ABSTRACT

STARMER, JOSHUA D. What can RNA hybrids tell us about translation?(Under the direction of Donald Bitzer, Mladen Vouk and Anne Stomp).

Molecular biologists have been observing interactions between messenger RNA (mRNA) molecules and other non-coding RNA molecules for quite some time. Here I revisit some of the classical hybridizations between the 16S ribosomal RNA (rRNA) and mRNA during initiation, as well as investigate the interactions between small interfering RNA (siRNA) molecules and mRNA. In reviewing rRNA-mRNA interactions, I observed that the majority of both bacterial and eukaryote genes can bind at the start codon. This novel result lead to a method for improving genome annotation as well as a new theory of translation initiation. The examination of siRNA-mRNA interactions lead to new criteria for predicting an siRNA's efficacy.

What can RNA hybrids tell us about translation?

by

Joshua D. Starmer

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Bioinformatics

Raleigh

2006

Approved By:

Dr. Jason Haugh	Dr. Jeffrey Thorne
Dr. Donald Bitzer Co-Chair of Advisory Committee	Dr. Paul Maggard
Dr. Mladen Vouk Co-Chair of Advisory Committee	Dr. Anne Stomp Co-Chair of Advisory Committee

For Frank.

Biography

Joshua Starmer was seven years old when he began playing the cello, although he did not find the instrument particularly interesting until his senior year in high-school. At that time three things occurred: 1) His current cello teacher, Brent Wissick, played him a recording of the Elgar cello concerto and the first movement melted his brain; 2) Henry Marr, one of his oldest brother's friends, asked him to play original music for cello and electric guitar (the two later formed the band 2manStan); 3) Josh played a concert at the Brewery on Hillsborough Street with the Miracle Legion. These three things suddenly made playing the cello a passion and he began to practice outside of lessons and rehearsals.

After graduating from high-school, Josh spent a year at the North Carolina School of the Arts studying cello with Robert Marsh. From there, Josh moved to the Oberlin Conservatory and studied music composition with Gary Nelson. While at Oberlin, Josh also learned how to program a computer, a skill that became useful later in life.

With his Bachelors of Music in hand, Josh moved to Charleston, South Carolina and had a steady gig at Saint Michael's episcopal church for two years and also played in the pit orchestras for theater productions. Charleston was also where Josh sat in on a college level introduction to biology course. The teacher, Robert Dillon, encouraged Josh to apply to graduate schools, and in particular, North Carolina State University.

In 2001, Josh moved to Raleigh, North Carlina and reunited with Henry Marr in the two bands *Oedipus Dick* and *The Sybaritic Gentlemen of Leisure.* Three years later Josh joined the Chapel Hill band *The Old Ceremony* and with them recorded two CDs, played to sold out audiences at Carborro's Cat's Cradle, and toured the east coast.

Acknowledgements

Many, many people helped me with this dissertation. Unfortunately, I am certainly going to forget to mention some of those who contributed the most. To those I offer my sincerest apologies.

For their advice, criticism, patience, and support, I would like to thank: Frank, Ellen, Jack, Mike, Rachel, Errol, Jessica, Raymond, David, Lalit, Chuanhua, Jeff, Ruben, Henry, Marti, Bibi, and Mary.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Before I began graduate school, before I had even taken a college level biology course, I became truly impressed by ribonucleic acid (RNA) while designing database software at a hospital during the day, and reading books on physics and biology that my father loaned me at night. One of these books was the 85 page *Origins of Life*, by physicist Freeman Dyson [18]. This book introduced me to both the key concepts in biology - namely reproduction, metabolism and evolution - and the notion that mathematical models could be applied to these concepts. In it, Dyson postulates that life originated from enzymatic proteins before DNA or RNA. However, Dyson gave special attention to RNA, opening my eyes to its flexibility. RNA, I discovered, had the potential to do it all.

Fast forward a few years - after finishing the bulk of my graduate coursework, I began looking for a topic for my dissertation. At that time, RNA interference (RNAi) was getting a good deal of press, and I remember the subject coming up quite often in Journal Club[1]. RNAi is the term for the inhibitory effects small, non-coding RNA molecules have on gene expression. Although the phenomenon of RNAi was observed in the '80s [39], it was not until the late '90s when Andrew Fire, Craig Mello, and their colleagues made the first major step toward understanding the mechanisms behind RNAi [22][2]. Their work resulted

[1] This was due largely to a functional genomics student, Jenora Waterman, and professor David Bird, who heads a *C. elegans* lab. I remember one day after class David Bird took me over to his lab to show off these tiny worms to me and being slightly underwhelmed with what I saw. Even when I squinted into the microscope, all I saw was what looked like a gray and black smudge. However, David Bird's enthusiasm for them impressed me. Since then, I've learned about the techniques for manipulating *C. elegans* and its genes, and have developed a good deal of respect for them as a model organism.

[2] On October 2, 2006, Fire and Mello were awarded the Nobel prize for their contributions to understanding RNAi's mechanism.

in the realization that when small-interfering RNA (siRNA) and micro-RNA (miRNA) molecules bind to messenger RNA (mRNA) sequences, they prevent the mRNA from being translated into protein, effectively silencing the gene. One very tantalizing aspect of RNAi, is that it allows researchers to silence a gene at a specific time. For example, an siRNA could be expressed via a promoter sequence that was only active during certain stages of development, silencing its target only during those stages.

Along with articles documenting the experimental successes with RNAi, I read articles describing bioinformatics approaches to identify miRNA targets. The interactions between miRNA and their mRNA target sequences were particularly amenable to bioinformatics analysis because the regularity of the base-paring between the two molecules[3]. At the time, however, there were no widely available tools for predicting how two separate RNA molecules might bind to each other. Bioinformatics publications examining interactions between miRNAs and potential targets coerced `Mfold`, a program that predicts the secondary structure of a single RNA molecule, into solving their problem [87]. Once I realized that researchers were using an indirect method for identifying miRNA targets, it became clear researchers needed a program specifically tailored to the problem of predicting how two separate RNA molecules might bind to each other.

Immediately, I set about finding an advisor to oversee my project. A friend, Errol Strain, suggested asking Jeffrey Thorne for advice. After hearing my ideas, Dr. Thorne recommended visiting the Bitzer-Vouk group because they had done quite a bit of work analyzing mRNA. In the fall of 2003 I attended one of the Bitzer-Vouk lab meetings to present my ideas and I found the group's energy and enthusiasm for problem solving contagious. I then began to attend the meetings regularly and after a few months Bitzer and Vouk agreed to be my advisors.

Since joining the Bitzer-Vouk group, I've worked on various projects, some of which generated publishable results. I began by reviewing the group's previous work and current projects. In the late '90s and early 2000s, the group identified what they called the *synchronization signal*. The synchronization signal indicates that, on average, the 16S rRNA tail binds to the coding region on an mRNA every three bases. Because the 16S rRNA is part of the essential protein synthesis machinery (the ribosome), and it must read an mRNA three bases at the time, this periodic binding pattern was thought to control the

[3] In contrast, protein-protein interactions require a fresh set of experimentally derived parameters to model binding for each individual protein.

position of the ribosome during translation.

If one accepted that the synchronization signal helped position the ribosome over the mRNA during translation, then it seemed reasonable to expect a stronger signal in organisms that lived at high temperatures. More specifically, I developed the hypothesis that organisms that lived at higher temperatures should have better signal to noise ratios (SNR) for the synchronization signal [4] in order to counter the increase in environmental noise caused by heat. What I found, however, was that while the SNR was positively correlated with GC content (the higher the GC content, the better the SNR ratio), neither the SNR nor the GC content correlated with optimal growth temperatures. These results were quite surprising since I remember being taught that hyperthermophiles had a higher percentage of GC base pairs in order to prevent the chromosomal DNA from unwinding [5]. Despite the somewhat disappointing results, I was able to use them in a conference paper and poster for GENSIPS 2005.

Another early project was to use the synchronization signal's SNR as a means for detecting exons in eukaryotic sequences. The idea was that exons would have a better SNR than introns due to evolutionary pressure to accurately translate the exon into amino-acid sequences. This theory was a bit of a stretch, because there is no direct experimental evidence suggesting that the 18S rRNA tail interacts with mRNA. However, there is a good deal of circumstantial evidence that these interactions are possible. For example, 16S and 18S rRNA have similar secondary and tertiary structures [85]. In addition, the tail sequence on a 18S rRNA is very similar to that of a 16S rRNA. While there is the obvious lack of an anti-Shine Dalgarno (aSD) element, $5'$-CCUCC-$3'$, in the 18S rRNA tail, all of the remaining bases are conserved in the 16S rRNA tail. This avenue of research, however, soon came to a dead end when I observed long, >400 nucleotide stretches within exons that had no signal at all. In these areas, the SNR was defined to be zero and thus, indistinguishable from introns.

Thinking I was making unreasonable assumptions about interactions between the 18S rRNA and the mRNA, I returned to prokaryotes as the objects of my analysis. At this time I tried to identify a correlation between the synchronization signal's SNR and protein concentrations. I also looked for correlations between Shine-Dalgarno binding data and

[4] For all of these studies, I defined the SNR ratio of the synchronization signal as the F-statistic for a sinusoid with a period of three bases fit to the data.

[5] The observation that GC content does not correlate with optimal growth temperature has been published by Glatier and Lobry [25]

protein expression levels. A previously published study demonstrated that the magnitude of the binding at SD sequences did not correlate with expression levels [58]. I thought that perhaps one reason for this result was that the analysis did not include the location of the SD sequence as a factor. I made this hypothesis based on a publication reporting correlations between SD binding and codon adaptation index (CAI) values for genes [60], as well as two experimental studies showing the location of the SD sequence and its magnitude affected protein concentration for a single gene [75, 11]. Despite my optimism, I failed to detect correlations.

At this point I backed off from studying the synchronization signal's SNR and started looking for more basic binding between the 16S rRNA and mRNA. After pursing several dead-ends, I wanted a sure win. To this end, I decided to reproduce the results in a study showing the presence of SD sequences in bacteria (and lack thereof in eukaryotes) using predictions of how the 16S tail would bind to sequences upstream from start codons [70]. As I had hoped, I was able to reproduce these results with an interesting difference. The difference occurred a few bases after the SD sequences and before the very $3'$ end of the tail passed the start codon. At this location a considerable number of genes were able to bind to the 16S rRNA tail. I made this discovery right before my laptop computer died, and when it did, I spent the next few weeks reading articles on codon bias in the first few codon positions [10, 59, 91, 92, 90]. Most of the codons in the second codon position associated with high expression rates contributed one or two consecutive bases to hybridization with the 16S rRNA tail. With a computer loaned to me by Michael Isen[6] I began to explore what would eventually be labeled *RS+1 binding*, the binding between the 16S tail and the first two codons.

Chapters 2 and 3 are primarily focused on characterizing RS+1 binding. Chapter 2 shows how identifying excessive binding potential at RS+1 improves computer-based genome annotation. In a nutshell, where there is excessive binding at RS+1, there is likely to be an error in the annotation. This chapter was published last spring in PLoS Computational Biology [88]. In Chapter 3, (which I recently submitted to PLoS Computational Biology) I first look to see if there is a relationship between RS+1 binding and SD binding. I did not find one [7], but I did find an interesting pattern of bases on the 16S tail that bind

[6] Mike Isen is a *Drosophila* researcher at the Lawrence Berkeley National Laboratory who helped found PloS (Public Library of Science). We met by a fairly random change when I was spending some time in Berkeley, CA. He was married to an art historian whom my girlfriend invited over for dinner one night.

[7] This result, however, was enough to get a paper with it accepted to GENSIPS2006 and the poster I made

at RS+1. The bases that bind at RS+1 are the ones that flank the aSD. These bases, while not implicated in doing much in terms of SD-aSD binding, are highly conserved in the 18 bacteria I studied and they are also conserved in eukaryotic 18S rRNA tail sequences. I then examined whether eukaryotes also have binding at RS+1, and they do. These results imply that it is possible that translation initiation requires two steps. The first step is ribosome recruitment. This step is taken care of by aSD-SD binding in prokaryotes [83, 38, 41] and “scanning” from the $5'$ cap in eukaryotes [50, 52]. The second step, which may be affected by binding at RS+1, is a fine tuning of the ribosome’s location over the coding portion of the mRNA.

In Chapter 4, we return to what brought me to the Bitzer-Vouk lab in the first place: RNAi. The tools for analyzing the binding potential between the 16S rRNA tail and the mRNA are general purpose tools and, without modification, can analyze the base pairing potential between siRNAs and their targets. In this chapter I show off the general purpose nature of my programs and demonstrate how they can show novel correlations between siRNAs and their efficacy. This chapter was recently submitted to RNA.

for this conference was awarded “Best Poster” at the 2006 NCSU Graduate Student Research Symposium.

Chapter 2

Predicting Shine-Dalgarno sequence locations exposes genome annotation errors.

2.1 Abstract

In prokaryotes, Shine-Dalgarno (SD) sequences, nucleotides upstream from start codons on messenger RNAs (mRNAs) that are complementary to ribosomal RNA (rRNA), facilitate the initiation of protein synthesis. The location of SD sequences relative to start codons and the stability of the hybridization between the mRNA and the rRNA correlate with the rate of synthesis. Thus, accurate characterization of SD sequences enhances our understanding of how an organism's transcriptome relates to its cellular proteome. We implemented the Individual Nearest Neighbor Hydrogen Bond model for oligo-oligo hybridization and created a new metric, Relative Spacing (RS), to identify both the location and the hybridization potential of SD sequences by simulating the binding between mRNAs and single-stranded 16S rRNA $3'$ tails. In 18 prokaryote genomes, we identified 2,420 genes out of 58,550 where the strongest binding in the translation initiation region included the start codon, deviating from the expected location for the SD sequence of 5 to 10 bases upstream. We designated these as RS+1 genes. Additional analysis uncovered an unusual bias of the start codon in that the majority of the RS+1 genes used GUG, not AUG. Furthermore, of the 624 RS+1 genes whose SD sequence was associated with a free energy

release of less than -8.4 kcal/mole (strong RS+1 genes), 384 were within 12 nucleotides upstream of in-frame initiation codons. The most likely explanation for the unexpected location of the SD sequence for these 384 genes is mis-annotation of the start codon. In this way, the new RS metric provides an improved method for gene sequence annotation. The remaining strong RS+1 genes appear to have their SD sequences in an unexpected location that includes the start codon. Thus, our RS metric provides a new way to explore the role of rRNA-mRNA nucleotide hybridization in translation initiation.

2.2 Introduction

In 1974 Shine and Dalgarno [83] sequenced the $3'$ end of *E. coli*'s 16S ribosomal RNA (rRNA) and observed that part of the sequence, $5'-$ACCUCC$-3'$, was complementary to a motif, $5'-$GGAGGU$-3'$, located $5'$ of the initiation codons in several messenger RNAs (mRNAs). They combined this observation with previously published experimental evidence and suggested that complementarity between the $3'$ tail of the 16S rRNA and the region $5'$ of the start codon on the mRNA was sufficient to create a stable, double-stranded structure that could position the ribosome correctly on the mRNA during translation initiation. The motif on the mRNAs, $5'-$GGAGGU$-3'$, and variations on it that are also complementary to parts of the $3'$ 16S rRNA tail, have since been referred to as the Shine-Dalgarno (SD) sequence. Shine and Dalgarno's theory was bolstered by Steitz and Jakes in 1975 [89] and eventually experimentally verified, in 1987, by Hui and de Boer [38] and Jacob et al. [41].

Since Shine and Dalgarno's publication, two different approaches have been used to identify and position SD sequences in prokaryotes: sequence similarity and free energy calculations.

Methods based on sequence similarity include searching upstream from start codons for sub-strings of the SD sequences that are at least three nucleotides long [93]. Identification errors can arise from this approach for several reasons [82]. A threshold of similarity does not exist that can clearly delineate actual SD sequences from spurious sites with a significant, but low, degree of similarity to the SD sequence. The lack of certainty has led to a number of observations in which gene sequences appear to partition themselves into two categories: those with obvious SD sequences and those without. The inability of sequence techniques to pinpoint the exact location of the SD sequence poses a problem because its location is believed to affect translation initiation [11, 75, 60, 52].

The second approach, using free energy calculations, is based on thermodynamic considerations of the proposed mechanism of 30S binding to the mRNA and overcomes the limitations of sequence analysis. Watson-Crick hybridization occurs between the 3′-terminal, single-stranded nucleotides of the 16S rRNA (the rRNA tail) and the SD sequence in the mRNA and has a significant effect on translation [38, 41]. The formation of hydrogen bonds between aligned, complementary nucleotides is the basis of Watson-Crick hybridization and results in a more stable, double-stranded structure with lower free energy than the participating single-stranded sequences. One long-standing implementation of this model, `Mfold` [110], quantifies the degree of hybridization and the stability of RNA secondary structure by calculating the change in energy (ΔG°) [40, 104, 62]. This method for estimating free energy has been adapted to identify SD sequences by repeatedly calculating the ΔG° values for progressive alignments of the rRNA tail with the mRNA in the region upstream of the start codon [93, 82, 70, 58]. All of these studies have observed a trough of negative ΔG° upstream of the start codon whose location is largely coincident with the SD consensus sequence. This second approach can both identify the SD sequence and pinpoint its exact location as that having the minimal ΔG° value. However, the exact location of the SD sequence is dependent on the nucleotide indexing scheme of the algorithm, i.e. on which nucleotide is designated as the "0" position.

To normalize indexing and to further extend free energy analysis through the start codon and into the coding region of genes, we created a new metric, *relative spacing* (RS). This metric localizes binding across the entire translation initiation region (TIR), relative to the rRNA tail, enabling us to characterize binding that involves the start codon as well as sequences downstream. RS is also independent of the length of the rRNA tail, and this property allows for comparison of binding locations between species.

By examining sequences downstream from start codons, we could explore mRNAs that lack any upstream region, the *leaderless* mRNAs [103, 20, 61, 65, 69, 102]. The lack of any 5′ untranslated leader in the mRNAs has prompted searches for other sequence motifs that could interact with the 16S rRNA. One of these, the downstream box hypothesis [86], has been disproved [68]. Thus, there is a continued search for an explanation for the highly conserved sequences 3′ of the initiation codon that have been observed in many leaderless mRNAs [86, 21, 102].

In this study we use the RS metric to identify the positions of minimal ΔG° troughs for genes of 18 species of prokaryotes as a test of its usefulness as a means to improve existing annotation tools, i.e. by identifying SD sequences. We observe 2420 genes where the strongest binding in the entire TIR takes place one nucleotide downstream from the start codon, at RS+1. Of these, 624 genes have unusually strong binding (less than -8.4 kcal/mol). We then determine if these 624 genes were mis-annotated and conclude that 384 are.

2.3 Results

The average ΔG° value at each position of the TIR for each species is shown in Figure 2.1, aligned according to RS. The ΔG° troughs upstream from RS 0 are consistent with previous experimental studies on the location of the SD sequence [75, 11], as well as with computational studies either simulating free energy changes [70, 77] or using information theory [84]. The ΔG° trough immediately after the first base in the initiation codon, at RS+1, is unexpected, but present in a significant portion of genes in all species examined. The histograms of Figure 2.2 show the distributions of RS positions of the strongest SD-like sequences (where $\Delta G^{\circ} < -3.4535$, see the Methods section for more details) in each TIR for all genes within a species. For all genes that contain an SD-like sequence, we will call genes where the lowest ΔG° value is at RS+1 *+1 genes*, and +1 genes where $\Delta G^{\circ} < -8.4$ kcal/mol *strong +1 genes*. Genes where the strongest SD-like sequence is between RS-20 and RS-1, inclusive, are designated *upstream genes*, and similarly, *downstream genes* are genes where the strongest SD-like sequence is between RS+1 and RS+20, keeping in mind that these designations do not imply that other SD-like sequences do not exist in the TIR, but only that they do not bind to the rRNA as well. If a trough of minimal free energy can be definitive of the SD sequence, a site whose location is presumed to be upstream from the coding region, the +1 genes are unexpected in that they exist within, not upstream from, the coding region. Our study focuses on the characterization of the sequence interactions that give rise to strong +1 genes and on possible explanations for their presence; we have reserved the downstream genes for future analysis.

We thought of four hypotheses to explain the unexpected RS+1 result. 1) The +1 site is an artifact of our model or implementation. 2) The +1 trough could result from known sequence bias around the start codon, assuming the start codon annotation is correct. 3) The start codon annotation could be incorrect: the presence of in-frame start codons downstream of the annotated start codons would be consistent with this interpretation. 4) If

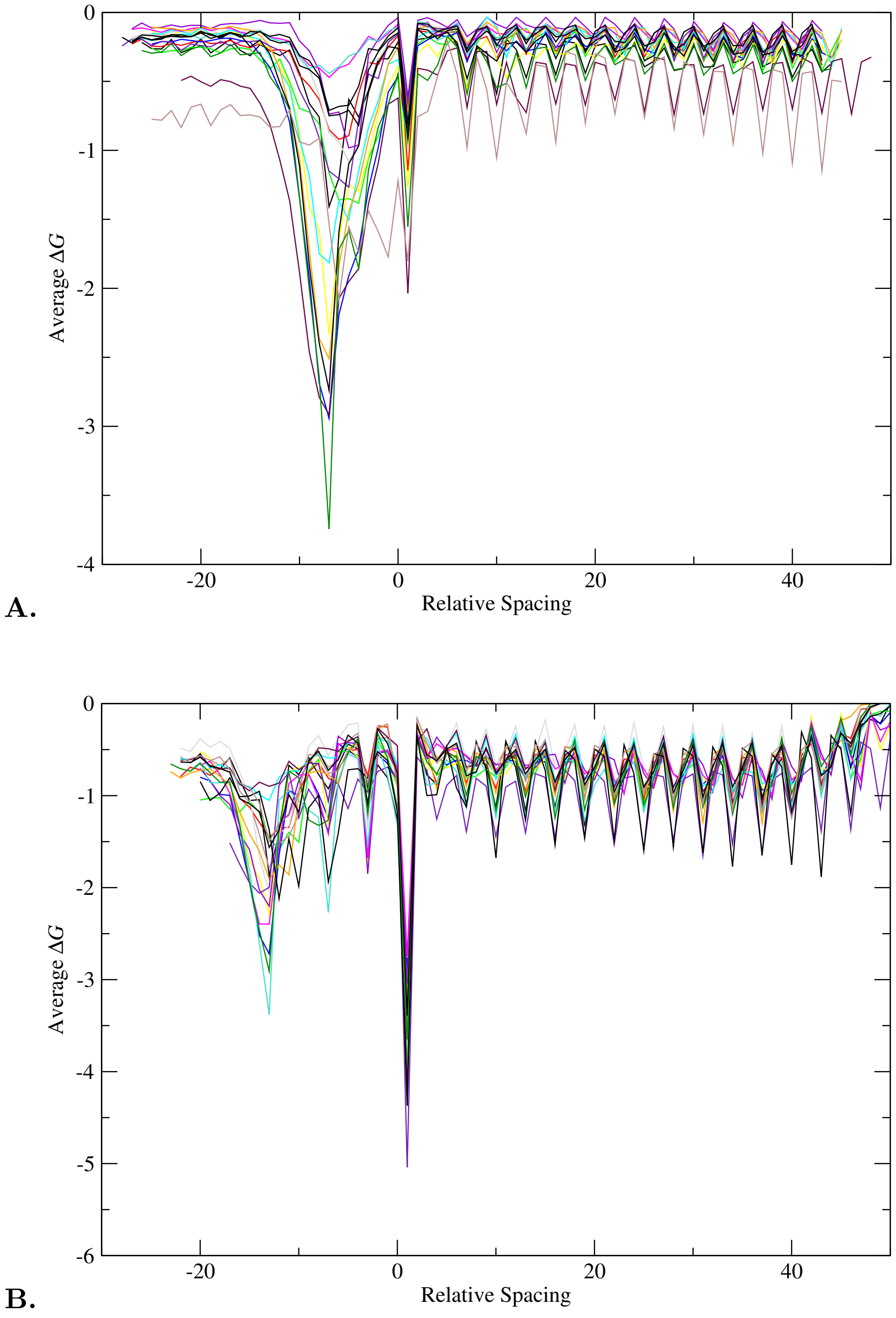

Figure 2.1: **Average ΔG° values in the TIRs for 18 organisms.** For all 18 genomes in our study, we calculated the average ΔG° value for each Relative Spacing (RS) position. Zero on the x-axis corresponds to the 5′ A residue in the rRNA sequence 5′−ACCUCC−3′ being positioned over the first base in the initiation codon. The dramatic drops in ΔG° prior to RS 0 show the presence of Shine-Dalgarno (SD) sequences. The sudden drop in ΔG° immediately after the first base in the initiation codon (at RS+1) shows that there is a significant binding potential between the 16S rRNA and the mRNA close to the initiation codon, an unexpected location. **A** was drawn from data generated by `free_scan` and **B** is from data generated from `RNAhybrid` [74]. Differences between the two graphs are discussed in the text.

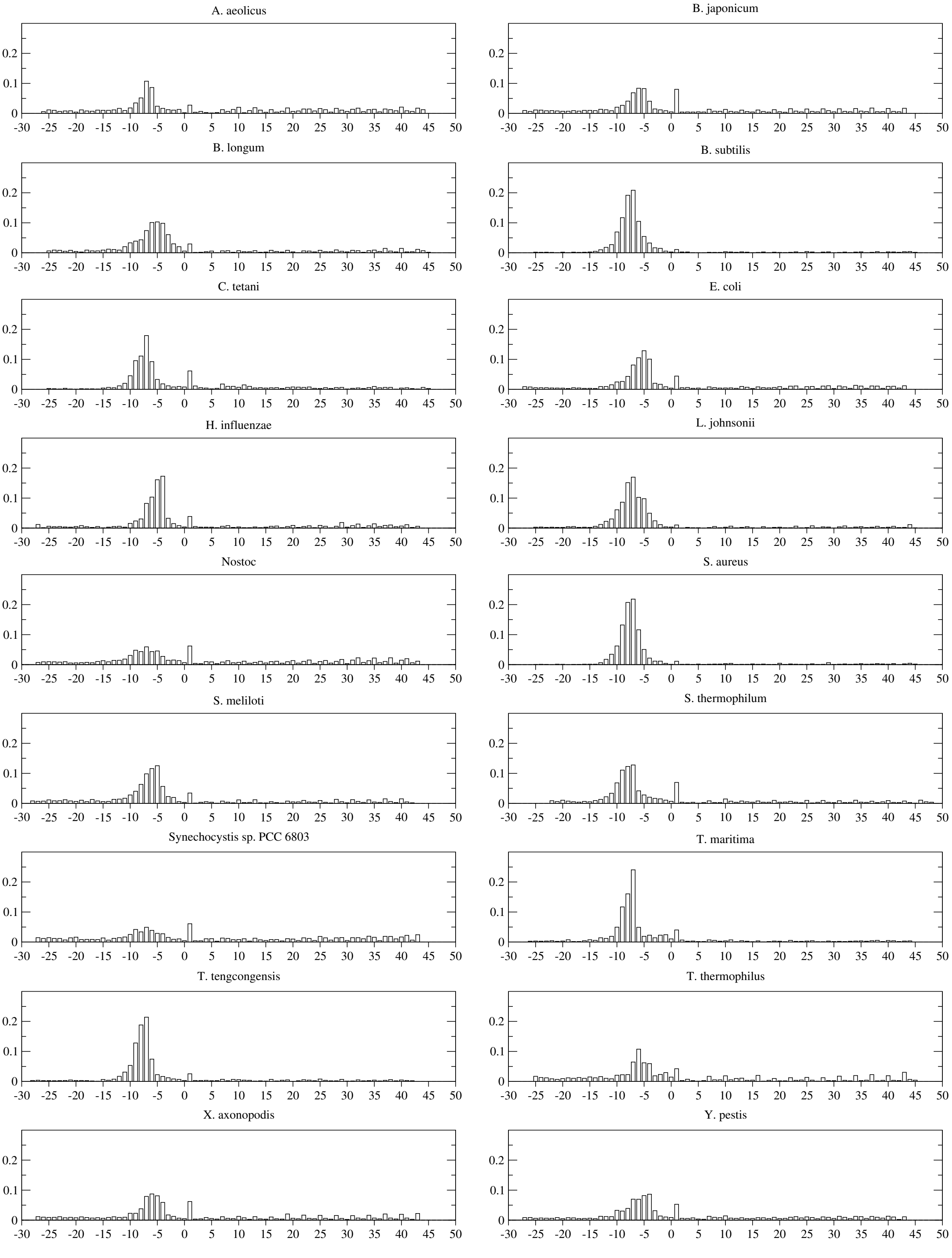

Figure 2.2: **Normalized histogram plots showing the RS for the lowest ΔG° values in the TIRs.** The x-axis shows the RS, or distance between the 5′ A residue in the rRNA sequence 5′−ACCUCC−3′ from the 3′ tail and the first base in the start codon. Negative numbers indicate that the 5′ A is upstream from the start codon, while positive numbers indicate that it is downstream. The y-axis is the fraction of genes in a genome where the lowest ΔG° value is at a particular RS.

there were sequence errors in the start codon, they could potentially change the free energy calculation for alignments in which the three nucleotides of the start codon participated. All four of these hypotheses were examined.

We were quickly able to dispose of our first hypothesis. The +1 site is not an artifact of the INN-HB model or its implementation. Both the INN and the INN-HB RNA secondary structure models are based on thermodynamics and use experimentally derived parameters. Implementations of INN models using dynamic programming have a well-established history of accurately predicting secondary structures for short RNA sequences [110, 24, 62] and SD sequence identification [70, 77, 60, 58, 82, 16]. The more recent INN-HB model improves secondary structure predictions in newer versions of `Mfold` [62]. While this study is the first use of the INN-HB model for SD sequence detection, it is not the first example of its use for oligo-oligo hybridization predictions [35]. With the exception of the +1 site, the results that our implementation of the INN-HB model generate are consistent with both experimental [11, 75] and computational studies [70, 96, 97, 57] of SD and coding sequences. Furthermore, analysis performed with `RNAhybrid` [74] is consistent with our results (see Figure 2.1). Based on this evidence, it is clear that the +1 site is not an artifact of the model we are using or its implementation.

The second hypothesis assumes that the significant negative free energy value at RS+1 results primarily from nucleotide biases in the first two codons of the coding region. Obviously there is extreme codon bias in the start codon for all genes and, therefore, for all species examined, as shown in Table 2.1. Studies of TIR sequences in *E. coli* have shown considerable bias in the second codon as well [92, 90, 91]. To examine this bias, sequence logos [80, 12] (http://weblogo.berkeley.edu/) were created for the region of mRNA that would be aligned with the rRNA tail for RS+1 (see Figure 2.3, *radC* (GeneID:948968), for an example of this alignment). Figure 2.4 is a sequence logo for *E. coli* genes that includes the first two codons. This logo was representative of the sequence logos for all 18 organisms (data not shown). For *E. coli*, the sequence logo gives two options for relatively abundant sequences that could bind to the rRNA tail: AUGA and GUGA. AUGA has a positive ΔG° value of 0.21 kcal/mole and cannot explain the trough of ΔG°. The alternate sequence, GUGA, has a negative ΔG° value of -1.88 kcal/mol. However, if all 570 *E. coli* genes whose start codons are GUG had this value, the total would be too small to cause the average value of the 4254 E. coli genes to be -0.79 kcal/mol. Using the same approach with the sequence logos for the remaining 17 organisms, sequence bias of the first two codons also failed to

Table 2.1: **Usage statistics for the three most common initiation codons, AUG, GUG, and UUG.** For all 18 organisms, AUG is the most commonly used start codon in upstream genes. The most commonly used start codon in strong +1 genes is GUG.

The total number of genes in each row may not add up to the total number of genes in an organism for two reasons: not all +1 genes were examined, only strong +1 genes, and a small set of genes do not use AUG, GUG, or UUG for start codons.

	Start Codon Usage - Upstream Genes			Start Codon Usage - Strong +1 Genes		
Organism	AUG	GUG	UUG	AUG	GUG	UUG
A. aeolicus	84% (554)	9% (57)	8% (50)	5% (1)	95% (19)	
B. japonicum	83% (2965)	16% (575)	1% (25)	2% (2)	98% (109)	
B. longum	82% (889)	12% (134)	6% (62)	33% (1)	33% (1)	33% (1)
B. subtilis	78% (2826)	8% (306)	13% (454)	8% (1)	83% (10)	8% (1)
C. tetani	80% (1216)	8% (115)	12% (185)	6% (5)	92% (73)	1% (1)
E. coli	90% (2041)	8% (193)	2% (37)		100% (28)	
H. influenzae	96% (918)	4% (34)	1% (7)		100% (2)	
L. johnsonii	86% (1269)	7% (98)	7% (107)	25% (1)	75% (3)	
Nostoc	84% (1398)	15% (254)	1% (12)	7% (3)	93% (38)	
S. aureus	85% (2049)	7% (158)	8% (198)		89% (8)	11% (1)
S. meliloti	88% (1777)	7% (140)	6% (113)		100% (8)	
S. thermophilum	57% (1275)	34% (752)	9% (199)	2% (3)	98% (121)	
Synechocystis	83% (721)	17% (150)			100% (15)	
T. maritima	71% (1038)	18% (269)	11% (157)	4% (2)	96% (50)	
T. tengcongensis	77% (1566)	12% (242)	11% (221)		100% (24)	
T. thermophilus	75% (793)	20% (216)	5% (48)	10% (4)	90% (37)	
X. axonopodis	82% (1446)	12% (214)	6% (104)	4% (1)	96% (23)	
Y. pestis	81% (1514)	11% (213)	8% (143)		96% (26)	4% (1)

explain the average negative free energy trough associated with the RS+1 alignment.

The third hypothesis assumes incorrect sequence annotation for the start codon in the strong +1 genes. To eliminate the possibility that a bias in a particular sequence annotation program caused the RS+1 site, we verified that the genomes in our study had been annotated using different tools (see Table 2.2). `GLIMMER` was used for half of the genomes, and the remaining genomes were annotated with `GeneMark`, `FrameD`, `ORPHEUS`, and `GeneLook`. Thus, if the RS +1 site can be explained as sequence annotation errors, these errors are being made by a variety of software packages.

One way to detect sequence annotation errors as the cause of the RS+1 site is to look for in-frame start codons downstream from the start codons annotated in GenBank. To investigate this potential explanation for strong +1 genes, 12-nucleotide-long sequences downstream from the annotated start codon were scanned for in-frame start codons. The results are shown in Table 2.3. The rationale for scanning 12 nucleotides downstream came from the observation that, in the majority of genes, the SD sequence is located within 10 nucleotides upstream from the start codon. As seen in Table 2.3, only a small percentage of the TIRs of upstream genes have in-frame start codons downstream from the annotated start *wecF*

```
                               v
rRNA:       3'-a U U C C U C C A c u a g-5'
                 | | | | | | | |
mRNA: 5'-c g c c A G G G A G G U c g c a u g a g-3'
                                       ^
                               |< -4  >|
```

argD

```
                         v
rRNA: 3'-a u u c c U C C A C U A G-5'
                   | | | | | | | |
mRNA: 5'-u a a c a G G G U G A U C a u g a g a u-3'
                                             ^
                         |<----  -10  ------>|
```

radC

```
                             v
rRNA:     3'-a U U C C U C C A C U a g
               | | | | | | | | | |
mRNA: 5'-a c a A A G G A G G U G A a g g u g a a-3'
                           ^
                           |+1|
```

Figure 2.3: **Examples from *E. coli* showing how relative spacing is calculated.** The complementary bases, plus G/U mismatches, that are predicted to bind together are capitalized. The predicted SD sequence consists of the capitalized letters in the mRNA. The location of the start codon is indicated with the hat character, ^, and the location of the $5'$ A residue in the rRNA sequence $5'$−ACCUCC−$3'$ is indicated with a v. The relative spacing is the distance between the $5'$ A and the first base in the start codon. If the SD is upstream from the start codon, then the relative spacing is given as a negative number. If the SD is downstream, it is given as a positive number. Both SD sequences for *wecF* (GeneID:2847677) and *argD* (GeneID:947864) come before the start codons (in these cases, the start codon is AUG). The relative spacing for *wecF* is -4 and for *argD* it is -10. *radC*'s SD sequence includes the start codon, GUG, and the relative spacing is +1.

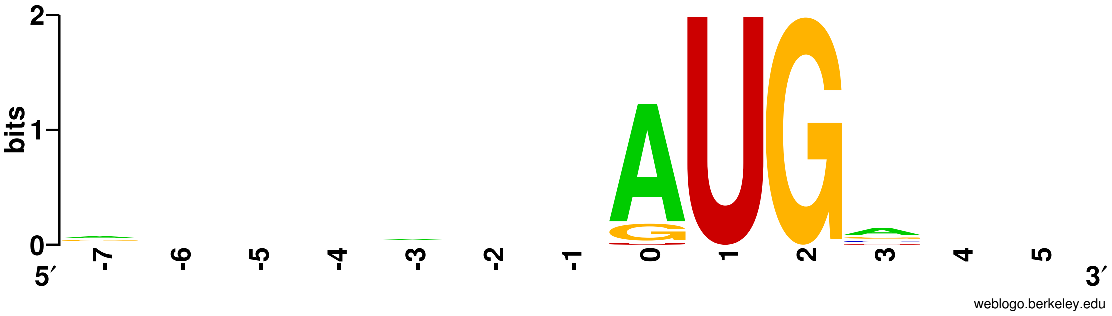

Figure 2.4: **A sequence logo for *E. coli*.** mRNA bases between positions -7 to 5 would need to bind to the rRNA tail for RS+1. For each position, the sequence logo displays amount of information content and the frequency of nucleotides. Positions that have no information content are blank, whereas those with information content contain a stack of nucleotide characters. The size of the nucleotide character in the stack is proportional to its frequency at that position.

Table 2.2: **A summary of the annotation programs used for the genomes in this study.** In addition to the program listed, all genomes used comparative ORF identification methods i.e. `BLASTP` and `BLASTX` applied to a non-redundant sequence database. The variety of annotation tools used to characterize ORFs suggests that the RS +1 site is not an artifact of any single tool.

[1] Both the original annotation and the reviewed REFSEQ used `GeneMark`.

Organism	Annotation Tool	Year Published
A. aeolicus [15]	comparative analysis only	1998
B. japonicum [44]	`GLIMMER`	2002
B. longum [79]	`ORPHEUS`	2002
B. subtilis [54]	`GeneMark`	1997
C. tetani [7]	`GLIMMER`	2003
E. coli [3]	comparative analysis only	1997
H. influenzae [23]	`GeneMark`[1]	1995
L. johnsonii [73]	`FrameD`	2004
Nostoc [45]	`GLIMMER`	2001
S. aureus [36]	`GLIMMER` and `ORPHEUS`	2004
S. meliloti [9]	`FrameD`	2001
S. thermophilum [98]	`GLIMMER` and `GeneLook`	2004
Synechocystis [43]	`GeneMark`	1996
T. maritima [67]	`GLIMMER`	1999
T. tengcongensis [1]	`GLIMMER`	2002
T. thermophilus [34]	`GeneMarkS`	2004
X. axonopodis [13]	`GLIMMER` and `GeneMark`	2002
Y. pestis [17]	`GLIMMER`	2002

site. In contrast, the majority of strong +1 genes have downstream, in-frame start codons that could serve as the actual start codons. This finding is consistent with the interpretation that at least a subset of strong +1 genes actually have errors in start codon annotation. All 28 strong +1 genes in *E. coli* contain a disagreement between the GenBank annotated start codons and the EcoGene database annotation, a database employing hand-curated annotation that is presumably more accurate [76]. These disagreements in annotation are probably the result of Blattner et al. selecting the start codon that will allow the open reading frame to be extended as far upstream as possible [3]. *E. coli*'s *radC* gene provides a useful example: assuming the GenBank annotation to be correct, the Relative Spacing metric identifies *radC* as a strong +1 gene. However, as can be seen in Figure 2.3, the initiation region sequence has an in-frame GUG six bases downstream from the annotated start codon. If the downstream GUG codon were the true start codon, then the gene would not be a strong +1 gene but would have its trough of minimal free energy in the regular, upstream SD position. Future experiments could differentiate these alternatives by examining the amino acid sequence of the gene's protein.

Another type of annotation error may explain the strong +1 genes that remain after accounting for those whose start codons are incorrectly located, the number of which, by species, are shown in Table 2.3. In *E. coli*, there are five strong +1 genes in which mis-annotation of their start codon position does not serve as an explanation of the unexpected position of their minimal free energy trough. In the GenBank database, all of these five genes are tagged as "hypothetical" or "putative", indicating that the assumption that they encode a polypeptide has not been verified. It is possible that they do not encode proteins. Therefore, at least in the case of *E. coli*, a strong case can be made for mis-annotation causing the RS+1 designation of these genes.

The fourth hypothesis proposes that sequence errors might account for the presence of a minimal free energy trough at the RS+1 alignment. To examine this idea further, Table 2.1 summarizes the frequencies of the three start codons in genes with minimal free energy troughs in the expected, upstream alignment (the upstream genes) versus strong +1 genes. It is immediately apparent that there is a significant bias in strong +1 genes towards the use of GUG start codons. One possible reason strong +1 genes preferentially utilize GUG as the start codon is that sequencing errors may have occurred, and that in actuality at least a portion of these genes used AUG as their start codons. The RS+1 trough would then, presumably, result from these sequencing errors. To test this hypothesis, GUG start

Table 2.3: **Downstream start codons.** The percentages of genes with in-frame start codons (AUG, GUG, or UUG) within 12 nucleotides of the annotated start site are shown for both upstream genes and strong +1 genes. Strong +1 genes are much more likely to have in-frame downstream start codons. The Adjusted Relative Spacing shows what the RS would be for strong +1 genes if the downstream start codon was the true start site, as well the number of initiation regions that would have that RS.

	Downstream Start Codons		Adjusted RS			
Organism	Upstream Genes	Strong +1 Genes	-1	-4	-7	-10
A. aeolicus	15%	70% (14 of 20)	0	1	13	0
B. japonicum	16%	50% (56 of 111)	21	18	12	5
B. longum	17%	33% (1 of 3)	0	1	0	0
B. subtilis	17%	50% (6 of 12)	0	0	6	0
C. tetani	11%	92% (73 of 79)	10	2	51	10
E. coli	15%	82% (23 of 28)	7	9	6	1
H. influenzae	10%	50% (1 of 2)	0	0	1	0
L. johnsonii	8%	50% (2 of 4)	0	0	1	1
Nostoc	14%	56% (23 of 41)	6	7	8	2
S. aureus	13%	56% (5 of 9)	0	0	5	0
S. meliloti	15%	12% (1 of 8)	0	0	0	1
S. thermophilum	17%	52% (64 of 124)	3	10	44	7
Synechocystis	17%	53% (8 of 15)	5	1	0	2
T. maritima	27%	85% (44 of 52)	4	2	36	2
T. tengcongensis	19%	88% (21 of 24)	4	3	14	0
T. thermophilus	21%	44% (18 of 41)	2	10	3	3
X. axonopodis	17%	38% (9 of 24)	2	5	1	1
Y. pestis	19%	48% (13 of 27)	5	5	2	1

Table 2.4: **Binding at the start codon for strong +1 genes compared to upstream binding.** In order to determine the differences in ΔG° between the strong binding at RS+1 and the most stable binding found within the canonical location for SD sequences, -10 to -4 RS, for these same genes, we calculated their averages, $\overline{\Delta G^\circ}$. The number of genes used to calculate each average, N, the number of strong +1 genes, is listed in the second column.

Organism	N = Strong +1 Genes	$\overline{\Delta G^\circ}$ -10 to -4 RS	$\overline{\Delta G^\circ}$ Strong RS+1
A. aeolicus	20	-0.44	-13.76
B. japonicum	111	-1.59	-10.38
B. longum	3	-5.33	-9.65
B. subtilis	12	-3.42	-10.78
C. tetani	79	-0.74	-10.97
E. coli	28	-0.77	-11.09
H. influenzae	2	0.00	-9.29
L. johnsonii	4	-3.21	-11.20
Nostoc	41	-1.21	-10.49
S. aureus	9	-0.25	-12.19
S. meliloti	8	-2.66	-9.86
S. thermophilum	124	-2.67	-12.37
Synechocystis	15	-1.81	-9.26
T. maritima	52	-2.17	-12.67
T. tengcongensis	24	-1.65	-10.74
T. thermophilus	41	-2.57	-12.95
X. axonopodis	24	-2.73	-9.88
Y. pestis	27	-1.03	-10.71

codons in strong +1 genes were changed to AUG start codons, and AUG start codons in all other genes were changed to GUG. Free energy values were calculated for these new sequences, and RS values were determined for each gene. For strong +1 genes, the RS values for the lowest ΔG° values were uniformly distributed (data not shown). In the case of the remaining genes, the changes resulted in many more of the initiation regions having their most stable binding at RS+1. However, the ΔG° value at RS+1 in these modified start codon sequences was only marginally stronger than the free energy trough still present at the upstream SD site. The small difference in energy values between the upstream SD site and the RS+1 site contrasts with that seen using the actual sequences of RS+1 genes. In those cases, the difference in energy values is quite large, as seen in Table 2.4.

Table 2.5 summarizes our results. It lists the total number of genes examined in each species, the number of upstream, downstream, +1, and strong +1 genes identified, as well as the number of strong +1 genes that do not appear to be artifacts of mis-annotation.

Table 2.5: **A summary of predicted rRNA-mRNA binding.** Upstream (US) genes are those where the strongest SD-like sequence ($\Delta G^{\circ} < -3.4535$) in the TIR takes place between RS-20 and RS-1, inclusive. Downstream (DS) genes are those where the strongest SD-like sequence in the TIR takes place between RS+1 and RS+20, inclusive. +1 Genes have their strongest SD-like sequence at RS+1. Strong +1 Genes are +1 genes that have $\Delta G^{\circ} < -8.4$ kcal/mol at RS+1. Unexplained Strong +1 Genes shows the number of strong +1 genes that do not have in-frame start codons within 12 nucleotides downstream from the annotated start codon. We predict that strong +1 genes that do have in-frame start codons just downstream are mis-annotated.

[1] These unexplained genes could be non-expressing open reading frames, as discussed in the text.

Organism	Genes	US Genes	DS Genes	+1 Genes	Strong +1 Genes	Unexplained Strong +1 Genes[1]
A. aeolicus	1529	661	267	38	20	6
B. japonicum	8317	3655	1573	579	111	55
B. longum	1727	1085	174	46	3	2
B. subtilis	4106	3600	184	45	12	6
C. tetani	2373	1516	461	141	79	6
E. coli	4254	2272	554	163	28	5
H. influenzae	1656	960	115	32	2	1
L. johnsonii	1821	1447	89	18	4	2
Nostoc	5366	1667	808	232	41	18
S. aureus	2739	2405	117	30	9	4
S. meliloti	3332	2030	340	103	8	7
S. thermophilum	3337	2226	543	229	124	60
Synechocystis	3167	871	475	135	15	7
T. maritima	1858	1464	190	74	52	8
T. tengcongensis	2588	2029	234	64	24	3
T. thermophilus	1982	1059	340	82	41	23
X. axonopodis	4312	1764	624	196	24	15
Y. pestis	4086	1870	654	182	27	14

2.4 Discussion

There is a long history of investigating SD sequences using approaches grounded in thermodynamics [93, 82, 58, 70, 77, 60]. As newer models are proposed and more accurate parameter values published, these methods have improved over the years. Here we present a new method that uses these previous approaches as a point of departure and, through both major and minor changes, enhances our ability to characterize SD sequences accurately.

Three major differences separate our method from prior methods. The primary difference is that we are examining both upstream and downstream sequences. Investigating downstream sequences allowed us to observe the large number of hybridization sites that include the start codon. The second main difference is our use of relative spacing as a means to compare hybridization locations among species. The third difference is our use of the INN-HB model instead of the INN model.

There are also many minor differences between our method and its predecessors. The most common are discrepancies in rRNA tail selection. We defined the 16S rRNA tails based on proposed secondary structures and conserved single-stranded 16S rRNA motifs. The sequences we used are the maximum number of single-stranded nucleotides available for hybridization based on accepted models of rRNA secondary structure. Osada et al. used the last 20 nucleotides of the 16S rRNA sequence without consideration of secondary structure models and the intramolecular helix formation that a significant portion of their $5'$ bases are involved [70]. On the other hand, Ma et al. enforce a 12 nucleotide limit on the length of the rRNA tails and truncate any that are longer [60]. Sakai et al. base their anti-SD motifs on the most frequent 7-mer found within 40 bases upstream of the start codon on the mRNA sequences [77], without reference to rRNA sequences.

As a result of these differences, our method improves SD sequence characterization. Table 2.6 shows the effect of using the INN-HB model in lieu of the INN model, used in Ma et al., as well as allowing for flexible tail lengths. For each organism common to both studies, we were able to identify more upstream SD sequences. Sakai et al. were unable to observe an upstream ΔG° trough indicative of SD sequences in *Synechocystis* [77]. Our method reveals the SD trough (see Figure 2.5 and Table 2.6). Comparison with Schurr et al.'s results [82] shows benefits to using the INN-HB model in conjunction with relative spacing and examining downstream sequences. Of the 38 genes they identified as having $\Delta G^\circ \geq 0$ kcal/mol, and thus no discernible binding site for the rRNA tail, we were able to

Table 2.6: **Model Comparisons.** The INN-HB model is able to identify a larger percentage of SD sequences in the 20 nucleotides upstream from the start codon than the INN model. When using the INN-HB model, the SD threshold is $\Delta G^\circ \leq -3.4535$ kcal/mol, which is the average value from binding GGAG, GAGG, and AGGA to the 16 rRNA tail. This is equivalent to using $\Delta G^\circ \leq -4.4$ kcal/mol as threshold for the INN model [60] (see text for more details). The third and fourth columns show the difference between using the same 12-nucleotide long rRNA tails that Ma et al. used, and using the longer tails used in our study.

[1] Despite limiting our examination to only genes with at least 100 codons, which is the procedure used in Ma et al., we ended up with slightly different data set sizes. Since the RefSeq versions for the genome files are the same, the source of these discrepancies is unknown.

	12-mer rRNA Tails		Full Length rRNA Tails
Organism	SD% with INN [60]	SD% with INN-HB	SD% with INN-HB
A. aeolicus	48.1% of 1,487	58.6% of 1,489[1]	59.2% of 1,489[1]
B. subtilis	89.4% of 3,624	94.3% of 3,629[1]	95.9% of 3,629[1]
E. coli	57.1% of 3,908	66.9% of 3,882[1]	68.1% of 3,882[1]
H. influenzae	53.7% of 1,533	65.5% of 1,527[1]	65.9% of 1,527[1]
Synechocystis	26.0% of 2,906	37.7% of 2,912[1]	39.3% of 2,912[1]
T. maritima	90.1% of 1,685	91.6% of 1,696[1]	92.7% of 1,696[1]
B. japonicum	na	59.0% of 7655	60.7% of 7655
B. longum	na	73.5% of 1644	76.9% of 1644
C. tetani	na	71.6% of 2373	74.4% of 2373
L. johnsonii	na	85.2% of 1672	90.8% of 1672
Nostoc	na	39.4% of 4660	40.5% of 4660
S. aureus	na	93.1% of 2387	95.5% of 2378
S. meliloti	na	76.9% of 3062	78.1% of 3062
S. thermophilum	na	83.9% of 3033	85.1% of 3033
T. tengcongensis	na	91.0% of 2264	91.6% of 2264
T. thermophilus	na	79.1% of 1835	82.4% of 1835
X. axonopodis	na	51.6% of 4022	53.5% of 4022
Y. pestis	na	60.5% of 3564	61.9% of 3564

identify eight as +1 genes, and two as having stronger than average SD sequences between five and ten bases upstream from the start codons. Of the eight +1 genes, two had in-frame start codons within 12 bases downstream from the annotated start codon. The remaining 28 genes were able to bind to the rRNA tail further downstream from the annotated start codon. These results show the benefit of our approach by providing more resolution of the TIR in genes that have unusual nucleotide sequences relative to previous methods.

Our method is also useful for detecting errors in sequence annotation. Table 2.5 shows that most of the strong +1 genes are probably mis-annotated. Only a few strong +1 genes remain that do not fit this explanation. Of the five that remain in *E. coli*, none are experimentally verified, and they have no assigned function, making it likely that they are not true genes, but only vestigial open reading frames (ORFs).

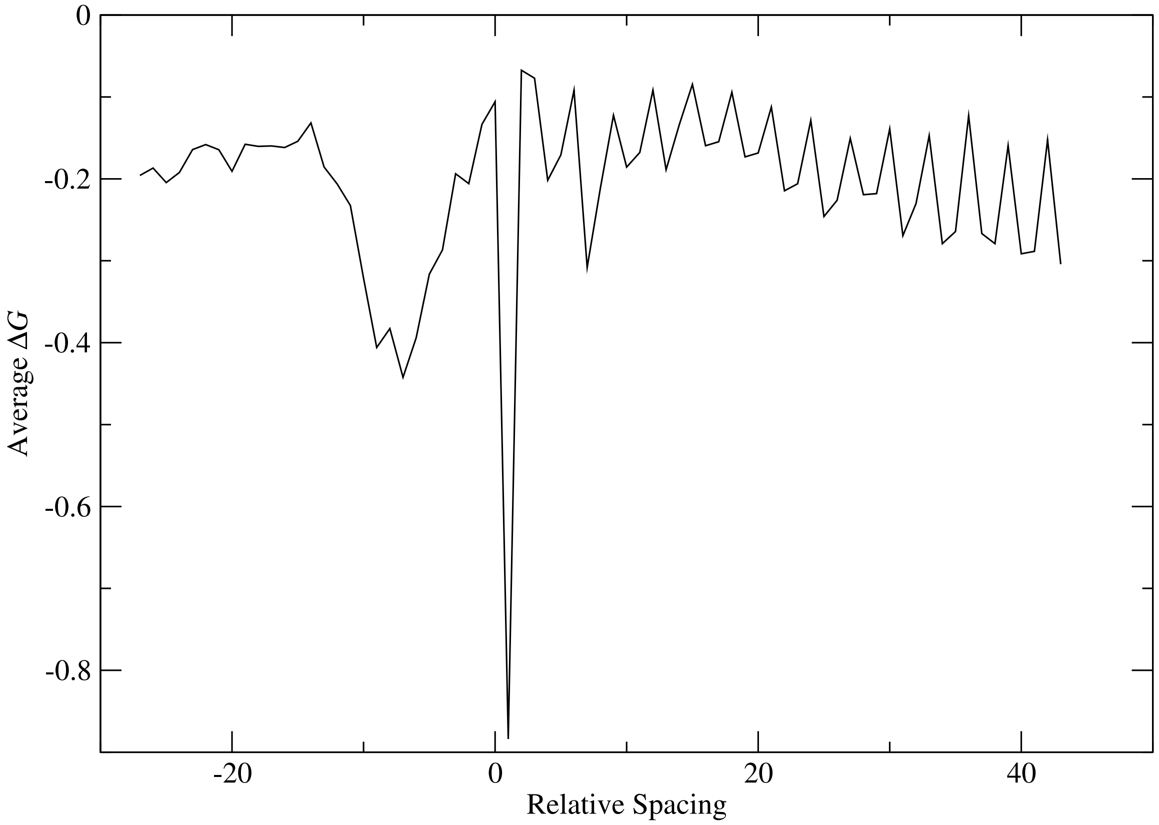

Figure 2.5: **Average ΔG° values in the TIR for *Synechocystis*.** The trough prior to RS 0 clearly shows the presence of an SD motif in many genes.

That said, it is harder to understand the strong +1 genes that do not appear to be the result of annotation errors in the 17 other organisms we studied. For example, *B. longum*'s strong +1 gene *rnpA* (GeneID:1023245), a ribonuclease P protein component, does not contain an in-frame start codon downstream from the annotated start site. *CTC02285* (GeneID:1060453), a strong +1 gene in *C. tetani* that codes for protein translation initiation factor 3 (IF3), is also without a downstream initiation codon. *B. japonicum* has many strong +1 genes without downstream start codons: *polE* (GeneID:1051409), which codes for the polymerase epsilon subunit, *cycK* (GeneID:1053038), *nah* (GeneID:1053188), and 52 others. Thus, while a large percentage of the strong +1 genes appear to be the result of sequence annotation errors, there remains a significant number that require an alternative explanation.

Two possible explanations for strong +1 genes that do not seem to be artifacts of annotation errors are: 1) the +1 site could stimulate translation initiation on leaderless genes, and 2) the binding site at RS+1 could be used as a translational stand-by site, i.e. sequences that hold the 16S rRNA close to the SD sequence [14]. In the former case, it is highly unlikely that the unexplained strong +1 genes in our study are leaderless because leaderless translation favors AUG start codons [20], in contrast to the strong +1 genes that favor GUG (see Table 2.1). In the latter case, it is unlikely that the +1 site functions as a translational standby site, because its location is too close to where the SD sequence should be; and for strong +1 genes, there does not appear to be an SD sequence.

Both ours and previous studies have also shown that many bacterial genes lack SD sequences upstream from proposed start codons (see Tables 2.5 2.6), suggesting the possibility of alternative mechanisms for recruiting ribosomes. Using Ma et al.'s criteria, only 68.1% genes in *E. coli* with more than 100 amino acids contained upstream SD sequences. The two cyanobacteria in our study, *Nostoc* and *Synechocystis*, both have relatively small percentages of upstream SD sequences. These two organisms are believed to be closely related to the free living predecessor of chloroplasts, which are thought to use SD sequences as well as alternative mechanisms to recruit ribosomes for translation (see Zerges [107] for a review). Furthermore, there is at least one example of a gene in *E. coli* that is efficiently translated without a canonical SD sequence [5], implying that these alternative mechanisms may exist in a variety of bacteria. One possible mechanism could be stem-loop structures within the TIR that form an SD-like sequence between loops. Boni et al. have shown that a disjoint SD sequence brought together by secondary structures is likely to function for the *E. coli* gene *rpsA* (GeneID:945536) [5]. It is also possible that there are viable substitutes for SD sequences. By generating a library of upstream sequences without canonical SD sequences and a low percentage of guanine bases, Kolev et al. were able to identify sequences in *E. coli* that would not bind to the 16S rRNA tail, but increased the efficiency of translation initiation beyond that of a consensus SD sequence [48].

We emphasize that our method is not for detecting start codons *de novo*, but for improving annotation accuracy once a candidate start codon is proposed by some other means. Our data suggests that we can identify unlikely start sites by examining the surrounding nucleotides, both upstream and downstream, and by using RS to characterize SD sequences. If the strongest binding between the TIR and the rRNA tail includes the candidate start codon, the true start codon may be in-frame and within 12 nucleotides downstream.

2.4.1 Conclusions

We have built on existing methods for characterizing SD sequences by developing software that utilizes the most recent nucleotide hybridization model, INN-HB, examining sequences that are both upstream and downstream from the start codon, and using Relative Spacing to indicate position. Our method has allowed us to identify both a larger percentage of SD sequences than previous methods and many potential annotation errors. Our method could be used to enhance genome annotation quality by accurately locating SD sequences with respect to proposed start codons. SD sequences that contain these start codons could indicate that a more likely start position is within 12 nucleotides downstream.

2.5 Materials and Methods

2.5.1 Genome Sequences

All genome sequences were downloaded from the National Center for Biotechnology Information (NCBI) GenBank database (http://www.ncbi.nlm.nih.gov/). Table 2.7 contains the names of the species whose sequences were analyzed, their RefSeq version numbers, the number of genes selected from each genome, their predicted 16S rRNA secondary structure, and the sequence of the rRNA tail used for the analysis.

Table 2.7: **A summary of the data, and its sources, used in this study.** All GenBank files were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/). All 16S rRNA secondary structures were downloaded from The Comparative RNA Web Site
(http://www.rna.icmb.utexas.edu/). The capitalized `A` in the 16S rRNA $3'$ tails is the nucleotide used to calculate Relative Spacing (RS).

[1] The structure was not used to define the $3'$ tail due to either the presence of wild card, `N`, characters or the lack of sequence altogether.

Organism	RefSeq Version	Genes	Secondary Structure	16S rRNA $3'$ Tail ($5'$ to $3'$)
A. aeolicus	NC_000918.1 GI:15282445	1529	d.16.b.A.aeolicus	`gaucAccucuuua`
B. japonicum	NC_004463.1 GI:27375111	8317	d.16.b.B.japonicum	`gaucAccuccuuu`
B. longum	NC_004307.2 GI:58036264	1727	NA	`gaucAccuccuuucu`
B. subtilis	NC_000964.2 GI:50812173	4106	d.16.b.B.subtilis	`gaucAccuccuuucu`
C. tetani	NC_004557.1 GI:28209834	2373	d.16.b.C.tetani[1]	`gaucAccuccuuucu`
E. coli	NC_000913.2 GI:49175990	4254	d.16.b.E.coli.K12	`gaucAccuccuua`
H. influenzae	NC_000907.1 GI:16271976	1656	d.16.b.H.influenzae[1]	`gaucAccuccuua`
L. johnsonii	NC_005362.1 GI:42518084	1821	NA	`gaucAccuccuuucu`
Nostoc	NC_003272.1 GI:17227497	5366	NA	`gaucAccuccuuu`
S. aureus	NC_002952.2 GI:49482253	2739	d.16.b.S.aureus	`gaucAccuccuuucu`
S. meliloti	NC_003047.1 GI:15963753	3332	NA	`gaucAccuccuu`
S. thermophilum	NC_006177.1 GI:51891138	3337	NA	`gaucAccuccuuucuaag`
Synechocystis	NC_000911.1 GI:16329170	3167	d.16.b.Synechocystis	`gaucAccuccuuu`
T. maritima	NC_000853.1 GI:15642775	1858	d.16.b.T.maritima	`gaucAccuccuuuc`
T. tengcongensis	NC_003869.1 GI:20806542	2588	NA	`gaucAccuccuu`
T. thermophilus	NC_005835.1 GI:46198308	1982	d.16.b.T.thermophilus.2[1]	`gaucAccuccuuucu`
X. axonopodis	NC_003919.1 GI:21240774	4312	NA	`gaucAccuccuuu`
Y. pestis	NC_004088.1 GI:22123922	4086	d.16.b.Y.pestis[1]	`gaucAccuccuua`

2.5.2 Selecting Criteria for Gene Sequences

For all genomes all gene sequences with `gene=` or `locus_tag=` tags were included in our data set, except those that also included a `transposon=` or `pseudo` tag.

We defined the translation initiation region (TIR) as 35 bases upstream and 35 bases downstream of the first base in the start codon. To this sequence, we added a number of additional nucleotides equivalent to the number of nucleotides in the species rRNA tail to the downstream sequence. For example, TIR sequences in a species whose rRNA tail length was 13 nucleotides would be 83 bases long (35 nucleotides upstream + 35 nucleotides downstream + 13 more downstream). Several observations determined this sequence window. In the majority of cases examined, SD sequences were within 10 nucleotides of the start codon. Although the hypothesis that a downstream box interacted with rRNA during translation initiation [86] was rejected [68], evidence from leaderless mRNAs suggests that sequences downstream and within 20 nucleotides of the start codon are involved [86, 21, 102]. Other studies that have analyzed initiation regions of mRNA sequences for negative free energy troughs [82, 70, 58, 60] have not examined bases downstream of the annotated start codon: downstream sequence analysis allowed for start codon annotation error detection.

2.5.3 Determining the $3'$ rRNA Tails for the 16S rRNAs

To determine the $3'$ tails for the 16S rRNAs, we downloaded predicted secondary structures from The Comparative RNA Web Site [8] (http://www.rna.icmb.utexas.edu/). We defined the $3'$ tail as the single-stranded terminal $3'$ nucleotides, and then, to verify consistency, compared these sequences with all annotated copies of the 16S rRNA in the genome.

If no secondary structure was available for an organism, we attempted to define the $3'$ tail from the genome sequence alone. First, we let the $3'$ end of the sequence define the $3'$ end of the tail. We then looked in the $5'$ direction for the first instance of the three letter motif, $5'{-}\mathrm{GAT}{-}3'$, because this motif was found consistently on the $5'$ end of the tails of 16S rRNAs with predicted structures. The location of this motif was then used to define the $5'$ end of the $3'$ tail.

When there was a conflict between the genome sequence and the secondary structure or between multiple sequences within a single genome, we chose the tail found in the secondary structure or, if there was no predicted secondary structure, the majority of the 16S rRNA genes.

Tails for all 18 organisms used in our study are listed in Table 2.7.

2.5.4 Quantifying the Helix Formation between the $3'$ 16S rRNA Tail and the mRNA Initiation Region with `free_scan`

For each gene in each organism, we predicted the change in the free energy, ΔG°, required to bring the two strands of nucleotides together and form a double helix structure using `free_scan`, a program we wrote. In the absence of catalytic enzymes, chemical reactions with ΔG° values greater than zero require additional energy from an external source and are unlikely to occur spontaneously. On the other hand, reactions with ΔG° values less than zero are likely to take place. This method has been used in many studies of SD sequences [94, 82, 58, 56, 70, 60, 49], as well as in the genome annotation program `GLIMMER` [16].

To calculate ΔG° at each position, `free_scan` begins by pairing the $5'$ end of the TIR with the $3'$ end of the rRNA tail and then pairs the mRNA and the rRNA in the $3'$ direction of the TIR and the $5'$ direction of the rRNA tail. `free_scan` calculates ΔG° using the Individual Nearest Neighbor - Hydrogen Bond (INN-HB) model [104], extended to allow for symmetrical internal loops (loops that contain an equal number of bases in both RNA strands):

$$\Delta G^\circ = \Delta G^\circ_{\text{init}} + \sum_j n_j \Delta G^\circ(\text{NN}) + m_{\text{term}-\text{AU}} \Delta G^\circ_{\text{term}-\text{AU}} + \Delta G^\circ_{\text{sym}} + \sum_k \text{Loop}_k.$$

In this formula, $\Delta G^\circ_{\text{init}}$ is the amount of free energy required to initiate a helix between the two strands of RNA; $\Delta G^\circ(\text{NN})$ is the free energy released by the hybridization of a particular nearest neighbor doublet, and n_j is its number of occurrences in the duplex. $m_{\text{term}-\text{AU}}$ is the number of terminal AU pairs, and $\Delta G^\circ_{\text{term}-\text{AU}}$ is the free energy penalty for having a terminal AU pair. Finally, $\Delta G^\circ_{\text{sym}}$ is the penalty for internal symmetry and Loop_k the penalty for the kth internal loop. `free_scan`'s hybridization parameter values for Watson-Crick binding are from Xia et al. [104], G/U mismatches from Mathews et al. [62], and loop penalties from Jaeger et al. [42]. `free_scan` uses a dynamic programming algorithm to determine the optimal number, location, and length of internal loops that minimize ΔG°. Bulges, where one of the two strands of RNA has intervening nucleotides between bases that bond with the other strand, as well as secondary structures involving only one of the two strands of RNA, are ignored due to uncertainty about how much space is available within the 30S ribosomal complex to accommodate these structures, as well

```
Alignment 1.    Binding Value = 0.0
rRNA: a u u c c u c c a C U a g
                        | |
mRNA: u a c c a g c a g G A g g u g...

Alignment 2.    Binding Value = 0.0
rRNA:   a u u c c u c c a c u a g

mRNA: u a c c a g c a g g a g g u g...
⋮

Alignment 6.    Binding Value = -16.5
rRNA:     a u U C C U C C A C U A g
              | | | | | | | | | |
mRNA:  ...g c A G G A G G U G A U g...

⋮

Alignment 71.  Binding Value = 0.0
rRNA:       a u u c c u c c a c u a g

mRNA: ..g c g c a a g u u u c a c u a
```

Figure 2.6: **An overview of how ΔG° values are calculated in each TIR.** For each base in each initiation region, we simulated the change in free energy required for the $3'$ 16S rRNA tail to hybridize with the mRNA. A minimum of two consecutive bases need to pair and, in order for the binding to occur spontaneously, require a change more negative than -4.08 kcal/mol [104], the value for $\Delta G^\circ_{\text{init}}$, In this example, the initiation region from *E. coli*'s gene *hcaF* (GeneID:946997), alignment 1 is set to zero because the change in free energy required to bring together a single complementary double is not favorable. Alignment 2 and 71 are set to zero because there are no complementary doublets. Alignment 6 is set to -16.5 because it requires -16.5 kcal/mol less than -4.08 kcal/mol to hybridize.

as the limitations they put on our ability to calculate RS. Dangling $5'$ or $3'$ ends are not considered because of ambiguities about what constitutes a dangling end on the mRNA sequences and on the $5'$ end of the 16S rRNA tail.

After the free energy value for the first alignment in the mRNA is calculated, `free_scan` shifts the rRNA tail downstream one base, and the second alignment is examined. This process, illustrated in Figure 2.6, was carried out for 71 alignments in the mRNA. We selected the initiation regions from each gene to allow for 35 ΔG° values to be computed before the start codon, one at the start codon, and 35 ΔG° values after.

Xia et al. created the INN-HB model [104] to improve the ΔG° estimates obtained using the prior Nearest Neighbor and Individual Nearest Neighbor (INN) models [28, 26, 27, 6, 24]. This improvement is obtained by adding a term to correctly count the number of hydrogen bonds that form in the terminal doublets in helices. The INN, in contrast, overestimates the stability of helices with terminal AU base pairs and underestimates the stability of helices with terminal GC base pairs [104].

To verify the accuracy of `free_scan`, we ran our analysis again using `RNAhybrid` [74] and plotted the average ΔG° value for each RS position (Figure 2.1). `RNAhybrid` uses free-energy parameters from Xia et al. [104] and Mathews et al. [62], but does not include $\Delta G^{\circ}_{\text{init}}$ or $m_{\text{term}-\text{AU}}\Delta G^{\circ}_{\text{term}-\text{AU}}$. We set the energy cut-off to -4.075225 kcal/mol and subtracted this value from `RNAhybrid`'s output in order to compensate for its lack of initiation penalty. We also turned off bulges and loops because these structures, when asymmetrical, are the alignment equivalent of inserting gaps, making it impossible to calculate RS. By forcing RNAhybrid to exclude internal loops we prevented it from correctly identifying many SD sequences that contain symmetrical loops. This factor, combined with the lack of penalties for terminal A/U pairs, explains the bulk of the differences between the output of `RNAhybrid` and `free_scan`. Figure 2.1 demonstrates that both programs show distinct binding at RS +1 in all 18 genomes. Thus, the RS +1 site is not an artifact of our particular INN-HB implementation.

We did not compare our results to `RNAcofold` because it uses a linker sequence to join the two sequences into a single strand of RNA prior to folding, and allows for intramolecular folding. These two conditions could cause potential binding sites to be overlooked. If the mRNA sequence being examined for binding sites formed a stem-loop structure with an SD sequence in the loop, then it would not be detected because of computational limitations in identifying pseudo-knot secondary structures.

To determine the effect of using the INN-HB model on the detection of SD regions, we did the following computational experiment. By limiting the TIR to the 20 bases proceeding the initiation codon and excluding all genes with fewer than 100 codons, we compared the number of SD sequences the INN-HB model was able to identify with previously published results that use the INN model [60]. The threshold ΔG° that Ma et al. use to define an SD sequence was -4.4 kcal/mol, which is the value predicted by the INN for the hybridization between three core SD sequences and the 16S rRNA tail:

$$\begin{array}{c} \texttt{mRNA:}\ 5'-\texttt{GGAG}-3' \\ \texttt{rRNA:}\ 3'-\texttt{CCUC}-5' \end{array}, \quad \begin{array}{c} \texttt{mRNA:}\ 5'-\texttt{GAGG}-3' \\ \texttt{rRNA:}\ 3'-\texttt{CUCC}-5' \end{array} \quad \text{and} \quad \begin{array}{c} \texttt{mRNA:}\ 5'-\texttt{AGGA}-3' \\ \texttt{rRNA:}\ 3'-\texttt{UCCU}-5' \end{array},$$

The INN-HB, however, does not assign all three hybridizations the same ΔG° value because the first two have 11 hydrogen bonds each, while the third only has 10 hydrogen bonds. The INN does not take this difference into account because all three hybridizations consist of 1 GG/CC doublet and 2 AG/UC doublets. With the updated parameters for both the doublets as well as helix initiation, combined with a penalty for terminal A/U pairings, the INN-HB predicts the ΔG° value -3.61 kcal/mol for the first two helices and -3.14 kcal/mol for the third helix. Thus, we defined our SD threshold to be the average ΔG° for all three helices: -3.4535. It is worth noting that the bulk of the difference between the thresholds calculated by the INN and the INN-HB is a result of their distinct helix initiation penalties ($\Delta G^{\circ}_{\text{init}} = 3.4$ kcal/mol for the INN and $\Delta G^{\circ}_{\text{init}} = 4.08$ kcal/mol for the INN-HB). Table 2.6 summarizes the comparison between the two models. Since we used an equivalent threshold to define sufficient binding for an SD sequence, we can conclude that INN-HB model is responsible for the increase in the number of SD sequences identified.

Our programs, `free_scan` and `free_align` are available at Source Forge: http://sourceforge.net/projects/free2bind.

2.5.5 Locating the SD Sequence and Determining SD Relative Spacing

We located the SD sequence by the position of the lowest ΔG° value calculated within the initiation region. If $\Delta G^{\circ} > -3.4535$ kcal/mol, then the gene was assumed not to have an SD sequence. This threshold is based on the work of Ma et al. [60] (see above).

The SD's Relative Spacing (RS) is the position of the $5'$ A in the rRNA sequence $5'$–ACCUCC–$3'$, relative to the first base in the start codon. This $5'$ A is the same base Chen et al. used to determine aligned spacing [11], which is another metric used to compare the locations of SD sequences. If the SD is upstream from the start codon, its RS is negative, while if it is downstream, its RS is positive. If the two are opposite one another, its RS is zero. See Figure 2.3 for RS examples taken from *E. coli.*

2.5.6 Defining Strong Binding

We defined strong binding as any binding between the mRNA and the 3' 16S rRNA tail that has $\Delta G^\circ \leq -8.4$ kcal/mol. This value is the ΔG° obtained from the optimal base-pairing between the rRNA and the original Shine-Dalgarno sequence, $5'{-}\mathrm{GGAGGU}{-}3'$.

Chapter 3

rRNA sequences conserved in prokaryotes and eukaryotes share an mRNA hybridization pattern.

3.1 Abstract

It has long been established that base pairing between the 3′ tail on the 16S ribosomal RNA (rRNA) and messenger RNA (mRNA) sequences facilitates both translation initiation. The core of the rRNA sequence that hybridizes with mRNA, 5′-CCUCC-3′, is conserved throughout the bacterial kingdom, as are the complementary bases on the mRNA sequences. The sequences flanking the rRNA core, however, are also well conserved, even though they are not thought to be as important in base pairing with mRNAs. We examined potential hybridization sites between rRNA and mRNA sequences in the translation initiation regions in 18 bacteria and found that the sequences flanking the rRNA core can bind to most mRNAs at their initiation codons. The core rRNA bases, however, almost never facilitate base pairing at the start codon. Since the flanking bases are well conserved in the 3′ tail of 18S rRNA found in eukaryotes, we performed the same analysis on 6 eukaryote genomes and observed the same phenomenon. We propose that the bases flanking the rRNA core sequence in the 16S rRNA may have a role in hybridizing with mRNA sequences that is independent of the core.

3.2 Introduction

Initiation of translation is a complex process that requires both the assembly and positioning of multiple components at the start codon. A major event in the assembly of the initiation complex is the binding of the small ribosomal subunit to the mRNA. In bacteria, it has long been established that the $3'$-terminal nucleotides of the 16S ribosomal RNA (16S tail) can form hydrogen bonds with messenger RNA (mRNA) sequences during translation initiation [83, 38, 41]. Shine and Dalgarno [83] hypothesized that this hybridization facilitates small ribosomal subunit binding to mRNA and positioning of the ribosomal complex appropriately for initiation at the start codon. Sequences $5'$ (upstream) of the start codon that are complementary to the 16S tail are named Shine-Dalgarno (SD) sequences because they contain all or a substantial part of the consensus sequence, $5'$-GGAGG-$3'$ [38, 41, 106]. The complementary sequence in the 16S tail is referred to as the anti-Shine-Dalgarno (aSD) sequence, whose core nucleotides, $5'$-CCUCC-$3'$, are highly conserved in prokaryotes.

Several experimental studies [38, 41] confirmed that Shine-Dalgarno sequence hybridization facilitates ribosome binding to mRNA and translation initiation. However, further investigations have called the necessity of the SD for translation initiation into question (see references as reviewed by Boni [4] and Nakamoto [66]). For example, there have been many studies that report translation of mRNAs without any $5'$ leader sequences, and therefore, no SD sequence [103, 20, 69]. In our recent study of 18 bacterial species [88], we found that an SD sequence occurred in only 68% of genes in E. coli, and, overall, in only 56% of the mRNAs with at least 100 codons.

In eukaryotes, an aSD sequence is completely absent in the principle rRNA of the small ribosomal subunit, 18S rRNA [30]. As might be expected, SD sequences are also absent from the upstream region of eukaryotic mRNAs. These early finding necessitated a new model for translation initiation in eukaryotes. As reviewed by Kozak [53], assembly of the eukaryotic small ribosomal subunit and the mRNA is accomplished by ribosomal binding to the mRNA 7-methylguanosine cap, a structure found on the $5'$-terminus of the majority of eukaryotic mRNAs. Following mRNA binding, the model proposes that the ribosome moves in the $3'$-direction along the mRNA scanning for the authentic AUG start codon. Kozak has proposed that the authentic start codon is identified by its nucleotide context, based on statistical analysis of nucleotide frequency and mutagenesis studies [51]. Ribosomal complex recognition of the correct start-codon context presumably positions the ribosome to begin translation.

Shultzaberger et al. [84] and Osada et al. [71] have extended the idea of start-codon context by using information theory approaches. Oasda et al. used measures of relative entropy to estimate the conservation of nucleotides and mutual information to determine the correlation between two bases. They surveyed $5'$-untranslated upstream regions ($5'$-UTRs) of at least 500 genes in each of 9 eukaryotic, 108 bacterial and 16 archaeal species. Their findings clearly identify the SD sequence nucleotides in bacterial $5'$-UTR and are in agreement with the Kozak consensus sequence in that high relative entropy was associated with the nucleotide position -3, a critical position in Kozak's consensus sequence. In addition, they found elevated values of relative entropy for nucleotides at position -1 or -2, and unexpectedly found high mutual correlation between nucleotides -1 and -2 in most species.

Another approach to identify specific nucleotides that could play a role in translation initiation is to examine the patterns of free energy resulting from binding the 16S tail to mRNA sequences [93, 82, 70, 58, 88]. Such an approach was used by Osada and co-workers [70] to examine the hybridization potential of the 16S tail sequences with $5'$-UTRs of genes of eubacterial and archaebacterial species and of the 18S rRNA tail sequence of yeast with the $5'$-UTRs of yeast genes. Their results indicated that free energy binding patterns of the archaebacteria were more similar to eubacteria than to yeast. This finding was interpreted as indicating different mechanisms for translation initiation between eukaryotes and the eubacteria and archaebacteria.

In our previous work [88] we examined the hybridization potential of the 16S tail and the translation initiation regions (TIRs) of mRNAs from 18 bacterial species. In contrast with the work of Osada and co-workers [70], who examined 50-nucleotide sequences upstream from the start codon, the TIRs we examined ranged from 35 nucleotides upstream from the start codon to 35 nucleotides downstream from the start codon. This was done to use the free energy approach to examine the sequence context of the start codon and its potential for 16S rRNA interaction. As expected, we were able to identify the region where most SD sequences are located, between 4 and 10 bases upstream of the start codons, as a distinct trough of negative free energy. An unexpected result, however, was that in 9% of the genes across all species, the start codon itself was part of a sequence that could hybridize with the 16S tail at least as well as a minimal SD sequence ($\Delta G^\circ \leq -3.4535$ kcal/mol [60, 88]). We labeled the location of this second free energy trough as RS+1 in reference to the location of the rRNA tail relative to the first base of the start codon [88] (See Material and Methods section for details).

The alignment of the RS+1 free energy trough has the 16S tail straddling the start codon of the mRNA. This suggested that the newly discovered binding site may arise from the sequence context of the start codon, and potentially play a role in translation initiation. In the investigation presented here, we show that binding at RS+1 is unrelated to binding at SD sequences, and that binding at RS+1, and the sequences required to do so, are conserved in both prokaryotes and eukaryotes.

3.3 Results

Since we first identified binding at RS+1 in eubacteria [88], GenBank (http://www.ncbi.nih.gov/) has released updated annotation for the genomes in our analysis. With the new annotation, we have re-analyzed the genomes, and while many RS+1 genes have been removed, or are now described as pseudo genes, the overall statistics are the same as the original results: 9% of the genes having SD-like binding at RS+1 (see Table 3.1). When we reduced our threshold to allow all sequences that could base pair with the 16S rRNA spontaneously ($\Delta G^\circ < 0$ kcal/mol), we found that 58% of the genes examined could hybridize at RS+1 (See fourth column in Table 3.1).

Our first question about the RS+1 binding site was whether or not it was related, in a statistical or predictive sense, to SD binding. For mRNAs that had both an SD sequence and a sequence at RS+1 that could bind to the 16S rRNA, we calculated the correlation coefficient between the magnitude of the SD binding and the RS+1 binding. Only four of the prokaryotes had correlation values that were determined to be non-zero. However, of these four (*B. japonicum, Nostoc, S. thermophilum* and *Synechocystis*), the largest correlation coefficient was only -0.07, indicating that the correlation accounted for less than one percent of the variation in the data. Thus, all correlation values were interpreted as non-significant. We also tested for an association between the presence of binding to an SD sequence and the presence of binding at RS+1. That is to say, we wanted to know if knowledge of the presence of binding at one site would provide any information about the presence of binding at the other. The analysis showed that all but three of the bacteria were not associated. Of the three that showed a statistical association (*B. longum, C. tetani* and *T. thermophilus*), the association was not very strong. For example, *T. thermophilus* had the strongest association

Table 3.1: **RS+1 hybridization statistics for 18 bacterial genomes.** The total number of protein coding genes is given in the second column. The third and fourth columns show what percentage of these genes have binding at RS+1 that meets or exceeds the given thresholds. $\Delta G^\circ \leq -3.4535$ is the cutoff for a minimal SD sequence [60, 88], and $\Delta G^\circ < 0$ is sufficient for the spontaneous formation of a helical structure. The RS+1 dataset, shown in the fifth column, is the set of set of all genes with binding at RS+1, excluding those that are predicted to be mis-annotated, and those that are predicted to use the binding at RS+1 as a secondary SD sequence, as described in the Materials and Methods section.

Organism	Number of genes	$\Delta G^\circ \leq -3.4535$	RS+1 dataset ($\Delta G^\circ < 0$)
A. aeolicus	1529	6% (91)	56% (851)
B. japonicum	8317	13% (1050)	56% (4624)
B. longum	1727	7% (115)	63% (1081)
B. subtilis	4105	6% (252)	67% (2757)
C. tetani	2373	10% (227)	62% (1460)
E. coli	4243	6% (168)	55% (2338)
H. influenzae	1656	3% (52)	52% (868)
L. johnsonii	1821	3% (63)	67% (1213)
Nostoc	5366	7% (379)	57% (3050)
S. aureus	2651	3% (86)	66% (1753)
S. meliloti	3332	6% (212)	52% (1745)
S. thermophilum	3337	23% (751)	62% (2075)
Synechocystis	3167	7% (226)	52% (1648)
T. maritima	1858	13% (244)	69% (1279)
T. tengcongensis	2588	6% (162)	58% (1494)
T. thermophilus	1982	16% (327)	65% (1291)
X. axonopodis	4312	8% (343)	52% (2251)
Y. pestis	4086	8% (334)	53% (2155)
Total	58450	9% (5082)	58% (33933)

with 85% of genes with RS+1 binding and 74% of genes without RS+1 binding having SD sequences.

To gain more insight into the source of binding at RS+1, we identified which mRNA bases at RS+1 were binding to the 16S tail. Our approach was to compute a consensus mRNA sequence for the RS+1 dataset for each of the 18 bacterial species in our study and align it with its respective 16S rRNA tail at RS+1. We then labeled the bases that were capable of hybridizing as shown in Figure 3.1. The results from all 18 species show that the bases in the 16S tails that are most likely to hybridize with bases in the consensus sequence, and thus contribute to the overall $-\Delta G^{\circ}$, are those that flank the core aSD sequence, 5′-CCUCC-3′.

With the nucleotides on the 16S rRNAs involved in binding at RS+1 identified, we wanted to determine if they were well conserved (see Table 3.2). The bases 5′ of the core aSD sequences, 5′-GAUCA-3′, are conserved with 100% fidelity among the 18 species. The bases on the 3′-side of the core are also relatively well conserved. The first two 3′ bases are UU in all 18 species. Considering the sixteen species whose 16S rRNA annotations extend one nucleotide beyond the two UU bases, in 13 species the next base is a U and in three species, the next base is an A. In the remaining two species, the 16S rRNA annotation extends more than three bases in the 3′-direction beyond the core. In these two species, a CU doublet is found. These results show significant conservation of flanking nucleotides across 18 bacterial species despite their limited role in binding to SD and SD-like sequences.

Relationship between bacterial 16S rRNA and eukaryotic 18S rRNA 3′-terminal sequences.

Similarities between the bases flanking the aSD core in the 16S rRNA and the 3′ tail in the eukaryotic 18S rRNA was first noted by Hagenbüchle et al. in 1978 [30]. In Table 3.3 the 3′-terminal sequences of 18S rRNAs from 6, widely diverse, eukaryotic species are given. Overall, when compared to the 16S rRNA tail sequences from bacterial species, the nucleotide sequences in eukaryotic 18S rRNA tails appears to conserve the flanking nucleotides in the 16S rRNA sequence and delete the aSD core nucleotides. This nucleotide conservation is almost 100%. The exceptions are the plant species, *Arabidopsis thaliana*, whose sequence differs by its 3′-terminal nucleotide base, and the roundworm, *C. elegans*, whose sequence differs by the last two 3′-terminal bases. The nucleotide sequences conserved

A. aeolicus

```
mRNA:  gAAAaaAuaUGAaa
        |||  |  |||
rRNA:  aUUUccUccACUag
```

B. japonicum

```
mRNA:   GGGcGccaUGAcc
        ||| |   |||
rRNA:   UUUcCuccACUag
```

B. longum

```
mRNA: GGGGAaaccaUGAcc
      |||||     |||
rRNA: UCUUUccuccACUag
```

B. subtilis

```
mRNA: GGGAAaaAaaUGAaa
      |||||  |  |||
rRNA: UCUUUccUccACUag
```

C. tetani

```
mRNA: GGAAAaaAuaUGAaa
      |||||  |  |||
rRNA: UCUUUccUccACUag
```

E. coli

```
mRNA:   aAAaauuaUGAaa
         ||     |||
rRNA:   aUUccuccACUag
```

H. influenzae

```
     aAAaauuaUGAaa
      ||     |||
     aUUccuccACUag
```

L. johnsonii

```
   GGAAAaaAaaUGAaa
   |||||  |  |||
   UCUUUccUccACUag
```

Nostoc

```
     AAAaauuaUGAaU
     |||     ||| |
     UUUccuccACUaG
```

S. aureus

```
   GGAAAaaAaaUGAaa
   |||||  |  |||
   UCUUUccUccACUag
```

S. meliloti

```
      GAGGccaUGAcC
      ||||   ||| |
      UUCCuccACUaG
```

S. thermophilum

```
gggGGGGGGGccaUGGcC
   |||||||   ||| |
gaaUCUUUCCuccACUaG
```

Synechocystis

```
  AuuaacuaUGAcU
  |       ||| |
  UuuccuccACUaG
```

T. maritima

```
 GuGAaaAaaUGAaa
 | ||  |  |||
 CuUUccUccACUag
```

T. tengcongensis

```
   AAaaAaaUGAaa
   ||  |  |||
   UUccUccACUag
```

T. thermophilus

```
GGGGGGGGcaUGGgg
||||||||  |||
UCUUUCCUccACUag
```

X. axonopodis

```
  GGCCCCCaUGAcC
  ||      ||| |
  UUuccuccACUaG
```

Y. pestis

```
  aAAaauuaUGAaU
   ||     ||| |
  aUUccuccACUaG
```

Figure 3.1: **Identification of the 16S tail nucleotides that base pair at RS+1 in bacteria.** For each organism we calculated the consensus sequences using the mRNAs in the RS+1 dataset. We then aligned the 16S rRNA tails to the consensus sequence at RS+1. Lines between the mRNA and the rRNA indicate possible base pairs (including G/U mismatches).

Table 3.2: **A summary of the data, and its sources, used in this study.** All GenBank files were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/). All 16S and 18S rRNA secondary structures were downloaded from The Comparative RNA Web Site (http://www.rna.icmb.utexas.edu/). The capitalized A in the 16S and 18S rRNA tails is the nucleotide used to calculate Relative Spacing (RS). The dashes in the 18S tails indicate gaps in the alignment with *E. coli*'s 16S tail.

[1] The structure was not used to define the $3'$ tail due to either the presence of wild card characters, N, or the lack of sequence altogether.

Prokaryotes

Organism	GenBank Files	Last Modified	16S rRNA Tail ($5' \rightarrow 3'$)	Secondary Structure
A. aeolicus	NC_000918.1 GI:15282445	02-DEC-2005	gaucAccuccuuua	d.16.b.A.aeolicus
B. japonicum	NC_004463.1 GI:27375111	24-MAR-2006	gaucAccuccuuu	d.16.b.B.japonicum
B. longum	NC_004307.2 GI:58036264	17-JAN-2006	gaucAccuccuuucu	NA
B. subtilis	NC_000964.2 GI:50812173	02-DEC-2005	gaucAccuccuuucu	d.16.b.B.subtilis
C. tetani	NC_004557.1 GI:28209834	03-APR-2006	gaucAccuccuuucu	d.16.b.C.tetani[1]
E. coli	NC_000913.2 GI:49175990	04-MAY-2006	gaucAccuccuua	d.16.b.E.coli.K12
H. influenzae	NC_000907.1 GI:16271976	13-DEC-2005	gaucAccuccuua	d.16.b.H.influenzae[1]
L. johnsonii	NC_005362.1 GI:42518084	06-FEB-2006	gaucAccuccuuucu	NA
Nostoc	NC_003272.1 GI:17227497	09-JAN-2006	gaucAccuccuuu	NA
S. aureus	NC_002952.2 GI:49482253	03-APR-2006	gaucAccuccuuucu	d.16.b.S.aureus
S. meliloti	NC_003047.1 GI:15963753	03-DEC-2005	gaucAccuccuu	NA
S. thermophilum	NC_006177.1 GI:51891138	03-DEC-2005	gaucAccuccuuucuaag	NA
Synechocystis	NC_000911.1 GI:16329170	15-JUN-2006	gaucAccuccuuu	d.16.b.Synechocystis
T. maritima	NC_000853.1 GI:15642775	03-DEC-2005	gaucAccuccuuuc	d.16.b.T.maritima
T. tengcongensis	NC_003869.1 GI:20806542	03-DEC-2005	gaucAccuccuu	NA
T. thermophilus	NC_005835.1 GI:46198308	04-APR-2006	gaucAccuccuuucu	d.16.b.T.thermophilus.2[1]
X. axonopodis	NC_003919.1 GI:21240774	17-JAN-2006	gaucAccuccuuu	NA
Y. pestis	NC_004088.1 GI:22123922	26-JAN-2006	gaucAccuccuua	d.16.b.Y.pestis[1]

Eukaryotes

Organism	GenBank Files	Last Modified	18S rRNA Tail ($5' \rightarrow 3'$)	Secondary Structure
A. thaliana	NC_00307[0,1,4,5,6]	04-NOV-2005	gaucA-----uug	d.16.e.A.thaliana
C. elegans	NC_0032[79-84]	06-FEB-2006	gaucA-----ucg	NA
D. melanogaster	NT_03377[7-9], NT_037436, NC_00435[3,4]	30-JAN-2006	gaucA-----uua	d.16.e.D.melanogaster
H. sapiens	RefSeq build 36v1	03-MAR-2006	gaucA-----uua	d.16.e.H.sapiens
M. musculus	RefSeq build 36v1	28-APR-2006	gaucA-----uua	d.16.e.M.musculus
S. cerevisiae	NC_0011[33-48]	14-JUN-2006	gaucA-----uua	d.16.e.S.cerevisiae

Table 3.3: **RS+1 hybridization statistics for 6 eukaryote genomes.** The majority of genes in these organisms show a potential for hybridization between the mRNA and the rRNA at RS+1.

Organism	Chromosomes	Genes	$\Delta G^\circ < 0$ at RS+1
A. thaliana	5	26479	93% (24634)
C. elegans	6	20031	66% (13197)
D. melanogaster	6	13530	84% (11412)
H. sapiens	24	22383	85% (18938)
M. musculus	21	23901	85% (20211)
S. cerevisiae	16	5850	84% (4910)
Total	78	112174	83% (93302)

between prokaryotes and eukaryotes are those that participate in hybridization at RS+1. This suggested to us that any function for binding at RS+1 may be conserved across both prokaryotic and eukaryotic organisms.

We examined the TIR sequences in the six eukaryotic species to determine if there was a potential binding site at RS+1 in eukaryotic genes. It was not clear to us if this would be confirmed as the spacing of the two regions of flanking sequences of the rRNA tail would be changed between prokaryotes (in which there are 5 nucleotides between the two, $3'$- and $5'$-flanking sequences) and eukaryotes (in which there are no intervening nucleotides). The average ΔG° value at each position of the TIR for the six eukaryotic species in our study is shown in Figure 3.2. In agreement with published work (see [50] for an early review) the lack of a negative ΔG° trough upstream of the start codon verifies the absence of an upstream SD-like sequence in the genes of eukaryotes. The negative trough of ΔG° at RS+1 however, is consistent with the RS+1 binding site in bacteria that results from hybridization between the 18S rRNA tail and the start codon and surrounding nucleotides. This hybridization site exists in the majority of eukaryotic genes that we examined as seen in Table 3.3.

To identify which bases in the 18S rRNA tails were involved in RS+1 binding, we used the same approach utilized with bacterial species. We calculated the consensus sequences for the genes with RS+1 binding for each eukaryotic species. We then aligned this consensus sequence with the 18S rRNA tail sequence and noted which nucleotides in the 18S rRNA tail contributed to the ΔG° value. Figure 3.3 shows that the bases in the 18S rRNA tails that bind at RS+1 follow a nearly identical pattern when compared to those in the 16S rRNAs. Thus, the flanking nucleotides in the 16S rRNAs are conserved in the 18S rRNAs, and despite the spacing differences within the tail sequence, binding at RS+1

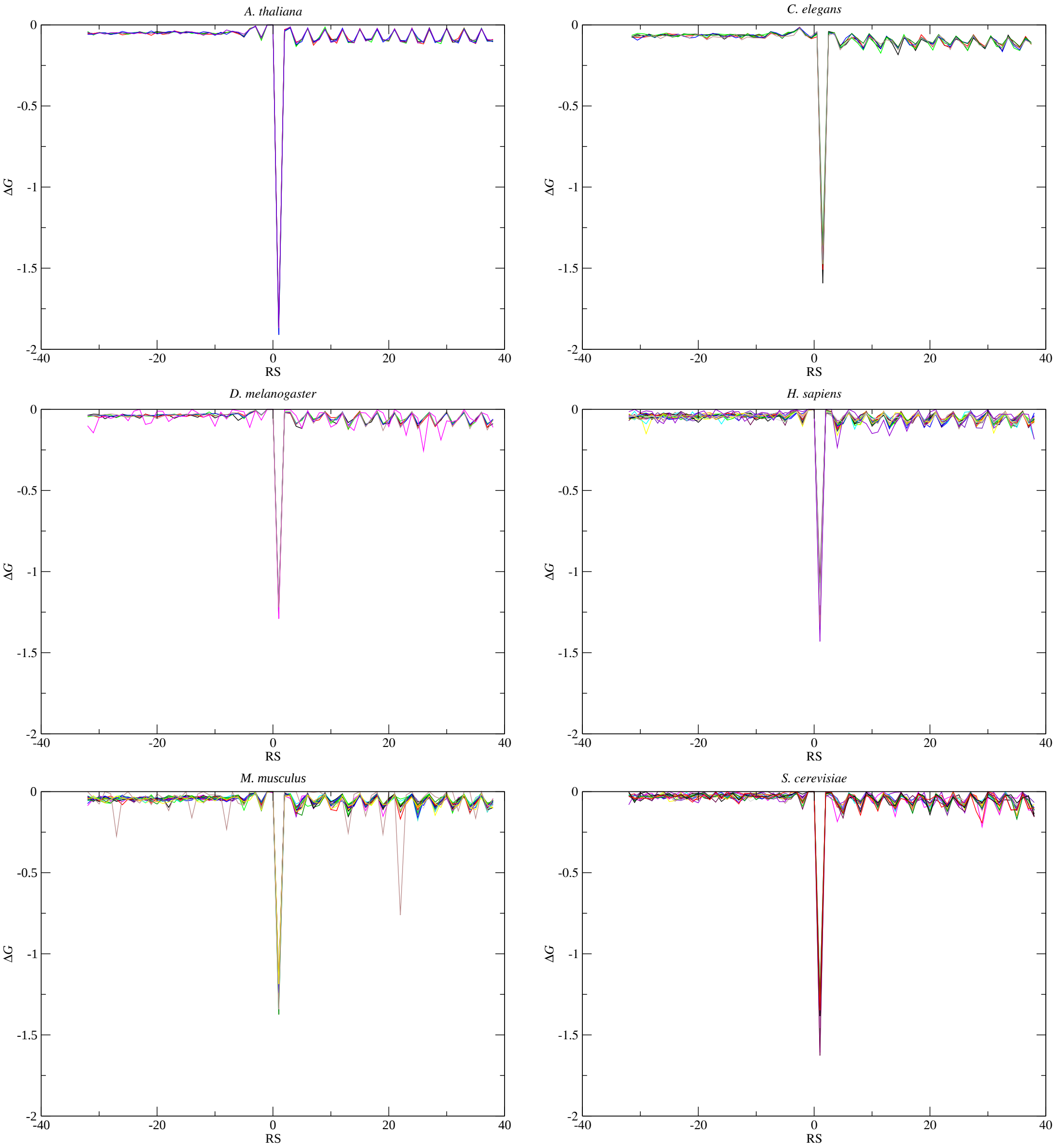

Figure 3.2: **Average ΔG° values in the TIRs for each chromosome in 5 eukaryotes.** For each organism, we calculated the average ΔG° value for each Relative Spacing (RS) position. Each line in each graph represents these averages for each chromosome. Zero on the x-axis corresponds to the 3′ A residue in the rRNA sequence 5′-GAUCA-3′ being positioned over the first base in the initiation codon. The drop in ΔG° immediately after the first base in the initiation codon, at RS+1, shows that there is a significant binding potential between the 16S rRNA and the mRNA close to the initiation codon. In the entire set, *M. musculus*'s X chromosome is the only one showing significant binding at locations other than RS+1.

```
A. thaliana      D. melanogaster M. musculus
mRNA: aAAUGGcg   aAAUGGcU       cGAUGGcg
       |||||      ||||| |        |||||
rRNA: gUUACUag   aUUACUaG       aUUACUaG

C. elegans       H. sapiens     S. cerevisiae
mRNA: aaAUGAaU   aGAUGGcg       aAAUGAcU
        |||| |    |||||          ||||| |
rRNA: gcUACUaG   aUUACUag       aUUACUaG
```

Figure 3.3: **Identification of the 18S tail nucleotides that base pair at RS+1 in eukaryotes.** For each eukaryote we created consensus sequences from the genes where $\Delta G^\circ < 0$ at RS+1 and aligned them with their 18S rRNA tails. The lines between sequences indicate that base pairing is possible between the two nucleotides (including G/U mismatches). With very few exceptions, the nucleotides in the 18S rRNA tails that can base pair with the mRNA consensus sequence are the same as those in *E. coli*'s and other prokaryote's 16S rRNA tails.

is conserved across prokaryotic and eukaryotic species.

3.4 Discussion

The present study is a characterization of the translation initiation region (TIR) in prokaryotic and eukaryotic mRNAs. In previous work we developed an algorithm that simulates base paring between the 16S rRNA tail and prokaryotic TIRs [88]. The algorithm predicts the affinities between the two RNA molecules at each position in the TIR, revealing both the location of the optimal base paring, as well as the locations of sub-optimal base pairing. This algorithm revealed two features that were common in prokaryotic mRNAs. In confirmation of previous studies [93, 82, 70, 58], it detected a trough of negative free energy in the 5′-UTR, between RS-10 and RS-5, corresponding to the SD region. A second, and previously unreported, trough of negative free energy, at RS+1, was also detected. This second trough, due to hybridization between 16S rRNA tail and the portion of the TIR that included the initiation codon, occurred in 58% of the 58,450 mRNA sequences examined across 18 prokaryotic species.

The existence of hybridization at RS+1 begs the question of what its role could be in translation. One possibility is that it somehow functions in conjunction with or in place of SD sequences to facilitate binding of the small ribosomal subunit. If this were the case, we would expect to find an association between the two hybridization sites. Our correlation and association tests, however, offer little support for this hypothesis.

Another approach that could illuminate a relationship between binding at RS+1 and SD sequences would be to analyze which nucleotides in the 16S rRNA tail interact with the mRNA at these two locations. The literature has established that the core aSD hybridizes to SD sequences. However, despite conservation across prokaryotic species, no functional role has been ascribed to the nucleotides that flank the core: the 3 to 5 nucleotides $3'$, and the 5 nucleotides $5'$. A relationship between the SD sequences and RS+1 binding might be reflected in the patterns of nucleotides from the 16S rRNA tail that participate in hybridization.

Our results indicate that binding at RS+1 arises largely from the nucleotides in the 16S rRNA tail that flank the core (see Figure 3.1). In 10 of the 18 species examined, the consensus sequences from genes with binding at RS+1 show that the core does not participate in the hybridization. In 6 of 18 species, only one nucleotide in the core, the U in the center, participates in binding at RS+1. In two other species examined, one or two of the C nucleotides from the $3'$ side of the core also participated. However, in all 18 species examined, a) the $5'$-UCA-$3'$ of the highly conserved $5'$-end of the tail, $5'$-GAUCA-$3'$ and, b) the two U's directly $3'$ to the core nucleotides, were found to participate in binding at RS+1. Yassin and co-workers [105] identified the A (A1535) at the $3'$-end of the highly conserved $5'$-end of the tail, $5'$-GAUCA-$3'$ as having a functional role through the ability of a mutation at this site to prevent cell growth. However, further characterization of this mutant strain to determine its specific effect on translation was not done.

Evidence that SD sequences and binding at RS+1 are not associated, and the different sets of nucleotides from the 16S rRNA tail that participate in hybridizing at these sites is consistent with the hypothesis that their biological roles are independent. However, the nature of the molecular process for which binding at RS+1 is useful is still unknown. That binding at RS+1 has a biological function is highly likely when considered in light of further observations from eukaryotic organisms. The bases flanking the core aSD, which are involved in binding at RS+1 in prokaryotes, are the highly conserved nucleotides in the eukaryotic 18S rRNA $3'$-terminal tails. This sequence conservation was noted at the time Shine and Dalgarno proposed a functional role for the 16S rRNA tail and immediately suggested that the $3'$-terminal tails of the prokaryotic 16S and eukaryotic 18S rRNAs might share an analogous function [30, 78]. However, it became quickly apparent that eukaryotic mRNAs lacked SD sequences and the 18S rRNA tail lacked the complementary core nucleotides [30]. More recent studies looking for 18S rRNA/mRNA hybridization have implicated other regions in the 18S rRNA [100, 63, 64, 95, 37], leaving the conservation of the 18S rRNA tail and its homology to the 16S rRNA tail as an unanswered puzzle.

The sequence conservation between the nucleotides flanking the core aSD and the nucleotides in the 18S rRNA tails, and the involvement of the flanking nucleotides in binding at RS+1 led us to look for the presence of binding at RS+1 in eukaryotic mRNA sequences. In the six species examined, including at least one representative of fungal, plant and animal kingdoms, binding at RS+1 is present in 66% to 93% of the gene sequences examined (see Figure 3.2 and Table 3.3). Further analysis to determine which mRNA nucleotides were binding to the 18S tail at RS+1 showed that in all 6 species, the start codon and one nucleotide immediately $5'$ to the start codon and/or one or two nucleotides $3'$ to the start codon, in the second codon, were participating in hybridization, i.e. the nucleotide "context" for the start codon (see Figure 3.3). There is a discernible pattern of more frequent nucleotides in these contexts: in plant (*A. thaliana*), yeast (*S. cerevisiae*) and insect (*D. melanogaster*), if there is a participating nucleotide $5'$ to the start codon is most likely to be an A; in mouse and humans it is a G; in nematode (*C. elegans*) hybridization at RS+1 with the nucleotide $5'$ to the start codon is rare. Both *C. elegans* and yeast are similar in that when hybridization involves nucleotides in the second codon, an A at the first nucleotide position of the second codon is most likely to participate. In mouse, humans, fruit fly and plant, the nucleotide in the similar position is likely to be a G. Finally, in fruit fly, yeast and nematode, the RS+1 hybridization pattern can involve the third nucleotide in the second codon and, if it participates, it is likely to be a U.

A number of researchers have investigated the non-random occurrence of nucleotides flanking the start codon in a variety of eukaryotes. Kochetov and co-workers [47] found that there were statistically significant differences in nucleotide frequencies in the start codon context of mRNAs encoding proteins of high abundance versus low abundance in mammalian cells. Kozak [51] has proposed an optimal context, $5'$-GCCACCAUGG-$3'$, for the start codons in vertebrates based on mutational and gene expression studies. Of these nucleotides, the most critical to expression are the A at position -3 to the start codon and the G at position +4. Our consensus sequence for mouse and humans also contains a G at position +4, however we cannot speak to the nucleotide at position -3 as the vertebrate tail alignment at RS+1 does not extend this far upstream. Our consensus sequence places a G immediately $5'$ to the start codon, whereas in Kozak's optimal context, a C is immediately $5'$ to the start codon (Position -1). In plants, Kawaguchi and Bailey-Serres [46] found a series of A's $5'$ to the start codon and a G immediately $3'$ to the start codon in highly translated mRNAs, in agreement with our consensus sequence for this species. In yeast, Hamilton and co-workers [31] determined the optimal context around the start codon in 50 highly expressed genes as $5'$-AAAAAAAUGUCU-$3'$. This is largely consistent with our consensus sequence with the exception of the nucleotide immediately $3'$ to the start codon in which they indicate a U and we found an A. The nucleotide similarities between our consensus sequences and the optimal start-codon, sequence-context for translation found by others is suggestive that binding at RS+1 may play a role in translation initiation. The magnitude of the binding at RS+1 is a function of the specific context in which the start codon is embedded. One possible interpretation is that the magnitude of the binding at RS+1 may correlated with translation initiation efficiency. However, experimental verification of computational results is needed to evaluate this interpretation and to determine the function of binding at RS+1.

3.5 Materials and Methods

3.5.1 Genome Sequences

All genome sequences were downloaded from the National Center for Biotechnology Information (NCBI) GenBank database (http://www.ncbi.nlm.nih.gov/). Table 3.2 contains the names of the species analyzed, their RefSeq version numbers, the sequence of the rRNA tail used for the analysis, and their predicted 16S rRNA secondary structure.

3.5.2 Selecting Criteria for Gene Sequences and the RS+1 Dataset

For all genomes we began with all gene sequences with `gene=` or `locus_tag=` tags in our dataset, with the following exceptions: genes that included a `transposon=` or `pseudo` tag, and genes that contained > and < characters to define the location of the sequence. The 58450 bacterial genes were then divided into four categories: No RS+1 binding ($\Delta G^\circ = 0$), Weak RS+1 binding ($0 > \Delta G^\circ > -3.4535$, where $\Delta G^\circ = -3.4535$ is the minimum value for an SD sequence, [60, 88]), SD-like RS+1 binding ($-3.4535 \geq \Delta G^\circ \geq -8.4$), and strong RS+1 binding ($\Delta G^\circ < -8.4$). We then excluded all genes with SD-like RS+1 binding or strong RS+1 binding that also had an in-frame downstream start codon (AUG, GUG, or UUG) within 12 nucleotides of the annotated initiation codon. This criteria was intended to eliminate mis-annotated genes, and genes with multiple start codons where the binding at RS+1 functions as secondary SD sequence. The final set of genes, called the RS+1 dataset, contained 33933 genes with RS+1 binding (see Table 3.1).

We defined the TIR so that we could calculate 35 ΔG° values before the start codon, one at the first base in the start codon, and 35 ΔG° values after the start codon. See Starmer et al. [88] for more details.

3.5.3 Determining the $3'$ rRNA Tails for the 16S rRNAs

Prokaryotic rRNA tails were selected using the same criteria used in Starmer et al. [88]. With the exception for *C. elegans*, all eukaryotic rRNA tails were determined from predicted secondary structures downloaded from The Comparative RNA Web Site [8] (http://www.rna.icmb.utexas.edu/). The rRNA tail for *C. elegans* was taken from Ellis et al. [19].

3.5.4 Calculating ΔG° and Relative Spacing

ΔG° and Relative Spacing (RS) were calculated in the same manner used by Starmer et al. [88]. In short, RS is the position of the $3'$ A in the rRNA sequence $5'-$GAUCA$-3'$, relative to the first base in the start codon. If the SD is upstream from the start codon, its RS is negative, while if it is downstream, its RS is positive. If the two are opposite one another, its RS is zero.

3.5.5 Correlation and association statistics

To test if the magnitude of binding at SD sequences was correlated with the magnitude of binding at RS+1, we computed the correlation coefficient for these two variables, using a Bonferroni cutoff of $0.05/18 = 0.0028$ to determine if the value was significantly different from 0.

To test for association between the presence of binding at and SD sequence and binding at RS+1, we generated χ^2 statistics, using a Bonferroni cutoff of $0.05/18 = 0.0028$ to determine its significance.

3.5.6 Determining consensus sequences

Given the set of sequences, for each base position in the consensus sequence we selected the most frequently occurring nucleotide,

3.6 Acknowledgments

The authors would like to thank C. Frank Starmer for advice on the statistical analysis, and Eric Miller for critical feedback.

Chapter 4

free2bind: a website for investigating RNA-RNA hybridizations.

4.1 Abstract

free2Bind (http://free2bind.dnsalias.org/free2bind/) is a web application for predicting the hybridization between two RNA sequences. free2bind offers two methods for processing the sequences: align mode and scanning mode. Align mode computes and displays the optimal hybridization between two RNA molecules. Scanning mode identifies all optimal hybridizations as a shorter RNA sequence is presented a sliding window of bases from a longer RNA sequence. free2bind can identify and quantify Shine-Dalgarno sequences, siRNA and miRNA target sequences, all possible secondary structure binding sites for a sub-sequence within an RNA molecule, and any other RNA-RNA hybridization. Here we demonstrate free2bind's utility by using it to find novel characteristics of siRNAs and their target sequences.

4.2 Introduction

In recent years it has become clear that non-coding RNA sequences exert an enormous amount of regulatory control over gene expression. Hybridizations between non-coding RNA and mRNAs affect all post-transcriptional stages. During translation, interactions between the ribosomal RNA (rRNA) and mRNA influence both the efficiency of initiation [83, 38, 41] as well as the reading frame [101, 55]. Base-pairing between transfer RNAs (tRNAs) and mRNA determines the efficiency of translation. Translation can also be repressed by micro-RNAs (miRNAs) binding to mRNA (see [33] for a review), and RNA-interference (RNAi) degrades mRNA when small-interfering-RNAs (siRNAs) anneal to it (see [32] for a review). The importance of RNA-RNA interactions has driven the creation of many tools for studying them, including `RNAHybrid` [74], `RNAcofold` [2], and `free_scan` [88].

Here we present free2bind (http://free2bind.dnsalias.org/free2bind/), a website that provides a unified interface to two tools for detecting RNA hybridization: `free_scan` and `free_align`, two components of the free2bind package (http://sourceforge.net/projects/free2bind). The combination of these two tools allows researchers to obtain both local and global perspectives on how RNA molecules interact with each other. Using the align option, one can calculate the single, optimal binding between the two molecules. The scanning option calculates all possible hybridizations between the two molecules, giving researchers a chance to identify both the optimal as well as significant suboptimal binding sites throughout the target sequence. The free2bind web application can be used for discovering and investigating the effectiveness of siRNA and miRNA targets, identifying and quantifying Shine-Dalgarno (SD) sequences, and investigating other mRNA-rRNA interactions. free2bind generates both numerical output, which can then be easily manipulated in a spreadsheet program, or graphical output, which provides users an intuitive means for inspecting potential binding sites between two RNA molecules.

4.3 Results

Investigating siRNA target sequences

Recently, Schubert et al. [81] and Overhoff et al. [72] described how mRNA secondary structures reduce the silencing effect of siRNA. Both groups made the hypothesis that siRNAs targeting sequences involved in secondary structures within the mRNA would be less effective at silencing the gene than those targeting exposed sequences. While both groups acknowledged that properties intrinsic to the siRNA, such as base composition, had a significant impact on its effectiveness, both showed secondary structures involving the target sequence limited the siRNA's ability to silence the gene.

Schubert et al. based their experiments on siRNAs that are highly effective at degrading their target, the vanilloid receptor subtype 1 (VR1) from *R. norvegicus* [29]. We used free2bind to investigate whether this set of siRNAs, which we refer to as the Grünweller set, shared characteristics. We used align mode to locate and quantify the target sequences for each mRNA. We then used the scanning mode to calculate the ΔG° values for all locations on the mRNA that could bind to the siRNA, revealing both the primary target sequence as well as alternative target sequences. We also used the scanning mode to characterize the binding potential between siRNA's target sequence and the remaining mRNA sequence, revealing its potential for secondary structure.

In Figure 4.1A, the top graph shows the ΔG° values for all of the locations on the VR1 mRNA that the siRNA, VsiRNA3, could bind to. Close to the center of the graph we can identify the location of the target sequence because its ΔG° value is dramatically lower than all others. The bottom graph in Figure 4.1A shows the ΔG° values for all of the subsequences in the mRNA that could bind to VsiRNA3's target sequence as part of its secondary structure. Of all the sequences that could bind to VsiRNA3's target sequence, only one has ΔG° value less than -10 kcal/mol, indicating that there is very limited competition between VsiRNA3 and the mRNA's secondary structure for binding to the target sequence.

Table 4.1 summarizes the data from Figure 4.1A as well as the data collected for the other five efficient siRNAs in the Grünweller dataset. The third and fourth columns quantify the siRNA target sequence's potential for inhibition by secondary structure. Secondary structure sites (SSSs) are mRNA sub-sequences that can readily bind to the target sequence ($\Delta G^{\circ} < -10$ kcal/mol).

To test the hypothesis that low SSS counts and smaller ΔG° values for secondary structures involving the siRNA target sequences are correlated with increased siRNA efficiency, we analyzed Overhoff et al.'s siRNA data [72] with free2bind. The Overhoff dataset provides both siRNA sequences as well as IC_{50} values, the concentration required to degrade 50% of the target mRNAs, indicating their relative efficiency. The siRNAs were also partitioned into two categories that described the predicted state of the siRNA target sequence: accessible and inaccessible. An analytical method detected whether a target sequence was involved in secondary structure. Briefly, they examined the mRNA sequence

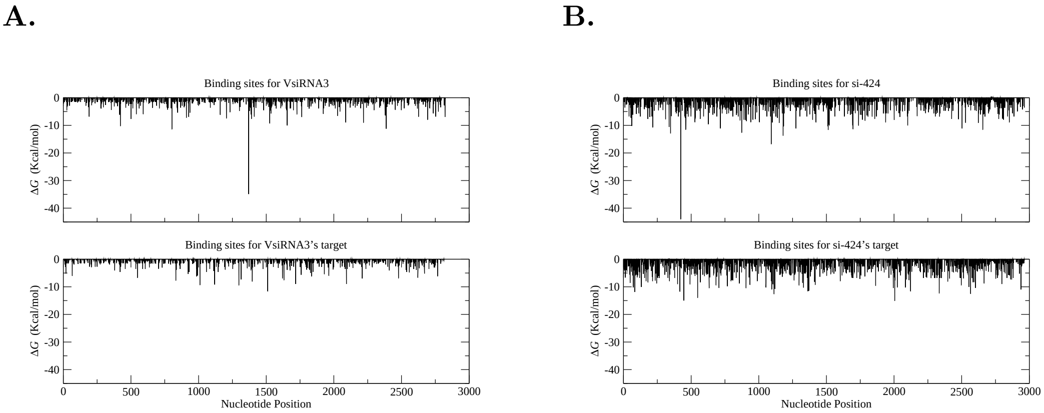

Figure 4.1: **Graphs of the ΔG° values calculated using free2bind's scanning mode.** In **A**, the plots are from the highly efficient siRNA, VsiRNA3 [29]. In **B**, the plots are from the relatively inefficient siRNA, si-424 [72]. The plots on the top show the ΔG° values for the locations where the siRNAs can bind to their target mRNAs. The plots on the bottom show the ΔG° values for the locations where the siRNA target sequences can bind to the mRNA, forming secondary structures.

Table 4.1: **Analysis of the siRNAs and their target sequences from Grünweller et al.** [29]**.** The second column lists the ΔG° value for the siRNA hybridizing with its target sequence. The third column lists the secondary structure site (SSS) count for the siRNA target sequence. An SSS is an mRNA subsequence that can readily bind to the siRNA target sequence ($\Delta G^\circ < -10$), implying that the siRNA target may be involved in secondary structure. The final column lists the lowest ΔG° value for the SSSs. This value indicates how competitive this SSS is with the siRNA for the target sequence.

Name	ΔG° at target	SSS count	Lowest ΔG°
VsiRNA1	-36.9	4	-15.4
VsiRNA2	-37.4	1	-12.5
VsiRNA3	-35.4	1	-11.6
VsiRNA4	-39.7	5	-12.5
VsiRNA5	-35.9	4	-13.4
VsiRNA6	-37.2	2	-10.7
Average	-37.1	2.8	-12.7

by using Mfold [108, 62] to predict folding of an 800 nucleotide sliding window, moving the window 700 nucleotides at a time. miRNA target sequences are labeled inaccessible or accessible, depending on the presence or absence of secondary structure.

Comparing Figure 4.1A to 4.1B illustrates differences between a highly effective siRNA and a relatively ineffectual siRNA. VsiRNA3, an efficient siRNA, has relatively few locations in the mRNA that are in the immediate vicinity of the target sequence. Compared with si-424, a poor siRNA, VsiRNA3 and its target sequence have far fewer potential binding sites.

Table 4.2 summarizes the results of our analysis of the Overhoff dataset with free2bind and shows that siRNA targets predicted to be accessible have lower SSS counts and smaller ΔG° values than the siRNA targets predicted to be inaccessible. The averages for both the SSS counts and the lowest SSS ΔG° value are significantly different between the targets predicted to be accessible and targets predicted to be inaccessible (both p-values = 0.03). Another very interesting difference between the two categories of siRNAs is between the IC_{50} values. While the IC_{50} values for the siRNAs predicted to be accessible are all relatively low, the IC_{50} values for the inaccessible siRNAs range from low to high.

Several of the siRNAs in the Overhoff dataset predicted to be inaccessible were distinctive because they did not follow the general trends. si-650's target sequence appears to have been mis-categorized as inaccessible because its IC_{50} measurement and our calculations for its SSSs are consistent with it being accessible. The target sequences for si-640, si-1153, and si-1437 were all predicted to be inaccessible, and our calculations for their SSSs are in accord with this prediction. However, they have relatively low IC_{50} measurements for this category. One possible explanation for the efficacy of these three siRNAs is that they posses intrinsic properties that allow them to overcome the hindrance of mRNA secondary structure. si-1452, which has a fairly high IC_{50} value, has a seemingly contradictory low SSS count. The high IC_{50}, however is likely to be the product of si-1452's ability to fold into its own stable secondary structure. We verified this using Mfold [109] to predict the folding for the siRNAs and found that si-1452 can fold itself into a stable hairpin structure ($\Delta G^{\circ} = -13$ kcal/mol). None of the remaining siRNAs with inaccessible target sequences were capable of forming such a stable hairpin structure (the lowest ΔG° value was -4.5 kcal/mol and the average was -2.3 kcal/mol), so it is unlikely that their own secondary structures inhibit silencing.

Table 4.2: **Analysis of the siRNAs and their target sequences from Overhoff et al.** [72]**.** The second column lists the ΔG° value for the siRNA hybridizing with its target sequence. The third column gives a '+' if Overhoff et al. predicted this siRNA target to be accessible (free of mRNA secondary structure), and '-' if it was predicted to be inaccessible (involved in mRNA secondary structure). The IC_{50} listed in the fourth column indicates the concentration required to reduce the amount of target mRNA by half. The fifth and sixth columns are as described in Table 4.1.

Name	ΔG° at target	Predicted accessibility	IC_{50}	SSS count	Lowest ΔG°
si-839	-42.1	+	5	7	-11.3
si-840	-40.6	+	4	5	-11.3
si-841	-41.6	+	4	5	-10.8
si-842	-40.5	+	3	4	-10.2
si-843	-41.3	+	3	4	-10.2
si-859	-46.8	+	4	14	-17.2
si-860	-45.6	+	2	15	-16.8
si-861	-47.4	+	10	22	-15.2
si-1546	-43.5	+	2	11	-20.0
si-1595	-38.2	+	4	4	-12.0
Average	-42.8	+	4.1	9.1	-13.5
si-424	-45.8	-	>100	26	-15.1
si-429	-44.7	-	>80	20	-15.1
si-640	-42.4	-	4	25	-21.4
si-650	-40.2	-	3	2	-13.6
si-1153	-47.4	-	7	20	-15.8
si-1437	-44.9	-	6	24	-15.6
si-1452	-48.2	-	40	10	-20.5
si-2289	-43.8	-	>100	12	-19.5
Average	-44.7	-	42.5	17.4	-17.1

4.4 Discussion

free2bind is a flexible program for identifying the location and magnitude of hybridizations between two RNA molecules. Here we show how it can characterize siRNAs and their target sequences, shedding light on why one siRNA is more efficient than another. By using a combination of free2bind's align and scanning modes we showed that si-650 was not categorized correctly. These tools also lead us to identify the secondary structure within si-1452. free2bind also identifies any other hybridization between two RNA molecules, including interactions between the 16S and 18S rRNAs and mRNA sequences before and during translation.

4.5 Acknowledgments

The authors wish to thank David Aylor and Rachel Knox for constructive comments and editing.

Chapter 5

Conclusion

When I talk about my research, especially what I wrote about in Chapters 2 and 3, the most common questions are “What about leaderless mRNAs?” and “What about IRESs?”. Both of these questions are, at their roots, asking whether RS+1 binding can take the place of the most common ribosome recruitment strategies in prokaryotes and eukaryotes. The first question concerns prokaryotic mRNA sequences without $5'$ untranslated regions; the start codon is the first three bases on the mRNA. Leaderless mRNAs clearly lack upstream SD sequences. The second question is about a somewhat contested phenomenon observed in eukaryote mRNAs. IRES stands for internal ribosomal entry site and refers to translation initiation sites that function more like those in prokaryotes. An IRES, it is thought, attracts the ribosome directly to the start codon and does not require a $5'$ cap (see [99] for a review).

For both of these questions, I typically give the same answer. I think that there are two related, but still very distinct, processes are taking place during initiation. The first process is ribosome recruitment, and it is most often taken care of by either SD-aSD binding in bacteria or scanning from a $5'$ cap in eukaryotes. The second process, in which RS+1 binding might be involved, is for fine tuning the ribosome’s position or for establishing the initial reading frame. While it may be possible that some mRNA use an ideal RS+1 sequence to perform both processes in the absence of one of the standard recruitment mechanisms, I don’t believe this is the case. I think that RS+1 binding only helps with the fine tuning and, in order for it to function correctly, can not bond strongly enough to either 16S or 18S rRNAs for recruitment purposes[1].

[1] One thing that is potentially tricky about analyzing the RS+1 site is that it could be like SD sequences,

I have several reasons for thinking that recruiting ribosomes is distinct from fine tuning the ribosome's position over the mRNA. The first was covered in Chapter 3: in bacteria, RS+1 binding does not compensate for missing SD sequences in leadered genes. If there were a relationship between the presence or the magnitude of the binding at an SD sequence and RS+1 binding, then I might be persuaded to think that one could compensate for the other. There isn't. In addition, I believe that recruitment is independent of fine tuning because placement of an SD sequence in relation to the start codon is flexible. As long as the SD sequence is not too close and not too far away from the start codon, it appears to do its job well; there isn't a particular location that is markedly better than another. On the other hand, the RS+1 binding site is consistent: it is always at RS+1.

The research I have conducted thus far leads to additional areas to explore. On the analytical side, I am interested in examining RS+1 binding in archaea as well as in chloroplast and mitochondrial mRNA sequences. Thus far it appears that RS+1 binding occurs in the majority of mRNA sequences on the planet. The remaining analytical work could confirm the universal nature of this binding and hopefully stimulate the experimental community's curiosity about the biological role of RS+1 binding. From an experimental perspective, I am interested in verifying whether the 16S or 18S rRNA tails are binding at RS+1. Once binding is established, I would like to determine what affect it has on translation efficiency.

As for continuing the work started in Chapter 4, I would like to see if there isn't more we can learn about how effective an siRNA is by looking at how it and its target can bind to mRNA sequences. This work would ideally be performed in a lab environment where each hypothesis can be tested.

requiring the strength of the binding with the ribosome to be within an optimal range. SD sequences can not be to strong[49] or too weak[75] if they are to function correctly.

Bibliography

[1] Q. Bao et al. A complete sequence of the *T. tengcongensis* genome. *Genome Res.*, 12(5):689–700, 2002.

[2] S. Bernhart, H. Tafer, U. Mückstein, C. Flamm, P. Stadler, and I. Hofacker. Partition function and base pairing probabilities of rna heterodimers. *Algorithms for Mol. Biol.*, 1(3), 2006.

[3] F. Blattner, G. Plunkett III, C. Bloch, N. Perna, V. Burland, M. Riley, J. Collado-Vides, J. Glasner, C. Rode, G. Mayhew, J. Gregor, N. Davis, H. Kirkpatric, M. Goeden, D. Rose, B. Mau, and Y. Shao. The complete genome sequence of *Escherichia coli* K-12. *Science*, 277:1453–1462, 1997.

[4] I. Boni. Diverse molecular mechanisms of translation initiation in prokaryotes. *Mol. Biol.*, 40(4):658–668, 2006.

[5] I. Boni, V. Artamonova, N Tzareva, and M. Dreyfus. Non-canonical mechanism for translational control in bacteria: synthesis of ribosomal protein S1. *EMBO J.*, 20(15):4222–4232, 2001.

[6] P. Borer, B. Dengler, I. Tinoco, and O. Uhlenbeck. Stability of ribonucleic acid double-stranded helices. *J. Mol. Biol.*, 86(4):843–853, 1974.

[7] H. Brüggemann, S. Bäumer, W. Frike, A. Wiezer, H. Liesegang, I. Decker, C. Herzberg, R. Martńez-Arias, R. Merkl, A. Henne, and G. Gottschalk. The genome sequence of *Clostridium tetani*, the causative agent of tetanus disease. *Proc. Natl. Acad. Sci.*, 100(3):1316–1321, 2003.

[8] J. Cannone, S. Subramanian, M. Schnare, J. Collett, L. D'Souza, Y. Du, B. Feng, N. Lin, L. Madabusi, K. Muller, N. Pande, Z. Shang, N. Yu, and R. Gutell. 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), 2002.

[9] D. Capela et al. Analysis of the chromosome sequence of the legume symbiont *Sinorhizobium meliloti* strain 1021. *Proc. Natl. Acad. Sci.*, 98(17):9877–9883, 2001.

[10] H. Chen, , L. Pomeroy-Cloney, M. Bjerknes, J. Tam, and J. Ernest. The influence of adenine-rich motifs in the 3′ portion of the ribosome binding site on human IFN-γ gene expression in *Escherichia coli*. *J. Mol. Biol.*, 240:20–27, 1994.

[11] H. Chen, M. Bjerknes, R. Kumar, and J. Ernest. Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon in *Escherichia coli* mRNAs. *Nucleic Acids Res.*, 22(23):4953–4957, 1994.

[12] G. Crooks, G. Hon, J. Chandonia, and S. Brenner. Weblogo: A sequence logo generator. *Genome Res.*, 14:1188–1190, 2004.

[13] A. da Silva et al. Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. *Nature*, 417:459–463, 2002.

[14] M. de Smit and J. van Duin. Translational standby sites: how ribosomes may deal with the rapid folding kinetics of mRNA. *J. Mol. Biol.*, 331:737–743, 2003.

[15] G. Deckert, P. Warren, T. Gaasterland, et al. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature*, 392(6674):353–358, 1998.

[16] A. Delcher, D. Harmon, S. Kasif, O. White, and S. Salzberg. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, 27(23):4636–4641, 1999.

[17] W. Deng et al. Genome sequence of *Yersinia pestis* KIM. *J. Bacteriol.*, 184(16):4601–4611, 2002.

[18] F. Dyson. *Origins of life*. Cambridge University Press, 1985.

[19] R. Ellis, J. Sulston, and A. Coulson. The rDNA of *C. elegans*: sequence and structure. *Nucleic Acids Res.*, 14(5):2345–2364, 1986.

[20] W. Van Etten and G. Janssen. An AUG initiation codon, not codon-anticodon complementarity, is required for the translation of unleadered mRNA in *Escherichia coli*. *Mol. Microbiol.*, 27(5):987–1001, 1998.

[21] M. Faxén, J. Plumbridge, and L. Isaksson. Codon choice and potential complementarity between mRNA downstream of the initiation codon and bases 1471-1480 in 16S ribosomal RNA affects expression of *glnS*. *Nucleic Acids Res.*, 19(19):5247–5251, 1991.

[22] A. Fire, S. Xu, M. Montgomer, S. Kostas, S. Driver, and C. Mello. Potent and specific genetic interfernce by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, 391:806–811, 1998.

[23] R. Fleischmann et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269:496–512, 1995.

[24] S. Freier, R. Kierzek, J. Jaeger, N. Sugimoto, M. Caruthers, T. Neilson, and D. Turner. Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci.*, 83:9373–9377, 1986.

[25] N. Galtier and J. Lobry. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J. Mol. Evol.*, 44:632–636, 1997.

[26] D. Gray. Derivation of nearest-neighbor properties from data on nucleic acid oligomers. I. simple sets of independent sequences and the influence of absent nearest neighbors. *Biopolymers*, 42(7):783–793, 1997.

[27] D. Gray. Derivation of nearest-neighbor properties from data on nucleic acid oligomers. II. thermodynamic parameters of DNA-RNA hybrids and DNA duplexes. *Biopolymers*, 42(7):795–810, 1997.

[28] D. Gray and I. Tinoco. A new approach to the study of sequence-dependent properties of polynucelotides. *Biopolymers*, 9(2):223–244, 1970.

[29] A. Grünweller, E. Wyszko, B. Bieber, R. Jahnel, V. Erdmann, and J. Kurreck. Comparison of different antisense strategies in mammalian cells using locked nucleic acids,

2'-O=methyl RNA, phosphorothioates and small interfering RNA. *Nucleic Acids Res.*, 31(12):3185–3193, 2003.

[30] O. Hagenbüchle, M. Santer, J. Steitz, and R. Mans. Conservation of the primary structure at the $3'$ end of 18S rRNA from eucaryotic cells. *Cell*, 13:551–563, 1978.

[31] R. Hamilton, C. Watanabe, and H. de Boer. Compilation and comparison of the sequence context around the AUG start codons in *Saccharomyces cerevisiae* mRNAs. *Nucleic Acids Res.*, 15(8):3581–3593, 1987.

[32] G. Hannon. RNA interference. *Nature*, 418:244–251, 2002.

[33] L. He and G. Hannon. MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genetics*, 5:522–531, 2004.

[34] A. Henne, H. Bruggemann, C. Raasch, A. Wiezer, T. Hartsch, H. Liesegang, A. Johann, T. Lienard, O. Gohl, R. Martinez-Arias, C. Jacobi, V. Starkuviene, S. Schlenczeck, S. Dencker, R. Huber, H. Klenk, W. Kramer, R. Merkl, G. Gottschalk, and H. Fritz. The genome sequence of the extreme thermophile *Thermus thermophilus. Nat. Biotechnol.*, 22:547–553, 2004.

[35] N. Hodas and D. Aalberts. Efficient computation of optimal oligo-RNA binding. *Nucleic Acids Res.*, 32(22):6636–6642, 2004.

[36] M. Holden, E. Feil, J. Lindsay, S. Peacock, et al. Complete genomes of two clinical *Staphylococus aureus* strains: Evidence for the rapid evolution of virulence and drug resistance. *Proc. Natl. Acad. Sci.*, 101(26):9786–9791, 2004.

[37] M. Hu, P. Tranque, G. Edelman, and V. Mauro. rRNA-complementarity in the $5'$ untranslated region of mRNA specifying the Gtx homeodomain protein: evidence that base-pairing to 18S rRNA affects translational efficiency. *Proc. Natl. Acad. Sci.*, 96:1339–1344, 1999.

[38] A. Hui and H. de Boer. Specialized ribosome system: preferential translation of a single mRNA species by a subpopulation of mutated ribosomes in *Escherichia coli. Proc. Natl. Acad. Sci.*, 84:4762–4766, 1987.

[39] J. Izant and H. Weintraub. Inhibition of thymidine kinase gene expression by antisense rna: a molecular approach to genetic analysis. *Cell*, 36:1007–1015, 1984.

[40] Jr. J. SantaLucia. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci.*, 95:1460–1465, 1998.

[41] W. Jacob, M. Santer, and A. Dahlberg. A single base change in the Shine-Dalgarno region of 16S rRNA of *Escherichia coli* affects translation of many proteins. *Proc. Natl. Acad. Sci.*, 84:4757–4761, 1987.

[42] J. Jaeger, D. Turner, and M. Zuker. Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci.*, 86:7706–7710, 1989.

[43] T. Kaneko et al. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.*, 3:109–136, 1996.

[44] T. Kaneko, Y. Nakamura, S. Sato, K. Minamisawa, T. Uchiumi, S. Sasamoto, A. Watanabe, K Idesawa, M. Iriguchi, K. Kawashima, M. Kohara, M. Matsumoto, S. Shimpo, H. Tsuruoka, T. Wada, M. Yamada, and S. Tabata. Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110. *DNA Res.*, 9:189–197, 2002.

[45] T. Kaneko, Y. Nakamura, C. Wolk, T. kuritz, S. Sasamoto, A. Watanabe, et al. Complete genomic sequence of filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. *DNA Res.*, 8:205–213, 2001.

[46] R. Kawaguchi and J. Bailey-Serres. mRNA sequence features that contribute to translation regulation in *Arabidopsis*. *Nucleic Acids Res.*, 33(2):955–965, 2005.

[47] A. Kochetov, I. Isecheako, D. Vorobiev, A. Kel, V. Babenko, L. Kisselev, and N. Kolchanov. Eukaryotic mRNAs encoding abundant and scarce proteins are statistically dissimilar in many structural features. *FEBS Lett.*, 440(3):351–355, 1998.

[48] V. Kolev, I. Ivanov, A. Berzai-Herranz, and I. Ivanov. Non-Shine-Dalgarno initiators of translation selected from combinatorial DNA libraries. *J. Mol. Microbiol. Biotechnol.*, 5(3):154–160, 2003.

[49] A. Komarova, L. Tchufitsova, E. Supina, and I. Boni. Extensive complementarity of the Shine-Dalgarno region and the $3'$-end of 16S rRNA is inefficient for translation *in vivo*. *Bioorg. Khim.*, 27(4):248–255, 2001.

[50] M. Kozak. How do eucaryotic ribosomes select initiation regions in messenger RNA. *Cell*, 15:1109–1123, 1978.

[51] M. Kozak. An analysis of 5′-noncoding sequences from 699 messenger RNAs. *Nucleic Acids Res.*, 15(20):8125–8148, 1987.

[52] M. Kozak. Initiation of translation in prokaryotes and eukaryotes. *Gene*, 234:187–208, 1999.

[53] M. Kozak. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene*, 361:13–37, 2005.

[54] F. Kunst et al. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature*, 390:249–256, 1997.

[55] B. Larsen, N. Wills, R. Gesteland, and J. Atkins. rRNA-mRNA base pairing stimulates a programmed -1 ribosomal frameshift. *J. Bacteriol.*, 176(22):6842–6851, 1994.

[56] K. Lee, C. Holland-Staley, and P. Cunningham. Genetic analysis of Shine-Dalgarno interaction: selection of alternative functional mRNA-rRNA combinations. *RNA*, 2:1270–1285, 1996.

[57] P. Lió, S. Ruffo, and M. Buiatti. Third codon g+c periodicity as a possible signal for an ïnternalšelective constraint. *J. Theor. Biol.*, 171:215–223, 1994.

[58] G. Lithwick and H. Margalit. Hierarchy of sequence-dependent features associated with prokaryotic translation. *Genome Res.*, 13:2665–2673, 2003.

[59] A. Looman, J. Bodlaender, L. Comstock, D. Eaton, P. Jhurani, H. de Boer, and P. von Knippenberg. Influence of the codon following the AUG initiation codon on the expression of a modified *lacZ* gene in *Escherichia coli*. *EMBO J.*, 6(8):2489–2492, 1987.

[60] J. Ma, A. Campbell, and S. Karlin. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J. Bacteriol.*, 184(20):5733–5745, 2002.

[61] J. Martin-Farmer and G. Janssen. A downstream CA repeat sequence increases translation from leadered and unleadered mRNA in *Escherichia coli*. *Mol. Microbiol.*, 31(4):1025–1038, 1999.

[62] D. Mathews, J. Sabina, M. Zuker, and D. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.

[63] O. Matveeva and S. Shabalina. Intermolecular mRNA-rRNA hybridization and the distribution of potential interaction regions in murine 18S rRNA. *Nucleic Acids Res.*, 21(4):1007–1011, 1993.

[64] V. Mauro and G. Edelman. rRNA-like sequences occur in diverse primary transcripts: implications for the control of gene expression. *Proc. Natl. Acad. Sci.*, 94:422–427, 1997.

[65] I. Moll, S. Grill, C. Gualerzi, and U. Bläsi. Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control. *Mol. Microbiol.*, 43(1):239–246, 2002.

[66] T. Nakamoto. A unified view of the initiation of protein synthesis. *Biochem. Biophys. Res. Commun.*, 341:275–278, 2006.

[67] K. Nelson et al. Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature*, 399:323–329, 1999.

[68] M. O'Connor, T. Asai, C. Squires, and A. Dahlberg. Enhancement of translation by the downstream box does not involve base pairing of mRNA with the penultimate stem sequence of 16S rRNA. *Proc. Natl. Acad. Sci.*, 96:8973–8978, 1999.

[69] S. O'Donnell and G. Janssen. The initiation codon affects ribosome binding and translational efficiency in *Escherichia coli* of *c*I mRNA with or without the $5'$ untranslated leader. *J. Bacteriol.*, 183(4):1277–1283, 2001.

[70] Y. Osada, R. Saito, and M. Tomita. Analysis of base-pairing potentials between 16S rRNA and $5'$ UTR for translation initiation in various prokaryotes. *Bioinformatics*, 15(7/8):578–581, 1999.

[71] Y. Osada, R. Saito, and M. Tomita. Comparative analysis of base correlations in 5′ untranslated regions of various species. *Gene*, 375:80–96, 2006.

[72] M. Overoff, M. Alken, R. Far, M. Lamaitre, B. Lebleu, G. Sczakiel, and I. Robbins. Local RNA target structure influences siRNA efficacy: a systematic global analysis. *J. Mol. Biol.*, 348:871–881, 2005.

[73] R. Primore, B. Berger, F. Desiere, D. Vilanova, C. Barretto, A. Pittet, M. Zwahlen, M. Rouvet, E. Altermann, R. Barrangou, B. Bollet, A. Mercenier, T. Klaenhammer, F. Arigoni, and M. Schell. The genome sequence of the probiotic intestinal bacterium *Lactobacillus johnsonii* NCC 533. *Proc. Natl. Acad. Sci.*, 101(8):2512–2517, 2004.

[74] M. Rehmsmeier, P. Steffen, M. Höchsmasnn, and R. Giegerich. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10:1507–1517, 2004.

[75] S. Ringquist, S. Shinedling, D. Barrick, L. Green, J. Binkley, G. Stormo, and L. Gold. Translation initiation in *Escherichia coli*: sequences within the ribosome-binding site. *Mol. Microbiol.*, 6(9):1219–1229, 1992.

[76] K. Rudd. EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res.*, 28(1):60–64, 2000.

[77] H. Sakai, C. Imamura, Y. Osada, R. Saito, T. Washio, and M. Tomita. Correlation between Shine-Dalgarno sequence conservation and codon usage of bacterial genes. *J. Mol. Evol.*, 52:164–170, 2001.

[78] D. Sargan, S. Gregory, and P. Butterworth. A possible novel interaction between the 3′-end of 18 S ribosomal RNA and the 5′-leader sequence of many eukaryotic messenger RNAs. *FEBS Lett.*, 147(2):133–136, 1982.

[79] M. A. Schell, M Karmirantzou, B. Snel, D. Vilanova, B. Berger, et al. The genome sequence of *Bifidobacterium longum* reflects its adaptation to the human gastrointestinal tract. *Proc. Natl. Acad. Sci.*, 99(22):14422–14427, 2002.

[80] T. Schneider and R. Stephens. Sequence logos: a new way to display consensus. *Nucleic Acids Res.*, 18:6097–6100, 1990.

[81] S. Schubert, A. Grünweller, V. Erdmann, and J. Kurreck. Local RNA target structure influences siRNA efficacy: systematic analysis of intentionally designed binding regions. *J. Mol. Biol.*, 348:883–893, 2005.

[82] T. Schurr, E. Nadir, and H. Margalit. Identification and characterization of *E. coli* ribosomal binding sites by free energy computation. *Nucleic Acids Res.*, 21(17):4019–4023, 1993.

[83] J. Shine and L. Dalgarno. The $3'$-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc. Natl. Acad. Sci.*, 71(4):1342–1346, 1974.

[84] R. Shultzaberger, R. Bucheimer, K. Rudd, and T. Schneider. Anatomy of *Escherichia coli* ribosome binding sites. *J. Mol. Biol.*, 313:215–228, 2001.

[85] C. Spahn, R. Beckmann, N. Eswar, P. Penczek, A. Sali, G. Blobel, and J. Frank. Structure of the 80S ribosome from *Saccharomyces cerevisiae*- tRNA-ribosome and subunit-subunit interactions. *Cell*, 107:373–386, 2001.

[86] M. Sprengart, Hans Fatscher, and E. Fuchs. The initiation of translation in *E. coli*: apparent base pairing between the 16srRNA and downstream sequences of the mRNA. *Nucleic Acids Res.*, 18(7):1719–1723, 1990.

[87] A Stark, J. Brennecke, R. Russell, and S. Cohen. Identification of *Drosophila* microRNA targets. *PLoS Biol.*, 1(3):397–409, 2003.

[88] J. Starmer, A. Stomp, M. Vouk, and D. Bitzer. Predicting Shine-Dalgarno sequence locations exposes genome annotation errors. *PLoS Computat. Biol.*, 2(5):e57, 2006.

[89] J. Steitz and K. Jakes. How ribosomes select initiator regions in mRNA: base pair formation between the $3'$ terminus of 16S rRNA and the mRNA during initiation of protein synthesis in *Escherichia coli. Proc. Natl. Acad. Sci.*, 72(12):4734–4738, 1975.

[90] C. Stenström, E. Holmgren, and L. Isaksson. Cooperative effects by the initiation codon and its flanking regions on translation initiation. *Gene*, 273:259–265, 2001.

[91] C. Stenström and L. Isaksson. Influences on translation and early elongation by the messenger RNA region flanking the initiation codon at the $3'$ side. *Gene*, 288:1–8, 2002.

[92] C. Stenström, H. Jin, L. Major, W. Tate, and L. Isaksson. Codon bias at the $3'-$side of the initiation codon is correlated with translation initiation efficiency in *Escherichia coli*. *Gene*, 263:273–284, 2001.

[93] G. Stormo, T. Schneider, and L. Gold. Characterization of translational initiation sites in *E. coli*. *Nucleic Acids Res.*, 10(9):2971–2996, 1982.

[94] T. A. Thanaraj and M. W. Pandit. An additional ribosome-binding site on mRNA of highly expressed genes and a bifunctional site on the colicin fragment of 16S rRNA from *Escherichia coli*: important determinants of the of the efficiency of translation-initiation. *Nucleic Acids Res.*, 17(8):2973–2985, 1989.

[95] P. Tranque, M. Hu, G. Edelman, and V. Mauro. rRNA complementarity within mRNAs: a possible basis for mRNA-ribosome interactions and translational control. *Proc. Natl. Acad. Sci.*, 95:12238–12243, 1998.

[96] E. Trifonov. Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16S rRNA nucleotide sequences. *J. Mol. Biol.*, 194:643–652, 1987.

[97] E. Trifonov. Recognition of correct reading frame by the ribosome. *Biochimie*, 74:357–362, 1992.

[98] K. Ueda, A. Yamashita, J. Ishikawa, M. Shimada, T. Watsuji, K. Morimura, H. Ikeda, M. Hattori, and T. Beppu. Genome sequence of *Symbiobacterium thermophilum*, an uncultivable bacterium that depends on microbial commensalism. *Nucleic Acids Res.*, 32(16):4937–4944, 2004.

[99] S. Vagner, B. Galy, and S. Pyronnet. Irresistible IRES: attracting the translation machinery to internal entry sites. *EMBO R.*, 2(10):893–898, 2001.

[100] S-D. Verrier and O. Jean-Jean. Complementarity between the mRNA $5'$ untranslated region and 18S ribosomal RNA can inhibit translation. *RNA*, 6:584–597, 2000.

[101] R. Weiss, D. Dunn, A. Dahlberg, J. Atkins, and R. Gesteland. Reading frame switch caused by base-pair formation between the $3'$ end of the 16S rRNA and the mRNA during elongation of protein synthesis in *Escherichia coli*. *EMBO J.*, 7(5):1503–1507, 1988.

[102] E. Winzeler and L. Shapiro. Translation of the leaderless *Caulobacter dnaX* mRNA. *J. Bacteriol.*, 179(12):3981–3988, 1997.

[103] C. Wu and G. Janssen. Translation of *vph* in *Streptomyces lividans* and *Escherichia coli* after removal of the 5′ untranslated leader. *Mol. Microbiol.*, 22(2):339–355, 1996.

[104] T. Xia, J. SantaLucia, Jr., M. Burkard, R. Kierzek, S. Schroeder, X. Jiao, C. Cox, and D. Turner. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, 37:14719–14735, 1998.

[105] A. Yassin, K. Fredrick, and A. Mankin. Deleterious mutations in small subunit ribosomal rna identify functional sites and potential targets for antibiotics. *Proc. Natl. Acad. Sci.*, 102(46):16620–16625, 2005.

[106] G. Yusupova, M. Yusupov, J. Cate, and H. Noller. The path of messenger rna through the ribosome. *Cell*, 106:233–241, 2001.

[107] W. Zerges. Translation in chloroplasts. *Biochimie*, 82:583–601, 2000.

[108] M. Zuker. On finding suboptimal foldings of an RNA molecule. *Science*, 244(4900):48–52, 1989.

[109] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, 31(13):306–3415, 2003.

[110] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, 9(1):133–148, 1981.