes and actual instances of these riboswitches than any *Bacillus*/*Clostridium* species. For example, *E. coli* has 6 riboswitches (3 TPP, 1 AdoCbl, 1 FMN, and 1 lysine) that regulate a total of 16 genes. It is interesting that the riboswitch complements of high G+C Gram-positive bacteria (Actinobacteria) seem to more closely resemble those of Proteobacteria than low G+C Gram-positive species.

Many "deeply-branched" bacteria such as *Deinococcus*/*Thermus* and *Thermotoga* species also appear to utilize a wide variety of riboswitches. However, no riboswitch sequences have been identified in *Aquifex* species, and riboswitches also seem to occur only rarely in Cyanobacteria, Spirochetes, and *Chlamydia* species. The sequence database sizes for many of these bacterial groups correspond to only a handful of complete genomes, therefore it seems likely that the currently observed frequencies of riboswitches in these groups will need to be revised as more genomic sequences become available.

As expected, we find all but the rarest riboswitches in genomic sequences from shotgun cloning projects that target environments supporting diverse phylogenetic distributions of bacteria. These sources of additional sequences have been helpful in some cases for rigorously defining our consensus structure models and adding statistical merit to our mutual information calculations (see below). Glycine and SAM-II riboswitches are unusually common in Sargasso Sea sequences. For SAM-II, at least, this probably reflects the abundance of α-Proteobacteria in this environment [295].

5.4 Riboswitch mechanisms

Aside from the *glmS* ribozyme, which uses a metabolite-dependent cleavage event at its 5' end to repress gene expression by an unknown mechanism [325], all known riboswitch classes utilize ligand-induced structural changes in their conserved aptamer cores to trigger changes in the conformations of nearby mRNA expression platform (EP) sequences that affect protein production. Biochemical evidence suggests that all riboswitches almost completely envelop their target metabolites, as seen in the x-ray crystal structures of purine riboswitches [22, 208, 255]. In a sense, it may be more accurate to say that the metabolite molecule nucleates a compact, folded aptamer state containing the paired elements and tertiary contacts predicted in an aptamer's

consensus structure rather than that it merely binds. When metabolite is not present, parts of the aptamer structure are not enforced, freeing some nucleotides to make alternate interactions with nearby sequences. Although EP structure changes are often studied *in vitro* in the context of complete leader RNA constructs at thermodynamic equilibrium (i.e. by in-line probing), the kinetics of transcription and ligand binding have recently been shown to dominate the co-transcriptional decision to follow one of these two distinct folding pathways to a different architecture in FMN and adenine riboswitches [89, 316, 317].

For most riboswitch aptamers, nucleotides that overlap the 5' or 3' strands of the P1 "switching" helix, which is enforced in the ligand-bound conformation, frequently base pair to downstream EP sequences to form an alternate helix in the absence of ligand [e.g. 269, 322]. Cobalamin riboswitches are an exception. They seem to predominantly alter the structures of their EPs with a ligand-induced pseudoknot pairing between a specific C-rich loop and sequences outside the aptamer core [201, 202, 299]. Microbial riboswitches almost always occur in 5' untranslated mRNA leader sequences, and the hand-off between two alternate EP conformations typically changes the efficiency of premature transcription termination upstream of protein coding sequences or ribosome binding to the Shine-Dalgarno sequence of the first gene's start codon.

Riboswitches with EPs that regulate transcription termination leverage the ability of stable GC-rich stems followed by polyuridine tails to cause RNA polymerase to terminate without the involvement of any additional protein factors [107, 329]. Glycine and adenine riboswitches with ON genetic logic (that activate gene expression when bound to their target metabolites) bury portions of the sequence required to form the stem of an intrinsic terminator in pairing interactions within their conserved aptamer structures [179, 180]. Most riboswitches operate as OFF switches that add an extra folding element to reverse this genetic logic. Instead of directly changing the terminator's

structure, aptamer folding in the presence of ligand disrupts a downstream antiterminator hairpin that normally sequesters sequences required to form the terminator stem under low metabolite concentrations. The same OFF mechanism (also known as transcription attenuation) is harnessed by a variety of other protein- and ribosome-mediated regulatory processes in many bacteria [188].

Riboswitches that regulate translation initiation in the presence of ligand molecules use a similar trade-off of paired stems to conditionally hide the ribosome binding site (RBS) of the downstream gene. In some cases, alternate mutually exclusive base-paired conformations analogous to transcription terminators and antiterminators can be predicted from sequence alone. Usually, an anti-RBS-sequestor stem-loop (the ON state) is disrupted by an interaction with aptamer sequences in the ligand-bound conformation such that a RBS-sequestor helix can form (the OFF state). In some cases translation is attenuated more directly because the RBS is buried within the 3' side of the P1 switching helix when ligand binds [234, 300]. Some sequences appear to combine multiple modes of regulation. They control the formation of transcription terminators that are positioned near enough the start codon of the downstream gene that its RBS is sequestered in the 3' side of the terminator hairpin when it forms. Thus ligand binding induces transcriptional attenutation and, if the mRNA does not terminate, translational attenuation [234].

Metabolite-dependent inhibition of ribosome binding has been proven *in vitro* for the *E coli* AdoCbl riboswitch upstream of the *btuB* gene [209]. However, most of the current *in vivo* evidence about the mechanisms of riboswitches that are believed to operate by RBS hiding is in the form of reporter assays using translational reporter gene fusions [e.g. 202] or computational predictions of RBS sequestor and anti-sequestor hairpins [e.g. 234]. These observations do not rule out that some or all of the gene expression changes for specific aptamers could be caused by other post-transcriptional

mechanisms, such as metabolite-dependent RNA processing by general or targeted ribonucleases. In this vein, AdoCbl riboswitches from *E. coli* and *B. subtilis* have been shown to be very weak non-canonical substrates for RNase P [6], even though the physiological importance of this specific case of cleavage for gene regulation is questionable. The *thiC* riboswitch in *E. coli* exerts regulation at both the transcriptional and translational levels, but its EP does not contain a canonical intrinsic transcription terminator hairpin [322]. This transcriptional regulation could be caused by changes in the accessibility of a binding site for nucleases or the Rho termination protein in the untranslated leader [232]. Finally, riboswitches may regulate splicing and polyadenylation in eukaryotes [268].

In order to analyze the distribution and variety of regulatory mechanisms in our collection of riboswitch sequences from different microbial groups, we developed a computational decision tree for classifying expression platforms into four categories (Figure 5.2). This scheme discriminates between riboswitches with mechanisms involving (1) transcription attenuation, (2) dual transcription and translation attenuation (3) translation attenuation, and (4) direct translation attenuation. These categories were inspired by careful bioinformatics studies of TPP [234], AdoCbl [299], FMN [298], and Lysine [236] riboswitch expression platforms. Note, that we do not predict the actual structures of RBS sequestering stem-loops or explicitly predict that expression platforms use downstream RBS sequesters. Rather, we assume that RBS sequesters are the most likely mechanism for EPs not classified into the other three categories by the decision tree. This category could also contain sequences that employ some of the other mechanisms described above. Overall, our computational assignments have an accuracy of 88% when compared to expert predictions for the phylogenetically diverse TPP data set [234].

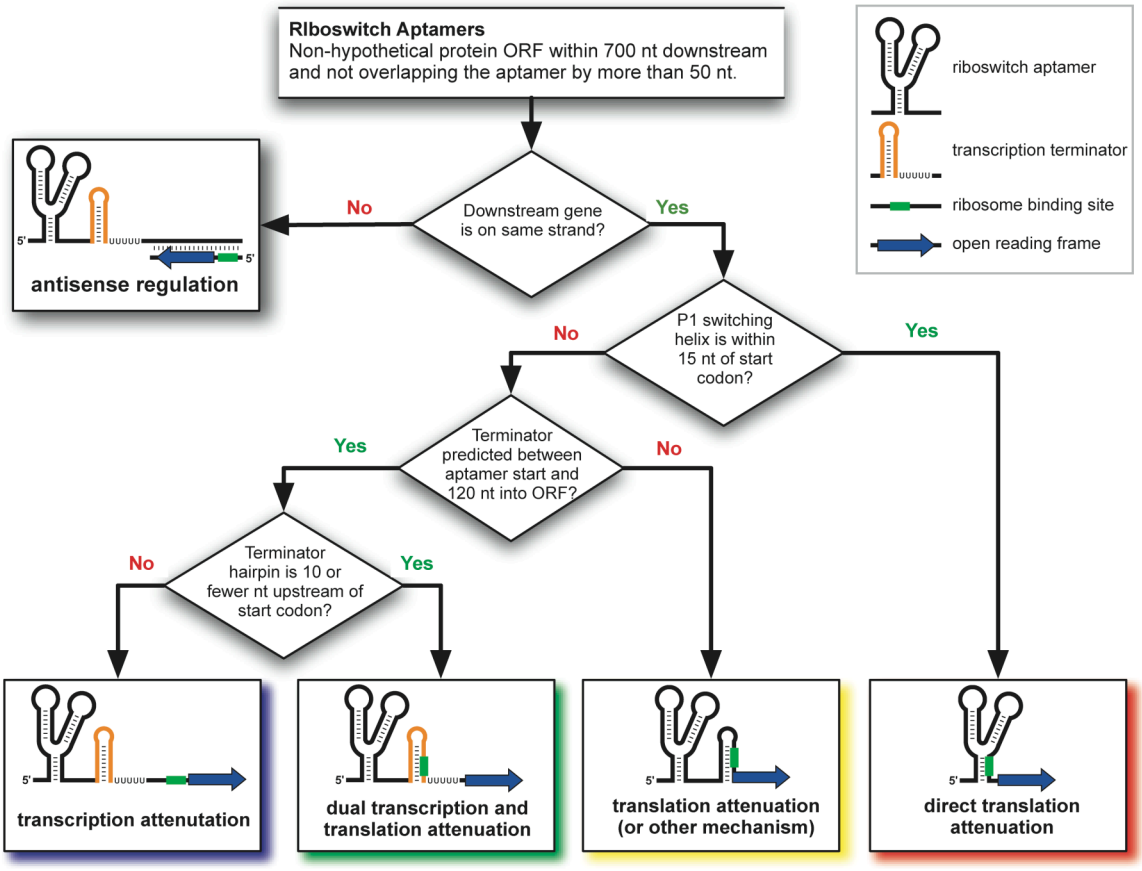


Figure 5.2 Riboswitch mechanism prediction scheme

The decision tree used to classify riboswitch mechanisms reported in Figure 5.3 is shown. All riboswitches pictured are OFF switches in their ligand-bound state where the P1 switching helix has formed. Refer to the text and methods for more details.

The most striking conclusion that can be drawn from this data (Figure 5.3) is that transcription attenuation dominates riboswitch mechanisms observed in low G+C Gram-positive bacteria, and translation attenuation appears to be the most common mechanism in most other bacterial groups. These trends have already been widely reported for many riboswitch classes, but, to our knowledge, an evolutionary or physiological rationale for this systematic difference has not yet been offered. Although there is some disagreement as to whether certain groups of bacteria utilize "standard" transcriptional terminators consisting of a stable hairpin followed by a U-tail [72, 290, 306], *E. coli* and other γ-Proteobacteria are known to use hundreds of intrinsic transcription terminators to define the ends of operons. Furthermore, leader peptide translation systems regulate several amino acid biosynthetic operons in *E. coli* via transcription attenuation, e.g. the *trp*, phe, *his*, *thr*, and *leu* operons [142, 151]. So it is unclear why transcription termination mechanisms are not also harnessed by riboswitches in γ-proteobacteria.

The phylogenetic distribution of predicted mechanisms for SAM-II riboswitches is unusual. It is the only riboswitch aptamer that appears to be most often associated with regulatory transcription terminators in α- and β-proteobacteria. This conserved RNA structure clearly functions as a very specific SAM aptamer *in vitro* [51] and occurs upstream of genes related to SAM and methionine biosynthesis. However, it should be noted that *in vivo* genetic control of transcription attenuation has not yet been demonstrated for this proposed SAM-II riboswitch. Transcription attenuation mechanisms may also be overrepresented in Fusobacteria, δ/ε-Proteobacteria, Thermatogae, and Chloroflexi species, although the small sample sizes in these groups makes these conclusions less certain.

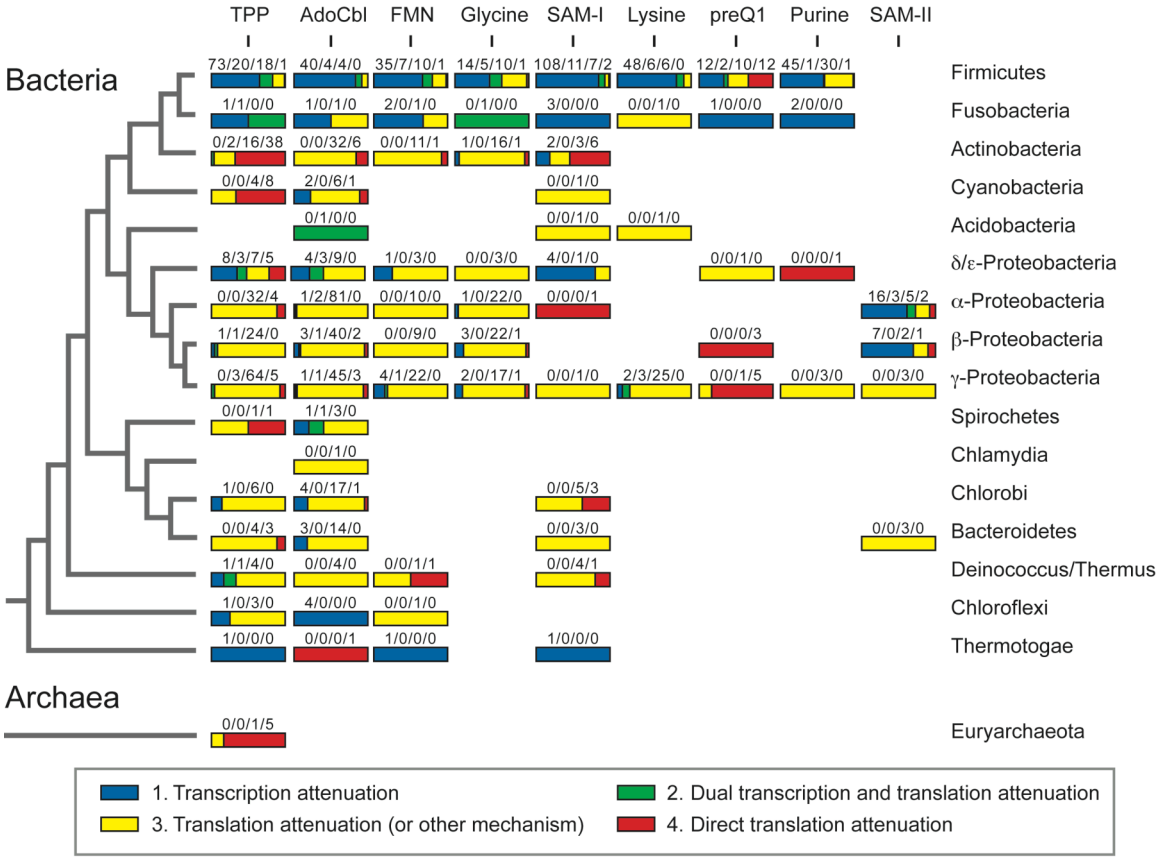


Figure 5.3 Riboswitch mechanisms

The mechanisms that riboswitches from different taxonomic groups use to regulate gene expression were computationally predicted on the basis of expression platform features (Figure 5.2). The fractions of riboswitch expression platforms in each of these four categories are displayed visually as shaded bars with the actual numbers observed shown above in the order given in the legend.

Expression platform mechanisms that rely on directly sequestering the RBS are most common for the TPP, preQ$_1$, and SAM-I riboswitches. In the first two cases, purine-rich conserved regions near the 3' ends of the riboswitch substitute for Shine-Dalgarno sequences. In the SAM-I riboswitch the RBS is incorporated into the 3' side of the P1 stem. Many other riboswitch classes also have purine-rich conserved regions near their 3' ends with consensus sequences close to ribosome binding sites. It is not obvious why direct regulation of translation attenuation should be rarer in these cases. Perhaps the RBS-like sequences in these aptamers are already inaccessible in the mRNA structures that form in the absence of ligand. Riboswitch regulation by direct translation attenuation appears to be most frequent in Actinobacteria and Cyanobacteria, except for the preQ$_1$ riboswitch where this mechanism is unusually prevalent, even in Firmicutes and Proteobacteria.

5.5 Evaluating structure models

Constructing an RNA secondary structure model requires proposing new base-paired stems and adjusting a sequence alignment to determine whether their existence is consistent across all representatives. This recursive refinement process has been used to create detailed comparative models of many functional RNA structures that have proven to be very accurate when compared to later molecular resolution three-dimensional structures. However, the presence of stretches of unvarying nucleotides within an RNA structure, the tolerance of real helical stems to some non-canonical base pairs, and the non-negligible frequency of sequencing errors in biological databases can introduce enough ambiguity that multiple structures may seem to agree with a sequence alignment and incorrect base-paired elements may be proposed. This problem is compounded if the multiple sequence alignment is incomplete and does not yet capture all of the variation that truly exists in an RNA family at each nucleotide position.

For example, during the initial characterization of the *gcvT* motif from *B. subtilis* (now known to be the glycine riboswitch), we proposed two mutually exclusive structures that were consistent with its sequence alignment and in-line probing results [19]. At the time, we favored a pairing of conserved sequences that proved to be incorrect. Covariance models trained on an alignment with the alternate pairing (now P3 and P3a) found additional glycine riboswitch sequences, and the expanded alignment clearly shows that there is more variability and covariation in these pairs [180].

This episode and inconsistencies in the structural models that have been proposed for some riboswitch aptamers motivated us to adopt a method for evaluating the statistical support for proposed base interactions. We chose to use mutual information (MI) scores [60] to mathematically formalize the interdependence between sequence alignment columns indicative of base interactions. MI is a normalized version of covariance that represents the amount of information (in bits) that you gain from knowing the identity of a base at one position in a sequence about what base occurs at a second position. The use of sequence covariation to predict secondary structures and tertiary interactions from RNA sequence alignments has a long history, and the nuances of calculating and interpreting MI scores have been comprehensively covered elsewhere [110]. Recently, an RNA structure prediction program has even been created that is entirely based on these principles [82].

Fundamentally, columns of interacting bases must be correctly aligned and there must be variation in each column (i.e. it cannot be completely conserved) in order to detect mutual information. Even when these preconditions are met, there are two major difficulties with directly comparing MI scores to determine which columns in a sequence alignment are truly covarying. First, sequence conservation derived from the shared evolutionary histories of sequence subsets in an alignment may result in a high residual background MI score between many columns whether or not they are functionally linked.

Second, alignments with fewer sequences will have more column pairs with elevated mutual information values simply by chance. Simulations addressing the expected magnitudes of these two sources of error in different data sets have been explored recently in the context of protein sequence alignments [92].

In order to gauge whether a MI score truly supports a proposed base interaction in an RNA alignment, we have developed a procedure for empirically estimating the statistical significance of MI scores (Figure 5.4). First, we eliminate redundant sequences and consensus gapped columns from an alignment to simplify calculations. Then, we infer a phylogenetic tree and estimate per-column mutation rates for the observed RNA sequence alignment according to a model that assumes independent evolution of each column. We next generate resampled sequence alignments with the same topology, branch lengths, and evolutionary rates in order to simulate a background distribution of MI scores between each pair of columns. These background MI score distributions represent the null hypothesis that there are no functional interactions between the columns. They are meant to implicitly correct for the evolutionary history and sample size of a real sequence alignment. Finally, We calculate a p-value for each pair of columns representing the probability that a *test* alignment has a higher MI score between these two columns than the *real* alignment.

5.6 Riboswitch structures

We have updated the consensus secondary structure models of the ten riboswitches classes (Figure 5.5) to reflect our expanded sequence alignments. The purine riboswitch consensus has been drawn in accordance with its molecular structure [22, 208, 255] and the double-pseudoknot of the GlcN6P-dependent (*glmS*) ribozyme has been drawn as observed in the molecular structure of the uncleaved RNA (D.J. Klein, A.R. Ferré-

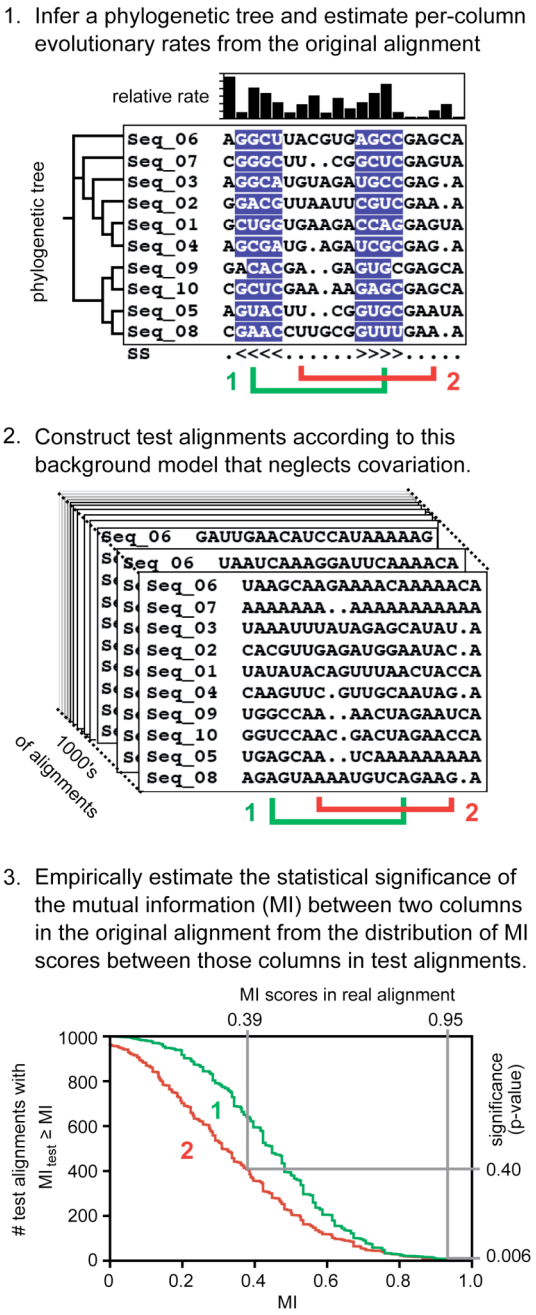


Figure 5.4 Procedure for estimating MI significance between alignment columns

See the test and methods for a complete description of the procedure used to estimate the statistical significance of mutual information (MI) scores between columns in a multiple sequence alignment in order to evaluate riboswitch secondary structures and predict new base interactions.

Figure 5.5 Structures of metabolite-binding riboswitch aptamers

The consensus secondary structure models based on expanded riboswitch sequence alignments are depicted according to the symbols defined in the legend. Each structure is further annotated with RNA structure motifs and the statistical significances (p-values) of the mutual information scores between base-paired alignment columns. New predictions of interacting bases from the MI analysis are numbered and starred. More detailed descriptions of the predicted base pairs are provided in Table 5.2.

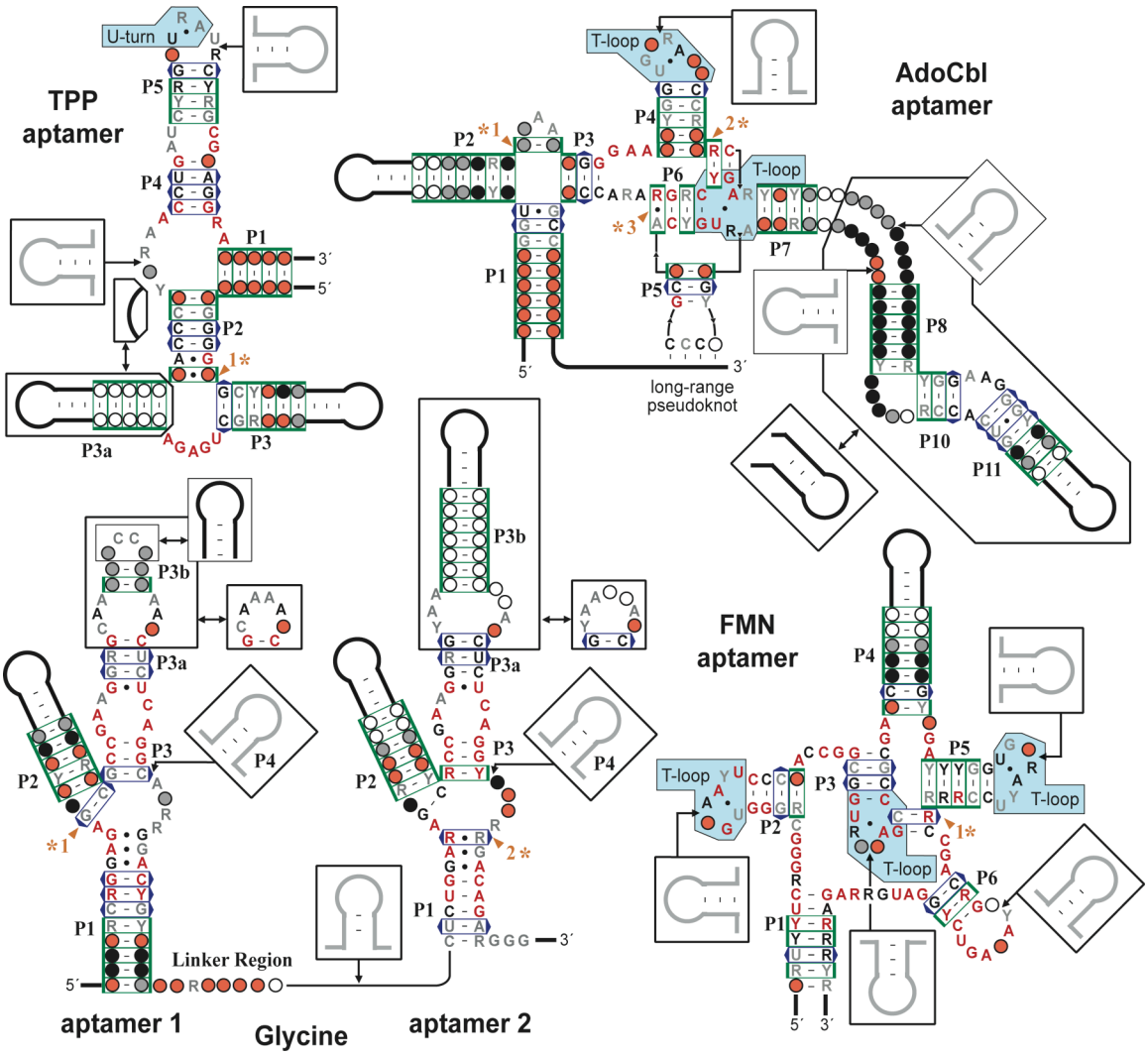


Figure 5.5 Structures of metabolite-binding riboswitch aptamers (page 1 of 2)

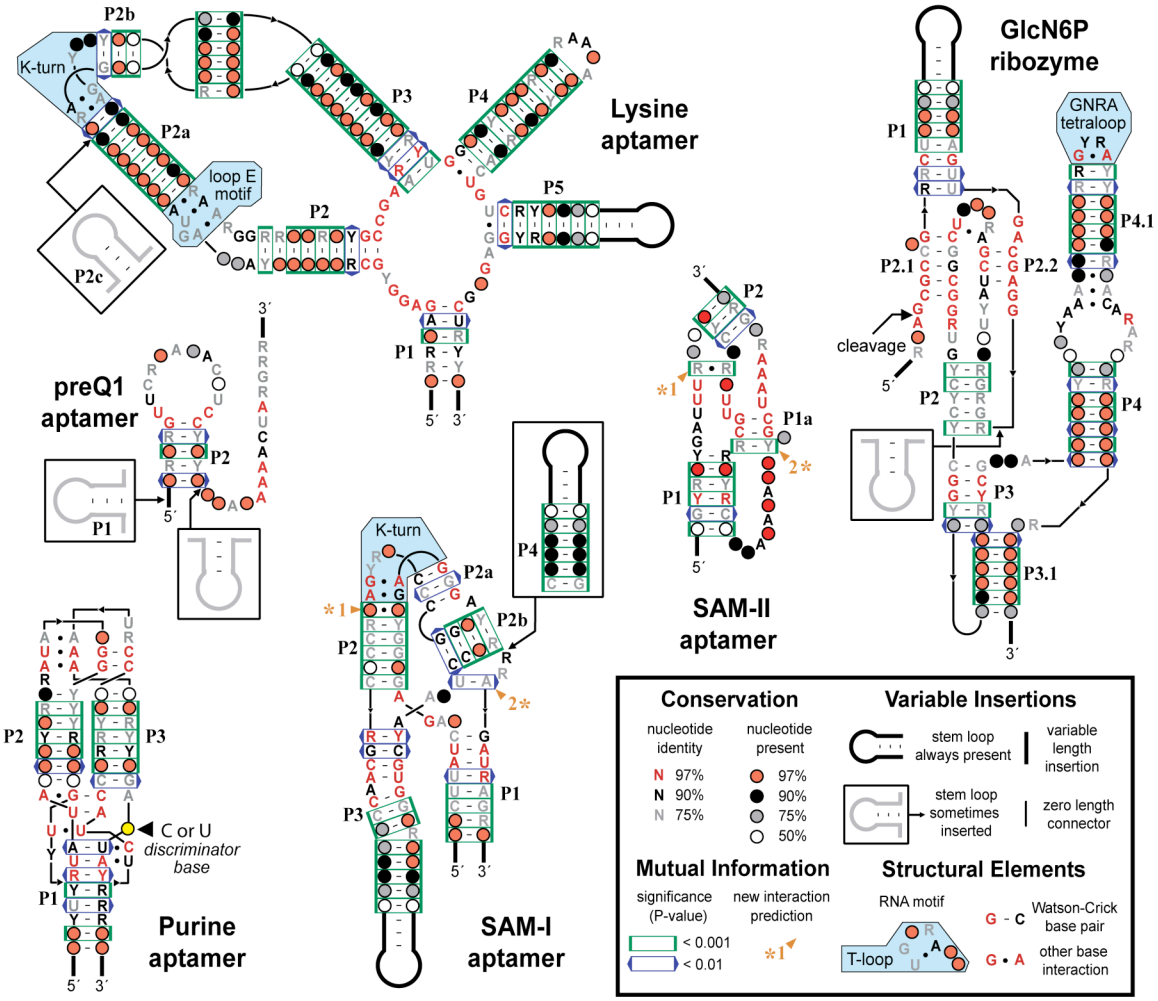


Figure 5.5 Structures of metabolite-binding riboswitch aptamers (page 2 of 2)

Aptamer	#	p-value	Observed Pairs	Compatible Interactions: Base Edges		Strands	Notes
TPP	**1**	< 0.0001	**AG** 42.5% * **UA** 31.9% * **UG** 8.7% **GG** 6.6% **AA** 5.1%	*cis* H/WC *trans* H/H	I_3 I_2	↑↑ ↑↑	Only in sequences with a P3a helix. Strands ↑↓ if one base has an unusual *syn* glycosidic bond conformation.
AdoCbl	**1**	< 0.0001	**UA** 38.9% * **AU** 11.5% * **CG** 10.1% * **UU** 7.9%	*cis* WC/WC	I_1	↑↓	Isolated pair closing an A-rich loop.
AdoCbl	**2**	< 0.0001	**GC** 53.9% * **AU** 38.6% * **GU** 4.5%(*)	*cis* WC/WC	I_1	↑↓	T-loop associated tertiary contact with adjacent C–G pair. R/Y strand bias.
AdoCbl	**3**	0.001	**AG** 70.6% * **GA** 22.8% * **AA** 3.7%	*cis* WC/WC *cis* H/H	I_3 I_2	↑↓ ↑↓	Noncanonical pair at the end of the P6 helix.
FMN	**1**	0.007	**CG** 71.1% * **UA** 24.4% * **UG** 3.1%(*)	*cis* WC/WC	I_1	↑↓	T-loop associated tertiary contact with adjacent G–C pair. Y/R strand bias.
Glycine	**1**	0.005	**GC** 83.5% * **AU** 7.7% *	*cis* WC/WC	I_1	↑↓	Pair may extend P2 helix after bulged nt. R/Y strand bias.
Glycine	**2**	0.002	**GG** 49.9% * **AA** 23.1% * **AU** 8.7% **AC** 5.9% * **AG** 5.4%	*cis* bifurcated	I_1	↑↓	Possible noncanonical pair in internal asymmetric bulge. GA and AG pairs could form adjacent to P1
SAM-I	**1**	0.0001	**GA** 55.5% * **CC** 12.2% * **GU** 8.1% **AA** 7.7% * **UA** 5.5% *	*trans* SE/H	I_1	↑↓	Continues P2 helix pairing adjacent to the K-turn.
SAM-I	**2**	0.001	**UA** 73.8% * **GC** 4.6% * **.C** 4.4% **UC** 4.2%	*cis* WC/WC	I_1	↑↓	Isolated pair bridging the P1 helix and the P2b pseudoknot.
SAM-II	**1**	0.001	**GG** 39.2% * **AA** 31.1% * **.G** 13.4% **GU** 11.7% *	*cis* bifurcated	I_1	↑↓	Isolated pair between the P2 pseudoknot and conserved loop sequences.
SAM-II	**2**	< 0.0001	**GC** 50.0% * **AU** 17.0% * **UA** 10.0% * **GU** 4.4%(*)	*cis* WC/WC	I_1	↑↓	May be part of a new helix (P1a) with two conserved adjacent pairs

Table 5.2 New base pair interaction predictions

For each numbered and starred prediction in Figure 5.5 the statistical significance (p-value) of the mutual information between the two alignment columns is shown, followed by the relative frequencies with which specific combinations of bases are observed in those columns. Base pair geometries and isostericity groups compatible with the starred pairs are summarized after [161]. These descriptions include the relative orientations of the glycosidic bonds across the pair (*cis* or *trans*), the edges of each base that interact (WC, Watson-Crick; H, Hoogsteen; SE, sugar edge; bifurcated, intermediate between two edges), and the relative backbone strand geometry (parallel or anti-parallel) assuming both glycosidic bonds are in default *anti* conformations.

D'Amaré, personal communication). We have revised other riboswitch structures to be consistent with new predictions of base interactions and structural motifs explained below. In all cases, we have maintained previous numbering schemes for the paired helical elements (designated P1, P2, P3, etc. beginning at the 5' end of each the aptamer), even when these stems do not occur in a majority of the sequences in the updated alignment. Newly discovered "variable stems" that do not have conserved sequences and are missing in most sequences have not been assigned numbers.

The results of our mutual information analysis are shown superimposed on the consensus riboswitch structures. Most base-paired helices in the ten riboswitch structures are supported by at least one contiguous base pair with a highly significant MI ($p < 0.001$), and almost all contain a base pair with at least a marginal MI significance ($p < 0.01$). However, we do not detect significant MI within the P2.1 and P2.2 stems observed in the crystal structure of the uncleaved GlcN6P-dependent ribozyme. The MI analysis also does not support an alternate P1.1 pseudoknot (not shown) proposed on the basis of biochemical experiments where the register of the regions involved in making the P2.1 pairing is slightly shifted [136, 184, 259]. Most of the predicted base pairs in the P2.1 and P2.2 helices are between highly conserved bases that may not vary enough to produce significant covariation with their pairing partners.

MI significance scores do resolve a conflict between two proposed pairing models that have been proposed for the highly conserved B12-box portion of the AdoCbl riboswitch. One model posits that a "facultative stem loop" forms by pairing nucleotides within the B12-box [299]. The other model proposes long range pairings between portions of the B12-box sequence and nucleotides more distant in the riboswitch's linear sequence [201]. We find only a single, marginally significant MI score that supports the formation of the "facultative stem loop", even though this region is correctly aligned to optimally discover such interactions (Figure 5.6). The MI analysis strongly supports

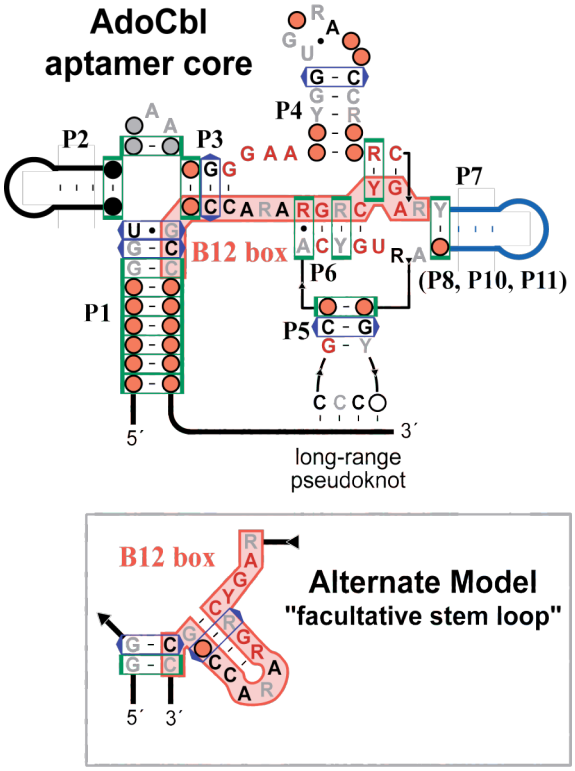


Figure 5.6 Comparison of alternate B12 box structure models

In addition to the structural model of the AdoCbl riboswitch aptamer presented here [201], an alternate model that folds the highly-conserved B12 box sequence (highlighted in red) into a "facultative stem loop" has been proposed [299]. The core of the AdoCbl riboswitch aptamer is shown with abbreviated peripheral helices and without the optional P8, P10, P11 domain for comparison with the alternate secondary structure model. The top model is supported by more base pairs with significant MI scores and the observation that an internal T-loop can form between P6 and P7 with pairing from the B12 box. Each diagram uses the symbols described in the legend of Figure 5.5.

several base pairs in the alternate proposed structure wherein portions of the conserved B12-box form the 3' sides of the short P3 and P6 helical stems.

5.7 Structural motifs in riboswitches

Many riboswitches contain common RNA structure motifs that are recognizable from their consensus sequences. A GNRA tetraloop [122] that favors a pyrimidine at its second position caps P4a of the GlcN6P ribozyme. A K-turn [147] is conserved between P2 and P2a in the SAM-I riboswitch aptamer [324]. The asymmetric bulge between helices P2a and P2b of the lysine riboswitch also adopts a K-turn in most sequences [105], but a number of riboswitch variants appear to dispense with this motif. In contrast, the sarcin-ricin motif [160] (a specific type of loop E motif) formed by the asymmetric bulge between the P2 and P2a helices of the lysine riboswitch is highly conserved.

We also find examples of other RNA structure motifs that have not previously been reported in riboswitches. The three terminal loops capping P2, P3, and P5 in the FMN riboswitch and the P4 loop and P6–P7 bulge in the AdoCbl riboswitch share a remarkable number of characteristics. They have (1) two closing G–C base pairs with a strand bias, (2) a possible U–A pair separated from the helical stem by two bulged nucleotides on the 3' side, and (3) a terminal GNR triloop sequence sometimes interrupted at a specific position by an intervening base-paired helix. These conserved features strongly suggest that they are examples of T-loop motifs (named for the T-loop of tRNA) where the U–A forms a key *trans* Watson-Crick/Hoogsteen pair [200]. Finally, sequence conservation in the UNR loop that closes the P5 stem in the TPP aptamer suggests that it forms a conserved U-turn [108]. As expected, there is a sharp reversal of backbone direction following this uridine, subsequent bases stack on the 3' side of the loop, and the uracil base can hydrogen bond with the phosphate group 3' of the third U-

turn nucleotide in the x-ray crystal structures of the *E. coli* [254] and *Arabidopsis thaliana* [281] *thiC* riboswitches.

5.8 Predictions of new base-base interactions

In addition to supporting almost all of the helical elements in riboswitch structures, our MI analysis predicts eleven additional base-pairing interactions (Figure 5.5 and Table 5.2, above). We caution that a significant MI score between two alignment columns represents a statistical correlation and does not necessarily imply direct hydrogen bonding between two nucleobases. We have screened out other MI pairs between adjacent nucleotides that probably represent favored base stacking patterns in helices and ignored column pairs with many gaps where MI scores can be dominated by correlations between the presence and absence of nucleotides rather than their base identities. It is also possible to observe high mutual information between two bases that do not interact if several separate structure motifs with their own specific sequence requirements can substitute for each other at a location in a functional RNA, as is seen for GNRA, UNCG, and CUUG tetraloops in 16S rRNA [110].

Furthermore, our estimates of MI significance rely on a phylogenetic tree reconstruction method that may not adequately model the evolution of RNA sequences, especially for the shorter riboswitch alignments. Even assuming our estimated p-values are completely accurate, there are 4950 possible combinations of columns in an alignment with 100 columns, and that would imply that 5 pairs with a MI significance of ≤ 0.01 should be observed by chance. Some columns that are *known* to be base paired do not have MI scores this significant. In light of this noisy background we have endeavored to concentrate on predictions that seem to have structural relevance. We cannot rule out alternate explanations from sequence alone, but believe that the new predictions below represent true base-pairing interactions.

The identities of interacting bases in a functional RNA are constrained during evolution. They can only mutate to other base pairs that preserve the local geometry of the sugar-phosphate backbone and any hydrogen bond acceptors or donors that are important for maintaining its structure or function. Generally, only one of the three planar edges of a nucleobase participates in any given interaction: the Watson-Crick face (WC), Hoogsteen face (H), or sugar edge (SE). A systematic study of RNA structures has produced isostericity matrices that tabulate which of the possible 16 base pairs should be interchangeable (in terms of C1'–C1' distances) when two nucleobases are interacting between different combinations of these three base edges and when the glycosidic bonds on both sides of the pair are *cis* or *trans* with respect to each other [161]. The pairs of bases conserved at some of the new correlated positions in riboswitches suggest unusual non-Watson-Crick interactions, and we use this isostericity framework to tentatively assign possible geometries to the newly predicted base pairs (Table 5.2).

In the TPP riboswitch, there is significant MI between the two bases directly 5' of P3 and 3' of P3a that could bridge this helical junction. This correlation was highly significant ($p < 0.0001$) in an alignment of all TPP riboswitch sequences. Re-examination of the alignment showed that the predominant A–G and U–A pairs mainly occurred in the 550 sequences that have the optional P3a stem-loop. In fact, we detect no correlation between these columns in the remaining 354 sequences that lack P3a. Exchange of U–A and A–G pairs is most consistent with a *cis* H/WC edge interaction between these two bases. These pairs are also isosteric in a *trans* H/H geometry, but this configuration only involves a single hydrogen bond, and there are four other isosteric nucleotide combinations that are not observed. Both pair geometries imply that either the sugar-phosphate backbones of the interacting bases are in a parallel orientation or that they are anti-parallel with one of the bases adopting a rare *syn* glycosidic bond rotation. We

believe that it may be necessary for these bases to assume an unusual geometry to accommodate the P3a helix at this location. The molecular resolution structures of TPP riboswitches that have been solved do not impinge on this predicted interaction as both are of constructs lacking P3a [254, 281].

Our MI analysis predicts three new base interactions in the AdoCbl riboswitch. A lone WC base pair ($p < 0.0001$) seems to enclose the conserved A-rich sequence between the P2 and P3 helices. A highly significant MI score ($p < 0.0001$) supports a WC pair with purine/pyrimidine strand bias between the nucleotide directly 3' of the P4 helix and a position within the two nucleotide 3' bulge of the P6–P7 T-loop motif. The adjacent nucleotides in this strand and the T-loop bulge could form a highly-conserved, cohelical C-G base pair. Similar long-range Watson-Crick base pairing interactions to these two bulged nucleotides are common with "type-II" T-loops [200]. The final new prediction in the AdoCbl riboswitch is a non-canonical G–A or A–G pair ($p = 0.0001$) that probably assumes a *cis* WC/WC geometry to continue base stacking with the P6 helix. These pairs are also isosteric in a *cis* H/H geometry, but this geometry seems less likely because it involves only a single hydrogen bond.

A strikingly similar T-loop interaction is predicted within the FMN riboswitch. The nucleotide directly 3' of its P5 helix can form a Watson-Crick pair ($p = 0.007$) with a pyrimidine/purine strand bias to the 3' bulge of the T-loop motif that caps P3, and an adjacent G–C base pair could form between highly conserved nucleotides in the strand and T-loop bulge. It is interesting that the stem-loops adjacent to the interacting strand have exactly five paired nucleotides and are capped by a *second* T-loop motif in both the AdoCbl and FMN riboswitches. Although their presence does not appear to be relevant for this interaction and these riboswitches recognize very different ligand molecules, it suggests that even more structural similarity exists between their overall tertiary folds.

MI analysis suggests two new base interactions in the glycine riboswitch. The first is a WC pair ($p = 0.005$) with purine/pyrimidine strand bias at the base of the P2 stem of the first aptamer. If this pair cohelically stacks with the P2 stem, then it would often require a bulged nt on the 5' side of the composite helix. The second interaction is a predicted G–G or A–A homopurine pair ($p = 0.002$) that might adopt a *cis* bifurcated geometry within the central bulge of the second aptamer. Bifurcated pairs hydrogen bond between an exocyclic functional group on one base and the edge of the other base and are consequently intermediate between two edge geometries (in this case possibly *cis* WC/WC and *trans* WC/H). If this pair forms, it suggests that the two bases on each strand between it and the P1 stem may form G–A and A–G pairs. Both of these putative interactions are maintained in the opposite aptamer of the glycine riboswitch. However, the nucleotides at the corresponding positions are less variable, which may explain why they were not detected twice by the MI analysis.

Two new base pairing contacts are predicted for the SAM-I riboswitch structure. The first occurs at the end of the P2 helix adjacent to the conserved G–A and A–G pairs of the K-turn motif. This pair has a highly significant MI score ($p = 0.0001$) and mainly varies from G–A to C–C, which is most compatible with a *trans* SE/H base interaction within this cohelical stacking context. Noncanonical pairs with this configuration are known to occur frequently adjacent to K-turns in other functional RNA structures [163]. The second predicted interaction ($p = 0.001$) was an unexpected long-range *cis* WC/WC base pair between the base directly upstream of the 5' side of the P2b pseudoknot and the base directly upstream of the P1 3' strand. After originally discovering these new interactions from sequence analysis, we were able to verify that both interactions occur with the predicted configurations in the x-ray crystal structure of a minimized version of the *Thermoanaerobacter tengcongensis metF* SAM-I riboswitch [192].

We also predict two new base interactions in the SAM-II riboswitch from MI analysis. A homopurine G–G or A–A pair ($p = 0.001$) could form between two positions in the bulge between P1 and the 5' strand of the P2 pseudoknot. This pair may adopt a *cis* bifurcated geometry. A Watson-Crick base pair ($p = 0.02$) may also exist between the last nucleotide in the central loop contained within P1 stem and a downstream position. This pair could be extended into a short helical element (P1a) if the adjacent, conserved C–G and G–C base pairs also form canonical WC pairs and an intervening base is bulged out. These new interactions appear to constrain the overall structure of this small riboswitch into an extended, triple-stranded configuration.

5.9 Conclusions

Metabolite-binding riboswitches are vital components of microbial, fungal, and plant genomes. Perhaps the most important impact of this large-scale reevaluation and refinement of riboswitch secondary structure models is that it will improve the accuracy of future computational searches for these regulatory elements as we move toward the automated annotation of structured RNAs in genomic sequences [100]. It is particularly important to train computational models on diverse riboswitch sequences to increase the coverage of homolog identification.

Evaluating the statistical significance of mutual information scores with an evolutionary background model can be used to screen possible interacting bases while developing an RNA secondary structure model. The new base pair predictions garnered from this analysis here have interesting implications for the three-dimensional architectures of some riboswitches. Specifically, they help to add some candidate tertiary interaction "staples" to constrain the (still) unrealistically non-compact secondary structure drawings of these riboswitches. The further identification of unrecognized RNA structure motifs and new sites tolerating large sequence insertions from our expanded

riboswitch alignments will also complement and inform ongoing high-resolution structure determination efforts.

Riboswitches are thought to be descended from an RNA World [33]. The widespread prevalence of many riboswitch classes strongly suggests that they were, at least, present in the last common ancestor of modern bacteria. One of the most interesting aspects of our detailed phylogenetic profiling is that it outlines gaps and holes in the known distributions of certain riboswitch classes. Some of these apparently vacant regulatory niches may be occupied by regulatory proteins or by structural variants of known riboswitch aptamers that have diverged in ways that we cannot detect with current RNA homology search techniques. We anticipate that other voids could harbor entirely new aptamer classes that recognize the same metabolite as a known riboswitch class. The discovery of SAM-II riboswitches in α-proteobacteria [51], which are almost devoid of SAM-I riboswitches, sets a precedent for this latter scenario. The recent discovery of a third SAM riboswitch in some lactic acid bacteria species [83], a subgroup of species within the low-G+C Gram-positive bacterial division normally dominated by SAM-I riboswitches, suggests that new riboswitch classes may occupy regulatory gaps that exist at an even finer taxonomic resolution.

5.10 Methods

Riboswitch identification

Covariance models [67] were created using the Infernal software package (version 0.55) [65] from manually curated seed sequence alignments adapted from various sources (Table 5.1). Heuristic filtering techniques were applied to accelerate CM searches [313] against the RefSeq database (version 12) [223] and microbial environmental shotgun sequences from an acid mine drainage community [289], the Sargasso Sea [295], and

Minnesota soil and whale fall sites [283]. These environmental sequences are available from GenBank with project accession numbers AADL00000000, AACY00000000, AAFX00000000, AAFY00000000, AAFZ00000000, and AAGA00000000. We also searched for the TPP riboswitch in the "plant" and "fungi" divisions of the RefSeq database (version 13).

We evaluated the regulatory potential of putative riboswitch aptamers by examining their genomic contexts. To aid in this process, we uniformly predicting gene functions in all sequences with the COG database [274] using RPS-BLAST and scoring matrices from the CDD database [181]. The plausibility of putative aptamer structures was further assessed by computationally aligning hits to the original CM with Infernal and manually screening the results for more divergent RNA structures. Using these two complementary criteria, we established trusted CM score cutoffs. Above these thresholds all hits in the RefSeq database were true riboswitches. Since gene context information is not available for most environmental sequences, we only included hits from these sequence sets that scored above the trusted threshold. We included additional low-scoring sequences from the RefSeq database when their genomic contexts and alignments strongly indicated that they were functional riboswitches. None of the newly included sequences dramatically altered the core structural models present in the seed alignments.

In order to verify that we were recovering known riboswitch sequences, we compared our final results to a list of TPP riboswitches compiled in an exhaustive comparative genomics analysis of thiamine metabolic genes and this regulatory RNA element [234]. For this aptamer, at least, our approach successfully identified every riboswitch that had been previously found and was present in our microbial sequence database. We also discovered a small number of new TPP riboswitches in front of

thiamine-related genes (e.g. *pnuC* in *Helicobacter pylori* and *thiM* in *Lactococcus lactis*) that had not been reported in the genomes used by the former study.

For the glycine riboswitch, we searched for hits to a single aptamer covariance model and a tandem model containing both the first and second aptamers. We verified that each single aptamer that is part of a tandem configuration was found by the single aptamer CM, and noted cases of lone aptamers. For consensus structure and mutual information calculations we considered only the tandem glycine aptamer alignment. We included the complete set of lone and tandem aptamer glycine riboswitches in the expression platform analysis.

Mechanism Classification

We predicted expression platforms for a subset of riboswitches in complete and unfinished microbial genomes. Aptamer sequences with more than 95% pairwise identity at reference columns (positions where ≥50% of the weighted sequences in the alignment do not contain a gap) were removed to avoid biasing statistics with duplicate sequences. We further screened out riboswitches with suspect gene annotations where >60 nt of an open reading frame (ORF) on the same strand overlapped the aptamer or >700 nt separated the aptamer and the nearest downstream ORF. Most of these situations result from incorrect start codon choices, overpredictions of hypothetical ORFs, or missing ORF predictions of real genes. The remaining sequences constituted our expression platform dataset, and we extracted sequences beginning at the 5' end of each aptamer and continuing through the first 120 nt of the downstream open-reading frame for further analysis.

To classify the mechanisms of these riboswitches, we first scanned expression platforms with the local RNA secondary structure prediction program RNall (version 1.1) [303] for intrinsic transcription terminators. Since many true riboswitch terminator

hairpins were longer than the default scanning size of 30 nt, we increased this parameter to 50 nt. We also required a higher U-tail weight of 4.0 (the default is 3.0) for the uridine-rich single strand directly 3' of the hairpin and required a stability lower than -8.3 kcal/mol for pairing between the U-tail region and sequences directly upstream of the hairpin (the default cutoff is -11.7 kcal/mol). Riboswitches with a terminator predicted in their expression platform presumably regulate transcription termination.

Riboswitches will regulate translation initiation if a ligand-induced change in structure alters the accessibility of the downstream gene's ribosome binding site. We therefore predict that a transcription terminator is also regulating translation if the distance between the terminator hairpin and the gene's start codon is no more than 10 nt. If the start codon is no more than 15 nt from the end of the conserved core of the riboswitch aptamer (usually the P1 paired element), then we predict that the change in the aptamer's structure upon metabolite binding directly regulates translation initiation. For all other expression platforms that do not fit these criteria, we assume that most contain a downstream unconserved stem-loop that can sequester the RBS, based on known examples, although they could employ other regulatory mechanisms.

We calibrated Rnall and distance parameters by comparing our predictions to the large and phylogenetically diverse dataset manually assembled for the TPP riboswitch that originally inspired this classification scheme [234]. Rnall correctly predicts 46 out of 52 terminators in this data set with only 3 predictions of terminators in sequences not manually evaluated as containing terminators, meaning it has a sensitivity of 88% and an accuracy of 94%. The three false positives resemble true terminators and may truly be functional, whereas the terminators that Rnall misses usually have large hairpins with poor thermodynamic stabilities. Overall, our computational procedure classifies 159 out of 180 TPP riboswitch expression platforms (88%) into the one correct category of the four mechanisms assigned in this control set.

Consensus structures

For most statistics, sequences in the full automated alignments for each riboswitch were weighted using Infernal's internal implementation of the GSC algorithm [87] to reduce biases from duplicate sequences before all calculations. Covariance models are not able to align pseudoknots, which are present in several riboswitch structures. Therefore, we repeated the entire alignment and analysis process using alternate covariance models that enforce pseudoknot pairings and remove incompatible helices to examine conservation and mutual information near these features. The original seed alignments and full automated alignments with newly predicted tertiary interactions annotated will be made available in Stockholm format on the web (http://bliss.biology.yale.edu).

Mutual Information Significance

Mutual information was calculated between column pairs according to standard formulas, treating gaps as a fifth character state in addition to the four RNA nucleotides [110]. In order to ascertain the statistical significance of MI values we resampled each riboswitch alignment according to an evolutionary model. A customized version of the program Rate4Site (version 2.01) [183] with modified output options was used to simultaneously estimate distances and per-column rates of evolution according to a gamma background model with at least 16 rate categories for a phylogenetic tree created with Jukes-Cantor distances and treating gaps as missing information. We purged duplicate sequences and removed columns with >50% gaps from riboswitch alignments prior to this analysis, and, if necessary, pruned alignments to the 300 most diverse sequences (as judged by pairwise distances in the full alignment). We used the resulting trees, rates, and distances (which assume independent evolution of all columns) to simulate 10,000 resampled alignments without covariation constraints starting from an arbitrary ancestral sequence. We re-inserted gaps into these derivative

alignments at the same positions as in the original alignment and used these alignments to estimate the background distribution of MI scores. The p-value significance of the MI between two columns is the fraction of the resampled alignments that have a greater MI score than the value observed between those two columns in the real alignment.

6 *E. coli* 6S RNA homologs are widespread in eubacteria

6.1 Introduction

Escherichia coli survives nutrient limitation by compacting its nucleoid, altering the promoter specificity of RNA polymerase (RNAP), and sequestering ribosomes in inactive 100S dimers to globally reduce and adapt gene expression [131]. Levels of the perhaps ~1000 growth-related genes expressed in an exponentially dividing bacterium generally decrease, while ~100 genes are specifically activated for stationary phase maintenance. Much of this adaptation is accomplished by increasing the population and activity of RNAP holoenzymes that direct transcription initiation with the stationary phase promoter-specific sigma factor (σ^S) relative to the housekeeping sigma factor (σ^{70}). Changes in cytoplasmic solute composition, an increase in σ^S levels, and expression of Rsd, an anti-σ^{70} factor, all contribute to this overall shift in promoter specificity [132].

E. coli 6S RNA participates in the transcriptional response to starvation by binding to σ^{70}-containing RNAP holoenzyme [308]. Its expression increases 11-fold during stationary phase to a maximum of ~10,000 copies per cell when >75% of σ^{70} holoenzymes are associated with 6S RNA. The molecular details of this recognition are unknown, but the extended hairpin structure proposed for 6S RNA resembles DNA template in an open promoter complex with RNAP [307]. 6S RNA is necessary for the repression of σ^{70}-dependent promoters that contain extended –10 sequences under nutrient limitation and concomitant activation of certain σ^S-dependent promoters [284]. Despite this widespread regulatory role, 6S RNA knockouts exhibit only subtle growth defects. Deletion of 6S RNA causes reduced viability compared to wild-type control cells after >20 days of continuous culture, and cells lacking 6S are at a competitive disadvantage when cocultured with wild-type cells after several days of growth [307].

Although *E. coli* 6S RNA was the first noncoding RNA to be sequenced more than 30 years ago [35], additional 6S RNA homologs have only been reported in *Pseudomonas aeruginosa* [301] and *Haemophilus influenzae* [34]. All currently known 6S RNA sequences identified by bioinformatics in the Rfam database are likewise restricted to species of γ-proteobacteria [99]. We have computationally identified numerous additional homologs of 6S RNA in >100 bacterial species representing diverse eubacterial lineages. A comparative analysis of 6S RNAs has allowed us to elaborate on how this RNA could mimic an open promoter to bind RNAP holoenzyme, to examine the evolution of two functionally divergent copies of 6S in some Gram-positive bacteria, and to predict that 6S RNA is cotranscribed with the reading frame for a protein that may regulate folate levels in many proteobacteria.

6.2 Identification of 6S RNA homologs

In the course of investigating new RNA motifs in *Bacillus subtilis* by genomic comparisons of intergenic regions [19], we rediscovered two noncoding RNAs, BsrA and BsrB. BsrA RNA had been isolated as a highly expressed transcript in total *B. subtilis* RNA separated on polyacrylamide gels [273]. The 201 nt BsrA RNA, encoded by the *aspS-yrvM* intergenic region, is slowly processed into a 190 nt RNA by the removal of 11 nt from its 5' end. At the same time, an abundant 203 nt transcript from the *yocI-yocJ* intergenic region of *B. subtilis* was recovered and named BsrB [9]. Preliminary biochemical investigations did not reveal the functions of BsrA or BsrB. We manually aligned BLAST hits between the *aspS-yrvM* intergenic region and sequences upstream of *yrvM* homologs in other Gram-positive bacteria. The common secondary structure model for this RNA family was essentially the same as that predicted previously for BsrA by thermodynamic calculations [273].

We used covariance models [67] trained on this alignment to search the complete and unfinished microbial genomes available in GenBank for more divergent BsrA homologs. Surprisingly, BsrB was present in the expanded collection of hits from Gram-positive organisms. These searches also uncovered convincing similarity to cyanobacterial genomes that overlapped annotations of an RNA named 6Sa. Reminiscent of BsrA and BsrB, 6Sa RNA was identified as an abundant noncoding transcript of unknown function with a size of 185 nt from *Synechococcus* sp. PCC6301 [309]. Note, however, that all genomic annotations of 6Sa RNA in cyanobacterial genomes are on the incorrect strand (e.g. *Nostoc* sp. PCC 7120, GenBank:NC_003272.1). The common secondary structure predicted for these three noncoding RNAs together was much different from the models first proposed for the BsrA and 6Sa RNAs based on thermodynamic folding (Figure 6.1). The suggestive name of 6Sa RNA, presence of this RNA family in distant bacterial lineages, and common ~200 nt length of these RNAs encouraged us to look for similarity between these noncoding RNAs and *Escherichia coli* 6S RNA.

Indeed, 6S RNA was originally isolated more than 35 years ago as a small stable RNA of 184 nt that formed a distinct band after polyacrylamide gel electrophoresis (PAGE) of *E. coli* total RNA [123], and its function was entirely cryptic until recently. Despite an average pairwise similarity between 6S RNA and the *B. subtilis* BsrA and BsrB RNAs of only 46%, conservation of key secondary structure and nucleotide sequence elements provide strong evidence that these noncoding RNAs are structural homologs (Figure 6.2). An improved covariance model, trained with known 6S sequences from the Rfam database [99] and with our original sequences from Gram-positive bacteria and cyanobacteria, identified 6S RNA sequences in almost every major group of Eubacteria.

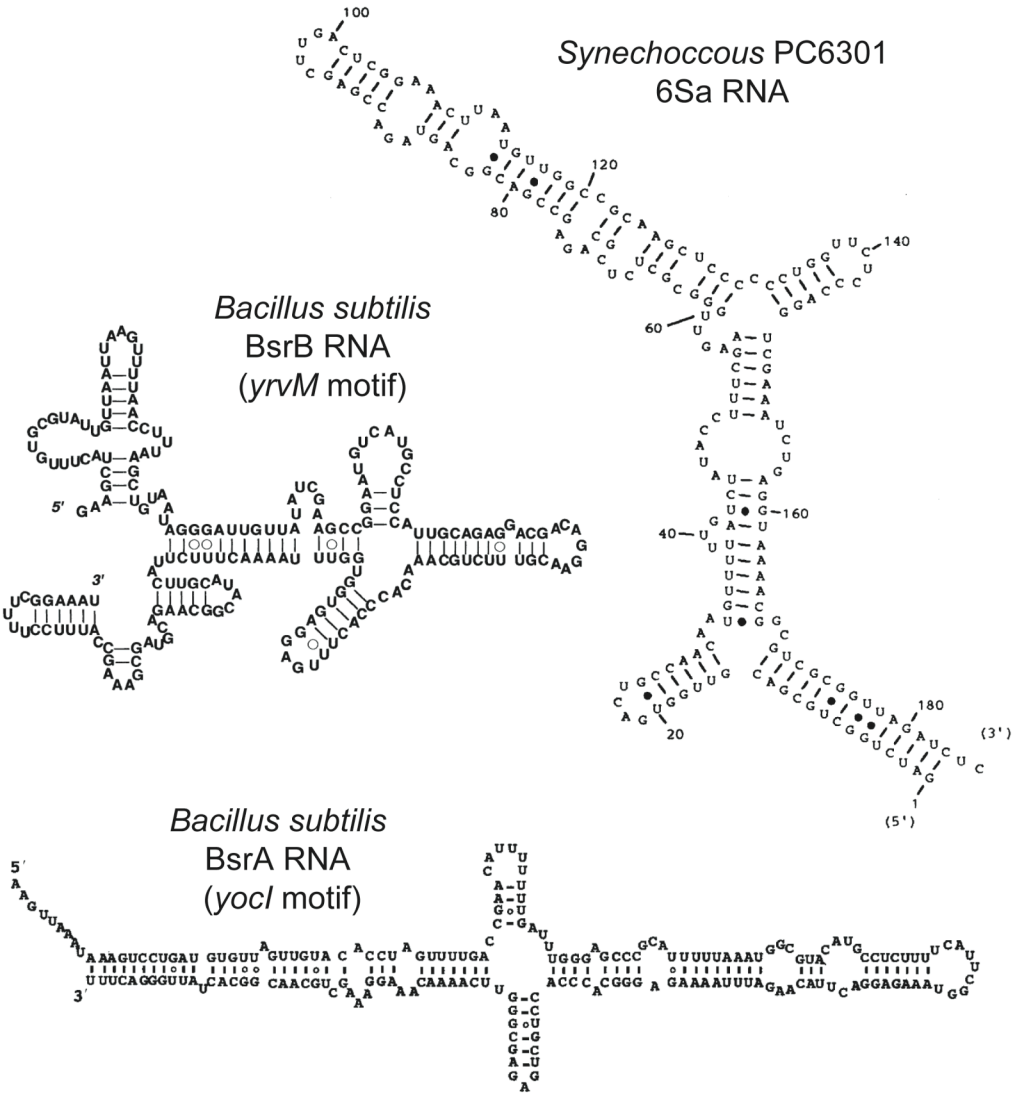


Figure 6.1 Previously published structure models for 6S RNA homologs

Structural models of *B. subtilis* BsrB RNA [273], BsrA RNA [9], and cyanobacterial 6Sa RNAs [309] are adapted from previous reports of these seemingly disparate, highly-expressed, noncoding RNAs. We initially encountered BsrB RNA and the reverse complement of BsrA RNA as putative regulatory motifs occurring upstream of *yrvM* and *yocI* genes in several related Gram-positive bacteria. All three of these RNAs can adapt a secondary structure similar to *E. coli* 6S RNA as shown in Figure 6.2, where BsrB and BsrA are 6Sa and 6Sb from *B. subtilis*, respectively.

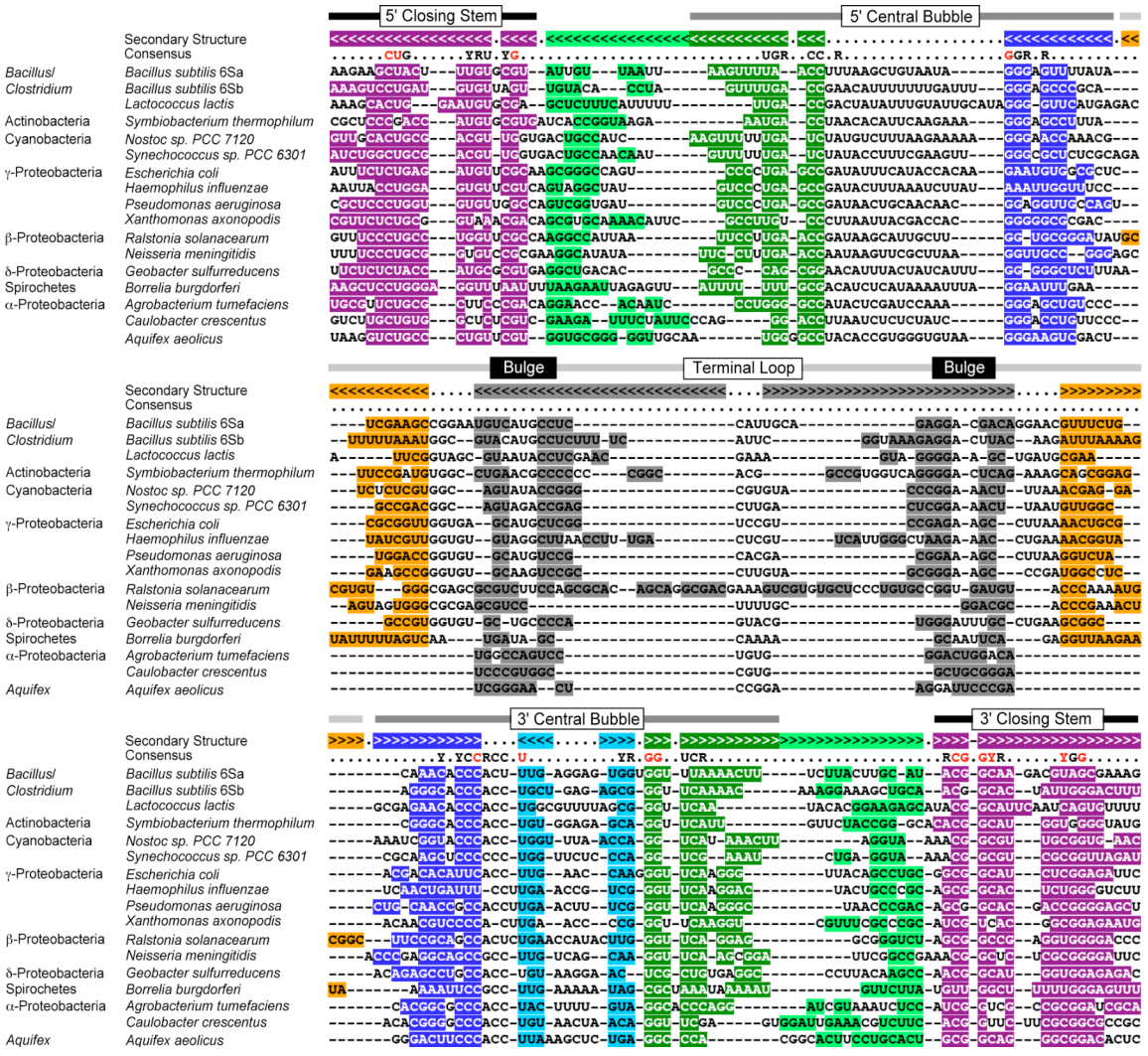


Figure 6.2 Representative 6S RNA sequence alignment

The three structural domains of 6S RNA (closing stem, central bubble, and terminal loop) and the conserved "bulge" present in the terminal loop sequences of certain bacterial lineages are labeled. Letters in the consensus line identify nucleotides that are conserved in ≥80% (gray) and ≥95% (black) of all 6S RNA sequences. Purine (R = A or G) and pyrimidine (Y = C or U) designations are used when a single nucleotide is not 80% conserved. Putative base pairing in individual sequences is highlighted with shaded backgrounds corresponding to paired angle brackets in the consensus secondary structure line.

6S RNA sequences from α-proteobacteria species were conspicuously absent from these expanded results. We initially investigated the presence of 6S RNA in this clade by isolating total RNA from *Caulobacter crescentus* cultures grown to mid-log and stationary phase and looking for ~200 nt bands on polyacrylamide gels stained with ethidium bromide. These experiments revealed a likely 6S candidate. An abundant ~180 nt RNA did not match the predicted sizes of annotated noncoding RNAs, and its expression increased during stationary phase (data not shown).

With this indication that our bioinformatics searches were not identifying all 6S homologs, we adopted a targeted strategy. Alignments of the other 6S RNA matches indicated that a conserved bulge in the terminal loop of 6S RNA was missing in certain lineages (Figure 6.2 and Figure 6.3). We used an alignment of only these sequences from β-proteobacteria, δ-proteobacteria, and spirochetes to create covariance models with the diverse terminal loops explicitly modeled as variable insertions and conducted unfiltered searches against selected α-proteobacterial genomes. Among the matches in *C. crescentus* and *Agrobacterium tumefactions* were sequences that clearly matched the consensus features of 6S RNA, but had terminal loops truncated to a single stem-loop. Incorporating these sequences and a similar loop-truncated 6S RNA from *Aquifex* into the multiple sequence alignment and repeating the search readily identified other high-scoring 6S homologs from α-proteobacteria with this variation.

Figure 6.2 is an alignment of 17 representative 6S RNA homolog sequences. Our final curated alignment contains 121 sequences, and covariance models built from this alignment find hundreds of additional 6S RNA sequences in microbial genomes and environmental sequences [295]. The curated seed alignment and an automated

Figure 6.3 6S RNA secondary structure and open promoter DNA template

(A) Consensus secondary structure model for 6S RNA. Nucleotide symbols and colors are the same as in Figure 6.2 (consensus line). Certain nucleotides whose identity is not conserved but are present in ≥60% of sequences are represented as empty circles. Solid lines represent variable regions of the structure. Three parallel insets show lineage specific terminal loop structures and sequence conservation, and the boxed nucleotides on the 3' side of the central bubble can alternately form the pictured base-paired stem in many sequences. Annotated nucleotide distances are the median lengths between conserved segments. ***(B)*** Schematic of DNA template in the open promoter complex with RNA polymerase (RP_o) as described elsewhere [197].

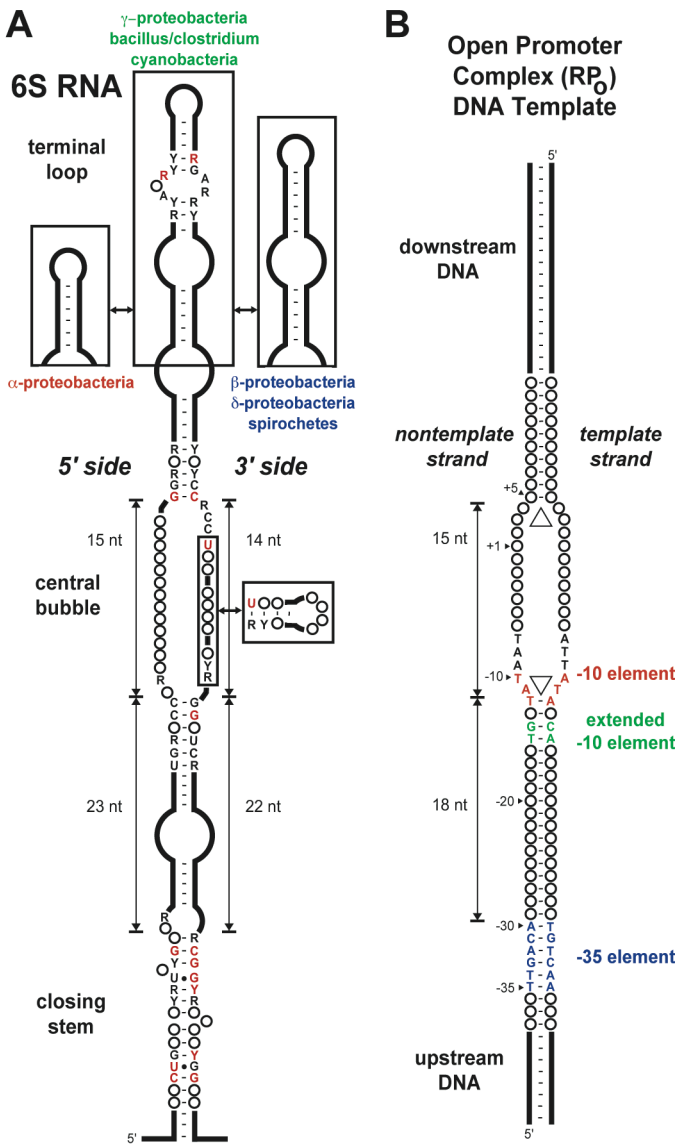


Figure 6.3 6S RNA secondary structure and open promoter DNA template

alignment of all matches are available online (http://bliss.biology.yale.edu). Additionally, the seed alignment has been submitted to the Rfam database [99] to update the model for the 6S RNA family (Rfam:RF00013).

6.3 Nomenclature

Independent descriptions of unrecognized 6S RNA homologs in different bacteria and the presence of multiple copies of 6S RNA within a single genome complicate 6S RNA nomenclature. We suggest using the *E. coli* 6S RNA and *ssrS* gene designations for all organisms. The *Synechococcus ssaA* gene and 6Sa RNA names can be directly replaced with *E. coli* equivalents in this naming scheme. We discriminate between multiple 6S sequences within one genome by appending a single letter to each name in order of 6S RNA gene distance from the genomic origin. Accordingly, *ssrSA* and *ssrSB* are updated synonyms for the *B. subtilis bsrB* and *bsrA* genes, and they encode 6Sa RNA and 6Sb RNA, respectively. These nomenclature recommendations are used throughout the remainder of this report.

6.4 Conserved features

The overall structure of 6S RNA (Figure 6.3*A*) can be divided into three conserved domains separated by variable stems:

(1) The *closing stem* is the outer boundary of the 6S RNA hairpin. It consists of a ≥15 nt long stem with conserved base pairs and bulges. The identities of both nucleotides in eight of these base pairs are constrained, including two G-U wobble base pairs. The one-nucleotide bulge interrupting the inner stretch of conserved bases on the 5' side is highly conserved, and a bulged nucleotide on the 3' side between the conserved base pairs is commonly present. The outer margin of the closing stem pairing is highly variable in individual sequences.

(2) The *central bubble* is a large internal loop. The 5' side of the bubble appears to be completely single-stranded and does not contain conserved sequences. However, its length is relatively constant, ranging from 12-21 nt with a median length of 15 nt. The poor secondary structure potential of this stretch of nucleotides is underscored by its abnormally low guanosine content of only 10.9% compared to an overall frequency of 25.1% in the entire alignment. In contrast, the 3' side of the central bubble includes four conserved bases on its inner side, and the remainder can fold into a short stem loop in many sequences. Two conserved G-C base pairs surround the central bubble on both sides. There is additional conservation of the identities of nucleotides forming three base pairs further along the outer stem, and a common purine-purine mismatch as the fourth base pair of the inner stem in certain groups of bacteria.

(3) The *terminal loop* is a lineage specific extension of the innermost base pairing elements of the 6S hairpin. In γ-proteobacteria, cyanobacteria, and the *Bacillus*/*Clostridium* group there is a variable penultimate base-paired stem and bulge separating the central bubble from a final conserved stem. This final stem contains a characteristic purine-rich asymmetric bulge with three nucleotides on the 5' side and two on the 3' side, surrounded by preferred base pairs. Spirochetes, β-proteobacteria, and δ-proteobacteria preserve this general arrangement but do not have the conserved bulge in their final stems. In α-proteobacteria, the terminal loop is truncated to a single, short stem without any apparent sequence conservation.

The most recently published secondary structure model for *E. coli* 6S RNA [308] must be modified slightly to match the conserved structure. Specifically, the previously proposed pairing of CAA to UUG on the inner side of the central bubble must be disrupted, and the optional stem-loop on the 3' side of the central bubble also can be formed by the *E. coli* variant.

Two features of 6S RNA conservation are unusual for a bacterial noncoding RNA. First, much of 6S primary sequence conservation occurs in canonical base-paired stems. Assuming that the central bubble's facultative stem-loop does not form and excluding the lineage specific conservation in the terminal loop, 88% of the highly-conserved positions (≥80% conservation of a specific nucleotide) are in putative base pairs. For comparison, only 47% and 59% of conserved nucleotides are paired within the aptamer domains of the metabolite-binding glycine [180] and cobalamin [201] riboswitches, respectively. Second, large unconstrained loops like the central bubble 5' strand are rare in structured bacterial RNAs. Typically, a putative single-stranded region of a functional RNA has conserved nucleotides that actually mediate the formation of a pseudoknot or tertiary packing. For example, all four of the single-stranded regions in the consensus minimal structure of bacterial RNase P RNAs of comparable size to the 5' central bubble strand of 6S RNA contain universally conserved nucleotides, and two combine with each other to form a pseudoknot helix [79].

6.5 6S RNA resembles open promoter DNA

The conserved features of 6S RNA homologs support the hypothesis that 6S RNA mimics the structure of DNA template in an open promoter complex (RP_o) with RNA polymerase [307] and suggest further possibilities for more detailed binding models (Figure 6.3). RNAP holoenzyme melts double-stranded DNA template around the –10 element from position –11 to +4 relative to the transcription initiation site at +1 so that the template strand can weave through the polymerase active site [197]. The 15 nucleotides of single-stranded nucleic acid in RP_o correspond remarkably well to the dimensions of the central bubble in 6S RNA homologs. The unstructured 5' side has 15 nt, and unwinding the optional 3' strand's stem-loop would free a total of 14 nt (median lengths).

Presenting a premelted promoter bubble may give 6S RNA a general affinity for core RNA polymerase (lacking a σ-factor). Similar DNA templates constructed with arbitrary nonpromoter sequences and ~10 bp single-stranded bubbles are capable of directing RNA synthesis with core RNAP [119]. Much of the affinity of RNAP for DNA templates is mediated by electrostatic interactions with the phosphate backbone that will be preserved by an RNA template. For example, single abasic substitutions at positions –11 to –7 on the nontemplate strand of fork junction DNA do not reduce its affinity for RNAP, although they do extenuate the subsequent formation of heparin-resistant complexes [75]. It is thought that RNAP recognizes the geometry of ss/ds DNA junctions in bubble templates. The observed conservation of two strong G-C base pairs flanking each side of the 6S central bubble might enforce its boundaries to favor these interactions. If the 6S central bubble binds core RNAP like an open promoter, then its surrounding base-paired stems will naturally follow the paths of upstream and downstream DNA template over basic surfaces in the polymerase structure [197].

It is possible that 6S RNA's closing stem replaces upstream DNA template so that its sequence conservation can interact with σ^{70}. The spacing between the central bubble and consensus elements in 6S RNA is broadly reminiscent of a typical DNA promoter (Figure 6.3*B*). The conserved UGR/UCR base pairs located directly outside the central bubble might engage σ^{70} like an extended –10 element. However, there is no corresponding sequence similarity between the 6S closing stem and the usual –35 TTGACA consensus box. It seems more likely that σ^{70} forms a novel distal contact with 6S RNA's closing stem conservation here. Unlike the other possible interactions we have described, this contact could directly contribute to the observed specificity of 6S RNA for σ^{70} holoenzyme and its inability to bind σ^{S} holoenzyme.

Orienting 6S RNA within RNAP with the closing stem in the direction of the DNA promoter is also appealing because it distinguishes the conserved 3' side of the central bubble as the DNA template strand mimic. This architecture positions its conserved RCCU sequence near the site where transcription initiates on a DNA promoter. In this context, the optional stem-loop might masquerade as the short DNA/RNA hybrid helix normally present within the transcription bubble during elongation. Its placement also resembles that of stem-loops formed within the nascent RNA during the process of intrinsic transcription termination [329]. Finally, this choice of template strand relegates the flexible 5' strand of the bubble to a role as nontemplate strand and suggests that its length (and not its sequence) is conserved because it does not traverse the active site of RNAP.

It is not clear how the lineage specific 6S RNA terminal loop could contribute to holoenzyme recognition. The purine-rich asymmetric bulge of the γ-proteobacterial loop type resembles a tertiary interaction motif far more than any other conservation in 6S RNA and might interact with a downstream site on RNA polymerase or fold back on 6S RNA. On the other hand, terminal loops from other bacterial groups appear to lack any sequence conservation and would therefore seem incapable of participating in specific interactions. We also note that there is no obvious evolutionary correlation between the type of terminal loop and the domain structure of RNA polymerase in different bacterial lineages [134].

Specific binding of bacterial DNA-dependent RNA polymerases to templates composed of ribonucleotides is not unprecedented. In fact, some RNAs are able to act as true promoters to direct the synthesis of complementary RNA transcripts. Certain RNA sequences selected from random copolymer mixtures are capable of autocatalytic replication by *E. coli* RNAP holoenzyme through unknown intermediates [314]. Also, a stem-loop derived from the peach latent mosaic viroid can initiate efficient *in vitro*

transcription by *E. coli* RNAP from one strand of its hairpin in a reaction thought to recapitulate the natural replication of this single-stranded RNA in plants [216].

Recently, mouse B2 RNAs have been shown to repress general transcription in heat-shocked cells by binding directly to RNA polymerase II [5, 73]. Although B2 RNA is similar in size and function to 6S RNA, these RNAs do not appear to be evolutionarily related. B2 RNAs are encoded by SINE elements that are thought to be derived from Ser-tRNA [55], and they do not have the consensus features of known 6S RNA homologs. Both RNAs halt transcription before initiation, but B2 RNA binds to a remote docking site on Pol II and stalls polymerase while it is engaged to DNA template at the active site [73] whereas 6S RNA probably directly competes with DNA template for RNAP binding. It will be interesting to compare the molecular mechanisms of these convergent solutions that widely inhibit transcription under stress conditions, particularly how each allows specific subsets of promoters to escape repression.

6.6 Structural probing

We subjected the 184 nt *E. coli* (6S-184) and *B. subtilis* (6Sb-201) 6S RNAs to in-line probing to verify that they folded into the structures predicted by comparative sequence analysis (Figure 6.4). In this assay, spontaneous transesterification of 5' radiolabeled RNA produces an RNA cleavage pattern that reflects the relative sampling of backbone conformations with the correct geometry for in-line attack of each ribose 2'-OH on the adjacent bridging phosphate [260]. Flexible regions of the RNA such as bulges and loops allow nucleotides to sample the in-line conformation and yield RNA degradation products (identified as bands upon autoradiography after PAGE), while base-paired regions are rigidly held in a structure that precludes in-line attack and are consequently resistant to degradation.

Figure 6.4 In-line probing of 6S RNA structures

(A) Sequence, secondary structure, and in-line probing data for *E. coli* 6S-184 RNA. Levels of spontaneous RNA cleavage at backbone linkages within the construct depicted to the right were measured by separating 5'-radiolabeled degradation products on a polyacrylamide gel. In-line probing gel lanes are: NR, no incubation; T1, partial digestion with RNase T1 (cleaves 3' of G nucleotides); $^{-}$OH, partial alkaline hydrolysis; P, spontaneous cleavage during a 40 hr in-line probing reaction incubated at 25°C. Pre identifies the full-length RNA. Bands corresponding to certain T1 cleavage products are identified as position markers. In the secondary structure model, shaded circles identify nucleotides whose 3' linkage undergoes a high level of spontaneous cleavage relative to most other linkages. Filled triangles mark the extent of the region where cleavage sites were mapped. Lowercase letters identify unnatural guanosine nucleotides added for efficient *in vitro* transcription with T7 RNAP. ***(B)*** Sequence, secondary structure, and in-line probing data for *B. subtilis* 6Sb-201 RNA. Details as in (*A*). Slow *in vivo* processing of 201 nt 6Sb RNA cleaves off 11 nucleotides (gray), resulting in the 190 nt form of 6Sb RNA.

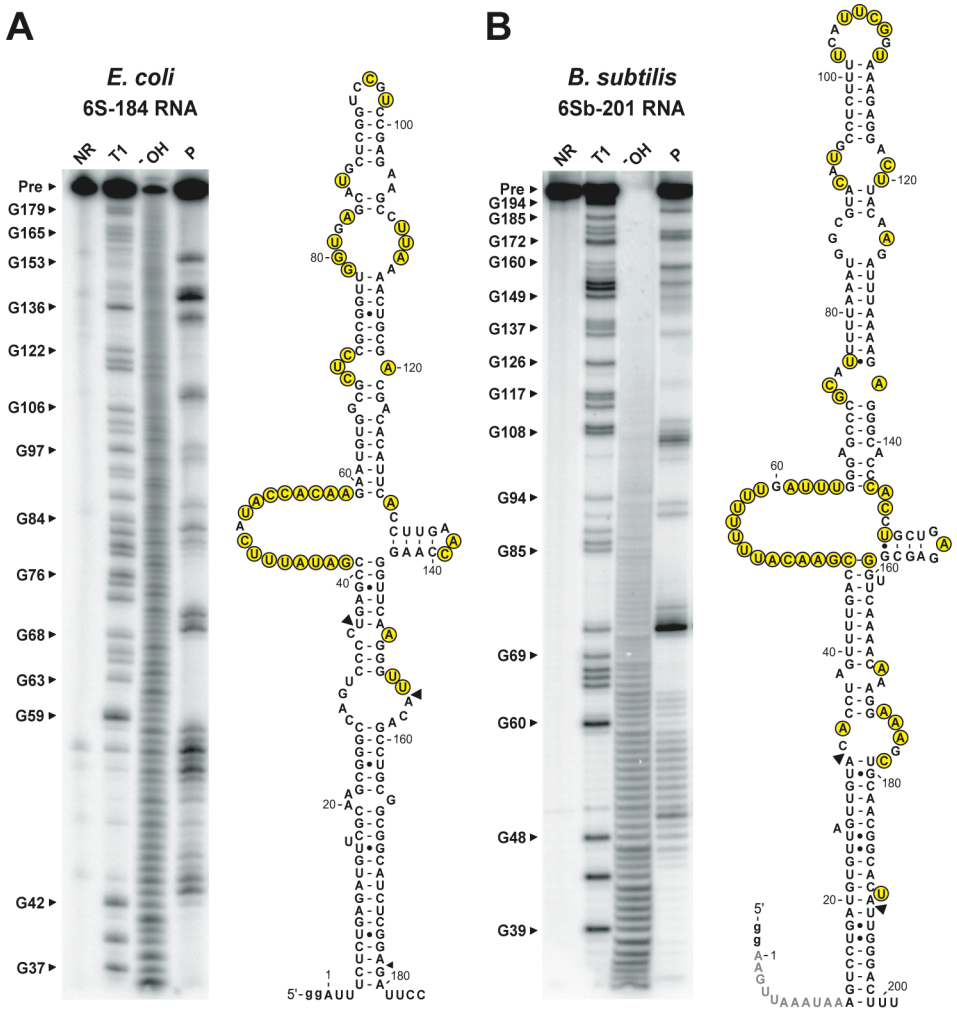


Figure 6.4 In-line probing of 6S RNA structures

Both RNAs produce in-line probing patterns that agree with their predicted secondary structures. Cleavage at positions 56-58 in the *E. coli* construct (Figure 6.4*A*) supports the omission, predicted by comparative sequence analysis, of three base pairs on the inner side of the central bubble in the previous structural model [308]. Reduced spontaneous cleavage of nucleotide linkages on the 3' side of the bubble indicates that the optional stem-loop forms in both 6S RNAs, although it is difficult to precisely map pairing in this region of the in-line probing gels. In contrast, the RNA backbone in the extended internal loop on the 5' side of the central bubble is consistently susceptible to spontaneous cleavage in both constructs. As predicted for mimicking an open promoter, this region is single-stranded in isolated 6S RNA and not involved in any higher-order structure. Overall, the agreement of the in-line probing patterns clearly indicates that the *E. coli* 6S and *B. subtilis* 6Sb RNAs adopt the same consensus structure derived from comparative sequence analysis.

6.7 Phylogenetic distribution

We constructed a distance-based phylogenetic tree from the curated multiple sequence alignment, restricting the analysis to majority ungapped positions and excluding highly-variable regions like the terminal loop (Figure 6.5). This tree supports an ancient origin and uninterrupted evolution for 6S RNA within the Eubacteria. The 6S RNA phylogenetic tree generally reproduces the standard bacterial taxonomy based on 16S ribosomal RNA [50], and there are no obvious cases of horizontal gene transfer. The clustering of 6S RNA terminal loop synapomorphies into branches corresponding to evolutionarily related bacteria — even though this portion of the sequence was not included in tree calculations — further supports the large-scale features of this tree topology. It also suggests that the expanded terminal loop with the conserved bulge present in *E. coli* 6S

Figure 6.5 Phylogenetic tree of 6S RNA homologs

An unrooted phylogenetic tree was constructed from the final seed alignment of 121 sequences using distance methods. Symbols represent the taxonomic classification of the genomes containing each 6S RNA sequence. They are *Bacillus*/*Clostridium* (filled squares), actinobacteria (filled diamond), spirochetes (shaded diamonds), cyanobacteria (filled triangles), α-proteobacteria (open squares), β-proteobacteria (shaded squares), γ-proteobacteria (filled circles), δ-proteobacteria (shaded circles), other proteobacteria (open diamonds), and *Aquifex* (open triangle). Groups with the same shading share terminal loop types as shown in Fig 2. Lowercase letters identify multiple 6S RNA sequences within one genome, and 6S RNA genes that are upstream of *E. coli ygfA* homologs are starred. Certain bacterial species are labeled with abbreviations as follows: Aae, *Aquifex aeolicus*; Aci, *Acinetobacter* sp. ADP1; Atu, *Agrobacterium tumefaciens*; Ban, *Bacillus anthracis*; Bbr, *Bordetella bronchiseptica*; Bbu, *Borrelia burgdorferi*; Bha, *Bacillus halodurans*; Bja, *Bradyrhizobium japonicum*; Bme, *Brucella melitensis*; Bsu, *Bacillus subtilis*; Cac, *Clostridium acetobutylicum*; Cbu, *Coxiella burnetii*; Ccr, *Caulobacter crescentus*; Cvi, *Chromobacterium violaceum*; Eco, *Escherichia coli*; Gsu, *Geobacter sulfurreducens*; Mag, *Magnetococcus* sp. MC-1; Mma, *Magnetospirillum magnetotacticum*; Nme, *Neisseria meningitidis*; Nos, *Nostoc* sp. PCC 7120; Oih, *Oceanobacillus iheyensis*; Pae, *Pseudomonas aeruginosa*; Pma, *Prochlorococcus marinus*; Sth, *Symbiobacterium thermophilum*; Tde, *Thiobacillus denitrificans*; Tte, *Thermoanaerobacter tengcongensis*; Rpr, *Rickettsia prowazekii*; Rso, *Ralstonia solanacearum*; Spy, *Streptococcus pyogenes*; Sy6, *Synechococcus* sp. PCC 6301; Vch, *Vibrio cholerae*; Xax, *Xanthomonas axonopodis*. Other species names are omitted for clarity.

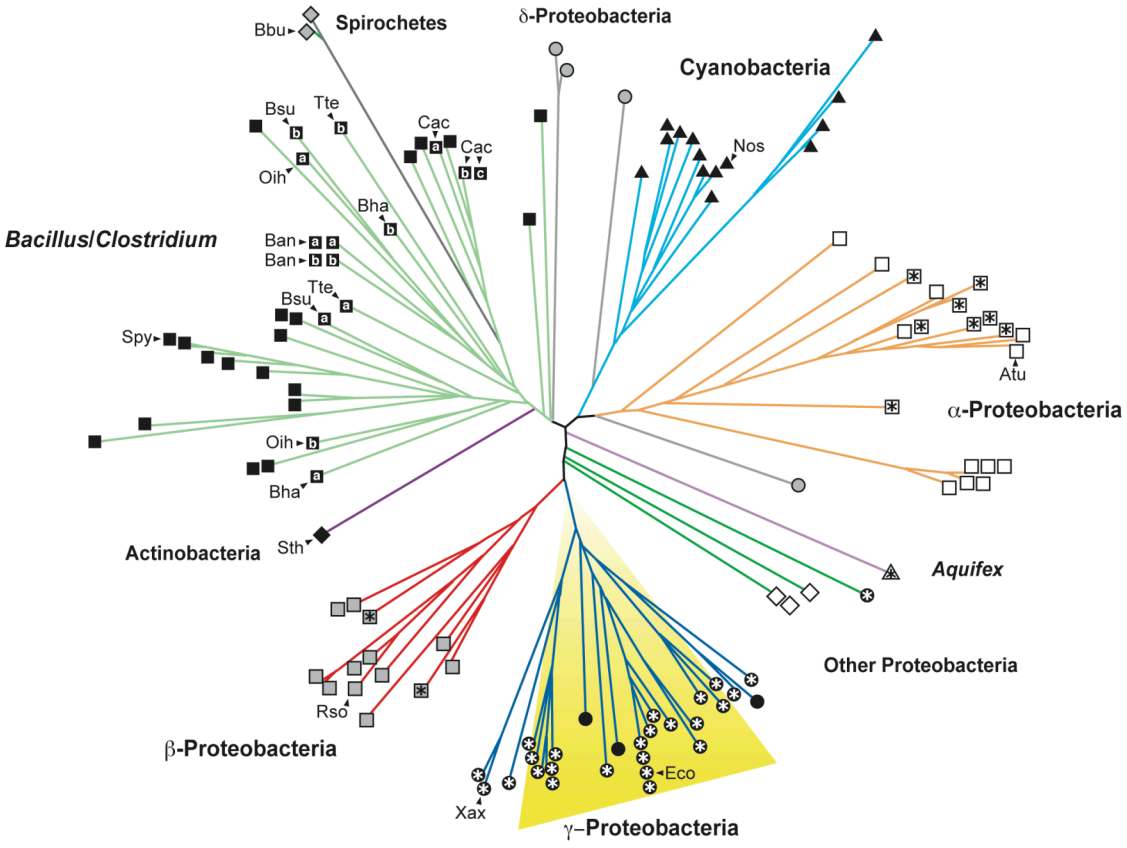


Figure 6.5 Phylogenetic tree of 6S RNA homologs

RNA is the ancestral state, since the alternative hypothesis that an identical structure evolved separately in the cyanobacteria, γ-proteobacteria, and low-GC Gram-positive bacteria (*Bacillus*/*Clostridium* group) is unlikely. Presumably, the terminal loop has atrophied or become modified in other lineages. As has been the case with identifying microbial RNase P RNAs [167], further targeted experimental and bioinformatic efforts may detect new 6S sequence variants in other genomes.

Generally, there is one copy of 6S RNA per microbial genome. However, two divergent 6S RNAs are present in several low-GC Gram-positive bacteria including *B. subtilis*, *Bacillus halodurans*, *Clostridium acetobutylicum*, *Oceanobacillus iheyensis*, and *Thermoanaerobacter tengcongensis*. The probable phylogenetic relationships of these multiple copies in the context of all 6S RNAs from this clade indicate that at least one gene duplication must have occurred within this lineage (Figure 6.5, e.g. Bsu 6Sa and 6Sb). The presumptive functional diversification in these select instances seems to have been accompanied by more widespread loss of the second 6S RNA copy in most branches. In contrast, the almost identical copies of 6S RNA in *Bacillus anthracis*, and *C. acetobutylicum* (which contains a total of three 6S homologs) are probably the result of very recent gene duplications. The only bacteria where we have identified multiple 6S RNA copies outside of the *Bacillus*/*Clostridium* group are *Magnetococcus* sp. MC-1 and *Magnetospirillum magnetotacticum* sp. MS-1. Each of these unfinished genomes encodes at least two divergent 6S sequences, and one *Magnetococcus* homolog is duplicated.

6.8 Growth phase dependent expression of *B. subtilis* 6S RNAs

We wondered how encoding two copies of 6S RNA could benefit some bacteria enough to be preserved during evolution. It had been previously reported that *B. subtilis* 6Sa RNA levels dramatically decrease during saturating growth after fresh inoculation [9].

Since this pattern is opposite the normal increase in *E. coli* 6S RNA during stationary phase, and the timing of *B. subtilis* 6Sb expression was unknown, we probed Northern blots of total RNA isolated after different intervals of growth for both 6S homologs (Figure 6.6).

The total levels of 6Sb RNA increase ~18-fold between early log and stationary phase growth (Figure 6.6*B*). Precursor 6Sb-201 RNA transiently accumulates relative to processed 190 nt RNA (6Sb-190) as overall 6Sa levels increase, peaking at 60% of the total 6Sb RNA in mid-log phase under these growth conditions [273]. In lag phase cells recovering from stationary phase in the culture used for inoculation (1 hr time point) more than 90% of the 6Sb RNA has been cleaved to 190 nt. We observe a peak in 6Sa expression during mid-log phase where 2-3 times as much RNA is present as in early log phase under our growth conditions (Figure 6.6*C*). After this point, the previously reported decrease in 6Sa RNA levels occurs, and stationary phase 6Sa RNA levels are reduced to at most one-eighth of the mid-log peak. Regardless, 6Sb RNA is the major 6S RNA species in *B. subtilis*. At the peak of 6Sa expression (5 hr) there is still roughly twice as much 6Sb RNA as determined by ethidium bromide staining of gels (data not shown).

We conclude that *B. subtilis* 6Sb RNA is the ortholog of *E. coli* 6S RNA. *B. subtilis* 6Sa has functionally diverged at least with respect to the timing of its expression during growth, perhaps to more finely tune the transcriptional response to the approach of nutrient limitation.

6.9 Conservation of a 6S RNA-*ygfA* operon

There is experimental evidence that *E. coli* 6S RNA is rapidly cleaved by an unknown mechanism from the 5' end of a transcript that includes the coding region for the *ygfA*

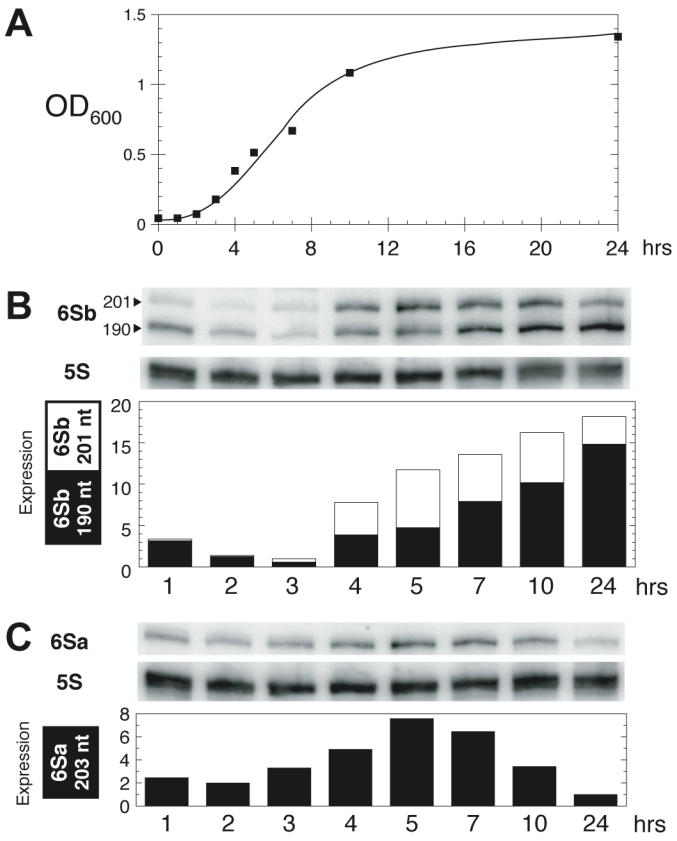


Figure 6.6 Expression of *B. subtilis* 6S RNAs during growth

(A) Growth curve for cultures from which total RNA was extracted to measure 6S RNA abundance. ***(B)*** Expression of 6Sb RNA. Northern blots were hybridized with radiolabeled probes specific for this 6S RNA and 5S RNA. Band intensities were quantitated, corrected for 5S RNA loading controls, and normalized to the 3 hr time point. Levels of the 201 nt precursor and 190 nt processed 6Sb RNA bands are displayed as unfilled and filled bars, respectively. ***(C)*** Expression of 6Sa RNA. Details as in (*B*) except RNA levels were normalized to the 24 hr time point.

gene [128]. We noticed that despite the marked divergence between the α- and γ-proteobacterial 6S RNA sequences there is a widespread occurrence of *ygfA* homologs directly downstream of *ssrS* in both groups as well as some β-proteobacteria (Figure 6.5). In other bacterial genomes there is no apparent conservation of the genes found adjacent to 6S RNA homologs. The conserved juxtaposition of 6S RNA and *ygfA* in *E. coli* and other proteobacteria implies that it has functional relevance, most likely as a way of linking *ygfA* and 6S RNA expression. It is interesting in this respect that microarray experiments indicate that *ygfA* expression increases 5- to 8-fold in *E. coli* cells growing as biofilms [231], which represent another condition where poor nutrients availability may limit growth.

YgfA proteins share sequence similarity with eukaryotic methenyltetrahydrofolate synthetases (MTHFSs). These enzymes have been implicated in folate degradation and convert 5-formyltetrahydrofolate, which is believed to be a stable storage form of reduced folate, into 5,10-methenyltetrahydrofolate. In human cell culture, increased MTHFS activity correlates with a decrease in cellular folate pools that cannot be overcome by increased folate concentrations in the growth medium [10]. Since folate derivatives shuttle one-carbon units from degradative pathways into the synthesis of key metabolic intermediates such as purines, thymidylate, SAM, and formylmethionine-tRNA, depletion of folate may be a way for cells to globally restrict metabolic flux. If MTHFS serves a similar function in prokaryotic folate regulation, then many proteobacteria may adapt to nutrient limitation in stationary phase by the concerted expression of YgfA and 6S RNA to slow one-carbon metabolism and RNA transcription.

6.10 Comparison of 6S RNA homologs

Previously studied noncoding RNAs of unknown function that we have recognized as 6S RNA homologs appear to differ in some respects from *E. coli* 6S RNA. Preliminary

experiments have investigated the expression, processing, and dispensability of *B. subtilis* 6Sa RNA [273] *B. subtilis* 6Sb RNA [9], and *Synechococcus* sp. PCC6301 6S RNA [309]. Notably, *Synechococcus* 6S RNA is abundant during exponential growth and later decreases in stationary phase. This timing is at odds with the normal regulation of *E. coli* 6S RNA and our observation that although *B. subtilis* expresses 6Sa RNA with an unusual timing, expression of the more abundant 6Sb RNA increases in stationary phase. A more detailed analysis will be necessary in all cases to determine whether regulation of other 6S RNAs is accomplished by producing transcripts from two promoter sites as in *E. coli* [146]. No growth defects have been detected in deletion mutants of *B. subtilis* 6Sa and *Synechococcus* 6S RNA under conditions where *E. coli* 6S RNA knockouts grow normally. In contrast, growth of *B. subtilis* 6Sb RNA deletion mutants is compromised during exponential phase, and mutant cultures are unable to reach densities as high as wild-type cultures during stationary phase. This is the first known instance where a defective 6S RNA has been reported to result in a phenotype that is potentially useful for genetic studies.

The diversity of observed 6S RNA processing suggests that it is not important for functional maturation. Rather, exonuclease trimming of neighboring unstructured RNA regions up to the stable 6S closing stem could be incidental, as this mechanism is common for other stable bacterial RNAs [168]. RNase E and/or RNase G cleave *E. coli* 6S RNA at its 5' end to produce a mixture of mature 6S RNA sequences with 5' ends at positions –1, +1, and +2 [146], and cleavage by an uncharacterized mechanism liberates its 3' end from the 5' untranslated region of an mRNA encoding the YgfA protein [128]. *B. subtilis* 6Sb RNA accumulates as a 201 nt transcript that is slowly cleaved to 190 nt by the loss of 11 nt from the 5' end by an unknown RNase. Only the 185 nt version of *Synechococcus* 6S RNA has been observed. Similarly, no intermediates have been observed for *B. subtilis* 6Sa, although it atypically retains an extra 3' stem-loop after its

closing stem. This hairpin is probably the remnant of an intrinsic transcription terminator, and we commonly observe terminators directly downstream of 6S RNA sequences in certain groups of bacteria (data not shown).

6.11 Conclusions

The absence of a strong deletion phenotype and lack of comparative sequence information have historically been obstacles to understanding the function of *E. coli* 6S RNA. We have shown that 6S RNA is an ancient and conserved regulator of RNA polymerase function. Only a small number of noncoding RNAs with critical cellular roles, including rRNAs, 5S RNA, tRNA, RNase P RNA, SRP RNA, tmRNA, and some riboswitches, are as widely distributed as 6S RNA across different bacterial groups. An analysis of 6S RNA sequences suggests more detailed models for how it might mimic an open DNA promoter. It also raises new questions about the purposes of multiple 6S homologs in some genomes and the significance of conserved cotranscription with downstream methylenetetrahydrofolate synthetase genes. The recognition of a 6S RNA homolog in *B. subtilis* whose deletion has been reported to cause a dramatic growth defect and knowledge of conserved regions within the structure of 6S RNA should greatly enable future genetic and molecular studies of its interactions with the transcriptional machinery.

Shortly after this study was published [20], two other groups separately reported discovering 6S RNAs in other bacterial species. The first group sequenced an abundant band in total RNA from the β-protoeobacterium *Bordetella pertussis* and found sequence homology to *E. coli* 6S RNA [285]. They then demonstrated that RNAs of a similar size from *B. subtilis* were also 6S homologs by coimmunoprecipitating them with antibodies to its housekeeping sigma factor (σ^A) or the α subunit of its RNAP. They were able to locate further 6S RNA homologs in Gram-positive species using these sequences as

BLAST queries. In line with our structural probing, they directly demonstrated that mutations that dramatically change the size of the central single-stranded bubble compromise 6S RNA function *in vivo* and *in vitro*. However, small insertions in the central bubble strands, and even a mutation that swaps the entire sequence from one side of the bubble to the other, appear to introduce only minor functional defects. The other group discovered a 6S RNA homolog in *Aquifex aeolicus* during large-scale small RNA cloning efforts aimed at finding the RNase P homolog in this species [320]. Interestingly, a majority of this 6S RNA appears to be nicked within the 5' side of the central bubble. The *B. pertussis* and *A. aeolicus* 6S RNA sequences isolated in these studies are identical to the ones we discovered with covariance model searches.

More recently it has been reported that 6S RNA limits the cellular response to elevated pH during stationary phase in *E. coli* by damping induction of the transcription factor *pspF* [286]. Known defects in *E. coli* 6S RNA null mutants are very subtle, and this is the first time that the presence of 6S has been linked to the expression of a specific gene that can affect survival. It appears contradictory that Δ6S mutants survive *better* than wild-type *E. coli* at elevated pH in this laboratory test (because 6S inhibits PspF expression and consequently the induction of factors for high pH survival). Current thinking is that, in more complex natural settings, 6S RNA must balance this bacterium's response to competing stresses and thereby increase its persistence.

6.12 Methods

Covariance model searches

Manual multiple alignments of 6S RNA sequences were used to construct covariance models [67] using the Infernal software package [65]. Filtering techniques were applied to accelerate searches of covariance models against sequence databases [311, 312].

Score thresholds that reliably predicted new 6S RNA homologs were determined by scoring the input sequences and examining marginally scoring matches for false-positives that overlapped conserved genes. A complete alignment of all 6S RNA homologs was generated by using Infernal to automatically align reliable matches to a covariance model trained on a seed alignment of 121 sequences. Sequences in the seed alignment were weighted before calculating the reported consensus sequence, length, and composition statistics to reduce biases from similar sequences with Infernal's internal implementation of the GSC algorithm [87].

Phylogenetic tree

We created a covariance model from the final seed alignment with 133 manually annotated consensus columns, encompassing conserved stems and the central bubble, and used it to automatically re-align this set of 121 sequences. The same consensus columns were extracted from the new alignment and input into the "dnadist" and "fitch" programs from the PHYLIP software package [74] to create an unrooted phylogenetic tree of 6S RNA homologs using the Jukes-Cantor distance method and all other parameters set to their default values .

In-line probing

DNA templates for the *in vitro* transcription of 6S RNAs were amplified by whole-cell PCR from *E. coli* strain MG1655 and *B. subtilis* strain 168 (BGSC #1A1; Bacillus Genetic Stock Center, Columbus, OH). Details of the in-line probing analysis have been reported elsewhere [260].

Northern Blotting

B. subtilis 168 was grown at 37°C in Difco™ nutrient broth (Becton, Dickinson and Company, Sparks, MD) starting from an overnight culture diluted to an initial OD_{600} of

0.02. At each time point 3 OD_{600} of cells were collected and stored at –80°C. Cell pellets were resuspended in 100 µl of 4 mg/ml lysozyme in TE buffer (10 mM Tris-HCl, pH 7.5 @ 25°C; 1 mM EDTA) and incubated for 10 minutes at 25°C before isolating total RNA with 1 ml of TRIzol reagent (Invitrogen) according to the manufacturer's instructions. Total RNA samples (2 µg) were heated at 90°C for 2 minutes in 1 x gel loading buffer (45 mM Tris-borate, 4 M urea, 10% sucrose (w/v), 5 mM EDTA, 0.05% SDS, 0.025% xylene cyanol FF, 0.025% bromophenol blue), separated on a denaturing 10% (8M urea) polyacrylamide gel, and transferred overnight to a nylon Hybond-N+ membrane (Amersham Biosciences). Blots were simultaneously probed at 37°C with 5' [^{32}P]-labeled oligonucleotides specific for *B. subtilis* 5S RNA (5'-AACGGGTGTGACCTCTTCGCT-ATCGCCA) and 6Sa RNA (5'- CGCTACGTCTTGCCGTATGCAAGTAAGAAA), or 5S RNA and 6Sb RNA (5'- TTCCTTTGTTTTGAACCCGCTCTCAGCAGG) in Rapid-hyb buffer (Amersham Biosciences) and analyzed with a PhosphorImager (Molecular Dynamics).

7 Conclusions and future directions

7.1 Introduction

The term "riboswitch" has gained a life of its own in scientific publications and on the internet in the last few years. Since improper and unclear examples of usage have appeared, we endeavor to explain its intended definition. While on this topic, it is important to introduce two other types of riboswitches that sense divalent cation concentrations and temperature. It is likely that new complex regulatory RNA structures will continue to be discovered in microbial genomes as additional sequences from underrepresented groups of bacteria and more sensitive motif detection algorithms become available. There is also considerable excitement that the large number of conserved elements that are being identified within introns and UTRs in eukaryotic genomes may harbor riboswitches or new kinds of structured *cis*-regulatory RNA elements. Finally, the complexity of riboswitch mechanisms and the widespread distribution of some classes in many contemporary genomes suggests that they may be derived from ancient metabolic ribozymes or regulatory sequences that existed in RNA World proto-organisms.

7.2 Definition and usage of the term "riboswitch"

The term "riboswitch" has been used to describe several different kinds of RNAs involved in genetic control since its first appearance in print [322]. It was originally coined to describe an mRNA sequence that is able to proactively sense a small molecule metabolite or a cellular condition and change its conformation in a way that affects gene expression without the involvement of other cellular factors. This designation is in contrast, for example, to an mRNA leader whose structure changes in response to binding of a regulatory protein that senses a small molecule. Here the RNA is only a

passive recipient of regulation rather than the master of its own fate. At the risk of appearing overly pedantic, we note there is more accurately an indirect switch in ribonucleic acid conformation rather than a "riboswitch" (as has been reported) during glucose-dependent regulation of an intrinsic transcription terminator by the RNA-binding antitermination protein GlcT [248, 249]. Granted, all riboswitches are dependent on RNA polymerase or ribosomes, but these macromolecular machines respond to riboswitch structural changes rather than determining them.

Several types of engineered RNA molecules that respond to exogenous small molecules have been properly referred to as synthetic riboswitches. For example, the insertion of theophylline aptamers in the 5' UTRs of reporter genes has been used create systems where the addition of theophylline either activates [58] or represses [270] translation in bacteria. A tetracycline-binding aptamer has been used to create an artificial riboswitch where ligand binding inhibits translation initiation in yeast [115], and pre-mRNA splicing can be controlled by riboswitches constructed out of theophylline aptamers overlapping the 3' splice junction in HeLa cell extract [145]. If allosteric hammerhead ribozymes that respond to small molecules [261] can be made to function in mammalian tissue culture and animal models [330], they too would be examples of synthetic riboswitches.

There are other important examples of natural noncoding RNA elements that function as riboswitches but do not respond to organic small molecules (see Section 7.3 below). However, very similar bacterial and synthetic regulatory systems should , strictly speaking, be disqualified from carrying the label. T-boxes are highly structured leader mRNA elements common in Gram-positive bacteria that regulate transcription attenuation with a mechanism similar to many riboswitches [104]. However, T-boxes rely on an auxiliary factor to sense amino acid availability — they bind to uncharged tRNA. Similarly, synthetic oligonucleotide-dependent hammerhead ribozymes are not

conventional riboswitches because they also sense the presence of a second macromolecule [217].

Do riboswitches truly operate like electronic "switches" at the molecular level? This question has recently been raised in the context of (1) observations that some riboswitch aptamers are unable to structurally rearrange and bind their target molecules *in vitro* when they have been pre-folded in the absence of ligand and (2) work showing that riboswitch responses depend on the kinetics of transcription and binding "on" rates rather than the equilibrium strength of binding. Thus, a riboswitch in an individual mRNA molecule that is being synthesized by RNA polymerase may act more like an electronic "fuse". That is, it might make a one-time decision to follow one of two alternative folding pathways depending on whether ligand binds as its aptamer sequence is synthesized by RNA polymerase. This view holds that a "ribofuse" will either trigger a transcription termination event or kinetically trap the mRNA leader in a state where the RBS is sequestered for its entire lifetime.

It is probably premature to conclude that all riboswitches make one-time co-transcriptional decisions that determine whether protein is ever expressed from an individual mRNA molecule. Particularly, there may be examples where an existing mRNA can be reactivated for translation after ligand dissociates. Even if this point of view is strictly correct at the stochastic single-molecule level, riboswitches function as "switches" at the bulk genetic level where increasing concentrations of ligand progressively turn downstream protein expression ON or OFF. In short, we believe that "riboswitch" is still the most succinct, accurate, and useful term for these autonomous RNA regulatory elements when it is applied consistently.

7.3 Other kinds of natural riboswitches

Although our work has concentrated on metabolite-binding regulatory RNA motifs, riboswitches that sense temperature and Mg^{2+} cations have also been reported. Several different RNA thermosensor families that are thought to operate autonomously have been characterized in a variety of bacterial groups [205]. One motif occurs within the amino acid coding sequences of *rpoH* genes in *E. coli* and other enterobacteria. This gene encodes σ^{32} (the heat-shock sigma factor), and the RNA element represses expression of this master regulatory protein at normal temperatures by forming a structure that prevents translational initiation. At elevated temperatures it unfolds to activate translation [203]. ROSE (Repression Of heat-Shock gene Expression) elements are noncoding mRNA leader motifs found in the 5' UTRs of multiple heat-shock genes. They were first described in *Bradyrhizobium japonicum* where they consist of four characteristic hairpins. Hairpins II, III, and IV appear to be present in ROSE-like motifs found in other α- and γ-proteobacterial instances. However, only the sequence of the short hairpin IV that sequesters the ribosome binding site and start codon seems to be widely conserved. Melting of this structure beginning at a characteristic bulged G has been demonstrated *in vitro* at physiological temperatures. Other possible thermosensing RNA motifs have been described upstream of the *Caulobacter crescentus dnaKJ* operon, the *Haemophilus ducreyi dnaK* ORF, and the Gram-positive *Streptomyces albus hsp18* gene.

A recently reported Mg^{2+}-sensing riboswitch that occurs in the 5' UTR of the *mgtA* gene in *E. coli* appears to be restricted to enterobacteria [52]. High Mg^{2+} levels have been proposed to preferentially stabilize a leader mRNA conformation that includes two upstream stem-loops on the basis of structural probing results. The second stem-loop appears to function as an unusual transcription termination site, based on *in vitro*

transcription data, to repress gene expression. A mutually exclusive antiterminator structure that overlaps both stem-loops is thought to form under low Mg^{2+} conditions to allow readthrough. Although the arrangement of these helical elements is conserved, the sequences that comprise putative Mg^{2+} riboswitches are quite different even among representatives from closely allied species.

From a molecular recognition and evolutionary standpoint, it is interesting to note that none of these gene control solutions is as conserved across very divergent organisms as a typical metabolite-binding riboswitch. To some extent *any* RNA structure changes in responsive to elevated temperatures or divalent cation concentrations. Therefore, it may be relatively easy for evolution to discover new sequence motifs that alter their conformation in response to these stresses in a way that usefully modulates the expression of nearby genes. This is one explanation for the narrow distribution and low sequence conservation of the *E. coli* Mg^{2+} riboswitch and most reported thermosensors (except perhaps the ROSE element).

We think that it is surprising that a natural RNA structure that responds to pH has not yet been discovered. Artificial RNAs that sense pH are common accidental outcomes of *in vitro* selections for allosteric hammerhead ribozymes triggered by small molecules in our laboratory when effector solutions are not properly buffered at neutral pH [150]. Some of the regulatory motifs of unknown function that we describe from *B. subtilis* could plausibly respond to pH changes, cation levels, or osmotic shock based on the predicted functions of downstream genes. Compared to the known examples of natural riboswitches that do not bind metabolites, these motifs seem to have unnecessarily complex RNA structures if their sole purposes are to sense ion and solute conditions. It is also possible that bacteria keep intracellular pH relatively constant using other, more rapid, mechanisms such that an internal riboswitch for sensing pH changes and altering gene expression is not needed.

7.4 Prospects for discovering more regulatory RNA motifs in bacteria

More sophisticated and thorough approaches that build on the comparative genomics methods described in Chapters 1 and 2 promise to identify still further *cis*-regulatory RNA motifs in bacteria. It seems clear that the possibilities for finding additional widespread riboswitches that occur in most bacterial groups and have relatively elaborate aptamer domains (e.g. TPP and AdoCbl) have been exhausted. The most recently discovered metabolite-binding riboswitches (SAM-II, S_{MK}, and $preQ_1$) have smaller aptamer domains and are restricted to narrower taxonomic distributions. Therefore, future approaches must concentrate on including more comparative information from specific bacterial groups and utilize methods that detect smaller conserved RNA structures.

One exciting development has been the creation of the covariance model motif-finding algorithm CMfinder, which predicts common RNA structures from unaligned sequences [328]. We have recently incorporated CMfinder into a computational pipeline much like the comparative approach that used MEME to identify conservation upstream of COGs in a diverse set of bacterial genomes [4]. We predict functional classifications with the Conserved Domain Database for all proteins in microbial genomes from a specific taxonomic group (e.g. actinobacteria or β-proteobacteria) and collect intergenic regions upstream of all proteins classified into each conserved domain in this clade. We use CMfinder to predict RNA motifs in these unaligned upstream sequences and then scan the entire set of IGRs for additional hits with RaveNnA to unite regulons and flag repeat elements. A final iteration of CMfinder on the resulting hits produces a refined secondary structure prediction. We display the genome contexts of each of these candidate RNA regulatory elements on web pages that allow annotation in a format similar to the BLISS database.

In a pilot study of low-G+C Gram-positive species this process identified most known riboswitches and orphan regulatory motifs that we have previously encountered (Z. Yao,, W.L. Ruzzo, Z. Weinberg, J.E. Barrick, R.R. Breaker, S. Neph, M. Tompa, unpublished results). This procedure makes the task of computationally investigating these motifs substantially easier compared to beginning with BLAST or MEME alignments because it predicts a set of most likely secondary structures that is quite accurate in most cases. In fact, we only needed to slightly manually adjust the computational alignments produced for two new putative regulatory structures that are conserved in ribosomal protein leaders within this taxonomic group before they were ready for submission to the Rfam database as a proof of principle.

Extension of this work to other bacterial groups has identified (1) a circularly permuted version of a known riboswitch that primarily appears in a different bacterial division than the original, (2) a new riboswitch aptamer that binds a very close molecular analogue of a metabolite recognized by a known riboswitch class, and, most surprisingly, (3) a new riboswitch that occurs in the *E. coli* genome with a conserved structure that includes a GAAA tetraloop and a tetraloop receptor. There are not yet sufficient numbers of completed genome sequences from many bacterial groups to expect any comparative method to have discovered lineage-specific *cis*-regulatory RNA motifs in these groups. It will be interesting to see what new "boutique" riboswitches or variants of known riboswitches these organisms might harbor and if yet more riboswitches remain undetected in model organism genomes.

7.5 Prospects for discovering eukaryotic riboswitches

Thiamine pyrophosphate riboswitches appear to control splicing of UTR introns in plants and fungi [154, 268]. This is the only riboswitch known to occur in eukaryotes. Arguments can be made both for and against whether a greater role for riboswitch-

mediated regulation should be expected in higher organisms. Eukaryotic organisms contain vast untranslated intron and UTR sequences that could provide more raw material for the evolution of new regulatory mechanisms and aptamers. The nuclear membrane contains pores that are large enough for small molecules to freely diffuse inside and interact with nascent RNA as it is transcribed and spliced. However, metabolic self-sufficiency and efficiency is probably less important in plant and animal cells than in bacteria. Specifically, eukaryotic organisms lack metabolic pathways for the *de novo* synthesis of many coenzymes or have dispensed with enzymes that require these expensive nutrients, and their cells are typically bathed in a relatively constant nutrient environment by a circulatory system. Plants do not use AdoCbl as a cofactor, for example [240]. On the other hand, plants and animals possess their own repertoires of small molecule hormones that could be sensed by undiscovered riboswitches.

Presently, one of the mysteries of eukaryotic genomes is the function of many nongenic sequences that are highly conserved throughout metazoans. Sometimes these elements are referred to as genomic "dark matter". They range from "ultraconserved elements" that extend for hundreds of nucleotides without indels [23] to "highly conserved elements" that extend over larger regions and are conserved between diverse organisms [256]. These elements can be detected and enumerated in many ways. For example, pair-wise BLASTZ comparisons allow the alignment of syntenic regions in two genomes [250], and MultiZ implements a threaded blockset aligner that simultaneously produces alignments of homologous regions from multiple genomes [28]. The PhastCons program uses a phylogenetic HMM that models the expected divergence in sequences from a species tree to score regions that are under purifying selection in a multiple sequence alignment created by MultiZ [256]. What is the significance of these elements? Some appear to cluster around developmental regulatory genes and may represent clusters of binding sites for proteins that regulate transcription [244]. However,

a common conclusion is that, in bulk, stable RNA structures are overrepresented in genomic dark matter.

Recently, two research groups have predicted functional RNA sequences in the highly conserved regions identified by PhastCons. The program RNAz integrates thermodynamic structure predictions and mutational evidence of covariation to attempt to separate true RNA structures from statistical noise [304, 305]. EvoFold continues in the intellectual footsteps of PhastCons and uses a phylogenetic stochastic context-free grammar (similar to a covariance model) to predict alignment regions that are most compatible with a noncoding RNA structure [215]. Both procedures recover known miRNAs, histone 3' UTR stem-loops, snoRNAs, tRNAs, and other noncoding RNAs in the human genome. They also predict 1000's of candidate structured RNA elements of varying complexity that have been mapped onto human chromosomes with the UCSC genome browser [124]. However, RNAz has been estimated to produce 20-30% false positive predictions, and 40-75% of EvoFold predictions may be attributable to chance depending on their lengths and background assumptions. In both cases the challenge for further characterizing the function of these putative RNAs is that (by design) there are relatively few mutations in these elements and consequently very little indication as to which of their conserved features are essential.

It is not obvious why functional RNAs should have to maintain a high level of primary sequence conservation. Recent efforts have shown (as expected) that looking at unaligned sequences with lower identity adjacent to sequence blocks that are alignable based on primary sequence between human and mouse elements reveals even more RNA structures [282]. This study used the program FOLDALIGN [117] to score common RNA structures in a pair of unaligned sequences. They found 1300 additional candidates for structured RNAs in a survey of ten human chromosomes, about half of which are estimated to be false positives.

In order to specifically target eukaryotic riboswitches, FOLDALIGN or CMfinder might be used to search introns, 5' UTRs, and 3' UTRs of orthologous genes from diverse organisms for common RNA structures. This computational pipeline would be directly analogous to how we have discovered *cis*-regulatory RNA elements in microbial genomes. However, preliminary work indicates that eukaryotes will present a new set of challenges. Not only will it be necessary to assemble a catalogue of new "irrelevant" explanations for such conservation like the one we outlined in Chapter 1 for microbes. Genome annotation information currently is far more rudimentary in many eukaryotes (e.g. many introns/exons are not annotated accurately), and fewer phylogenetically well-spaced genomes are available. Perhaps sequencing genomes from basal vertebrates and more diverged species will aid future efforts to define highly structured regulatory RNA elements that may lurk in the "junk" DNA of these organisms.

7.6 Riboswitches and the RNA World

Several lines of evidence support the existence of a primordial RNA World wherein RNA served a dual role as the primary genetic and functional molecule [88]. The RNA World hypothesis was initially proposed to explain why protein enzymes employ seemingly ubiquitous coenzyme molecules that contain pieces of RNA bases and/or sugar-phosphate backbone that are unrelated to their chemical reactivities [24, 25, 139, 219, 315]. The discovery of self-splicing introns, RNase P RNA, and natural self-cleaving ribozymes showed that RNA could catalyze chemical transformations. *In vitro* selections have explored the catalytic potential of RNA, sometimes with the explicitly stated goal of reproducing chemical activities that would be necessary for metabolism in an RNA World [21]. They have succeeded in (re)creating ribozymes capable of aminoacylating tRNA [159], accelerating peptide bond formation [332], catalyzing steps in nucleotide synthesis [291], and even acting as a processive RNA replicase [137].

Today we also understand that the ribosome's active site consists solely of RNA [206], that spliceosomal snRNAs are probably derived from group II self-splicing introns and can catalyze at least one step of splicing [292], and that telomerase RNA serves as a template for nucleotide polymerization [44]. All of these systems are evidence that proteins have gradually replaced RNA function in all but a few remaining pieces of legacy equipment. This ongoing nature of this process is most clearly illustrated by differences among RNase P RNAs. Whereas the *E. coli* RNA can catalyze pre-tRNA cleavage *in vitro* in the absence of its protein subunit under certain conditions, the archaeal and human RNA components seem to require several accessory proteins to display any catalytic activity [79]. The most extreme replacement of ancient ribozymes may be protein enzymes that only retain coenzyme molecules as evidence that an ancient ribozyme once catalyzed a reaction

Riboswitches may be another relic from the RNA World [33]. The widespread taxonomic distribution of certain classes of riboswitches like TPP and AdoCbl leaves little doubt that they were at least present in the last common ancestor of bacteria. Their structural complexity makes it highly unlikely that the same aptamer family evolved independently in each of these groups, and the observed cases of horizontal riboswitch transfer cannot account for their wide distribution. Pushing back even further in time, it is possible that riboswitches were present in the common ancestor of all kingdoms of life due to an RNA World origin but that they were subsequently lost from the genomes of "higher" organisms for some of the reasons described in Section 7.5. Riboswitches would certainly have filled an important functional niche in riboorganisms by providing sophisticated gene control elements, and it may be significant that most known riboswitches sense the very cofactor molecules that are also though to be molecular fossils from this evolutionary epoch.

In fact, it is plausible that contemporary riboswitches evolved from ancient SAM-synthesizing and SAM-utilizing ribozymes (Figure 7.1). Coenzymes may have first been trapped by a covalent chemical bond within the structure of a larger RNA molecule. Imagine a primitive ribozyme that was modified by attaching the amino acid methionine to a terminal adenosine. This nucleotide would have an identical chemical structure to the coenzyme *S*-adenosylmethionine (SAM). Current protein enzymes use SAM to transfer the active methyl group to another molecule during the biosynthesis of hemes or for methylating RNA bases in tRNA and rRNA. The rest of our RNA chain could fold into a complex three-dimensional structure that positioned this tethered coenzyme in an active site where it could accelerate similar metabolic reactions. After millions of years, the bond between this terminal SAM nucleotide and the ribozyme might no longer be necessary. SAM might now be recruited via a small RNA adaptor that could base pair to the ribozyme where it had formed a hairpin between its own strands before. Eventually, even these interactions would atrophy as the structure around the active site evolved to directly bind SAM. Now we have a arrived at a metabolic ribozyme that acts like current protein enzymes and very specifically recognizes and grasps a reactive coenzyme.

At some point a copy of this metabolic ribozyme might be co-opted for regulation. Nucleotides in its active site that were necessary for efficient catalytic activity might mutate and its activity would atrophy. It would no longer be a ribozyme, but it could still bind the coenzyme molecule. Eventually, other mutations would accumulate and a dead-ribozyme variant might fortuitously develop the capacity to change shape upon binding metabolite in a context where it regulated production of that metabolite. Proto-cells with this early riboswitch would out-replicate their neighbors because they more efficiently policed metabolic waste. Later, a protein enzyme might evolve that used the coenzyme to do the same chemical reaction. Now the remaining relatives of our proto-riboswitch

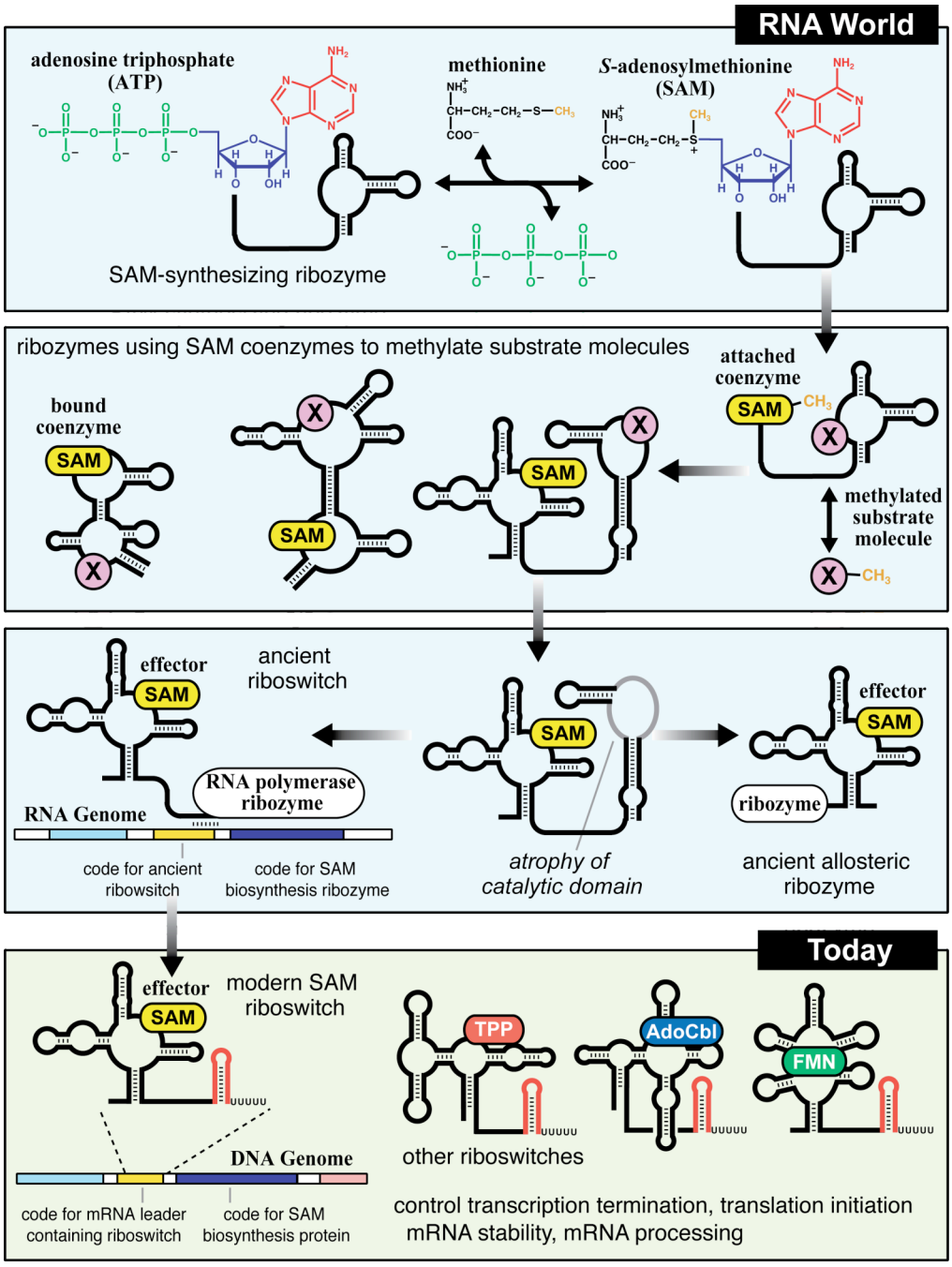


Figure 7.1 Possible pathway for the evolution of modern SAM riboswitches

This figure was adapted from [33]. Refer to the text for a detailed description.

that functioned as ribozymes would be relegated to the scrap heap. The ancestors of modern cells might maintain the riboswitch, and it could adapt to fulfill new regulatory roles as the current incarnations of RNA polymerase and the ribosome matured.

This is a hypothetical scenario, but we note that there are modern cases where metabolic protein enzymes have apparently been co-opted to perform gene control roles [113, 267]. Perhaps the most compelling data possible for this imagined origin would be resurrecting the catalytic activity of a riboswitch aptamer by randomizing its sequence and employing an *in vitro* selection procedure. This approach might be prone to the criticism that many riboswitch targets are highly reactive coenzyme molecules, and that a riboswitch-derived ribozyme might therefore be expected to only confer specificity toward certain reactants rather than accelerating the rate of the reaction by participating in the chemistry. However, to some extent, this is exactly how many modern protein enzymes employ these cofactors. One of the most exciting outcomes of this research program might be the discovery of a ribozyme that catalyzes a metabolic reaction in contemporary cells. We have found an extremely complex RNA structure in extremophilic organisms that may be a candidate for such a ribozyme [224]. This conserved "OLE RNA" element appears to be transcribed between two protein-coding genes as part of a longer mRNA molecule. The purpose of the operon that contains the OLE RNA motif is unknown, but it may be related to membrane lipid biogenesis.

It is also possible that some riboswitch classes are recent inventions: the fruits of selection experiments that are ongoing in every modern organism that makes thousands of different messenger RNA molecules. Genomes could be "filled with aptamers" that have only recently evolved from these unconstrained noncoding sequences and fortuitously been of use [93]. The existence of multiple SAM aptamers brings this possibility into sharp focus. It seem likely that some of the smaller and taxonomically restricted riboswitch classes like the S_{MK} box may have evolved more recently.

7.7 Conclusions

Riboswitches are widespread genetic control elements in contemporary bacteria that may be relics from an ancient RNA world. Despite the regulatory mysteries that have been solved by applying this new paradigm, where the messenger RNA is master of its own fate, there remains much to be understood about the mechanisms of genetic regulation in bacteria and eukaryotes. The recently appreciated prominence of miRNA and siRNA control of developmental processes in mammals and plants and gene control by small antisense RNAs in bacteria show that RNA can be a facile regulator simply through Watson-Crick base pairing when auxiliary protein machinery has evolved. However natural ribozymes and now riboswitches show that RNA's penchant for forming more complex structures on its own has not been wasted by modern organisms. With the ubiquity of RNA in gene expression processes and the expanses of unused RNA in eukaryotic genomes it is likely that there are still surprises in store for RNA biology.

Bibliography

[1] Manatee. The Institute for Genomic Research. (2006). http://manatee.sourceforge.net

[2] Mycoplasma Sequencing Project. Broad Institute of Harvard and MIT. (2006). http://www.broad.mit.edu

[3] Abreu-Goodger, C., and Merino, E. (2005). RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements. Nucleic Acids Res. *33*, W690-W692.

[4] Abreu-Goodger, C., Ontiveros-Palacios, N., Ciria, R., and Merino, E. (2004). Conserved regulatory motifs in bacteria: riboswitches and beyond. Trends Genet. *20*, 475-479.

[5] Allen, T. A., Von Kaenel, S., Goodrich, J. A., and Kugel, J. F. (2004). The SINE-encoded mouse B2 RNA represses mRNA transcription in response to heat shock. Nat. Struct. Mol. Biol. *11*, 816-821.

[6] Altman, S., Wesolowski, D., Guerrier-Takada, C., and Li, Y. (2005). RNase P cleaves transient structures in some riboswitches. Proc. Natl. Acad. Sci. USA *102*, 11284-11289.

[7] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic Local Alignment Search Tool. J. Mol. Biol. *215*, 403-410.

[8] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., 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.

[9] Ando, Y., Asari, S., Suzuma, S., Yamane, K., and Nakamura, K. (2002). Expression of a small RNA, BS203 RNA, from the *yocI-yocJ* intergenic region of the *Bacillus subtilis* genome. FEMS Microbiol. Lett. *207*, 29-33.

[10] Anguera, M. C., Suh, J. R., Ghandour, H., Nasrallah, I. M., Selhub, J., and Stover, P. J. (2003). Methenyltetrahydrofolate synthetase regulates folate turnover and accumulation. J. Biol. Chem. *278*, 29856-29862.

[11] Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E. G. H., Margalit, H., and Altuvia, S. (2001). Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. Curr. Biol. *11*, 941-950.

[12] Babitzke, P. (2004). Regulation of transcription attenuation and translation initiation by allosteric control of an RNA-binding protein: the *Bacillus subtilis* TRAP protein. Curr. Opin. Microbiol. *7*, 132-139.

[13] Bachellier, S., Clement, J. M., and Hofnung, M. (1999). Short palindromic repetitive DNA elements in enterobacteria: a survey. Res. Microbiol. *150*, 627-639.

[14] Baichoo, N., and Helmann, J. D. (2002). Recognition of DNA by Fur: a reinterpretation of the Fur box consensus sequence. J. Bacteriol. *184*, 5826-5832.

[15] Baichoo, N., Wang, T., Ye, R., and Helmann, J. D. (2002). Global analysis of the *Bacillus subtilis* Fur regulon and the iron starvation stimulon. Mol. Microbiol. *45*, 1613-1629.

[16] Bailey, T. L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. *2*, 28-36.

[17] Bailey, T. L., and Gribskov, M. (1998). Combining evidence using p-values: application to sequence homology searches. Bioinformatics *14*, 48-54.

[18] Bao, Z., and Eddy, S. R. (2002). Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. *12*, 1269-1276.

[19] Barrick, J. E., Corbino, K. A., Winkler, W. C., Nahvi, A., Mandal, M., Collins, J., Lee, M., Roth, A., Sudarsan, N., Jona, I.*, et al.* (2004). New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. Proc. Natl. Acad. Sci. USA *101*, 6421-6426.

[20] Barrick, J. E., Sudarsan, N., Weinberg, Z., Ruzzo, W. L., and Breaker, R. R. (2005). 6S RNA is a widespread regulator of eubacterial RNA polymerase that resembles an open promoter. RNA *11*, 774-784.

[21] Bartel, D. P., and Unrau, P. J. (1999). Constructing an RNA World. Trends. Biol. Sci. *24*, M9-M13.

[22] Batey, R. T., Gilbert, S. D., and Montange, R. K. (2004). Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. Nature *432*, 411-415.

[23] Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., and Haussler, D. (2004). Ultraconserved elements in the human genome. Science *304*, 1321-1325.

[24] Benner, S. A., and Ellington, A. D. (1991). RNA World. Science *252*, 1232-1232.

[25] Benner, S. A., Ellington, A. D., and Tauer, A. (1989). Modern metabolism as a palimpsest of the RNA World. Proc. Natl. Acad. Sci. USA *86*, 7054-7058.

[26] Bik, E. M., Gouw, R. D., and Mooi, F. R. (1996). DNA fingerprinting of *Vibrio cholerae* strains with a novel insertion sequence element: a tool to identify epidemic strains. J. Clin. Microbiol. *34*, 1453-1461.

[27] Birney, E., Andrews, D., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T.*, et al.* (2006). Ensembl 2006. Nucleic Acids Res. *34*, D556-561.

[28] Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D.*, et al.* (2004). Aligning

multiple genomic sequences with the threaded blockset aligner. Genome Res. *14*, 708-715.

[29] Bompfunewerer, A. F., Flamm, C., Fried, C., Fritzsch, G., Hofacker, I. L., Lehmann, J., Missal, K., Mosig, A., Muller, B., Prohaska, S. J., *et al.* (2005). Evolutionary patterns of non-coding RNAs. Theor. Biosci. *123*, 301-369.

[30] Borchardt, R. T., Huber, J. A., and Wu, Y. S. (1974). Potential inhibitor of *S*-adenosylmethionine-dependent methyltransferases. 2. Modification of the base portion of *S*-adenosylhomocysteine. J. Med. Chem. *17*, 868-873.

[31] Borchardt, R. T., and Wu, Y. S. (1975). Potential inhibitors of *S*-adenosylmethionine-dependent methyltransferases. 3. Modifications of the sugar portion of *S*-adenosylhomocysteine. J. Med. Chem. *18*, 300-304.

[32] Borchardt, R. T., and Wu, Y. S. (1976). Potential inhibitors of *S*-adenosylmethionine-dependent methyltransferases. 5. Role of the asymmetric sulfonium pole in the enzymatic binding of *S*-adenosyl-L-methionine. J. Med. Chem. *19*, 1099-1103.

[33] Breaker, R. R. (2005). Riboswitches and the RNA World. In The RNA World, R. F. Gesteland, T. R. Cech, and J. F. Atkins, eds. (Woodbury, NY, Cold Spring Harbor Laboratory Press), pp. 89-108.

[34] Brosius, J. (1996). More *Haemophilus* and *Mycoplasma* genes. Science *271*, 1302-1302.

[35] Brownlee, G. G. (1971). Sequence of 6S RNA of *E. Coli*. Nature New Biol. *229*, 147-149.

[36] Brunel, C., and Romby, P. (2000). Probing RNA structure and RNA-ligand complexes with chemical probes. Method Enzymol. *318*, 3-21.

[37] Burden, S., Lin, Y. X., and Zhang, R. (2005). Improving promoter prediction for the NNPP2.2 algorithm: a case study using *Escherichia coli* DNA sequences. Bioinformatics *21*, 601-607.

[38] Burgstaller, P., and Famulok, M. (1994). Isolation of RNA aptamers for biological cofactors by *in vitro* selection. Angew. Chem. Int. Edit. Engl. *33*, 1084-1087.

[39] Campbell, J. W., Morgan-Kiss, R. M., and Cronan, J. E., Jr. (2003). A new *Escherichia coli* metabolic competency: growth on fatty acids by a novel anaerobic beta-oxidation pathway. Mol. Microbiol. *47*, 793-805.

[40] Cannone, J. J., Subramanian, S., Schnare, M. N., Collett, J. R., D'Souza, L. M., Du, Y., Feng, B., Lin, N., Madabusi, L. V., Muller, K. M.*, 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.

[41] Carothers, J. M., Oestreich, S. C., Davis, J. H., and Szostak, J. W. (2004). Informational complexity and functional activity of RNA structures. J. Am. Chem. Soc. *126*, 5130-5137.

[42] Cate, J. H., Gooding, A. R., Podell, E., Zhou, K. H., Golden, B. L., Kundrot, C. E., Cech, T. R., and Doudna, J. A. (1996). Crystal structure of a group I ribozyme domain: principles of RNA packing. Science *273*, 1678-1685.

[43] Cech, T. R. (1990). Self-splicing of group I introns. Annu. Rev. Biochem. *59*, 543-568.

[44] Cech, T. R., Nakamura, T. M., and Lingner, J. (1997). Telomerase is a true reverse transcriptase. A review. Biochemistry (Moscow) *62*, 1202-1205.

[45] Chaturvedi, S., and Bhakuni, V. (2003). Unusual structural, functional, and stability properties of serine hydroxymethyltransferase from *Mycobacterium tuberculosis*. J. Biol. Chem. *278*, 40793-40805.

[46] Chen, S. L., and Shapiro, L. (2003). Identification of long intergenic repeat sequences associated with DNA methylation sites in *Caulobacter crescentus* and other alpha-proteobacteria. J. Bacteriol. *185*, 4997-5002.

[47] Cheo, D. L., Bayles, K. W., and Yasbin, R. E. (1991). Cloning and characterization of DNA damage-inducible promoter regions from *Bacillus subtilis*. J. Bacteriol. *173*, 1696-1703.

[48] Christiansen, L. C., Schou, S., Nygaard, P., and Saxild, H. H. (1997). Xanthine metabolism in *Bacillus subtilis*: characterization of the *xpt-pbuX* operon and evidence for purine- and nitrogen-controlled expression of genes involved in xanthine salvage and catabolism. J. Bacteriol. *179*, 2540-2550.

[49] Ciria, R., Abreu-Goodger, C., Morett, E., and Merino, E. (2004). GeConT: gene context analysis. Bioinformatics *20*, 2307-2308.

[50] Cole, J. R., Chai, B., Marsh, T. L., Farris, R. J., Wang, Q., Kulam, S. A., Chandra, S., McGarrell, D. M., Schmidt, T. M., Garrity, G. M., and Tiedje, J. M. (2003). The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. Nucleic Acids Res. *31*, 442-443.

[51] Corbino, K. A., Barrick, J. E., Lim, J., Welz, R., Tucker, B. J., Puskarz, I., Mandal, M., Rudnick, N. D., and Breaker, R. R. (2005). Evidence for a second class of *S*-adenosylmethionine riboswitches and other regulatory RNA motifs in alpha-proteobacteria. Genome Biol. *6*, R70.

[52] Cromie, M. J., Shi, Y., Latifi, T., and Groisman, E. A. (2006). An RNA sensor for intracellular Mg^{2+}. Cell *125*, 71-84.

[53] D'Ascenzo, M. D., Collmer, A., and Martin, G. B. (2004). PeerGAD: a peer-review-based and community-centric web application for viewing and annotating prokaryotic genome sequences. Nucleic Acids Res. *32*, 3124-3135.

[54] de Hoon, M. J., Makita, Y., Nakai, K., and Miyano, S. (2005). Prediction of transcriptional terminators in *Bacillus subtilis* and related species. PLoS Comput. Biol. *1*, e25.

[55] Deininger, P. L., and Daniels, G. R. (1986). The recent evolution of mammalian repetitive DNA elements. Trends Genet. *2*, 76-80.

[56] Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. Nucleic Acids Res. *27*, 4636-4641.

[57] Derre, I., Rapoport, G., and 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.

[58] Desai, S. K., and Gallivan, J. P. (2004). Genetic screens and selections for small molecules based on a synthetic riboswitch that activates protein translation. J. Am. Chem. Soc. *126*, 13247-13254.

[59] Deshpande, N., Addess, K. J., Bluhm, W. F., Merino-Ott, J. C., Townsend-Merino, W., Zhang, Q., Knezevich, C., Xie, L., Chen, L., Feng, Z*., et al.* (2005). The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. Nucleic Acids Res. *33*, D233-237.

[60] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). Biological Sequence Analysis (Cambridge, UK, Cambridge UP).

[61] Eddy, S. R. (1996). Hidden Markov models. Curr. Opin. Struct. Biol. *6*, 361-365.

[62] Eddy, S. R. (1996). RNAbob. Version 2.1. Distributed by the author. Dept. of Genetics, Washington University School of Medicine. St. Louis, Missouri.

[63] Eddy, S. R. (1998). Profile hidden Markov models. Bioinformatics *14*, 755-763.

[64] Eddy, S. R. (2003). HMMER. Version 2.3. Distributed by the author. Dept. of Genetics, Washington University School of Medicine. St. Louis, Missouri. http://hmmer.wustl.edu/

[65] Eddy, S. R. (2003). INFERNAL. Version 0.55. Distributed by the author. Dept. of Genetics, Washington University School of Medicine. St. Louis, Missouri.

[66] Eddy, S. R. (2006). RNA genes track settings. http://genome.ucsc.edu/cgi-bin/hgTrackUi?g=rnaGene&db=hg16

[67] Eddy, S. R., and Durbin, R. (1994). RNA sequence analysis using covariance models. Nucleic Acids Res. *22*, 2079-2088.

[68] Emilsson, G. M., Nakamura, S., Roth, A., and Breaker, R. R. (2003). Ribozyme speed limits. RNA *9*, 907-918.

[69] Ennifar, E., Nikulin, A., Tishchenko, S., Serganov, A., Nevskaya, N., Garber, M., Ehresmann, B., Ehresmann, C., Nikonov, S., and Dumas, P. (2000). The crystal structure of UUCG tetraloop. J. Mol. Biol. *304*, 35-42.

[70] Enserink, M. (2002). Microbial genomics. TIGR begins assault on the anthrax genome. Science *295*, 1442-1443.

[71] Epshtein, V., Mironov, A. S., and Nudler, E. (2003). The riboswitch-mediated control of sulfur metabolism in bacteria. Proc. Natl. Acad. Sci. USA *100*, 5052-5056.

[72] 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.

[73] Espinoza, C. A., Allen, T. A., Hieb, A. R., Kugel, J. F., and Goodrich, J. A. (2004). B2 RNA binds directly to RNA polymerase II to repress transcript synthesis. Nat. Struct. Mol. Biol. *11*, 822-829.

[74] Felsenstein, J. (2004). PHYLIP (Phylogeny Inference Package) Version 3.62. Distributed by the author. Dept. of Genome Sciences, University of Washington, Seattle.

[75] Fenton, M. S., and Gralla, J. D. (2003). Effect of DNA bases and backbone on σ70 holoenzyme binding and isomerization using fork junction probes. Nucleic Acids Res. *31*, 2745-2750.

[76] Ferbeyre, G., Bourdeau, V., Pageau, M., Miramontes, P., and Cedergren, R. (2000). Distribution of hammerhead and hammerhead-like RNA motifs through the GenBank. Genome Res. *10*, 1011-1019.

[77] Finn, R. D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R.*, et al.* (2006). Pfam: clans, web tools and services. Nucleic Acids Res. *34*, D247-251.

[78] Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., and et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science *269*, 496-512.

[79] Frank, D. N., and Pace, N. R. (1998). Ribonuclease P: Unity and diversity in a tRNA processing ribozyme. Annu. Rev. Biochem. *67*, 153-180.

[80] Freyhult, E., Bollback, J. P., and Gardner, P. P. (2006). Exploring genomic dark matter: homology search for non-coding RNA. *in preparation*.

[81] Freyhult, E., Moulton, V., and Ardell, D. H. (2006). Visualizing bacterial tRNA identity determinants and antideterminants using function logos and inverse function logos. Nucleic Acids Res. *34*, 905-916.

[82] Freyhult, E., Moulton, V., and Gardner, P. (2005). Predicting RNA structure using mutual information. Appl. Bioinformatics *4*, 53-59.

[83] Fuchs, R. T., Grundy, F. J., and Henkin, T. M. (2006). The S_{MK} box is a new SAM-binding RNA for translational regulation of SAM synthetase. Nat. Struct. Mol. Biol.

[84] Fuller, R. S., Funnell, B. E., and Kornberg, A. (1984). The dnaA protein complex with the E. coli chromosomal replication origin (*oriC*) and other DNA sites. Cell *38*, 889-900.

[85] Gautheret, D., Konings, D., and Gutell, R. R. (1995). G.U base pairing motifs in ribosomal RNA. RNA *1*, 807-814.

[86] Gelfand, M. S., Mironov, A. A., Jomantas, J., Kozlov, Y. I., and Perumov, D. A. (1999). A conserved RNA structure element involved in the regulation of bacterial riboflavin synthesis genes. Trends Genet. *15*, 439-442.

[87] Gerstein, M., Sonnhammer, E. L. L., and Chothia, C. (1994). Volume changes in protein evolution. J. Mol. Biol. *236*, 1067-1078.

[88] Gesteland, R. F., Cech, T. R., and Atkins, J. F., eds. (2005). The RNA World, 3rd edn (Woodbury, NY, Cold Spring Harbor Laboratory Press).

[89] Gilbert, S. D., Stoddard, C. D., Wise, S. J., and Batey, R. T. (2006). Thermodynamic and kinetic characterization of ligand binding to the purine riboswitch aptamer domain. J. Mol. Biol. *359*, 754-768.

[90] Gish, W. (1996-2006). WU-BLAST. Version 2.0. Distributed by the author. Dept. of Genetics, Washington University School of Medicine. St. Louis, Missouri. http://blast.wustl.edu

[91] Glasner, J. D., Rusch, M., Liss, P., Plunkett, G., 3rd, Cabot, E. L., Darling, A., Anderson, B. D., Infield-Harm, P., Gilson, M. C., and Perna, N. T. (2006). ASAP: a resource for annotating, curating, comparing, and disseminating genomic data. Nucleic Acids Res. *34*, D41-45.

[92] Gloor, G. B., Martin, L. C., Wahl, L. M., and Dunn, S. D. (2005). Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry *44*, 7156-7165.

[93] Gold, L., Brody, E., Heilig, J., and Singer, B. (2002). One, two, infinity: Genomes filled with aptamers. Chem. Biol. *9*, 1259-1264.

[94] Gold, L., Polisky, B., Uhlenbeck, O., and Yarus, M. (1995). Diversity of oligonucleotide functions. Annu. Rev. Biochem. *64*, 763-797.

[95] Gordon, J. J., Towsey, M. W., Hogan, J. M., Mathews, S. A., and Timms, P. (2006). Improved prediction of bacterial transcription start sites. Bioinformatics *22*, 142-148.

[96] Gorodkin, J., Heyer, L. J., Brunak, S., and Stormo, G. D. (1997). Displaying the information contents of structural RNA alignments: the structure logos. Comput. Appl. Biosci. *13*, 583-586.

[97] Gourse, R. L., Gaal, T., Bartlett, M. S., Appleman, J. A., and Ross, W. (1996). rRNA transcription and growth rate-dependent regulation of ribosome synthesis in *Escherichia coli*. Annu. Rev. Microbiol. *50*, 645-677.

[98] Griffiths-Jones, S. (2005). RALEE—RNA ALignment editor in Emacs. Bioinformatics *21*, 257-259.

[99] Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S. R. (2003). Rfam: an RNA family database. Nucleic Acids Res. *31*, 439-441.

[100] 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-D124.

[101] Grillo, G., Licciulli, F., Liuni, S., Sbisa, E., and Pesole, G. (2003). PatSearch: a program for the detection of patterns and structural motifs in nucleotide sequences. Nucleic Acids Res. *31*, 3608-3612.

[102] Grundy, F. J., and Henkin, T. M. (1991). The *rpsD* gene, encoding ribosomal protein S4, is autogenously regulated in *Bacillus subtilis*. J. Bacteriol. *173*, 4595-4602.

[103] Grundy, F. J., and Henkin, T. M. (1998). The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in Gram-positive bacteria. Mol. Microbiol. *30*, 737-749.

[104] Grundy, F. J., and Henkin, T. M. (2004). Regulation of gene expression by effectors that bind to RNA. Curr. Opin. Microbiol. *7*, 126-131.

[105] Grundy, F. J., Lehman, S. C., and Henkin, T. M. (2003). The L box regulon: Lysine sensing by leader RNAs of bacterial lysine biosynthesis genes. Proc. Natl. Acad. Sci. USA *100*, 12057-12062.

[106] Grundy, F. J., Yousef, M. R., and Henkin, T. M. (2005). Monitoring uncharged tRNA during transcription of the *Bacillus subtilis glyQS* gene. J. Mol. Biol. *346*, 73-81.

[107] Gusarov, I., and Nudler, E. (1999). The mechanism of intrinsic transcription termination. Mol. Cell. *3*, 495-504.

[108] Gutell, R. R., Cannone, J. J., Konings, D., and Gautheret, D. (2000). Predicting U-turns in ribosomal RNA with comparative sequence analysis. J. Mol. Biol. *300*, 791-803.

[109] Gutell, R. R., Cannone, J. J., Shang, Z., Du, Y., and Serra, M. J. (2000). A story: unpaired adenosine bases in ribosomal RNAs. J. Mol. Biol. *304*, 335-354.

[110] Gutell, R. R., Power, A., Hertz, G. Z., Putz, E. J., and Stormo, G. D. (1992). Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. Nucleic Acids Res. *20*, 5785-5795.

[111] Gutell, R. R., Weiser, B., Woese, C. R., and Noller, H. F. (1985). Comparative anatomy of 16S-like ribosomal RNA. Prog. Nucleic Acid Res. Mol. Biol. *32*, 155-216.

[112] Haba, G., Jamieson, G. A., Mudd, A. H., and Richard, H. H. (1959). *S*-adenosylmethionine: The relation of configuration at the sulfonium center to enzymatic reactivity. J. Am. Chem. Soc. *81*, 3975.

[113] Hall, D. A., Zhu, H., Zhu, X., Royce, T., Gerstein, M., and Snyder, M. (2004). Regulation of gene expression by a metabolic enzyme. Science *306*, 482-484.

[114] Hampel, K. J., and Tinsley, M. M. (2006). Evidence for preorganization of the *glmS* ribozyme ligand binding pocket. Biochemistry *45*, 7861-7871.

[115] Hanson, S., Bauer, G., Fink, B., and Suess, B. (2005). Molecular analysis of a synthetic tetracycline-binding riboswitch. RNA *11*, 503-511.

[116] Hauser, L., Larimer, F., Land, M., Shah, M., and Uberbacher, E. (2004). Analysis and annotation of microbial genome sequences. Genet. Eng. (NY) *26*, 225-238.

[117] Havgaard, J. H., Lyngso, R. B., Stormo, G. D., and Gorodkin, J. (2005). Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. Bioinformatics *21*, 1815-1824.

[118] Helmann, J. D. (1995). Compilation and analysis of *Bacillus subtilis* σ^A-dependent promoter sequences: evidence for extended contact between RNA polymerase and upstream promoter DNA. Nucleic Acids Res. *23*, 2351-2360.

[119] Helmann, J. D., and deHaseth, P. L. (1999). Protein-nucleic acid interactions during open complex formation investigated by systematic alteration of the protein and DNA binding partners. Biochemistry *38*, 5959-5967.

[120] Henkin, T. M. (1994). tRNA-directed transcription antitermination. Mol. Microbiol. *13*, 381-387.

[121] Hertz, G. Z., and Stormo, G. D. (1996). *Escherichia coli* promoter sequences: analysis and prediction. Method Enzymol. *273*, 30-42.

[122] Heus, H. A., and Pardi, A. (1991). Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. Science *253*, 191-194.

[123] Hindley, J. (1967). Fractionation of ^{32}P-labelled ribonucleic acids on polyacrylamide gels and their characterization by fingerprinting. J. Mol. Biol. *30*, 125-136.

[124] Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F.*, et al.* (2006). The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. *34*, D590-598.

[125] Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. Monatsh. Chem. *125*, 167-188.

[126] Hofacker, N. L., Fekete, M., and Stadler, P. F. (2002). Secondary structure prediction for aligned RNA sequences. J. Mol. Biol. *319*, 1059-1066.

[127] Horton, P. B., and Kanehisa, M. (1992). An assessment of neural network and statistical approaches for prediction of *E. coli* promoter sites. Nucleic Acids Res. *20*, 4331-4338.

[128] Hsu, L. M., Zagorski, J., Wang, Z., and Fournier, M. J. (1985). *Escherichia coli* 6S RNA gene is part of a dual-function transcription unit. J. Bacteriol. *161*, 1162-1170.

[129] Huang, Z., and Szostak, J. W. (2003). Evolution of aptamers with a new specificity and new secondary structures from an ATP aptamer. RNA *9*, 1456-1463.

[130] Huynen, M., Snel, B., Lathe, W., and Bork, P. (2000). Exploitation of gene context. Curr. Opin. Struct. Biol. *10*, 366-370.

[131] Ishihama, A. (1999). Modulation of the nucleoid, the transcription apparatus, and the translation machinery in bacteria for stationary phase survival. Genes to Cells *4*, 135-143.

[132] Ishihama, A. (2000). Functional modulation of *Escherichia coli* RNA polymerase. Annu. Rev. Microbiol. *54*, 499-518.

[133] Ishii, T., Yoshida, K., Terai, G., Fujita, Y., and Nakai, K. (2001). DBTBS: a database of *Bacillus subtilis* promoters and transcription factors. Nucleic Acids Res. *29*, 278-280.

[134] Iyer, L. M., Koonin, E. V., and Aravind, L. (2004). Evolution of bacterial RNA polymerase: implications for large-scale bacterial phylogeny, domain accretion, and horizontal gene transfer. Gene *335*, 73-88.

[135] Jack, D. L., Storms, M. L., Tchieu, J. H., Paulsen, I. T., and Saier, M. H. (2000). A broad-specificity multidrug efflux pump requiring a pair of homologous SMR-type proteins. J. Bacteriol. *182*, 2311-2313.

[136] Jansen, J. A., McCarthy, T. J., Soukup, G. A., and Soukup, J. K. (2006). Backbone and nucleobase contacts to glucosamine-6-phosphate in the *glmS* ribozyme. Nat. Struct. Mol. Biol.

[137] Johnston, W. K., Unrau, P. J., Lawrence, M. S., Glasner, M. E., and Bartel, D. P. (2001). RNA-catalyzed RNA polymerization: Accurate and general RNA-templated primer extension. Science *292*, 1319-1325.

[138] Joris, B., Hardt, K., and Ghuysen, J.-M. (1994). Induction of beta-lactamase and low-affinity penicillin binding protein 2' synthesis in Gram-positive bacteria. New Compr. Biochem. *27*, 505-515.

[139] Joyce, G. F. (2002). The antiquity of RNA-based evolution. Nature *418*, 214-221.

[140] Jucker, F. M., and Pardi, A. (1995). GNRA tetraloops make a U-Turn. RNA *1*, 219-222.

[141] Kanhere, A., and Bansal, M. (2005). A novel method for prokaryotic promoter prediction based on DNA stability. BMC Bioinformatics *6*, 1.

[142] Keller, E. B., and Calvo, J. M. (1979). Alternative secondary structures of leader RNAs and the regulation of the *trp*, *phe*, *his*, *thr*, and *leu* operons. Proc. Natl. Acad. Sci. USA *76*, 6186-6190.

[143] Kenner, J., and Nomura, M. (1996). Regulation of ribosome synthesis. In *Escherichia Coli* and *Salmonella*: Cellular and Molecular Biology, F. C. Neidhardt, J. L. Ingraham, and R. C. Curtiss, III, eds. (ASM Press), pp. 1417-1431.

[144] Kiga, D., Futamura, Y., Sakamoto, K., and Yokoyama, S. (1998). An RNA aptamer to the xanthine/guanine base with a distinctive mode of purine recognition. Nucleic Acids Res. *26*, 1755-1760.

[145] Kim, D. S., Gusti, V., Pillai, S. G., and Gaur, R. K. (2005). An artificial riboswitch for controlling pre-mRNA splicing. RNA *11*, 1667-1677.

[146] Kim, K. S., and Lee, Y. (2004). Regulation of 6S RNA biogenesis by switching utilization of both sigma factors and endoribonucleases. Nucleic Acids Res. *32*, 6057-6068.

[147] Klein, D. J., Schmeing, T. M., Moore, P. B., and Steitz, T. A. (2001). The kink-turn: a new RNA secondary structure motif. EMBO J. *20*, 4214-4221.

[148] Klein, R. J., and Eddy, S. R. (2003). RSEARCH: finding homologs of single structured RNA sequences. BMC Bioinformatics *4*.

[149] Kochhar, S., and Paulus, H. (1996). Lysine-induced premature transcription termination in the *lysC* operon of *Bacillus subtilis*. Microbiology *142 (Pt. 7)*, 1635-1639.

[150] Koizumi, M., Soukup, G. A., Kerr, J. N. Q., and Breaker, R. R. (1999). Allosteric selection of ribozymes that respond to the second messengers cGMP and cAMP. Nat. Struct. Biol. *6*, 1062-1071.

[151] Kolter, R., and Yanofsky, C. (1982). Attenuation in amino acid biosynthetic operons. Annu. Rev. Genet. *16*, 113-134.

[152] Krasilnikov, A. S., and Mondragon, A. (2003). On the occurrence of the T-loop RNA folding motif in large RNA molecules. RNA *9*, 640-643.

[153] Kruger, E., and Hecker, M. (1998). The first gene of the *Bacillus subtilis clpC* operon, *ctsR*, encodes a negative regulator of its own operon and other class III heat shock genes. J. Bacteriol. *180*, 6681-6688.

[154] Kubodera, T., Watanabe, M., Yoshiuchi, K., Yamashita, N., Nishimura, A., Nakai, S., Gomi, K., and Hanamoto, H. (2003). Thiamine-regulated gene expression of *Aspergillus oryzae thiA* requires splicing of the intron containing a riboswitch-like domain in the 5 '-UTR. FEBS Lett. *555*, 516-520.

[155] Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessieres, P., Bolotin, A., Borchert, S.*, et al.* (1997). The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. Nature *390*, 249-256.

[156] Laferriere, A., Gautheret, D., and Cedergren, R. (1994). An RNA pattern-matching program with enhanced performance and portability. Comput. Appl. Biosci. *10*, 211-212.

[157] Laserson, U., Gan, H. H., and Schlick, T. (2005). Predicting candidate genomic sequences that correspond to synthetic functional RNA motifs. Nucleic Acids Res. *33*, 6057-6069.

[158] Lathe, W. C., Suyama, M., and Bork, P. (2002). Identification of attenuation and antitermination regulation in prokaryotes. Genome Biol. *3*, preprint0003.0001-0003.0060.

[159] Lee, N., Bessho, Y., Wei, K., Szostak, J. W., and Suga, H. (2000). Ribozyme-catalyzed tRNA aminoacylation. Nat. Struct. Biol. *7*, 28-33.

[160] Leontis, N. B., Stombaugh, J., and Westhof, E. (2002). Motif prediction in ribosomal RNAs: lessons and prospects for automated motif prediction in homologous RNA molecules. Biochimie *84*, 961-973.

[161] Leontis, N. B., Stombaugh, J., and Westhof, E. (2002). The non-Watson-Crick base pairs and their associated isostericity matrices. Nucleic Acids Res. *30*, 3497-3531.

[162] Leontis, N. B., and Westhof, E. (1998). The 5S rRNA loop E: chemical probing and phylogenetic data versus crystal structure. RNA *4*, 1134-1153.

[163] Lescoute, A., Leontis, N. B., Massire, C., and Westhof, E. (2005). Recurrent structural RNA motifs, isostericity matrices and sequence alignments. Nucleic Acids Res. *33*, 2395-2409.

[164] 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.

[165] Letunic, I., Copley, R. R., Pils, B., Pinkert, S., Schultz, J., and Bork, P. (2006). SMART 5: domains in the context of genomes and networks. Nucleic Acids Res. *34*, D257-260.

[166] Lewin, B. (1999). Transcription. In Genes VII (Oxford, Oxford UP), pp. 233-271.

[167] Li, Y., and Altman, S. (2004). In search of RNase P RNA from microbial genomes. RNA *10*, 1533-1540.

[168] Li, Z., Pandit, S., and Deutscher, M. P. (1998). 3' Exoribonucleolytic trimming is a common feature of the maturation of small, stable RNAs in *Escherichia coli*. Proc. Natl. Acad. Sci. USA *95*, 2856-2861.

[169] Lim, J., Winkler, W. C., Nakamura, S., Scott, V., and Breaker, R. R. (2006). Molecular-recognition characteristics of SAM-binding riboswitches. Angew. Chem. Int. Edit. Engl. *45*, 964-968.

[170] Lin, D. C., and Grossman, A. D. (1998). Identification and characterization of a bacterial chromosome partitioning site. Cell *92*, 675-685.

[171] Lorsch, J. R., and Szostak, J. W. (1994). *In vitro* selection of RNA aptamers specific for cyanocobalamin. Biochemistry *33*, 973-982.

[172] Lowe, T. M., and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. *25*, 955-964.

[173] Lu, Y., Turner, R. J., and Switzer, R. L. (1996). Function of RNA secondary structures in transcriptional attenuation of the *Bacillus subtilis pyr* operon. Proc. Natl. Acad. Sci. USA *93*, 14462-14467.

[174] Lupas, A. (1996). Prediction and analysis of coiled-coil structures. Method Enzymol. *266*, 513-525.

[175] 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.

[176] Madigan, M. T., Martinko, J. M., and Parker, J. (2003). Brock Biology of Microorganisms, 10th edn (Upper Saddle River, NJ, Pearson Education, Inc.).

[177] Mahillon, J., and Chandler, M. (1998). Insertion sequences. Microbiol. Mol. Biol. Rev. *62*, 725-774.

[178] Mandal, M., Boese, B., Barrick, J. E., Winkler, W. C., and Breaker, R. R. (2003). Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. Cell *113*, 577-586.

[179] Mandal, M., and Breaker, R. R. (2004). Adenine riboswitches and gene activation by disruption of a transcription terminator. Nat. Struct. Mol. Biol. *11*, 29-35.

[180] Mandal, M., Lee, M., Barrick, J. E., Weinberg, Z., Emilsson, G. M., Ruzzo, W. L., and Breaker, R. R. (2004). A glycine-dependent riboswitch that uses cooperative binding to control gene expression. Science *306*, 275-279.

[181] Marchler-Bauer, A., Anderson, J. B., Cherukuri, P. F., DeWeese-Scott, C., Geer, L. Y., Gwadz, M., He, S., Hurwitz, D. I., Jackson, J. D., Ke, Z.*, et al.* (2005). CDD: a Conserved Domain Database for protein classification. Nucleic Acids Res. *33*, D192-196.

[182] Marchler-Bauer, A., Anderson, J. B., DeWeese-Scott, C., Fedorova, N. D., Geer, L. Y., He, S., Hurwitz, D. I., Jackson, J. D., Jacobs, A. R., Lanczycki, C. J.*, et al.* (2003). CDD: a curated Entrez database of conserved domain alignments. Nucleic Acids Res. *31*, 383-387.

[183] Mayrose, I., Graur, D., Ben-Tal, N., and Pupko, T. (2004). Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol. Biol. Evol. *21*, 1781-1791.

[184] McCarthy, T. J., Plog, M. A., Floy, S. A., Jansen, J. A., Soukup, J. K., and Soukup, G. A. (2005). Ligand requirements for *glmS* ribozyme self-cleavage. Chem. Biol. *12*, 1221-1226.

[185] McDaniel, B. A. M., Grundy, F. J., Artsimovitch, I., and Henkin, T. M. (2003). Transcription termination control of the S box system: direct measurement of *S*-adenosylmethionine by the leader RNA. Proc. Natl. Acad. Sci. USA *100*, 3083-3088.

[186] McGinnis, S., and Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. *32*, W20-25.

[187] Meng, Q., Turnbough, C. L., Jr., and Switzer, R. L. (2004). Attenuation control of *pyrG* expression in *Bacillus subtilis* is mediated by CTP-sensitive reiterative transcription. Proc. Natl. Acad. Sci. USA *101*, 10943-10948.

[188] Merino, E., and Yanofsky, C. (2005). Transcription attenuation: a highly conserved regulatory strategy used by bacteria. Trends Genet. *21*, 260-264.

[189] Mira, A., Pushker, R., Legault, B. A., Moreira, D., and Rodriguez-Valera, F. (2004). Evolutionary relationships of *Fusobacterium nucleatum* based on phylogenetic analysis and comparative genomics. BMC Evol Biol *4*, 50.

[190] Miranda-Rios, J., Navarro, M., and Soberon, M. (2001). A conserved RNA structure (*thi* box) is involved in regulation of thiamin biosynthetic gene expression in bacteria. Proc. Natl. Acad. Sci. USA *98*, 9736-9741.

[191] Mironov, A. S., Gusarov, I., Rafikov, R., Lopez, L. E., Shatalin, K., Kreneva, R. A., Perumov, D. A., and Nudler, E. (2002). Sensing small molecules by nascent RNA: A mechanism to control transcription in bacteria. Cell *111*, 747-756.

[192] Montange, R. K., and Batey, R. T. (2006). Structure of the *S*-adenosylmethionine riboswitch regulatory mRNA element. Nature *441*, 1172-1175.

[193] Moreno-Hagelsieb, G., and Collado-Vides, J. (2002). A powerful non-homology method for the prediction of operons in prokaryotes. Bioinformatics *18 Suppl 1*, S329-336.

[194] Moszer, I., Jones, L. M., Moreira, S., Fabry, C., and Danchin, A. (2002). SubtiList: the reference database for the *Bacillus subtilis* genome. Nucleic Acids Res. *30*, 62-65.

[195] Munch, R., Hiller, K., Barg, H., Heldt, D., Linz, S., Wingender, E., and Jahn, D. (2003). PRODORIC: prokaryotic database of gene regulation. Nucleic Acids Res. *31*, 266-269.

[196] Munch, R., Hiller, K., Grote, A., Scheer, M., Klein, J., Schobert, M., and Jahn, D. (2005). Virtual Footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes. Bioinformatics *21*, 4187-4189.

[197] Murakami, K. S., Masuda, S., Campbell, E. A., Muzzin, O., and Darst, S. A. (2002). Structural basis of transcription initiation: an RNA polymerase holoenzyme-DNA complex. Science *296*, 1285-1290.

[198] Muth, G. W., Ortoleva-Donnelly, L., and Strobel, S. A. (2000). A single adenosine with a neutral pK_a in the ribosomal peptidyl transferase center. Science *289*, 947-950.

[199] Mwangi, M. M., and Siggia, E. D. (2003). Genome wide identification of regulatory motifs in *Bacillus subtilis*. BMC Bioinformatics *4*.

[200] Nagaswamy, U., and Fox, G. E. (2002). Frequent occurrence of the T-loop RNA folding motif in ribosomal RNAs. RNA *8*, 1112-1119.

[201] Nahvi, A., Barrick, J. E., and Breaker, R. R. (2004). Coenzyme B_{12} riboswitches are widespread genetic control elements in prokaryotes. Nucleic Acids Res. *32*, 143-150.

[202] Nahvi, A., Sudarsan, N., Ebert, M. S., Zou, X., Brown, K. L., and Breaker, R. R. (2002). Genetic control by a metabolite binding mRNA. Chem. Biol. *9*, 1043-1049.

[203] Nakahigashi, K., Yanagi, H., and Yura, T. (1995). Isolation and sequence analysis of *rpoH* genes encoding □^{32} homologs from gram negative bacteria: conserved mRNA and protein segments for heat shock regulation. Nucleic Acids Res. *23*, 4383-4390.

[204] Narberhaus, F. (1999). Negative regulation of bacterial heat shock genes. Mol. Microbiol. *31*, 1-8.

[205] Narberhaus, F., Waldminghaus, T., and Chowdhury, S. (2006). RNA thermometers. FEMS Microbiol. Rev. *30*, 3-16.

[206] Nissen, P., Hansen, J., Ban, N., Moore, P. B., and Steitz, T. A. (2000). The structural basis of ribosome activity in peptide bond synthesis. Science *289*, 920-930.

[207] Nissen, P., Ippolito, J. A., Ban, N., Moore, P. B., and Steitz, T. A. (2001). RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. Proc. Natl. Acad. Sci. USA *98*, 4899-4903.

[208] Noeske, J., Richter, C., Grundl, M. A., Nasiri, H. R., Schwalbe, H., and Wohnert, J. (2005). An intermolecular base triple as the basis of ligand specificity and affinity in the guanine- and adenine-sensing riboswitch RNAs. Proc. Natl. Acad. Sci. USA *102*, 1372-1377.

[209] Nou, X. W., and Kadner, R. J. (2000). Adenosylcobalamin inhibits ribosome binding to *btuB* RNA. Proc. Natl. Acad. Sci. USA *97*, 7190-7195.

[210] Osteras, M., Stanley, J., and Finan, T. M. (1995). Identification of Rhizobium-specific intergenic mosaic elements within an essential 2-component regulatory system of *Rhizobium* species. J. Bacteriol. *177*, 5485-5494.

[211] Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., de Crecy-Lagard, V., Diaz, N., Disz, T., Edwards, R.*, et al.* (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. *33*, 5691-5702.

[212] Panina, E. M., Vitreschak, A. G., Mironov, A. A., and Gelfand, M. S. (2003). Regulation of biosynthesis and transport of aromatic amino acids in low-GC Gram-positive bacteria. FEMS Microbiol. Lett. *222*, 211-220.

[213] Patte, J. C., Akrim, M., and Mejean, V. (1998). The leader sequence of the Escherichia coli *lysC* gene is involved in the regulation of LysC synthesis. FEMS Microbiol. Lett. *169*, 165-170.

[214] Pearson, W. R. (2000). Flexible sequence similarity searching with the FASTA3 program package. In Bioinformatics Methods and Protocols, S. Misener, and S. A. Krawetz, eds. (Totowa, NJ, Humana Press), pp. 185-219.

[215] Pedersen, J. S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E. S., Kent, J., Miller, W., and Haussler, D. (2006). Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput. Biol. *2*, e33.

[216] Pelchat, M., Grenier, C., and Perreault, J. P. (2002). Characterization of a viroid-derived RNA promoter for the DNA-dependent RNA polymerase from *Escherichia coli*. Biochemistry *41*, 6561-6571.

[217] Penchovsky, R., and Breaker, R. R. (2005). Computational design and experimental validation of oligonucleotide-sensing allosteric ribozymes. Nat. Biotechnol. *23*, 1424-1433.

[218] Peterson, J. D., Umayam, L. A., Dickinson, T., Hickey, E. K., and White, O. (2001). The Comprehensive Microbial Resource. Nucleic Acids Res. *29*, 123-125.

[219] Poole, A. M., Jeffares, D. C., and Penny, D. (1998). The Path from the RNA World. J. Mol. Evol. *46*, 1-17.

[220] Posnick, L. M., and Samson, L. D. (1999). Influence of *S*-adenosylmethionine pool size on spontaneous mutation, dam methylation, and cell growth of *Escherichia coli*. J. Bacteriol. *181*, 6756-6762.

[221] Price, A. L., Jones, N. C., and Pevzner, P. A. (2005). *De novo* identification of repeat families in large genomes. Bioinformatics *21 Suppl 1*, i351-358.

[222] Price, M. N., Huang, K. H., Alm, E. J., and Arkin, A. P. (2005). A novel method for accurate operon predictions in all sequenced prokaryotes. Nucleic Acids Res. *33*, 880-892.

[223] Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. *33*, D501-504.

[224] Puerta-Fernandez, E., Barrick, J. E., Roth, A., and Breaker, R. R. (2006). Identification of a new, non-coding RNA in extremophilic eubacteria. *in preparation*.

[225] Ravnum, S., and Andersson, D. I. (2001). An adenosyl-cobalamin (coenzyme-B12)-repressed translational enhancer in the *cob* mRNA of *Salmonella typhimurium*. Mol. Microbiol. *39*, 1585-1594.

[226] Read, T. D., Peterson, S. N., Tourasse, N., Baillie, L. W., Paulsen, I. T., Nelson, K. E., Tettelin, H., Fouts, D. E., Eisen, J. A., Gill, S. R.*, et al.* (2003). The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. Nature *423*, 81-86.

[227] Reader, J. S., Metzgar, D., Schimmel, P., and de Crecy-Lagard, V. (2004). Identification of four genes necessary for biosynthesis of the modified nucleoside queuosine. J. Biol. Chem. *279*, 6280-6285.

[228] Reents, H., Munch, R., Dammeyer, T., Jahn, D., and Härtig, E. (2006). The Fnr regulon of *Bacillus subtilis*. J. Bacteriol. *188*, 1103-1112.

[229] Reese, M. G. (2001). Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. Comput. Chem. *26*, 51-56.

[230] Reese, M. G. (2006). NNPP. Version 2.2. Berkeley Drosophila Genome Project. http://www.fruitfly.org/seq_tools/promoter.html

[231] Ren, D., Bedzyk, L. A., Thomas, S. M., Ye, R. W., and Wood, T. K. (2004). Gene expression in *Escherichia coli* biofilms. Appl. Microbiol. Biotechnol. *64*, 515-524.

[232] Richardson, J. P. (1990). Rho-dependent transcription termination. Biochim. Biophys. Acta *1048*, 127-138.

[233] Riley, M., Abe, T., Arnaud, M. B., Berlyn, M. K., Blattner, F. R., Chaudhuri, R. R., Glasner, J. D., Horiuchi, T., Keseler, I. M., Kosuge, T.*, et al.* (2006). *Escherichia coli* K-12: a cooperatively developed annotation snapshot--2005. Nucleic Acids Res. *34*, 1-9.

[234] Rodionov, D. A., Vitreschak, A. G., Mironov, A. A., and Gelfand, M. S. (2002). Comparative genomics of thiamin biosynthesis in procaryotes: New genes and regulatory mechanisms. J. Biol. Chem. *277*, 48949-48959.

[235] Rodionov, D. A., Vitreschak, A. G., Mironov, A. A., and Gelfand, M. S. (2003). Comparative genomics of the vitamin B_{12} metabolism and regulation in prokaryotes. J. Biol. Chem. *278*, 41148-41159.

[236] Rodionov, D. A., Vitreschak, A. G., Mironov, A. A., and Gelfand, M. S. (2003). Regulation of lysine biosynthesis and transport genes in bacteria: yet another RNA riboswitch? Nucleic Acids Res. *31*, 6748-6757.

[237] Rodionov, D. A., Vitreschak, A. G., Mironov, A. A., and Gelfand, M. S. (2004). Comparative genomics of the methionine metabolism in Gram-positive bacteria: a variety of regulatory systems. Nucleic Acids Res. *32*, 3340-3353.

[238] Rollins, S. M., Grundy, F. J., and Henkin, T. M. (1997). Analysis of *cis*-acting sequence and structural elements required for antitermination of the *Bacillus subtilis tyrS* gene. Mol. Microbiol. *25*, 411-421.

[239] Roth, A., Nahvi, A., Lee, M., Jona, I., and Breaker, R. R. (2006). Characteristics of the *glmS* ribozyme suggest only structural roles for divalent metal ions. RNA *12*, 607-619.

[240] Roth, J. R., Lawrence, J. G., and Bobik, T. A. (1996). Cobalamin (coenzyme B_{12}): synthesis and biological significance. Annu. Rev. Microbiol. *50*, 137-181.

[241] Rutberg, B. (1997). Antitermination of transcription of catabolic operons. Mol. Microbiol. *23*, 413-421.

[242] Sacerdot, C., Caillet, J., Graffe, M., Eyermann, F., Ehresmann, B., Ehresmann, C., Springer, M., and Romby, P. (1998). The *Escherichia coli* threonyl-tRNA synthetase gene contains a split ribosomal binding site interrupted by a hairpin structure that is essential for autoregulation. Mol. Microbiol. *29*, 1077-1090.

[243] Salgado, H., Moreno-Hagelsieb, G., Smith, T. F., and Collado-Vides, J. (2000). Operons in *Escherichia coli*: genomic analyses and predictions. Proc. Natl. Acad. Sci. USA *97*, 6652-6657.

[244] Sandelin, A., Bailey, P., Bruce, S., Engstrom, P. G., Klos, J. M., Wasserman, W. W., Ericson, J., and Lenhard, B. (2004). Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics *5*, 99.

[245] Sankoff, D. (1985). Simultaneous solution of the RNA folding, alignment and protosequence Problems. Siam J. Appl. Math. *45*, 810-825.

[246] Sassanfar, M., and Szostak, J. W. (1993). An RNA motif that binds ATP. Nature *364*, 550-553.

[247] Sazani, P. L., Larralde, R., and Szostak, J. W. (2004). A small aptamer with strong and specific recognition of the triphosphate of ATP. J. Am. Chem. Soc. *126*, 8370-8371.

[248] Schilling, O., Langbein, I., Muller, M., Schmalisch, M. H., and Stulke, J. (2004). A protein-dependent riboswitch controlling *ptsGHI* operon expression in *Bacillus subtilis*: RNA structure rather than sequence provides interaction specificity. Nucleic Acids Res. *32*, 2853-2864.

[249] Schmalisch, M. H., Bachem, S., and Stulke, J. (2003). Control of the *Bacillus subtilis* antiterminator protein GlcT by phosphorylation. Elucidation of the phosphorylation chain leading to inactivation of GlcT. J. Biol. Chem. *278*, 51108-51115.

[250] Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D., and Miller, W. (2003). Human-mouse alignments with BLASTZ. Genome Res. *13*, 103-107.

[251] Seetharaman, S., Zivarts, M., Sudarsan, N., and Breaker, R. R. (2001). Immobilized RNA switches for the analysis of complex chemical and biological mixtures. Nat. Biotechnol. *19*, 336-341.

[252] Sekowska, A., Danchin, A., and Risler, J. L. (2000). Phylogeny of related functions: the case of polyamine biosynthetic enzymes. Microbiol. (UK) *146*, 1815-1828.

[253] Seliverstov, A. V., Putzer, H., Gelfand, M. S., and Lyubetsky, V. A. (2005). Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria. BMC Microbiol. *5*, 54.

[254] Serganov, A., Polonskaia, A., Phan, A. T., Breaker, R. R., and Patel, D. J. (2006). Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. Nature *441*, 1167-1171.

[255] Serganov, A., Yuan, Y. R., Pikovskaya, O., Polonskaia, A., Malinina, L., Phan, A. T., Hobartner, C., Micura, R., Breaker, R. R., and Patel, D. J. (2004). Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. Chem. Biol. *11*, 1729-1741.

[256] Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S.*, et al.* (2005). Evolutionarily

conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034-1050.

[257] Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., and Chandler, M. (2006). ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. *34*, D32-36.

[258] Smit, A. F. S., Hubley, R., and Green, P. (1996-2004). RepeatMasker Open-3.0 http://www.repeatmasker.org

[259] Soukup, G. A. (2006). Core requirements for *glmS* ribozyme self-cleavage reveal a putative pseudoknot structure. Nucleic Acids Res. *34*, 968-975.

[260] Soukup, G. A., and Breaker, R. R. (1999). Relationship between internucleotide linkage geometry and the stability of RNA. RNA *5*, 1308-1325.

[261] Soukup, G. A., and Breaker, R. R. (2000). Allosteric Ribozymes. In Ribozyme Biochemistry and Biotechnology, G. Krupp, and R. J. Gaur, eds. (Natick, MA, Eaton Publishing), pp. 149-170.

[262] Sponer, J., Mokdad, A., Sponer, J. E., Spackova, N., Leszczynski, J., and Leontis, N. B. (2003). Unique tertiary and neighbor interactions determine conservation patterns of *cis* Watson-Crick A/G base-pairs. J. Mol. Biol. *330*, 967-978.

[263] Springer, M., and Portier, C. (2003). More than one way to skin a cat: translational autoregulation by ribosomal protein S15. Nat. Struct. Biol. *10*, 420-422.

[264] Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H.*, et al.* (2002). The Bioperl toolkit: Perl modules for the life sciences. Genome Res. *12*, 1611-1618.

[265] Stein, L. (2001). Genome annotation: from sequence to biology. Nat. Rev. Genet. *2*, 493-503.

[266] Steitz, T. A., and Steitz, J. A. (1993). A general two-metal-ion mechanism for catalytic RNA. Proc. Natl. Acad. Sci. USA *90*, 6498-6502.

[267] Streaker, E. D., and Beckett, D. (2006). The biotin regulatory system: kinetic control of a transcriptional switch. Biochemistry *45*, 6417-6425.

[268] Sudarsan, N., Barrick, J. E., and Breaker, R. R. (2003). Metabolite-binding RNA domains are present in the genes of eukaryotes. RNA *9*, 644-647.

[269] Sudarsan, N., Wickiser, J. K., Nakamura, S., Ebert, M. S., and Breaker, R. R. (2003). An mRNA structure in bacteria that controls gene expression by binding lysine. Genes Dev. *17*, 2688-2697.

[270] Suess, B., Fink, B., Berens, C., Stentz, R., and Hillen, W. (2004). A theophylline responsive riboswitch based on helix slipping controls gene expression *in vivo*. Nucleic Acids Res. *32*, 1610-1614.

[271] Sussman, D., Nix, J. C., and Wilson, C. (2000). The structural basis for molecular recognition by the vitamin B_{12} RNA aptamer. Nat Struct Biol *7*, 53-57.

[272] Suzek, B. E., Ermolaeva, M. D., Schreiber, M., and Salzberg, S. L. (2001). A probabilistic method for identifying start codons in bacterial genomes. Bioinformatics *17*, 1123-1130.

[273] Suzuma, S., Asari, S., Bunai, K., Yoshino, K., Ando, Y., Kakeshita, H., Fujita, M., Nakamura, K., and Yamane, K. (2002). Identification and characterization of novel small RNAs in the *aspS-yrvM* intergenic region of the *Bacillus subtilis* genome. Microbiol. (UK) *148*, 2591-2598.

[274] Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N.*, et al.* (2003). The COG database: an updated version includes eukaryotes. BMC Bioinformatics *4*, 41.

[275] Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. *28*, 33-36.

[276] Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997). A genomic perspective on protein families. Science *278*, 631-637.

[277] Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., and Koonin, E. V. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. *29*, 22-28.

[278] Terai, G., Takagi, T., and Nakai, K. (2001). Prediction of co-regulated genes in *Bacillus subtilis* on the basis of upstream elements conserved across three closely related species. Genome Biol. *2*, research0048.0041 - research0048.0012.

[279] Thoeny, P. (2003). TWiki™—an enterprise collaboration platform. Release 01-Feb-2003. Distributed by the author. http://www.twiki.org

[280] Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*, 4673-4680.

[281] Thore, S., Leibundgut, M., and Ban, N. (2006). Structure of the eukaryotic thiamine pyrophosphate riboswitch with its regulatory ligand. Science *312*, 1208-1211.

[282] Torarinsson, E., Sawera, M., Havgaard, J. H., Fredholm, M., and Gorodkin, J. (2006). Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. Genome Res. *16*, 885-889.

[283] Tringe, S. G., von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., Chang, H. W., Podar, M., Short, J. M., Mathur, E. J., Detter, J. C.*, et al.* (2005). Comparative metagenomics of microbial communities. Science *308*, 554-557.

[284] Trotochaud, A. E., and Wassarman, K. M. (2004). 6S RNA function enhances long-term cell survival. J. Bacteriol. *186*, 4978-4985.

[285] Trotochaud, A. E., and Wassarman, K. M. (2005). A highly conserved 6S RNA structure is required for regulation of transcription. Nat. Struct. Mol. Biol. *12*, 313-319.

[286] Trotochaud, A. E., and Wassarman, K. M. (2006). 6S RNA regulation of *pspF* transcription leads to altered cell survival at high pH. J. Bacteriol. *188*, 3936-3943.

[287] Tucker, B. J., and Breaker, R. R. (2005). Riboswitches as versatile gene control elements. Curr. Opin. Struct. Biol. *15*, 342-348.

[288] Tufte, E. R. (1990). Envisioning Information (Cheshire, CT, Graphics Press).

[289] Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., and Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature *428*, 37-43.

[290] Unniraman, S., Prakash, R., and Nagaraja, V. (2002). Conserved economics of transcription termination in eubacteria. Nucleic Acids Res. *30*, 675-684.

[291] Unrau, P. J., and Bartel, D. P. (1998). RNA-catalysed nucleotide synthesis. Nature *395*, 260-263.

[292] Valadkhan, S. (2005). snRNAs as the catalysts of pre-mRNA splicing. Curr. Opin. Chem. Biol. *9*, 603-608.

[293] Vallenet, D., Labarre, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., Lajus, A., Pascal, G., Scarpelli, C., and Medigue, C. (2006). MaGe: a microbial genome annotation system supported by synteny results. Nucleic Acids Res. *34*, 53-65.

[294] Van Lanen, S. G., Reader, J. S., Swairjo, M. A., de Crécy-Lagard, V., Lee, B., and Iwata-Reuyl, D. (2005). From cyclohydrolase to oxidoreductase: discovery of nitrile reductase activity in a common fold. Proc. Natl. Acad. Sci. USA *102*, 4264-4269.

[295] Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D. Y., Paulsen, I., Nelson, K. E., Nelson, W.*, et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. Science *304*, 66-74.

[296] Vitreschak, A. G., Lyubetskaya, E. V., Shirshin, M. A., Gelfand, M. S., and Lyubetsky, V. A. (2004). Attenuation regulation of amino acid biosynthetic operons in proteobacteria: comparative genomics analysis. FEMS Microbiol. Lett. *234*, 357-370.

[297] Vitreschak, A. G., Mironov, A. A., and Gelfand, M. S. (2001). Computer prediction of RNA secondary structure. The RNApattern program: searching for RNA secondary structure by the pattern rule. In Proc. 3rd Int. Conf. 'Complex Systems: Control and modeling problems' (Samara), pp. 623-625.

[298] Vitreschak, A. G., Rodionov, D. A., Mironov, A. A., and Gelfand, M. S. (2002). Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation. Nucleic Acids Res. *30*, 3141-3151.

[299] Vitreschak, A. G., Rodionov, D. A., Mironov, A. A., and Gelfand, M. S. (2003). Regulation of the vitamin B_{12} metabolism and transport in bacteria by a conserved RNA structural element. RNA *9*, 1084-1097.

[300] Vitreschak, A. G., Rodionov, D. A., Mironov, A. A., and Gelfand, M. S. (2004). Riboswitches: the oldest mechanism for the regulation of gene expression? Trends Genet. *20*, 44-50.

[301] Vogel, D. W., Hartmann, R. K., Struck, J. C. R., Ulbrich, N., and Erdmann, V. A. (1987). The sequence of the 6S RNA gene of *Pseudomonas aeruginosa*. Nucleic Acids Res. *15*, 4583-4591.

[302] Vogel, J., Bartels, V., Tang, T. H., Churakov, G., Slagter-Jager, J. G., Huttenhofer, A., and Wagner, E. G. (2003). RNomics in *Escherichia coli* detects new sRNA species and indicates parallel transcriptional output in bacteria. Nucleic Acids Res. *31*, 6435-6443.

[303] Wan, X. F., and Xu, D. (2005). Intrinsic terminator prediction and its application in *Synechococcus* sp. WH8102. J. Comput. Sci. Tech. *20*, 465-482.

[304] Washietl, S., Hofacker, I. L., Lukasser, M., Huttenhofer, A., and Stadler, P. F. (2005). Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. Nat. Biotechnol. *23*, 1383-1390.

[305] Washietl, S., Hofacker, I. L., and Stadler, P. F. (2005). Fast and reliable prediction of noncoding RNAs. Proc. Natl. Acad. Sci. USA *102*, 2454-2459.

[306] Washio, T., Sasayama, J., and Tomita, M. (1998). Analysis of complete genomes suggests that many prokaryotes do not rely on hairpin formation in transcription termination. Nucleic Acids Res. *26*, 5456-5463.

[307] Wassarman, K. M. (2002). Small RNAs in bacteria: diverse regulators of gene expression in response to environmental changes. Cell *109*, 141-144.

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

[309] Watanabe, T., Sugiura, R., and Sugita, M. (1997). A novel small stable RNA, 6Sa RNA, from the cyanobacterium *Synechococcus* sp. strain PCC6301. FEBS Lett. *416*, 302-306.

[310] Weinberg, Z. (2006). RaveNnA. Version 0.2f Distributed by the author. http://bliss.biology.yale.edu/~zasha/ravenna/.

[311] Weinberg, Z., and Ruzzo, W. L. (2004). Exploiting conserved structure for faster annotation of non-coding RNAs without loss of accuracy. Bioinformatics *20 (Suppl. 1)*, i334-i341.

[312] Weinberg, Z., and Ruzzo, W. L. (2004). Faster genome annotation of non-coding RNA families without loss of accuracy. In Proc. Eighth Annu. Int. Conf. on Comp. Mol. Biol. (RECOMB) (ACM Press), pp. 243-251.

[313] Weinberg, Z., and Ruzzo, W. L. (2006). Sequence-based heuristics for faster annotation of non-coding RNA families. Bioinformatics *22*, 35-39.

[314] Wettich, A., and Biebricher, C. K. (2001). RNA species that replicate with DNA-dependent RNA polymerase from *Escherichia coli*. Biochemistry *40*, 3308-3315.

[315] White, H. B. (1976). Coenzymes as fossils of an earlier metabolic state. J. Mol. Evol. *7*, 101-104.

[316] Wickiser, J. K., Cheah, M. T., Breaker, R. R., and Crothers, D. M. (2005). The kinetics of ligand binding by an adenine-sensing riboswitch. Biochemistry *44*, 13404-13414.

[317] Wickiser, J. K., Winkler, W. C., Breaker, R. R., and Crothers, D. M. (2005). The speed of RNA transcription and metabolite binding kinetics operate an FMN riboswitch. Mol Cell *18*, 49-60.

[318] Wiggs, J. L., Bush, J. W., and Chamberlin, M. J. (1979). Utilization of promoter and terminator sites on bacteriophage T7 DNA by RNA polymerases from a variety of bacterial orders. Cell *16*, 97-109.

[319] Wilkinson, S. R., and Been, M. D. (2005). A pseudoknot in the 3' non-core region of the *glmS* ribozyme enhances self-cleavage activity. RNA *11*, 1788-1794.

[320] Willkomm, D. K., Minnerup, J., Huttenhofer, A., and Hartmann, R. K. (2005). Experimental RNomics in *Aquifex aeolicus*: identification of small non-coding RNAs and the putative 6S RNA homolog. Nucleic Acids Res. *33*, 1949-1960.

[321] Wilson, D. S., and Szostak, J. W. (1999). *In vitro* selection of functional nucleic acids. Annu. Rev. Biochem. *68*, 611-647.

[322] Winkler, W., Nahvi, A., and Breaker, R. R. (2002). Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. Nature *419*, 952-956.

[323] Winkler, W. C., Cohen-Chalamish, S., and Breaker, R. R. (2002). An mRNA structure that controls gene expression by binding FMN. Proc. Natl. Acad. Sci. USA *99*, 15908-15913.

[324] Winkler, W. C., Grundy, F. J., Murphy, B. A., and Henkin, T. M. (2001). The GA motif: an RNA element common to bacterial antitermination systems, rRNA, and eukaryotic RNAs. RNA *7*, 1165-1172.

[325] Winkler, W. C., Nahvi, A., Roth, A., Collins, J. A., and Breaker, R. R. (2004). Control of gene expression by a natural metabolite-responsive ribozyme. Nature *428*, 281-286.

[326] Winkler, W. C., Nahvi, A., Sudarsan, N., Barrick, J. E., and Breaker, R. R. (2003). An mRNA structure that controls gene expression by binding *S*-adenosylmethionine. Nat. Struct. Biol. *10*, 701-707.

[327] Wuyts, J., Perriere, G., and Van De Peer, Y. (2004). The European ribosomal RNA database. Nucleic Acids Res. *32*, D101-103.

[328] Yao, Z., Weinberg, Z., and Ruzzo, W. L. (2006). CMfinder—a covariance model based RNA motif finding algorithm. Bioinformatics *22*, 445-452.

[329] Yarnell, W. S., and Roberts, J. W. (1999). Mechanism of intrinsic transcription termination and antitermination. Science *284*, 611-615.

[330] Yen, L., Svendsen, J., Lee, J.-S., Gray, J. T., Magnier, M., Baba, T., D'Amato, R. J., and Mulligan, R. C. (2004). Exogenous control of mammalian gene expression through modulation of RNA self-cleavage. Nature *431*, 471-476.

[331] Zengel, J. M., and Lindahl, L. (1994). Diverse mechanisms for regulating ribosomal protein synthesis In *Escherichia coli*. Prog. Nucleic Acid Res. Mol. Biol. *47*, 331-370.

[332] Zhang, B. L., and Cech, T. R. (1997). Peptide bond formation by *in vitro* selected ribozymes. Nature *390*, 96-100.

[333] Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. *31*, 3406-3415.