Predicting Shine–Dalgarno Sequence Locations Exposes Genome Annotation Errors

J. Starmer[1*], A. Stomp[2], M. Vouk[3], D. Bitzer[3]

1 Bioinformatics Program, North Carolina State University, Raleigh, North Carolina, United States of America, **2** Department of Forestry, North Carolina State University, Raleigh, North Carolina, United States of America, **3** Department of Computer Science, North Carolina State University, Raleigh, North Carolina, United States of America

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 five to ten 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/mol (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.

Citation: Starmer J, Stomp A, Vouk M, Bitzer D (2006) Predicting Shine–Dalgarno sequence locations exposes genome annotation errors. PLoS Comput Biol 2(5): e57. DOI: 10.1371/journal.pcbi.0020057

Introduction

In 1974 Shine and Dalgarno [1] sequenced the 3′ end of *Escherichia 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 [2] and eventually experimentally verified, in 1987, by Hui and de Boer [3] and Jacob et al. [4].

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 [5]. Identification errors can arise from this approach for several reasons [6]. 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 [7–10].

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 [3,4]. 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

Editor: Vivian Hook, University of California San Diego, United States of America

Received December 2, 2005; **Accepted** April 10, 2006; **Published** May 19, 2006

A previous version of this article appeared as an Early Online Release on April 10, 2006 (DOI: 10.1371/journal.pcbi.0020057.eor).

DOI: 10.1371/journal.pcbi.0020057

Abbreviations: INN, individual nearest neighbor; INN-HB, individual nearest neighbor hydrogen bond; mRNA, messenger RNA; rRNA, ribosomal RNA; RS, relative spacing; SD, Shine–Dalgarno; TIR, translation initiation region

* To whom correspondence should be addressed. E-mail: jdstarme@ncsu.edu

Synopsis

More than 30 years ago researchers first discovered a sequence of messenger RNA (mRNA) nucleotides in bacteria that ribosomes recognize as a signal for where to begin protein synthesis. Today, genome annotation software takes advantage of this finding and uses it to help identify the location of start codons. Because these sequences, now referred to as Shine–Dalgarno (SD) sequences, are always upstream from start codons, annotation programs look for them in the region 5′ to these candidate sites. In a comprehensive analysis of 18 bacterial genomes, the authors show that when looking for SD sequences, it sometimes pays off to analyze unlikely locations. By examining the region that immediately surrounds the start codon for SD sequences, the authors identify many mis-annotated genes and in so doing offer a method to help check for these in future annotation projects.

lower free energy than the participating single-stranded sequences. One long-standing implementation of this model, Mfold [11], quantifies the degree of hybridization and the stability of RNA secondary structure by calculating the change in energy ($\Delta G°$) [12–14]. This method for estimating free energy has been adapted to identify SD sequences by repeatedly calculating the $\Delta G°$ values for progressive alignments of the rRNA tail with the mRNA in the region upstream of the start codon [5,6,15,16]. All of these studies have observed a trough of negative $\Delta G°$ 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 $\Delta G°$ 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 [17–22]. 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 [23], has been disproved [24]. 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 [22,23,25].

In this study we use the RS metric to identify the positions of minimal $\Delta G°$ 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 2,420 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.

Results

The average $\Delta G°$ value at each position of the TIR for each species is shown in Figure 1, aligned according to RS. The $\Delta G°$ troughs upstream from RS 0 are consistent with previous experimental studies on the location of the SD sequence [7,8], as well as with computational studies either simulating free energy changes [15,26] or using information theory [27]. The $\Delta G°$ 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 show the distributions of RS positions of the strongest SD-like sequences (where $\Delta G° < -3.4535$, see the Materials and 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 $\Delta G°$ value is at RS+1, *+1 genes,* and +1 genes where $\Delta G° < -8.4$ 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 with as low a $\Delta G°$ value to the rRNA. 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 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 individual nearest neighbor–hydrogen bond (INN-HB) model or its implementation. Both the individual nearest neighbor (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 [11,14,28,] and SD sequence identification [6,9,15,16,26,29]. The more recent INN-HB model improves secondary structure predictions in newer versions of Mfold [14]. 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 [30]. With the exception of the +1 site, the results that our implementation of the INN-HB model generate are consistent with both experimental [7,8] and computational studies [15,31–33] of SD and coding sequences. Furthermore, analysis

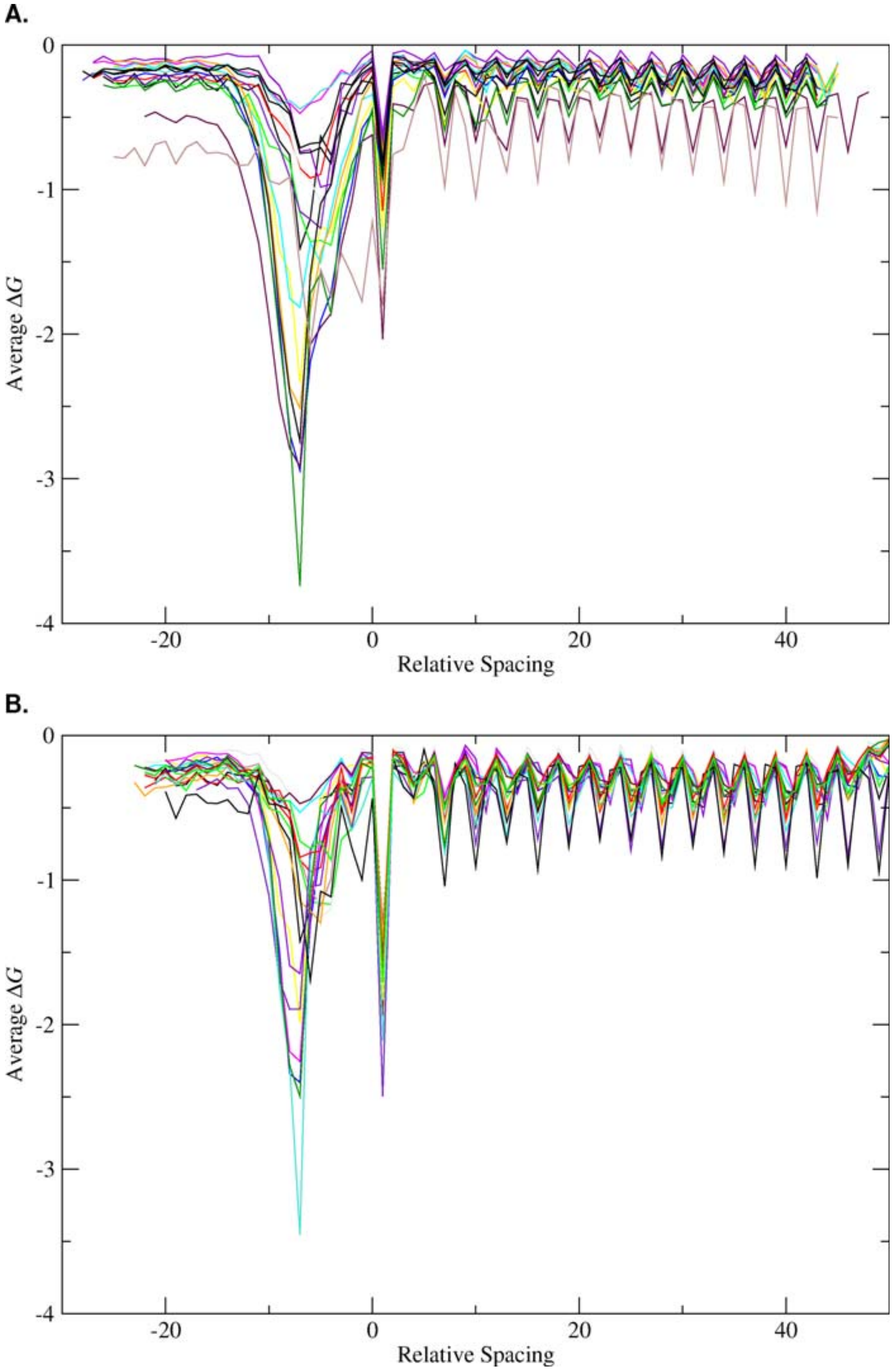


Figure 1. Average ΔG ° Values in the TIRs for 18 Organisms

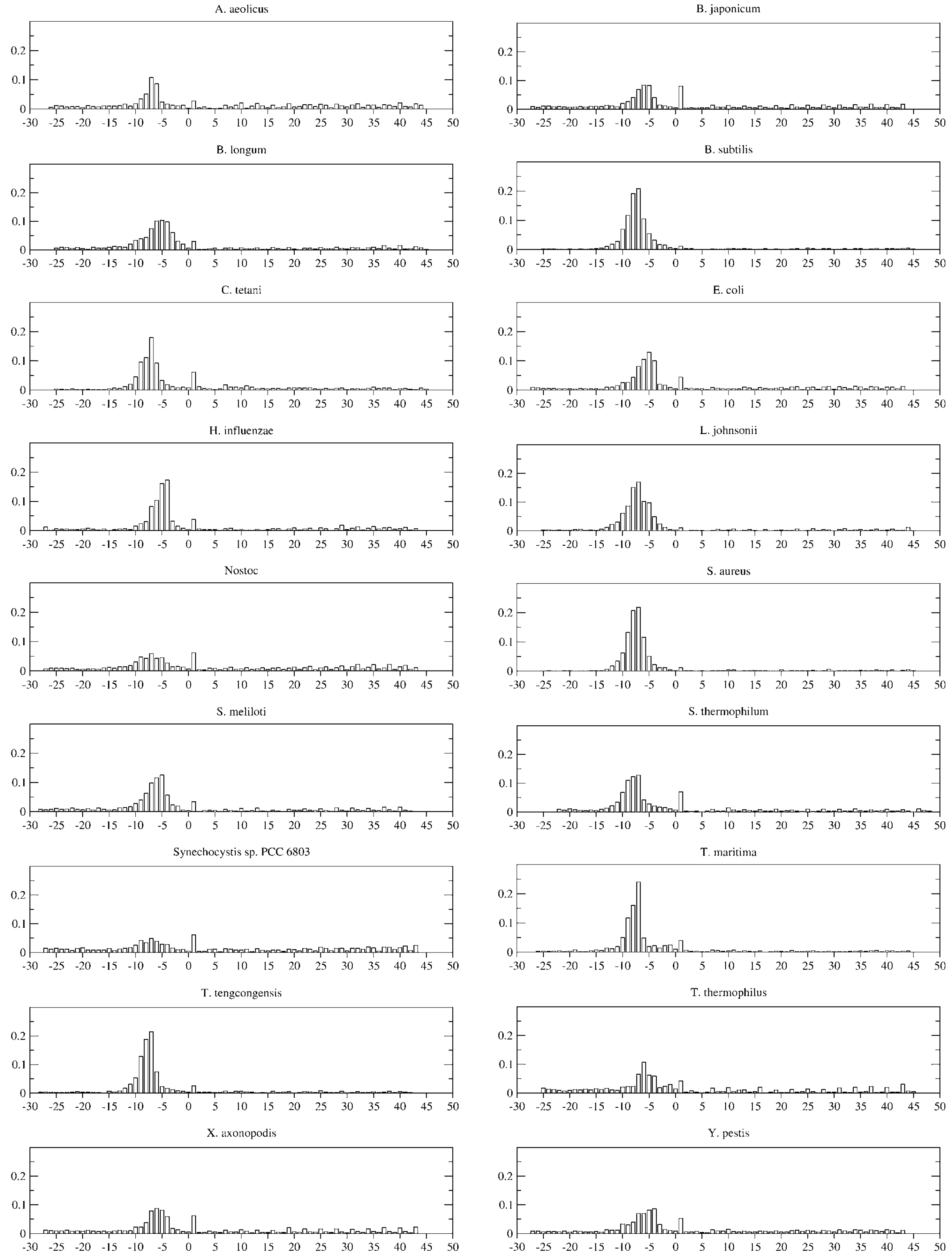
For all 18 genomes in our study, we calculated the average ΔG ° value for each RS position. Zero on the *x*-axis corresponds to the 5′ A residue in the rRNA sequence 5′–ACCUCC–3′ being positioned over the first base in the initiation codon. The dramatic drops in ΔG ° prior to RS 0 show the presence of 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 [34]. Differences between the two graphs are discussed in the text.

DOI: 10.1371/journal.pcbi.0020057.g001

Figure 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.

DOI: 10.1371/journal.pcbi.0020057.g002


A. aeolicus
B. japonicum
B. longum
B. subtilis
C. tetani
E. coli
H. influenzae
L. johnsonii
Nostoc
S. aureus
S. meliloti
S. thermophilum
Synechocystis sp. PCC 6803
T. maritima
T. tengcongensis
T. thermophilus
X. axonopodis
Y. pestis
0
0.1
0.2
-30
-25
-20
-15
-10
-5
0
5
10
15
20
25
30
35
40
45
50

performed with RNAhybrid [34] is consistent with our results (see Figure 1). Based on this evidence, it is clear that the +1 site is not an artifact of the model we are using or of 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 1. Studies of TIR sequences in *E. coli* have shown considerable bias in the second codon, too [35–37]. To examine this bias, sequence logos [38,39] (http://weblogo.berkeley.edu/) were created for the region of mRNA that would be aligned with the rRNA tail for RS+1 (see Figure 3, *radC,* for an example of this alignment). Figure 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 (unpublished data). 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/mol 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 4,254 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 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). 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 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 3, only a small percentage of the TIRs of upstream genes have in-frame start codons downstream from the annotated start 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 [40]. These disagreements in annotation are probably the result of Blattner et al. selecting the start codon that will allow the open reading frame (ORF) to be extended as far upstream as possible [41]. *E. coli*'s *radC* gene provides a useful example: assuming the GenBank annotation to be correct, the RS metric identifies *radC* as a strong +1 gene. However, as can be seen in Figure 3, the initiation region sequence has an in-frame GUG six bases downstream from the annotated start codon. If the down-

Table 1. Usage Statistics for the Three Most Common Initiation Codons: AUG, GUG, and UUG

Organism	Start Codon Usage—Upstream Genes			Start Codon Usage−Strong +1 Genes		
	AUG	GUG	UUG	AUG	GUG	UUG
A. aeolicus	84% (554)	9% (57)	8% (50)	5% (1)	95% (19)	
B. japonicum	83% (2,965)	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% (2,826)	8% (306)	13% (454)	8% (1)	83% (10)	8% (1)
C. tetani	80% (1,216)	8% (115)	12% (185)	6% (5)	92% (73)	1% (1)
E. coli	90% (2,041)	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% (1,446)	12% (214)	6% (104)	4% (1)	96% (23)	
Y. pestis	81% (1,514)	11% (213)	8% (143)		96% (26)	4% (1)

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.

DOI: 10.1371/journal.pcbi.0020057.t001

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 3. Examples from *E. coli* Showing How RS 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 RS 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 RS is given as a negative number. If the SD is downstream, it is given as a positive number. Both SD sequences for *wecF* and *argD* come before the start codons (in these cases, the start codon is AUG). The RS for *wecF* is –4 and for *argD* it is –10. *radC*'s SD sequence includes the start codon, GUG, and the RS is +1.

DOI: 10.1371/journal.pcbi.0020057.g003

stream 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 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 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 toward 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 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 $\Delta G°$ values were uniformly distributed (unpublished data). 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 $\Delta G°$ 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 4.

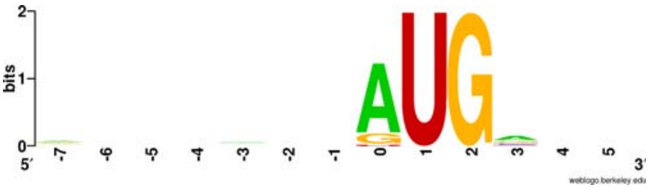


Figure 4. A Sequence Logo for *E. coli*

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.

DOI: 10.1371/journal.pcbi.0020057.g004

Table 2. A Summary of the Annotation Programs Used for the Genomes in This Study

Organism	Annotation Tool	Year Published
A. aeolicus [55]	Comparative analysis only	1998
B. japonicum [56]	GLIMMER	2002
B. longum [57]	ORPHEUS	2002
B. subtilis [58]	GeneMark	1997
C. tetani [59]	GLIMMER	2003
E. coli [41]	comparative analysis only	1997
H. influenzae [60]	GeneMark[a]	1995
L. johnsonii [61]	FrameD	2004
Nostoc [62]	GLIMMER	2001
S. aureus [63]	GLIMMER and ORPHEUS	2004
S. meliloti [64]	FrameD	2001
S. thermophilum [65]	GLIMMER and GeneLook	2004
Synechocystis [66]	GeneMark	1996
T. maritima [67]	GLIMMER	1999
T. tengcongensis [68]	GLIMMER	2002
T. thermophilus [69]	GeneMarkS	2004
X. axonopodis [70]	GLIMMER and GeneMark	2002
Y. pestis [71]	GLIMMER	2002

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.

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

DOI: 10.1371/journal.pcbi.0020057.t002

Table 3. Downstream Start Codons

Organism	Downstream Start Codons		Adjusted RS			
	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

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 RS column 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.
DOI: 10.1371/journal.pcbi.0020057.t003

Table 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.

Discussion

There is a long history of investigating SD sequences using approaches grounded in thermodynamics [5,6,9,15,16,26]. 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 that, 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 RS 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 in [15]. 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 [9]. Sakai et al. base their anti-SD motifs on the most frequent 7-mer found within 40 bases upstream

Table 4. Binding at the Start Codon for Strong +1 Genes Compared with Upstream Binding

Organism	N = Strong +1 Genes	$\overline{\Delta G}$ °	
		−10 to −4 RS	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

To determine the differences in $\Delta G°$ 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, $\overline{\Delta G}$ °. The number of genes used to calculate each average, N, the number of strong +1 genes, is listed in the second column.
DOI: 10.1371/journal.pcbi.0020057.t004

Table 5. A Summary of Predicted rRNA–mRNA Binding

Organism	Genes	US Genes	DS Genes	+1 Genes	Strong +1 Genes	Unexplained Strong +1 Genes[a]
A. aeolicus	1,529	661	267	38	20	6
B. japonicum	8,317	3,655	1,573	579	111	55
B. longum	1,727	1,085	174	46	3	2
B. subtilis	4,106	3,600	184	45	12	6
C. tetani	2,373	1,516	461	141	79	6
E. coli	4,254	2,272	554	163	28	5
H. influenzae	1,656	960	115	32	2	1
L. johnsonii	1,821	1,447	89	18	4	2
Nostoc	5,366	1,667	808	232	41	18
S. aureus	2,739	2,405	117	30	9	4
S. meliloti	3,332	2,030	340	103	8	7
S. thermophilum	3,337	2,226	543	229	124	60
Synechocystis	3,167	871	475	135	15	7
T. maritima	1,858	1,464	190	74	52	8
T. tengcongensis	2,588	2,029	234	64	24	3
T. thermophilus	1,982	1,059	340	82	41	23
X. axonopodis	4,312	1,764	624	196	24	15
Y. pestis	4,086	1,870	654	182	27	14

US (upstream) 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.
DS (downstream) 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.
[a]These unexplained genes could be non-expressing ORFs, as discussed in the text.
DOI: 10.1371/journal.pcbi.0020057.t005

of the start codon on the mRNA sequences [26], without reference to rRNA sequences.

As a result of these differences, our method improves SD sequence characterization. Table 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

Table 6. Model Comparisons

Organism	12-mer rRNA Tails		Full Length rRNA Tails
	SD% with INN [9]	SD% with INN-HB	SD% with INN-HB
A. aeolicus	48.1% of 1,487	58.6% of 1,489[a]	59.2% of 1,489[a]
B. subtilis	89.4% of 3,624	94.3% of 3,629[a]	95.9% of 3,629[a]
E. coli	57.1% of 3,908	66.9% of 3,882[a]	68.1% of 3,882[a]
H. influenzae	53.7% of 1,533	65.5% of 1,527[a]	65.9% of 1,527[a]
Synechocystis	26.0% of 2,906	37.7% of 2,912[a]	39.3% of 2,912[a]
T. maritima	90.1% of 1,685	91.6% of 1,696[a]	92.7% of 1,696[a]
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

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 [9] (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.
[a]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 dataset sizes. The RefSeq versions for the genome files are the same, but the source of these discrepancies is unknown.
NA, Not available.
DOI: 10.1371/journal.pcbi.0020057.t006

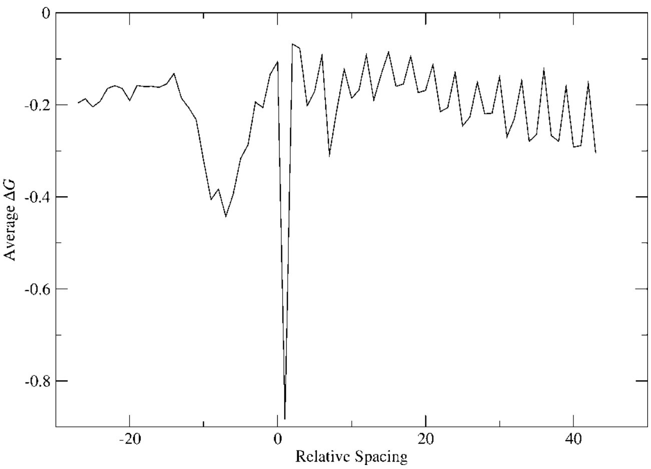


Figure 5. Average ΔG ° Values in the TIR for *Synechocystis*
The trough prior to RS 0 clearly shows the presence of an SD motif in many genes.
DOI: 10.1371/journal.pcbi.0020057.g005

upstream ΔG ° trough indicative of SD sequences in *Synechocystis* [26]. Our method reveals the SD trough (see Figure 5 and Table 6). Comparison with Schurr et al.'s results [6] shows benefits to using the INN-HB model in conjunction with RS and examining downstream sequences. Of the 38 genes they identified as having $\Delta G ° \geq 0$ kcal/mol, and thus no discernible binding site for the rRNA tail, we were able to 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 farther 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 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 ORFs.

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,* a ribonuclease P protein component, does not contain an in-frame start codon downstream from the annotated start site. *CTC02285,* a strong +1 gene in *Clostridium tetani* that codes for protein translation initiation factor 3 (IF3), is also without a downstream initiation codon. *Bradyrhizobium japonicum* has many strong +1 genes without downstream start codons: *polE,* which codes for the polymerase epsilon subunit, *cycK, nah,* and 52 others. Thus, while a large percentage of the strong +1 genes appears 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 standby site, i.e., sequences that hold the 16S rRNA close to the SD sequence [42]. 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 [18], in contrast to the strong +1 genes that favor GUG (see Table 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 5 and 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 [43] 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 [44], 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 disjointed SD sequence brought together by secondary structures is likely to function for the *E. coli* gene *rpsA* [44]. 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 which increased the efficiency of translation initiation beyond that of a consensus SD sequence [45].

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.

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 RS 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.

Table 7. A Summary of the Data and Its Sources Used in This Study

Organism	RefSeq Version	Genes	Secondary Structure	16S rRNA 3′ Tail (5′ to 3′)
A. aeolicus	NC_000918.1 GI:15282445	1,529	d.16.b.A.aeolicus	gaucAccuccuuua
B. japonicum	NC_004463.1 GI:27375111	8,317	d.16.b.B.japonicum	gaucAccuccuuu
B. longum	NC_004307.2 GI:58036264	1,727	NA	gaucAccuccuuucu
B. subtilis	NC_000964.2 GI:50812173	4,106	d.16.b.B.subtilis	gaucAccuccuuucu
C. tetani	NC_004557.1 GI:28209834	2,373	d.16.b.C.tetani[a]	gaucAccuccuuucu
E. coli	NC_000913.2 GI:49175990	4,254	d.16.b.E.coli.K12	gaucAccuccuua
H. influenzae	NC_000907.1 GI:16271976	1,656	d.16.b.H.influenzae[a]	gaucAccuccuua
L. johnsonii	NC_005362.1 GI:42518084	1,821	NA	gaucAccuccuuucu
Nostoc	NC_003272.1 GI:17227497	5,366	NA	gaucAccuccuuu
S. aureus	NC_002952.2 GI:49482253	2,739	d.16.b.S.aureus	gaucAccuccuuucu
S. meliloti	NC_003047.1 GI:15963753	3,332	NA	gaucAccuccuu
S. thermophilum	NC_006177.1 GI:51891138	3,337	NA	gaucAccuccuuucuaag
Synechocystis	NC_000911.1 GI:16329170	3,167	d.16.b.Synechocystis	gaucAccuccuuu
T. maritima	NC_000853.1 GI:15642775	1,858	d.16.b.T.maritima	gaucAccuccuuuc
T. tengcongensis	NC_003869.1 GI:20806542	2,588	NA	gaucAccuccuu
T. thermophilus	NC_005835.1 GI:46198308	1,982	d.16.b.T.thermophilus.2[a]	gaucAccuccuuucu
X. axonopodis	NC_003919.1 GI:21240774	4,312	NA	gaucAccuccuuu
Y. pestis	NC_004088.1 GI:22123922	4,086	d.16.b.Y.pestis[a]	gaucAccuccuua

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 RS.

[a] The structure was not used to define the 3′ tail due to either the presence of the wild-card character, 'N', or the lack of sequence altogether.

DOI: 10.1371/journal.pcbi.0020057.t007

Materials and Methods

Genome sequences. All genome sequences were downloaded from the National Center for Biotechnology Information (NCBI) GenBank database (http://www.ncbi.nlm.nih.gov). Table 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.

Selecting criteria for gene sequences. For all genomes, all gene sequences with gene= or locus_tag= tags were included in our dataset, except those that also included a transposon= or pseudo tag.

We defined the 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 [23] was rejected [24], evidence from leaderless mRNAs suggests that sequences downstream and within 20 nucleotides of the start codon are involved [22,23,25]. Other studies that have analyzed initiation regions of mRNA sequences for negative free energy troughs [6,9,15,16] have not examined bases downstream of the annotated start codon: downstream sequence analysis allowed for start codon annotation error detection.

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 [46] (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′–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 7.

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 to 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 [6,9,15,16,47–49], as well as in the genome annotation program GLIMMER [29].

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 INN-HB model [13], extended to allow for symmetrical internal loops (loops that contain an equal number of bases in both RNA strands):

$$\Delta G^\circ = \Delta G_{\text{init}}{}^\circ + \sum_j n_j \Delta G^\circ(\text{NN}) + m_{\text{term-AU}} \Delta G_{\text{term-AU}}{}^\circ + \Delta G_{\text{sym}}{}^\circ + \sum_k \text{Loop}_k \qquad (1)$$

In this formula, $\Delta G_{\text{init}}{}^\circ$ is the amount of free energy required to initiate a helix between the two strands of RNA; ΔG° (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. $m_{\text{term-AU}}$ is the number of terminal AU pairs, and $\Delta G_{\text{term-AU}}{}^\circ$ is the free energy penalty for having a terminal AU pair. Finally, $\Delta G_{\text{sym}}{}^\circ$ 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. [13], G/U mismatches from Mathews et al. [14], and loop penalties from Jaeger et al. [50]. 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 as the limitations they put on our ability to calculate RS. Dangling 5′ or 3′ ends are not considered

```
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 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 for the binding to occur spontaneously require a change more negative than –4.08 kcal/mol [13], the value for $\Delta G_{\text{init}}^\circ$, In this example, the initiation region from *E. coli's* gene *hcaF*, 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.

DOI: 10.1371/journal.pcbi.0020057.g006

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 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 [13] to improve the ΔG° estimates obtained using the prior INN models [28,51–54]. 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 [13].

To verify the accuracy of free_scan, we ran our analysis again using RNAhybrid [34] and plotted the average ΔG° value for each RS position (Figure 1). RNAhybrid uses free energy parameters from Xia et al. [13] and Mathews et al. [14], but does not include $\Delta G_{\text{init}}^\circ$ or $m_{\text{term-AU}}\Delta G_{\text{term-AU}}^\circ$. We set the energy cutoff to –4.075225 kcal/mol and subtracted this value from RNAhybrid's output 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 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 preceding 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 [9]. The threshold ΔG° that Ma et al. used 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:

$$\begin{array}{ll}\texttt{mRNA:5'-GGAG-3'} & \texttt{mRNA:5'-GAGG-3'}\\ \texttt{rRNA:3'-CCUC-5''} & \texttt{rRNA:3'-CUCC-5'}\end{array}$$

$$\text{and}\quad \begin{array}{l}\texttt{mRNA:5'-AGGA-3'}\\ \texttt{rRNA:3'-UCCU-5''}\end{array} \tag{2}$$

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 one GG/CC doublet and two AG/UC doublets. With the updated parameters for both the doublets as well as the helix initiation penalty, 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 ($\Delta G_{\text{init}} = 3.4^\circ$ kcal/mol for the INN and $\Delta G_{\text{init}} = 4.08^\circ$ kcal/mol for the INN-HB). Table 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.

Locating the SD sequence and determining SD RS. 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. [9] (see above).

The SD's 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 [7], 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 3 for RS examples taken from *E. coli.*

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′–GGAGGU–3′.

Supporting Information

Accession Numbers

Accession numbers from the National Center for Biotechnology Information (NCBI) GenBank database (http://www.ncbi.nlm.nih.gov) for genes mentioned in this paper are: *radC* (948968); *rnpA* (1023245); *CTC02285* (1060453); *polE* (1051409); *cycK* (1053038); *nah* (1053188); *rpsA* (945536); *wecF* (2847677); *argD* (947864); and *hcaF* (946997).

Acknowledgments

The authors would like to thank Eric Miller and Errol Strain for critical feedback and advice, and Bibiana Obler for editing.

Author contributions. JS, AS, MV, and DB conceived and designed the experiments. JS performed the experiments. JS, AS, and DB analyzed the data. JS and AS wrote the paper.

Funding. The authors received no specific funding for this study.

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

References

1. Shine J, Dalgarno L (1974) The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: Complementarity to nonsense triplets and ribosome binding sites. Proc Natl Acad Sci U S A 71: 1342–1346.
2. Steitz J, Jakes K (1975) 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 U S A 72: 4734–4738.
3. Hui A, de Boer H (1987) Specialized ribosome system: Preferential translation of a single mRNA species by a subpopulation of mutated ribosomes in *Escherichia coli*. Proc Natl Acad Sci U S A 84: 4762–4766.
4. Jacob W, Santer M, Dahlberg A (1987) A single base change in the Shine–Dalgarno region of 16S rRNA of *Escherichia coli* affects translation of many proteins. Proc Natl Acad Sci U S A 84: 4757–4761.
5. Stormo G, Schneider T, Gold L (1982) Characterization of translational initiation sites in *E. coli*. Nucleic Acids Res 10: 2971–2996.
6. Schurr T, Nadir E, Margalit H (1993) Identification and characterization of *E. coli* ribosomal binding sites by free energy computation. Nucleic Acids Res 21: 4019–4023.
7. Chen H, Bjerknes M, Kumar R, Ernest J (1994) Determination of the optimal aligned spacing between the Shine–Dalgarno sequence and the translation initiation codon in *Escherichia coli* mRNAs. Nucleic Acids Res 22: 4953–4957.
8. Ringquist S, Shinedling S, Barrick D, Green L, Binkley J, et al. (1992) Translation initiation in *Escherichia coli*: Sequences within the ribosome-binding site. Mol Microbiol 6: 1219–1229.
9. Ma J, Campbell A, Karlin S (2002) Correlations between Shine–Dalgarno sequences and gene features such as predicted expression levels and operon structures. J Bacteriol 184: 5733–5745.
10. Kozak M (1999) Initiation of translation in prokaryotes and eukaryotes. Gene 234: 187–208.
11. Zuker M, Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 9: 133–148.
12. SantaLucia J (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc Natl Acad Sci 95: 1460–1465.
13. Xia T, SantaLucia J Jr, Burkard M, Kierzek R, Schroeder S, et al. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick base pairs. Biochemistry 37: 14719–14735.
14. Mathews D, Sabina J, Zuker M, Turner D (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 288: 911–940.
15. Osada Y, Saito R, Tomita M (1999) Analysis of base-pairing potentials between 16S rRNA and 5′ UTR for translation initiation in various prokaryotes. Bioinformatics 15: 578–581.
16. Lithwick G, Margalit H (2003) Hierarchy of sequence-dependent features associated with prokaryotic translation. Genome Res 13: 2665–2673.
17. Wu C, Janssen G (1996) Translation of *vph* in *Streptomyces lividans* and *Escherichia coli* after removal of the 5′ untranslated leader. Mol Microbiol 22: 339–355.
18. Etten WV, Janssen G (1998) An AUG initiation codon, not codon–anticodon complementarity, is required for the translation of unleadered mRNA in *Escherichia coli*. Mol Microbiol 27: 987–1001.
19. Martin-Farmer J, Janssen G (1999) A downstream CA repeat sequence increases translation from leadered and unleadered mRNA in *Escherichia coli*. Mol Microbiol 31: 1025–1038.
20. Moll I, Grill S, Gualerzi C, Bläsi U (2002) Leaderless mRNAs in bacteria: Surprises in ribosomal recruitment and translational control. Mol Microbiol 43: 239–246.
21. O'Donnell S, Janssen G (2001) 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: 1277–1283.
22. Winzeler E, Shapiro L (1997) Translation of the leaderless *Caulobacter dnaX* mRNA. J Bacteriol 179: 3981–3988.
23. Sprengart M, Fatscher H, Fuchs E (1990) The initiation of translation in *E. coli*: Apparent base pairing between the 16srRNA and downstream sequences of the mRNA. Nucleic Acids Res 18: 1719–1723.
24. O'Connor M, Asai T, Squires C, Dahlberg A (1999) 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 U S A 96: 8973–8978.
25. Faxén M, Plumbridge J, Isaksson L (1991) 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: 5247–5251.
26. Sakai H, Imamura C, Osada Y, Saito R, Washio T, et al. (2001) Correlation between Shine–Dalgarno sequence conservation and codon usage of bacterial genes. J Mol Evol 52: 164–170.
27. Shultzaberger R, Bucheimer R, Rudd K, Schneider T (2001) Anatomy of *Escherichia coli* ribosome binding sites. J Mol Biol 313: 215–228.
28. Freier S, Kierzek R, Jaeger J, Sugimoto N, Caruthers M, et al. (1986) Improved free-energy parameters for predictions of RNA duplex stability. Proc Natl Acad Sci U S A 83: 9373–9377.
29. Delcher A, Harmon D, Kasif S, White O, Salzberg S (1999) Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27: 4636–4641.
30. Hodas N, Aalberts D (2004) Efficient computation of optimal oligo-RNA binding. Nucleic Acids Res 32: 636–6642.
31. Trifonov E (1987) 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.
32. Trifonov E (1992) Recognition of correct reading frame by the ribosome. Biochimie 74: 357–362.
33. Lió P, Ruffo S, Buiatti M (1994) Third codon g+c periodicity as a possible signal for an internal selective constraint. J Theor Biol 171: 215–223.
34. Rehmsmeier M, Steffen P, Höchsmasnn M, Giegerich R (2004) Fast and effective prediction of microRNA/target duplexes. RNA 10: 1507–1517.
35. Stenström C, Jin H, Major L, Tate W, Isaksson L (2001) Codon bias at the 3′– side of the initiation codon is correlated with translation initiation efficiency in *Escherichia coli*. Gene 263: 273–284.
36. Stenström C, Holmgren E, Isaksson L (2001) Cooperative effects by the initiation codon and its flanking regions on translation initiation. Gene 273:: 259–265.
37. Stenström C, Isaksson L (2002) Influences on translation and early elongation by the messenger RNA region flanking the initiation codon at the 3′ side. Gene 288: 1–8.
38. Schneider T, Stephens R (1990) Sequence logos: A new way to display consensus. Nucleic Acids Res 18: 6097–6100.
39. Crooks G, Hon G, Chandonia J, Brenner S (2004) Weblogo: A sequence logo generator. Genome Res 14: 1188–1190.
40. Rudd K (2000) EcoGene: A genome sequence database for *Escherichia coli* K-12. Nucleic Acids Res 28: 60–64.
41. Blattner F, Plunket GP III, Bloch C, Perna N, Burland V, et al. (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277: 1453–1462.
42. de Smit M, van Duin J (2003) Translational standby sites: How ribosomes may deal with the rapid folding kinetics of mRNA. J Mol Biol 331: 737–743.
43. Zerges W (2000) Translation in chloroplasts. Biochimie 82: 583–601.
44. Boni I, Artamonova V, Tzareva N, Dreyfus M (2001) Non-canonical mechanism for translational control in bacteria: Synthesis of ribosomal protein S1. EMBO J 20: 4222–4232.
45. Kolev V, Ivanov I, Berzai-Herranz A, Ivanov I (2003) Non-Shine–Dalgarno initiators of translation selected from combinatorial DNA libraries. J Mol Microbiol Biotechnol 5: 154–160.
46. Cannone J, Subramanian S, Schnare M, Collett J, D'Souza L, et al. (2002) The comparative RNA web (CRW) site: An online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics 3: 2.
47. Thanaraj TA, Pandit MW (1989) 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 efficiency of translation–initiation. Nucleic Acids Res 17: 2973–2985.
48. Lee K, Holland-Staley C, Cunningham P (1996) Genetic analysis of Shine–Dalgarno interaction: Selection of alternative functional mRNA-rRNA combinations. RNA 2: 1270–1285.
49. Komarova A, Tchufitsova L, Supina E, Boni I (2001) Extensive complementarity of the Shine–Dalgarno region and the 3′-end of 16S rRNA is inefficient for translation in vivo. Bioorg Khim 27: 248–255.
50. Jaeger J, Turner D, Zuker M (1989) Improved predictions of secondary structures for RNA. Proc Natl Acad Sci U S A 86: 7706–7710.
51. Gray D, Tinoco I (1970) A new approach to the study of sequence-dependent properties of polynucelotides. Biopolymers 9: 223–244.
52. Gray D (1997) 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: 783–793.
53. Gray D (1997) Derivation of nearest-neighbor properties from data on nucleic acid oligomers. II. Thermodynamic parameters of DNA-RNA hybrids and DNA duplexes. Biopolymers 42: 795–810.
54. Borer P, Dengler B, Tinoco I, Uhlenbeck O (1974) Stability of ribonucleic acid double-stranded helices. J Mol Biol 86: 843–853.
55. Deckert G, Warren PV, Gaasterland T, Young WG, Lenox AL, et al. (1998) The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. Nature 392: 353–358.
56. Kaneko T, Nakamura Y, Sato S, Minamisawa K, Uchiumi T, et al. (2002) Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110. DNA Res 9: 189–197.
57. Schell MA, Karmirantzou M, Snel B, Vilanova D, Berger B, et al. (2002) The genome sequence of *Bifidobacterium longum* reflects its adaptation to the human gastrointestinal tract. Proc Natl Acad Sci U S A 99: 14422–14427.
58. Kunst F, Ogasawara K, Moszer I, Albertini AM, Alloni G, et al. (1997) The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. Nature 390: 249–256.
59. Brüggemann H, Bäumer S, Frike W, Wiezer A, Liesegang H, et al. (2003) The genome sequence of *Clostridium tetani*, the causative agent of tetanus disease. Proc Natl Acad Sci U S A 100: 1316–1321.
60. Fleischmann R, Adams MD, White O, Clayton RA, Kirkness EF, et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269: 496–512.
61. Primore R, Berger B, Desiere F, Vilanova D, Barretto C, et al. (2004) The

genome sequence of the probiotic intestinal bacterium *Lactobacillus johnsonii* NCC 533. Proc Natl Acad Sci U S A 101: 2512–2517.

62. Kaneko T, Nakamura Y, Wolk C, Kuritz T, Sasamoto S, et al. (2001) Complete genomic sequence of filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. DNA Res 8: 205–213.
63. Holden M, Feil E, Lindsay J, Peacock S, Day NP, et al. (2004) Complete genomes of two clinical *Staphylococus aureus* strains: Evidence for the rapid evolution of virulence and drug resistance. Proc Natl Acad Sci U S A 101: 9786–9791.
64. Capela D, Barloy Hubler F, Gouzy J, Bothe G, Ampe F, et al. (2001) Analysis of the chromosome sequence of the legume symbiont *Sinorhizobium meliloti* strain 1021. Proc Natl Acad Sci U S A 98:: 9877–9883.
65. Ueda K, Yamashita A, Ishikawa J, Shimada M, Watsuji T, et al. (2004) Genome sequence of *Symbiobacterium thermophilum*, an uncultivable bacterium that depends on microbial commensalism. Nucleic Acids Res 32: 4937–4944.
66. Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu A, et al. (1996) 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.
67. Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, et al. (1999) Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature 399:: 323–329.
68. Bao Q, Tian Y, Li W, Xu Z, Xuan Z, et al. (2002) A complete sequence of the *T. tengcongensis* genome. Genome Res 12: 689–700.
69. Henne A, Bruggemann H, Raasch C, Wiezer A, Hartsch T, et al. (2004) The genome sequence of the extreme thermophile *Thermus thermophilus*. Nat Biotechnol 22: 547–553.
70. da Silva AC, Ferro JA, Reinach FC, Farah CS, Furlan LR, et al. (2002) Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. *Nature* 417: 459–463.
71. Deng W, Burland V, Plunkett G III, Boutin A, Mayhew GF, et al. (2002) Genome sequence of *Yersinia pestis* KIM. J Bacteriol 184: 4601–4611.