

# Anatomy of *Escherichia coli* Ribosome Binding Sites

Ryan K. Shultzaberger<sup>\*†</sup>, R. Elaine Bucheimer<sup>‡</sup>, Kenneth E. Rudd<sup>§</sup>  
and Thomas D. Schneider<sup>† ¶</sup>

version = 3.21 of rbs.tex 2001 Jan 22

During translational initiation in prokaryotes the 3' end of the 16S rRNA binds to a region just upstream of the initiation codon. The relationship between this 'Shine-Dalgarno' (SD) region and the binding of ribosomes to translation start points has been well studied, but a unified mathematical connection between the SD, the initiation codon and the spacing between them has been lacking. Using information theory we constructed a model that treats these three components uniformly by assigning to the SD and the initiation region (IR) a conservation in bits of information, and by assigning to the spacing an uncertainty, also in bits. To build the model we first aligned the SD region by maximizing the information content there. The ease of this process confirmed the existence of the SD pattern within a set of 4122 reviewed and revised *Escherichia coli* gene starts. This large data set allowed us to graphically show by sequence logos that the spacing between the SD and the initiation region affects both the SD site conservation and its pattern. We used the aligned SD, the spacing, and the initiation region to model ribosome binding and to identify gene starts that do not conform to the ribosome binding site model. 569 experimentally proven starts are more conserved (have higher information content) than the full set of revised starts, which probably reflects an experimental bias against the detection of gene products that have inefficient ribosome binding sites. Models were cyclically refined by removing nonconforming weak sites. After this procedure, models derived from either the original or the revised gene start annotation were similar. Therefore, this information theory based technique provides a method for easily constructing biologically sensible ribosome binding site models. Such models should be useful for refining gene start predictions of any sequenced bacterial genome.

---

<sup>\*</sup>University of Maryland, College Park, Maryland, 20742

<sup>†</sup>National Cancer Institute at Frederick, Laboratory of Experimental and Computational Biology

<sup>‡</sup>Univ. of Virginia School of Medicine, Charlottesville, VA 22908

<sup>§</sup>Department of Biochemistry and Molecular Biology (R-629), University of Miami School of Medicine, P. O. Box 016129, Miami, FL 33101-6129

<sup>¶</sup>*Corresponding author*, P. O. Box B, Frederick, MD 21702-1201. (301) 846-5581 (-5532 for messages), fax: (301) 846-5598, email: toms@ncifcrf.gov, <http://www.lecb.ncifcrf.gov/~toms/>

Keywords: “Ribosome”, “Shine-Dalgarno”, “information theory”, “sequence logo”, “sequence walker”.

1

## Introduction

Ribosomes play a central role in cells by reading mRNA to synthesize proteins<sup>1</sup>. The entire high resolution atomic structure of the 50S<sup>2</sup> and 30S<sup>3</sup> ribosomal subunits have recently been determined, but a full understanding of translation will also require quantitative mathematical descriptions. Because codons are three bases long, translational initiation must be directed to within one base on the mRNA. This requires a pattern in the mRNA known as a ribosome binding site, which includes the initiation codon. The completion of entire genome sequences and the identification of likely genes within them now allows inspection of most ribosome binding sites and allows statistics of the patterns to be determined in greater detail than was previously possible<sup>4, 5, 6, 7</sup>.

In eukaryotes, ribosomes recognize the 7-methyl guanine cap to help identify the translation initiation codon<sup>8</sup>. Prokaryotes, however, lack this marker and instead have a contact between the 3' end of the 16S rRNA in the 30S ribosomal subunit and a region upstream of the initiation codon, referred to as the Shine-Dalgarno region (SD)<sup>9, 10</sup>. In addition, the initiation codon and the region downstream are protected in toeprint experiments<sup>11</sup>. Since there may be specific contacts besides the initiation codon, this entire region, excluding the SD, will be referred to as the initiation region (IR).

The Shine-Dalgarno has strong effects on translation<sup>10, 12, 13</sup>, and one of its most intriguing features is the variable spacing between it and the initiation region. Preferential binding of the 16S rRNA at certain spacings has been shown<sup>14, 15, 16, 17, 11</sup>. We investigated how this spacing affects the sequence conservation of the SD and IR and the patterns being bound for the majority of ribosome binding sites in *Escherichia coli*.

Nucleic acid and protein sequences can be described by information theory, an approach that was originally applied to quantify the movement of data in communication systems<sup>18, 19</sup>. In communication, information measured in bits defines the minimum number of choices needed to represent some data. The advantage of the measure, over all other measures, is that information from independent sources can be added together, and bits provide a universal scale. In molecular biology, the amount of information indicates the degree of sequence conservation among a set of aligned sequences. It is a quantitative measure that has proven to be more useful than consensus sequences for understanding a variety of genetic systems<sup>6, 20, 21, 22, 23</sup>. The average information computed from a set of related sequences<sup>6</sup> describes the overall conservation at each position in the alignment and this can be shown with a sequence logo graphic<sup>24</sup>. The individual information present in a single sequence<sup>25</sup> measures how representative that sequence is in comparison with the other members of the sequence family and these can be shown with sequence walker graphics<sup>26</sup>.

Although the spacing between the SD contact and the IR is variable, the conventional

---

<sup>1</sup>Abbreviations used: SD, Shine-Dalgarno; IR, Initiation Region; GS, Gap Surprisal

‘rigid’ model functioned reasonably well<sup>6,7</sup>, but as more sites were added to the model, the information content of the SD region dropped, suggesting that the model was not sufficient to explain the variation among sites. We therefore investigated a ribosome binding model where the spacing between the SD and the IR was allowed to vary. This flexible model provides a better representation of the SD region than a rigid model does.

We describe four main results. First, multiple alignment of the regions upstream of *Escherichia coli* genes by maximizing information identified the SD pattern without reference to the 16S rRNA sequence. Secondly, ribosome binding could be modelled using a unified mathematical representation for the aligned SD, the initiation region and the distribution of spacings. Thirdly, the second law of thermodynamics sets zero as the theoretical lower bound for the information of binding sites<sup>27,25</sup>, so we could iteratively remove sites with negative information to heighten the model’s predictive capabilities. Finally, further characterization of the Shine-Dalgarno model allowed us to observe how the SD pattern varies with distance from the initiation region.

## Results

### Characteristics of flexible ribosome binding site models

Several different *Escherichia coli* ribosome binding site models are presented, and each is used for a different purpose. Models may be *rigid*, in which all parts are fixed relative to a zero coordinate, or *flexible*, in which case the model contains two rigid parts (SD and IR) separated by a variable distance. These models are further characterized as being either *unrefined* or *refined*. Refinement refers to a cyclic process in which an individual information model is made from the current set of binding sites and then sites that have negative information content are removed from the set. This process is repeated until only positive sites remain. (See Materials and Methods.)

We used three databases in this work. The protein-coding feature locations in the complete-genome Genbank entry U00096 have not been updated since the original publication<sup>28</sup>, so our first database was the alternate set of gene intervals present in EcoGene12<sup>29</sup>. This revised database, which contains 4122 known and putative translation start sites in *E. coli*, is the result of an intense and continuous effort to improve the annotation and prediction of *E. coli* genes. Second, we used the Verified subset of this database, which is composed of protein start sites confirmed by N-terminal protein sequencing. The third database is the original *E. coli* annotation (reference<sup>28</sup>, GenBank U00096). To create a reliable baseline model, we refined the Verified set. In contrast, a ribosome model built from the refined EcoGene12 database is probably the most representative of all genes. We also refined the Blattner database to determine if we could automatically derive a model comparable to EcoGene12.

The Verified model is derived from ribosome binding sites for proteins that have been well-studied and/or detected as spots on 2D gels and probably lacks many sites that show lower binding affinity. Despite this bias, the Verified model is useful since it is composed only of sites proven to be actual ribosome binding sites. For example the range of SD to IR

spacing from  $-18$  to  $-4$  was established by observing spacings utilized within the Verified set. The EcoGene12 model is based on the full set of proven and predicted gene starts and thus is representative of all ribosome binding sites, including weak sites responsible for low level protein expression. Although EcoGene12 may contain a few predicted sites that turn out to be incorrect, we consider it to be the most accurate model, and therefore we used it as our benchmark model.

The rigid model sequence logo made from all EcoGene12 translation start sites (Fig. 1A) shows the expected strong conservation for the initiation region at bases 0, +1 and +2 and a low region of conservation from bases  $-12$  to  $-6$  for the Shine-Dalgarno. When the SD was re-aligned to maximize the information<sup>30</sup>, its information present rose from  $1.53 \pm 0.03$  to  $4.96 \pm 0.04$  bits. (We report here the mean  $R_{sequence}$  and standard error of this mean from the individual information distribution<sup>25</sup>.) The range of re-alignment ( $-18$  to  $-4$ ) was selected to allow for all spacings observed in the Verified model. The SD was realigned using only sequences from translation start sites, and this was done independently of the 16S rRNA sequence, yet the sequence logo closely complements the 16S rRNA 3' end. This flexible model has an SD - IR spacing of  $-18$  to  $-4$  bases, with a peak of occurrence at  $-9$  (Fig. 1B). When the model was tightened by using the exclusionary refinement process (Fig. 1C), there was again an increase in the information present in the Shine-Dalgarno logo to  $5.23 \pm 0.04$  bits. In contrast, the refined Verified SD has  $5.77 \pm 0.10$  bits with an SD - IR spacing of  $-18$  to  $-4$  bases, with a peak of occurrence at  $-10$  (Fig. 1D).

←Fig 1

When the SD region was re-aligned to maximize information in the Blattner model (Fig. 2A), the SD information rose from  $0.91 \pm 0.03$  to  $3.87 \pm 0.05$  bits. This model has an SD - IR spacing of  $-16$  to  $-2$  bases, with a peak of occurrence at  $-8$  (Fig. 2B). Refinement of the Blattner model also showed a further increase in the SD information to  $5.01 \pm 0.04$  bits. Interestingly, upon refinement the SD - IR spacing shifted to  $-18$  to  $-4$  and the peak shifted to  $-9$  bases (Fig. 2C). This is the same range seen in the well characterized Verified model (Fig. 1D).

←Fig 2

For all models, a Gaussian distribution with the same mean and standard deviation as the respective SD - IR spacing distributions was plotted along with the spacing histogram. In all cases, the histogram did not match the Gaussian plot.

Using the individual information method<sup>25</sup> all sites in the Verified set were evaluated by the rigid and flexible EcoGene12 refined models over the range of 30 bases upstream to 14 bases downstream of the first base of the initiation codon. This is the range required to identify a site with the maximum SD - IR spacing of 18 bp. Previous information theory based ribosome evaluations with a rigid model have been reasonably accurate<sup>7</sup>, but since this model does not take into account variable spacing it is limited in its analysis of ribosome binding. The rigid EcoGene12 model picked up more upstream non-sites (sites other than those annotated in that data set) than the flexible model (141 versus 87 respectively). By contrast, the two models identified nearly the same number of Verified start points (565 versus 567). The average site strength assessed by the rigid model was  $9.52 \pm 0.13$  bits and with the flexible model it was  $10.20 \pm 0.13$  bits, indicating that the flexible model generally assessed the Verified sites more strongly.

## Shine-Dalgarno as a function of spacing

To better understand the function of the Shine-Dalgarno, we examined its sequence logo at every SD - IR spacing in the EcoGene12 set (Fig. 3). The shape and pattern of the SD remained fairly constant, but the information present fluctuated. There was a constant increase in the information as the spacing was increased from  $-4$  to  $-9$  and a decrease in information for  $-9$  to  $-18$ . This is apparent by the increase of the size of the bases surrounding the central G. The information present in the SD at each alignment relative to the IR is only weakly related to the conservation of information in the IR ( $r = -0.17$ ) (Fig. 4). When the total flexible site information, as calculated from equation (1) (see Materials and Methods), was examined for all positions a similar increase and decrease in information was observed as with the SD region alone. When the refined Blattner sites were split into spacing classes, similar results were obtained (data not shown).

For all spacings of  $-4$  to  $-11$  there is an A with low conservation at position  $-3$ , and it is also present from  $-16$  to  $-18$ , indicating that conservation at this position is an effect of the initiation contact and not the SD (Fig. 3).

The minimum SD - IR spacing of  $-4$  has been observed in *nadB*<sup>31</sup> but appears to be used infrequently. Binding of regions with spacings more than 18 bases is known, but is rare and due to RNA structural effects such as hairpins that bring the SD closer to the IR<sup>12</sup>.

The nucleotide correlation program **diana**<sup>32</sup> was used to determine if there was any correlation between bases in the SD and IR for each of the spacing classes. There was no observed correlation between bases at any spacing (data not shown), further suggesting independence between the SD and the IR. In addition, no correlation was observed between parts of the refined EcoGene12 SD (Fig. 1C) when all classes are combined.

## Correlation between the refined Blattner and EcoGene12 models

To test the refined EcoGene12 model's accuracy and corrective capabilities, we scanned it across several proven ribosome binding sites and also across several sites predicted by Blattner that have been corrected in EcoGene12 (Fig. 5). First we wanted to see how the model performed on well studied genes. When our model was applied to the *lacZ* and *lacI* initiation regions, it concurred with Blattner's locations (Fig. 5A,B). In the case of *lacI*, which starts at an 8.1 bit GTG in the context atGTGa, a second weaker 5.5 bit site is seen at the out of frame ATG just upstream. Interestingly, ribosomes binding to this site would terminate immediately at the TGA. Using two of Blattner's sites that have been corrected based on N-terminal protein sequencing, we tested whether our model locates the correct binding sites. In *mhpD* (Fig. 5C), we saw a 12.8 bit site at the correct location 6 bases downstream from Blattner's prediction. Our model did not predict any site ( $R_i > 0$ ) at the Blattner location. In *yhbL* (Fig. 5D) there is a predicted weak 4.5 bit site at the incorrect position, but experimentally the start site was proven to be 9 bases downstream and our model favored this location (13.7 bits). As expected, in both cases the correct site was found in the same reading frame as the predicted site. When the refined Blattner model was scanned over these same sites the same predictions were made, indicating that the refined

Blattner model is comparable to the refined EcoGene12 model.

To further investigate the effect of refinement, we scanned both the Blattner unrefined and refined models over all of the EcoGene12 sites for regions 100 bases upstream and 100 bases downstream of each of the 3900 refined EcoGene12 start points. The unrefined Blattner model found 17254 non-sites and the refined model found considerably fewer non-sites (11439). This large number of sites detected may represent weak ribosome binding sites, untranscribed regions or may be false positive artifacts of this model. Alternatively, some of these sites may be occluded by RNA secondary structure. Since the unrefined Blattner model contains many non-sites, it has a lower information content and therefore picks up more non-sites than the refined Blattner model. The unrefined Blattner model identified as many EcoGene12 sites as the refined Blattner model (3883 and 3823 respectively).

To generalize Fig. 5C and D, we scanned both the refined Blattner model and the refined EcoGene12 model over the 26 sites in the Blattner annotation that have been corrected in the EcoGene12 dataset based on experimental verification. The Blattner model identified the experimentally reported start site as the strongest site in 21 of the 26 cases. In 4 of the 5 other cases, the model assessed the Blattner annotation more strongly, but also found a site at the confirmed start point. For one gene (*gntK*), the model predicted a site that matched neither Blattner’s annotation nor the experimental result. When this same analysis was done using the refined EcoGene12 model, the correct site was predicted in 22 of the 26 sites and three of Blattner’s annotations were favored. As with the refined Blattner model, for the *gntK* gene no site was predicted at either the Blattner or EcoGene12 locations. As exemplified by Fig. 5D, in approximately half of the 26 corrected sites both models predicted strong sites at the verified locations and these were accompanied by weaker sites at Blattner’s locations. In the 3 (EcoGene12) or 4 (Blattner) cases where the verified start was weaker than the Blattner site for either model, the difference in site strength between the site at the Blattner location and the site at the Verified location was generally only 1 to 2 bits (except for one case where the difference was around 5 bits). These results show that refined models can be used to improve ribosome binding site predictions.

Can we create a valid ribosome model from the large lists of gene start points determined from open reading frames that are presented as annotations for complete genome sequences? To test for relatedness, we compared various models using the Euclidean distance between  $R_{iw}(b, l)$  matrices (Materials and Methods, equation (5)). The distance between the unrefined Blattner SD matrix and the refined EcoGene12 SD matrix was 16.9 bits and the distance between the corresponding IR matrices was 4.8 bits. In contrast, when the refined Blattner was compared to the refined EcoGene12 model, there was a much smaller difference: for the SD matrix there was a distance of 1.1 bits and for the IR matrix there was a distance of 2.0 bits. Refining the Blattner model brought it close to the refined EcoGene12 model, which is representative of the bulk of *E. coli* ribosome binding sites.

When the individual information distributions for all models were compared, there was a general increase in the strength of sites from the unrefined to the refined to the Verified model (Fig. 6, Table 1). This effect may occur not only because the refinement process

⇐Fig 6  
⇐Table  
1

may tend to neglect weaker sites, as these may often be harder to characterize biochemically. The sets overlap reasonably well since 507 of the 569 Verified sites are found in the refined Blattner set, and all but 6 of the Verified sites are found in the refined EcoGene12 model.

## Discussion

To create a large refined set of ribosome binding sites, we first removed the initiation codon and downstream open reading frame by embedding the SD region into random sequence. This allowed us to use multiple alignment to focus the SD region by maximizing its information content<sup>30</sup>. The SD emerged easily (Fig. 1B), mathematically demonstrating the existence of this feature in the majority of *E. coli* ribosome binding sites. Furthermore, the general pattern matches the 3' end of the 16S rRNA well, independently confirming that these are complementary to each other.

In contrast with the notion of an SD consensus sequence, the SD is variable and its pattern depends on how far the sequence is from the IR (Fig. 3). Despite this variability, the information content of the SD is relatively constant at various spacings, smoothly increasing and decreasing in a range of only 2 bits from  $-18$  to  $-5$  (Fig. 4). Surprisingly, the SD-IR spacing contributes more variation to the total information than either the SD or the IR. Furthermore, the variation of the SD information works in the same direction as the gap surprisal; they do not compensate for each other but instead work together. This sets up a maximal range of variation for efficiency of translational initiation. These observations are consistent with the SD-anti-SD helix formed between the rRNA and the mRNA as being a reasonably consistent 'object' whose placement relative to the IR is important.

In all cases (Fig. 1, Fig. 2) the spacing distribution between the SD and the initiation region was similar to but differed from a Gaussian distribution. There is a predominance of  $-9$  and  $-8$  spacings (and  $-10$  for the refined Verified set). This implies that the spacing is not a simple harmonic oscillator under thermal Brownian motion influence<sup>33</sup>; there must be other physical constraints. What these are may become apparent only when crystal structures of initiating ribosomes have been determined, but a clue that the meaning is related to the placement of the SD-anti-SD helix comes from the shape of the SD sequence logos.

Unlike the rectangular block that a consensus sequence would make on an information graph, the SD sequence logo smoothly rises and declines with position (Fig. 1C). This is consistent with the idea that mismatches at the center of an RNA-RNA hybrid should be more disruptive than mismatches towards the ends. However, the situation may be more complicated. Sequence logos for duplex DNA binding proteins also rise and decline with position<sup>34, 35, 21, 36</sup>. One intriguing explanation is that the formation of the mRNA-rRNA hybrid is followed by binding of a ribosomal protein or RNA<sup>37, 3</sup> into the resulting major or minor groove as a step during translational initiation. Such a model accounts for the shape of the sequence logo because proteins tend to evolve contacts on one face of a helix, and such contacts become progressively more difficult to form when they are close to the back face<sup>35</sup>. The proximity of protein S1 to the SD<sup>11, 38, 39</sup> suggests it as a candidate for

this process, but other proteins such as S7, S18 and S21<sup>40</sup>, and various 16S rRNA positions<sup>41, 42</sup> that crosslink to the mRNA<sup>43</sup> could be involved. To allow us to judge the validity of this model, we added a dashed sine wave to the sequence logos. The peaks of this wave are separated by 11 bp, which is the distance between two major grooves of A-form RNA<sup>44</sup>. Preferred spacings of the SD (Fig. 3) are consistent with this model, but there is clearly a greater degree of flexibility than in DNA-protein interactions. However, tight packing is observed throughout the 70S subunit<sup>45, 2</sup> and there is close packing in the 30S<sup>3</sup>, so it is likely that the fully assembled initiation region is also tightly packed. This suggests a mechanism for initiation in which the binding of the SD to the mRNA allows the resulting helix to be smaller than unpaired single strands would be. The smaller helix could pack against other components of the ribosome, reducing the volume further and completing initiation, perhaps by creating sufficient space in the A site for the next tRNA. Even a non-specific RNA phosphate backbone binding into the minor groove between the SD and the mRNA<sup>37, 3</sup> could account for the shape of the aligned SD sequence logo. IF3 appears to recognize codon-anticodon complementarity at the initiation codon rather than direct recognition of the codon itself<sup>46</sup>. Because complementarity usually creates a more compact structure than a mismatch, this effect is also consistent with a packing model for initiation. Finally, this tight packing model may account for why the SD-IR spacing is more narrow than a Gaussian distribution.

The concept of individual information<sup>25</sup> allows us to consistently apply an information measure of the gap distribution to individual sequences, thereby creating a flexible search tool. The problem of how to compute the information content of flexible binding sites was recognized previously<sup>6</sup>. If two sequence elements have a variable distance between them, then the uncertainty in position decreases the overall information content. For example, GC, with an information content of  $2 \times 2 = 4$  bits, is found every 16 bases in equiprobable DNA, while GNC is found at the same frequency. A shorthand notation for the set containing both of these is G1EC, in which ‘1E’ means to search for G followed by C with an extendible spacing of 1 or 0 bases<sup>5</sup>. Because it contains the search for both GC and GNC, G1EC occurs approximately every 8 bases in equiprobable random DNA. So although the G and C contribute 4 bits of information, the variable spacing removes one bit and the site is therefore effectively only 3 bits. With G3EC there are 4 possible search patterns, GC, GNC, GNNC and GNNNC; this removes  $\log_2 4 = 2$  bits. Interestingly G15EC has 16 search patterns and this removes 4 bits giving the, at first sight, odd result that the information content is 0 bits. However, in a sequence M bases long, G15EC will occur roughly M times because of overlapping cases, so the result is consistent. It is interesting to note that there can be sites with negative information by this method: in a sequence of length M, G31EC will occur roughly 2M times, giving an apparent information of  $-1$  bit. The reason for this odd effect is that there are many overlapping sites. We interpret zero or negative information to mean that the two components are independent.

We have extended these computations by using Shannon’s uncertainty measure to consistently assess the contribution when different spacings occur with different frequencies. Because frequencies are not probabilities, a small sample correction was also applied<sup>6</sup>.



Fortunately, the negative information effect does not occur for ribosome binding sites. In the refined EcoGene12 model (Fig. 1C) the SD contains  $5.80 \pm 0.04$  bits, the initiation region contains  $6.72 \pm 0.04$  bits and the uncertainty of the distance between them (gap uncertainty,  $H_{\text{gap}}$ ) is  $3.25 \pm 0.02$  bits, giving a total information content <sup>6</sup> of  $R_s(\text{SD}) + R_s(\text{IR}) - H_{\text{gap}} = 9.28 \pm 0.06$  bits. This is similar to the refined rigid EcoGene12 model which has  $8.92 \pm 0.05$  bits, but quite different from the flexible Verified model at  $10.35 \pm 0.16$  bits. We suggest that the difference occurs because strong sites tend to be experimentally identified first and some nonfunctional sites may still contaminate the refined EcoGene12 model. The latter effect can be observed in Fig. 6 where the unrefined EcoGene12 has examples of sites below zero, while the refined EcoGene12 set does not have any sites below zero bits, by definition. While there are no sites below zero in the refined Verified set (because we removed the 13 that we found) the lower end of the distribution curve is smaller than that for the refined EcoGene12. Further, the shape of the Verified distribution is a more Gaussian-like curve, trailing smoothly down to nearly zero at zero bits <sup>25</sup>, while the refined EcoGene12 distribution still has members near zero bits and is therefore discontinuous. It is not known if these very weak sites are functional.

The Verified sites that we removed during refinement (gene at U00096 coordinates and orientations: *uppS* 194903 +, *gsk* 499349 +, *fes* 612038 +, *dbpA* 1407535 +, *topB* 1844984 -, *guaB* 2632090 -, *xseA* 2632252 +, *trmD* 2743359 -, *pcm* 2867542 -, *cysI* 2888122 -, *dnaN* 3879949 -, *aceK* 4216175 + and *arcA* 4637875 -) presumably initiate differently than the majority of sites. Surprisingly, this set does not contain *infC*, which codes for IF3. In the absence of this initiation factor the ribosome can use the AUU start of *infC* (1798662 -) for initiation, forming a regulatory feedback loop <sup>46</sup>. By the Verified model the AUU containing IR is  $-4.8$  bits but this is compensated by a GS of 2.3 bits at the optimal spacing of  $-9$  bases and an SD of 9.6 bits to give a total of 2.5 bits. This anomalous site was automatically removed during refinement of EcoGene12 because the G at the third base of start codons is otherwise invariant and rare bases are more heavily weighted against in larger datasets <sup>25</sup>. By the EcoGene12 model, the *infC* IR is  $-8.4$  bits with an SD at a  $-9$  spacing of 8.7 bits for a total of  $-2.0$  bits. This model predicts that mutating the start codon from AUU to AUG should bring the IR up to 5.5 bits to give a strong 11.9 bit site.

Other mechanisms may be needed to explain the anomalous Verified sites. Only two excluded cases in the Verified set have GTG starts, which are known to be weaker than ATG starts <sup>13</sup>. With fewer examples in the data set, marginal GTG starts could have been removed because of statistical noise.

Another way to explain the Verified site anomalies is that RNA secondary structures might bring an SD closer to the IR <sup>12</sup> and so influence translation <sup>47, 11</sup>. As shown in Fig. 7, this mechanism might be involved in *fes* (612038) in which a 4 base helix ( $-4.5$  kcal/mole) <sup>48</sup> may bring a 3.4 bit SD to position  $-11$  with respect to the IR and *pcm* (2867542) in which a 5 base helix ( $-7.8$  kcal/mole) brings a 3.5 bit SD to position  $-9$  with respect to the IR. The other sites do not appear to use this mechanism.

The relatively large number of initiation regions that do not conform to the majority model (*i.e.* the rejected Verified sites) suggests the possibility that there are even more

⇐Fig 7

alternative mode(s) of translation initiation. We are left with a number of likely and proven genes that fail to have ribosome binding sites that conform to our model. A combination of computational and experimental approaches will be needed to identify alternative models among the rejected sites. Of course, one simple possibility is that apparent anomalies can be caused by sequencing errors <sup>26</sup>.

An empirical observation for human splice junctions is that, in addition to the thermodynamic bound at zero bits <sup>25</sup>, sites with less than 2.4 bits are non-functional <sup>49</sup>. We suspect that such a fuzzy non-zero bound may also apply to the majority of ribosome binding sites. However, unlike the case with splice junctions, experimental data are not currently available to suggest what a natural bound may be that delineates a functional from a nonfunctional ribosome site. For this reason we used the zero bound for cyclic refining.

The process of subtracting the gap uncertainty is similar to the accounting of gaps provided by hidden Markov models (HMM) <sup>50</sup> except that the gap size we use is variable and the frequencies of different gap distances are accounted for. It may be possible to extend the model given here to a full information theory based HMM, but this was not attempted.

The information theory approach allowed us to build models that represent the vast majority of ribosome sites without having to assume that some sequences were not sites. In contrast, training with a neural net <sup>51</sup> requires this assumption because data on where ribosomes *do not* bind are sparse. Because the data set is so large it could be split into spacing classes (Fig. 3), effectively dissecting the ribosome binding sites. The resulting models revealed that the weakly conserved “A at -3” <sup>5,6</sup> correlates with the IR and not with the SD. This is consistent with the presence of an A at -3 relative to the translational start in eukaryotic mRNAs, which do not have an SD <sup>8</sup>. The function of this conservation is unknown but crosslink experiments place it close to U1381 on the 16S rRNA <sup>42</sup> and to the S7 protein <sup>40,3</sup>.

The effort required to generate a data set as exemplified by EcoGene12 is enormous. The refining process described here gives results comparable to EcoGene12, so we believe it will be useful for gene analysis in other species. Ideally, all organisms would have models consisting of biochemically supported sites, rather than sites that were chosen by computer algorithms that do not model the SD. However, as shown here, it is possible to use information theory methods to help produce a reasonably clean identification of gene starts. This technique may be useful to better characterize medically important disease organisms.

## Materials and Methods

### Databases

To create our models we drew from three databases. One database that we used was the 4122 sites in the EcoGene12 database, which represent the majority of *E. coli* genes <sup>29</sup>.

The second database was a carefully compiled list of 569 experimentally supported sites, referred to as the Verified database <sup>29</sup>. This database is a subset of the EcoGene12 database and provided us with an initial comparison model which was used to determine the allowed SD - IR spacings and the general pattern of the Shine-Dalgarno. Rudd (2000) has catalogued

from the biomedical literature 717 *E. coli* proteins whose N-termini have been directly determined by protein sequencing. The Verified proteins that have cleaved signal peptides were omitted since these N-terminal protein sequences do not verify the translation start codons as definitively as the 569 Verified proteins that are uncleaved or only have the initiator methionine residue cleaved.

The third database was the 4290 gene starts presented by Blattner *et al.* (1997) and extracted from their complete *E. coli* GenBank entry, U00096 (version M52, September 02, 1997). This database will be referred to as the Blattner database. The refining method described below was performed on these databases.

## Creating a Ribosome Model

Our ribosome models have two rigid binding elements connected by a flexible bond that allows the distances between the elements to vary. If both elements do not find a suitable contact at a good distance, then the model will not bind. The first binding element is represented by a sequence logo made from translation start codons, which we will refer to as the initiation region (IR) (Fig. 1A). The range of this model is from  $-3$  to  $+14$ , representing the predominantly A conservation at  $-3$  through the downstream mRNA protected by the ribosome<sup>6, 17, 11</sup>. To create this logo, standard Delila tools were used as previously documented<sup>24, 35</sup>. The programs **dbbk**, **catal**, **dbinst** and **delila** were used to extract predicted translation start points from the complete *E. coli* genome<sup>28</sup>. **Alist** was used to align the extracted sites and to generate a color listing. **Encode**, **rseq** and **dalvec** were used to prepare the fragments for the **makelogo** program, which generates sequence logos.

For the SD there is only a low sequence conservation over the range of  $-11$  to  $-7$  in the rigid logo (Fig. 1A), so we used multiple alignment to build on the rigid model to create a flexible ribosome model. The  $-20$  to  $-4$  range of the sequences was embedded into random sequence using the **markov** and **embed** programs to avoid alignment by the IR in the next step. Then, by using the **malign** program, this region was realigned to maximize the information present<sup>30</sup>. The resulting alignment was then represented by a logo using the previously described method. This realigned set displays the compliment of the anti-Shine-Dalgarno found on the 16S rRNA (Fig. 1B) and was used as our Shine-Dalgarno model. The zero coordinate of this model was shifted to the coordinate of the large central G using **instshift**. Since this base position contains the maximum information, presumably it is the most stable position to use and it will appear as the most significant base in a sequence walker (Fig. 5). By this definition, our  $+4$  base corresponds to the  $SD_{ref}$  reference point defined by Chen *et al.* (1994),<sup>14</sup> and our spacing measures are “aligned spacing” according to these authors. This measures the distance from a fixed point on the 16S rRNA to the initiation codon, as they advocate.

Once we had both an SD and an IR model, we made a histogram of the distances between their zero coordinates for all sites in the database. This was done using the **diffinst** program and these distances were then presented in a histogram using the **genhis** and **genpic** programs. **Ri**<sup>25</sup> was then used to generate an  $R_{iw}(b, l)$  weight matrix for both the

SD and IR for further analysis of the individual information conserved in ribosome binding sites.

To evaluate the individual information <sup>25</sup> of a ribosome binding site using a flexible model, we calculate three values. This information will be referred to as Flexible Site Information and is characterized by the following equation:

$$\text{Flexible Site Information} = R_i(\text{SD}) + R_i(\text{IR}) - \text{GS} \quad (\text{bits/site}) \quad (1)$$

$R_i(\text{SD})$  is the individual information of the Shine-Dalgarno region, while  $R_i(\text{IR})$  is the individual information of the initiation region. The SD sites were assessed for the region  $-12$  to  $+4$  and the IR sites were assessed for the region  $-3$  to  $+14$  because this is the region covered by footprints <sup>7</sup>.

By aligning the SD relative to the IR, we obtained the distribution of spacing distances. Any probability distribution has an uncertainty measured in bits:

$$H = - \sum_i p_i \log_2 p_i = \sum_i p_i (-\log_2 p_i) \quad (\text{bits}) \quad (2)$$

<sup>18, 19</sup>. Rewriting the uncertainty as shown on the right hand side shows that it is composed of an average of the surprisal function <sup>52</sup>:

$$u_i = -\log_2 p_i \quad (\text{bits/spacing}). \quad (3)$$

To calculate the Gap Surprisal (GS) the relative occurrence of each binding distance is calculated by dividing the occurrences at that position by the total number of occurrences. This value is the relative frequency of binding at each spacing. The GS equation is therefore:

$$\text{GS} = -\log_2 \frac{\text{occurrence of a spacing}}{\text{total occurrences}} + e(n) \quad (\text{bits/spacing}). \quad (4)$$

The GS factor accounts for the variable spacing.  $e(n)$  is a small-sample correction for GS based on the total occurrences ( $n$ ), required because we have substituted a frequency for the probability  $p_i$  <sup>6, 25</sup>.

Because GS is positive when there is more than one spacing possibility, has the same units as the SD and IR and is presumably independent of them <sup>16</sup>, we subtracted it from the SD and IR individual information to obtain equation (1).

The program used to compute equation (1) is **biscan**. Biscan finds pairs of SD and IR that fall within the range of the spacing histogram and then the flexible site information is calculated for each coupling using the distribution of distances from the genhis histogram.

Further information about the programs is available at <http://www.lecb.ncifcrf.gov/~toms/> and a web-based server with guest-access is available at <http://www.lecb.ncifcrf.gov/~toms/delilaserver.html>

## Cyclic refinement

The 5' ends of genes are often incorrectly placed in sequence database feature tables. To obtain a reliable ribosome model containing a minimum number of misplaced sites, a cyclic refinement method was used. To do this we looked at the flexible site information for all sites in the set. We removed all sites whose information content was less than zero (this is the theoretical boundary for binding because of the second law of thermodynamics<sup>25</sup>) and we rebuilt the model with the corrected set. Following every round of refining, the Shine-Dalgarno region was realigned using `malign` as previously described<sup>30</sup>, using 1000 alignments, maximizing information in a window from  $-20$  to  $-4$  and allowing the sequences to shift from  $-8$  to  $+6$ . This range was chosen to match the known binding range of the Verified model. If more than one Shine-Dalgarno site was found upstream of an initiation start site, then the SD which gave the strongest flexible site information was used in the model. This process was repeated until no sites remained in the data set with a negative flexible site information. This was done for the EcoGene12 set, the Verified set and for Blattner's set. The EcoGene12 required 10 rounds of refining and lost 222 sites; the Verified set required 2 rounds and we dropped 13 sites and Blattner's set required 20 rounds of refining and we dropped 781 sites. Each round took approximately 2 hours on a Sun Ultra60 Sparc workstation.

## Dissecting the SD

To generate the SD as a function of spacing (Fig. 3), a logo was made for each of the 15 observed spacing groups. For example, 6 sites were observed to have an SD - IR spacing of  $-4$  bases so a logo was made for those sites (upper left corner of the figure). This was repeated for the range of  $-18$  to  $-4$  for the refined EcoGene12 model. The graph of information present in the Shine-Dalgarno region versus spacing (Fig. 4), is the  $R_{sequence}$  for the range of  $-12$  to  $+4$  around the central G in the SD portion of the logo. The range  $-12$  to  $+3$  was used for the spacing of  $-4$ , because of interference with the zero position of the initiation region. The range for the IR information curve was  $-3$  to  $+14$ . To examine relatedness between nucleotides within the separated SD - IR spacing binding sites, the program **diana** was used<sup>32</sup>.

## Sequence Walkers

To make the sequence walkers<sup>26</sup> (Fig. 5), **biscan** generated features which were then mapped by **live**, **mergemarks** and **lister**. The refined EcoGene12 model was used for this analysis. Mfold 3.1<sup>48</sup> was used to fold RNA sequences and the structures were displayed along with the walkers using **mfoldseq** and **mfoldfea**.

## Comparing Matrices

To compare two weight matrices, we calculated the Euclidean distance between them

using the following equation:

$$\text{Distance} = \sqrt{\sum_l \sum_b (\text{Ri}_1(\text{b}, l) - \text{Ri}_2(\text{b}, l))^2} \quad (\text{bits}). \quad (5)$$

Here the difference is taken between the individual information of each base (**b**) at each position (**l**) between matrices 1 and 2. The difference is then squared and summed for all positions and the square root of this value is taken, giving the distance in bits. The program used to do this is **diffrib1**.

## Acknowledgements

We would like to thank Peter Rogan for reporting to us the results of cyclic rigid information theory based refining of human splice junctions, Karen Lewis, Xiao Ma, Shu Ouyang, Denise Rubens, Brent Jewett, Ilya Lyakhov, Peter Rogan, Frank Boellmann and Eric Miller for their comments. K.E.R. was supported by NIH grant GM58560.

## References

1. Green, R. & Noller, H. F. (1997). Ribosomes and translation. *Annu Rev Biochem*, **66**, 679–716.
2. Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, **289**, 905–920.
3. Wimberly, B. T., Brondersen, D. E., Clemons Jr., W. M., Morgan-Warren, R. J., Carter, A. P., Vornheln, C., Hartsch, T. & Ramakrishnan, V. (2000). Structure of the 30S ribosomal subunit. *Nature*, **407**, 327–339.
4. Gold, L., Pribnow, D., Schneider, T., Shinedling, S., Singer, B. S. & Stormo, G. (1981). Translational initiation in prokaryotes. *Annu. Rev. Microbiol.* **35**, 365–403.
5. Stormo, G. D., Schneider, T. D. & Gold, L. M. (1982). Characterization of translational initiation sites in *E. coli*. *Nucleic Acids Res.* **10**, 2971–2996.
6. Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* **188**, 415–431.
7. Rudd, K. E. & Schneider, T. D. (1992). Compilation of *E. coli* ribosome binding sites. In *A Short Course in Bacterial Genetics: A Laboratory Manual and Handbook for Escherichia coli and Related Bacteria*, (Miller, J. H., ed.), pp. 17.19–17.45, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
8. Kozak, M. (1999). Initiation of translation in prokaryotes and eukaryotes. *Gene*, **234**, 187–208.

9. 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. USA*, **71**, 1342–1346.
10. Calogero, R. A., Pon, C. L., Canonaco, M. A. & Gualerzi, C. O. (1988). Selection of the mRNA translation initiation region by *Escherichia coli* ribosomes. *Proc. Natl Acad. Sci. USA*, **85**, 6427–6431.
11. Hartz, D., McPheeters, D. S. & Gold, L. (1991). Influence of mRNA on determinants on translational initiation in *Escherichia coli*. *J. Mol. Biol.* **218**, 83–97.
12. Gold, L. (1988). Posttranscriptional regulatory mechanisms in *Escherichia coli*. *Annu. Rev. Biochem.* **57**, 199–233.
13. Barrick, D., Villanueva, K., Childs, J., Kalil, R., Schneider, T. D., Lawrence, C. E., Gold, L. & Stormo, G. D. (1994). Quantitative analysis of ribosome binding sites in *E. coli*. *Nucleic Acids Res.* **22**, 1287–1295.
14. Chen, H., Bjerknes, M., Kumar, R. & Jay, E. (1994). Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of *Escherichia coli* mRNAs. *Nucleic Acids Res.* **22**, 4953–4957.
15. Rinke-Appel, J., Junke, N., Brimacombe, R., Lavrik, I., Dokudovskaya, S., Dontsova, O. & Bogdanov, A. (1994). Contacts between 16S ribosomal RNA and mRNA, within the spacer region separating the AUG initiator codon and the Shine-Dalgarno sequence; a site-directed cross-linking study. *Nucleic Acids Res.* **22**, 3018–3025.
16. Ringquist, S., Shinedling, S., Barrick, D., Green, L., Binkley, J., Stormo, G. D. & Gold, L. (1992). Translation initiation in *Escherichia coli*: sequences within the ribosome-binding site. *Mol Microbiol*, **6**, 1219–1229.
17. Hartz, D., McPheeters, D. S., Green, L. & Gold, L. (1991). Detection of *Escherichia coli* ribosome binding at translational initiation sites in the absence of tRNA. *J. Mol. Biol.* **218**, 99–105.
18. Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27**, 379–423, 623–656. <http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html>.
19. Pierce, J. R. (1980). *An Introduction to Information Theory: Symbols, Signals and Noise*. second edition, Dover Publications, Inc., New York.
20. Schneider, T. D. (1994). Sequence logos, machine/channel capacity, Maxwell's demon, and molecular computers: a review of the theory of molecular machines. *Nanotechnology*, **5**, 1–18. <http://www.lecb.ncifcrf.gov/~toms/paper/nano2/>.

21. Hengen, P. N., Bartram, S. L., Stewart, L. E. & Schneider, T. D. (1997). Information analysis of Fis binding sites. *Nucleic Acids Res.* **25** (24), 4994–5002. <http://www.lecb.ncifcrf.gov/~toms/paper/fisinfo/>.
22. Shultzaberger, R. K. & Schneider, T. D. (1999). Using sequence logos and information analysis of Lrp DNA binding sites to investigate discrepancies between natural selection and SELEX. *Nucleic Acids Res.* **27** (3), 882–887. <http://www.lecb.ncifcrf.gov/~toms/paper/lrp/>.
23. Zheng, M., Doan, B., Schneider, T. D. & Storz, G. (1999). OxyR and SoxRS regulation of *fur*. *J. Bact.* **181**, 4639–4643.
24. Schneider, T. D. & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100. <http://www.lecb.ncifcrf.gov/~toms/paper/logopaper/>.
25. Schneider, T. D. (1997). Information content of individual genetic sequences. *J. Theor. Biol.* **189** (4), 427–441. <http://www.lecb.ncifcrf.gov/~toms/paper/ri/>.
26. Schneider, T. D. (1997). Sequence walkers: a graphical method to display how binding proteins interact with DNA or RNA sequences. *Nucleic Acids Res.* **25**, 4408–4415. <http://www.lecb.ncifcrf.gov/~toms/paper/walker/>, erratum: NAR 26(4): 1135, 1998.
27. Schneider, T. D. (1991). Theory of molecular machines. II. Energy dissipation from molecular machines. *J. Theor. Biol.* **148**, 125–137. <http://www.lecb.ncifcrf.gov/~toms/paper/edmm/>.
28. Blattner, F. R., Plunkett III, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B. & Shao, Y. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
29. Rudd, K. E. (2000). EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res.* **28**, 60–64.
30. Schneider, T. D. & Mastronarde, D. (1996). Fast multiple alignment of ungapped DNA sequences using information theory and a relaxation method. *Discrete Applied Mathematics*, **71**, 259–268. <http://www.lecb.ncifcrf.gov/~toms/paper/malign>.
31. Flachmann, R., Kunz, N., Seifert, J., Gutlich, M., Wientjes, F. J., Laufer, A. & Gassen, H. G. (1988). Molecular biology of pyridine nucleotide biosynthesis in *Escherichia coli*. Cloning and characterization of quinolinate synthesis genes *nadA* and *nadB*. *Eur. J. Biochem.* **175**, 221–228.
32. Stephens, R. M. & Schneider, T. D. (1992). Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *J. Mol. Biol.* **228**, 1124–1136. <http://www.lecb.ncifcrf.gov/~toms/paper/splice/>.



33. Schneider, T. D. (1991). Theory of molecular machines. I. Channel capacity of molecular machines. *J. Theor. Biol.* **148**, 83–123.  
<http://www.lecb.ncifcrf.gov/~toms/paper/ccmm/>.
34. Papp, P. P., Chatteraj, D. K. & Schneider, T. D. (1993). Information analysis of sequences that bind the replication initiator RepA. *J. Mol. Biol.* **233**, 219–230.
35. Schneider, T. D. (1996). Reading of DNA sequence logos: prediction of major groove binding by information theory. *Meth. Enzym.* **274**, 445–455.  
<http://www.lecb.ncifcrf.gov/~toms/paper/oxyr/>.
36. Wood, T. I., Griffith, K. L., Fawcett, W. P., Jair, K.-W., Schneider, T. D. & Wolf, R. E. (1999). Interdependence of the position and orientation of SoxS binding sites in the transcriptional activation of the class I subset of *Escherichia coli* superoxide-inducible promoters. *Molec. Microbiol.* **34**, 414–430.
37. Clemons Jr, W. M., May, J. L., Wimberly, B. T., McCutcheon, J. P., Capel, M. S. & Ramakrishnan, V. (1999). Structure of a bacterial 30S ribosomal subunit at 5.5Å resolution. *Nature*, **400**, 833–840.
38. Boni, I. V., Isaeva, D. M., Musychenko, M. L. & Tzareva, N. V. (1991). Ribosome-messenger recognition: mRNA target sites for ribosomal protein S1. *Nucleic Acids Res.* **19**, 155–162.
39. Sorensen, M. A., Fricke, J. & Pedersen, S. (1998). Ribosomal protein S1 is required for translation of most, if not all, natural mRNAs in *Escherichia coli* *in vivo*. *J. Mol. Biol.* **280**, 561–569.
40. Dontsova, O., Kopylov, A. & Brimacombe, R. (1991). The location of mRNA in the ribosomal 30S initiation complex; site-directed cross-linking of mRNA analogues carrying several photo-reactive labels simultaneously on either side of the AUG start codon. *EMBO J.* **10**, 2613–2620.
41. Greuer, B., Thiede, B. & Brimacombe, R. (1999). The cross-link from the upstream region of mRNA to ribosomal protein S7 is located in the C-terminal peptide: experimental verification of a prediction from modeling studies. *RNA*, **5**, 1521–1525.
42. Bhangu, R., Juzumiene, D. & Wollenzien, P. (1994). Arrangement of messenger RNA on *Escherichia coli* ribosomes with respect to 10 16S rRNA cross-linking sites. *Biochemistry*, **33**, 3063–3070.
43. Baranov, P. V., Kubarenko, A. V., Gurvich, O. L., Shamolina, T. A. & Brimacombe, R. (1999). The Database of Ribosomal Cross-links: an update. *Nucleic Acids Res.* **27**, 184–185.

44. Arnott, S., Hukins, D. W. & Dover, S. D. (1972). Optimised parameters for RNA double-helices. *Biochem Biophys Res Commun*, **48**, 1392–1399.
45. Nissen, P., Hansen, J., Ban, N., Moore, P. B. & Steitz, T. A. (2000). The structural basis of ribosome activity in peptide bond synthesis. *Science*, **289**, 920–930.
46. Meinnel, T., Sacerdot, C., Graffe, M., Blanquet, S. & Springer, M. (1999). Discrimination by *Escherichia coli* initiation factor IF3 against initiation on non-canonical codons relies on complementarity rules. *J. Mol. Biol.* **290**, 825–837.
47. de Smit, M. H. & van Duin, J. (1990). Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. *Proc. Natl Acad. Sci. USA*, **87**, 7668–7672.
48. Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **288**, 911–940.
49. Rogan, P. K., Faux, B. M. & Schneider, T. D. (1998). Information analysis of human splice site mutations. *Human Mutation*, **12**, 153–171.  
<http://www.lecb.ncifcrf.gov/~toms/paper/rfs/>.
50. Baldi, P., Brunak, S., Chauvin, Y. & Krogh, A. (1996). Naturally occurring nucleosome positioning signals in human exons and introns. *J. Mol. Biol.* **263**, 503–510.
51. Stormo, G. D., Schneider, T. D., Gold, L. & Ehrenfeucht, A. (1982). Use of the ‘Perceptron’ algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Res.* **10**, 2997–3011.
52. Tribus, M. (1961). *Thermostatistics and Thermodynamics*. D. van Nostrand Company, Inc., Princeton, N. J.

<b>Individual Information Distribution Values</b>				
model	Mean (bits)	Stdev (bits)	SEM (bits)	n
Blattner unrefined	6.83	4.84	0.07	4290
Blattner refined	8.82	3.63	0.06	3509
EcoGene12 unrefined	8.81	3.99	0.06	4122
EcoGene12 refined	9.28	3.58	0.06	3900
Verified	10.35	3.73	0.16	569

Table 1: Comparing Individual Information Distribution Values

We report the mean, standard deviation, standard error of the mean and number of sites for each model. These values correspond to the distributions in Fig. 6.

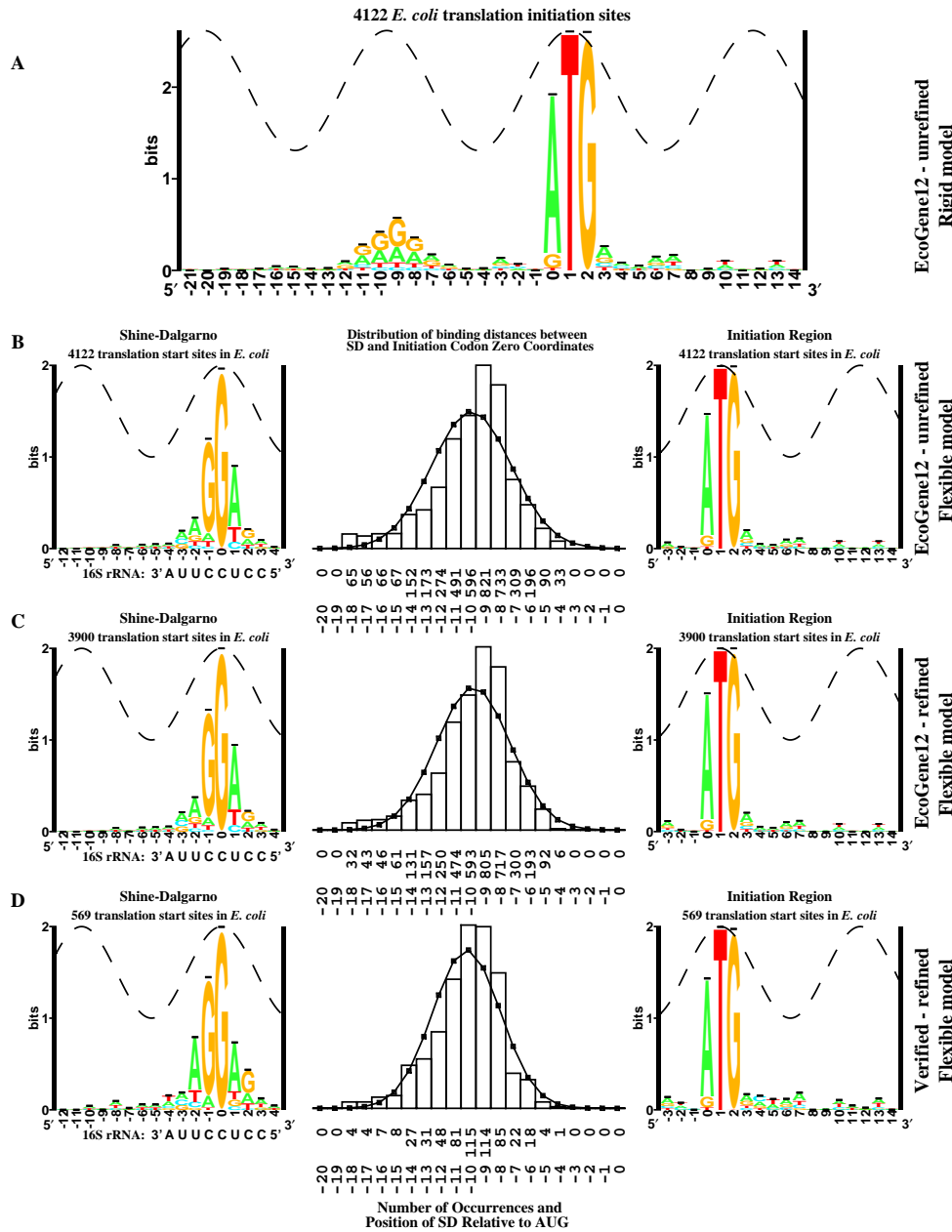


Figure 1: Representing EcoGene12 ribosome binding sites with sequence logos. For all logos the size of each stack of letters corresponds to the total sequence conservation at that position, measured in bits<sup>6</sup>. The height of each letter corresponds to the relative frequency of that base at that position<sup>24</sup>. The sine wave represents the 11 base twist of A-form RNA<sup>44</sup>. Part A is the rigid model of the entire EcoGene12 set<sup>29</sup> while parts B-D represent flexible ribosome models. The histogram between each pair of logos represents the distribution of distances between the Shine-Dalgarno and the initiation region zero coordinates. A Gaussian distribution with the same mean and standard deviation is shown for comparison. All logos on the left of the page represent the Shine-Dalgarno alignment and all logos on the right represent the initiation region alignment. The sequence shown under each SD logo is the anti-Shine-Dalgarno sequence found on the 3' end of the 16S rRNA. B) is the EcoGene12 unrefined flexible model. C) is the EcoGene12 refined flexible model. D) is the Verified refined flexible model.

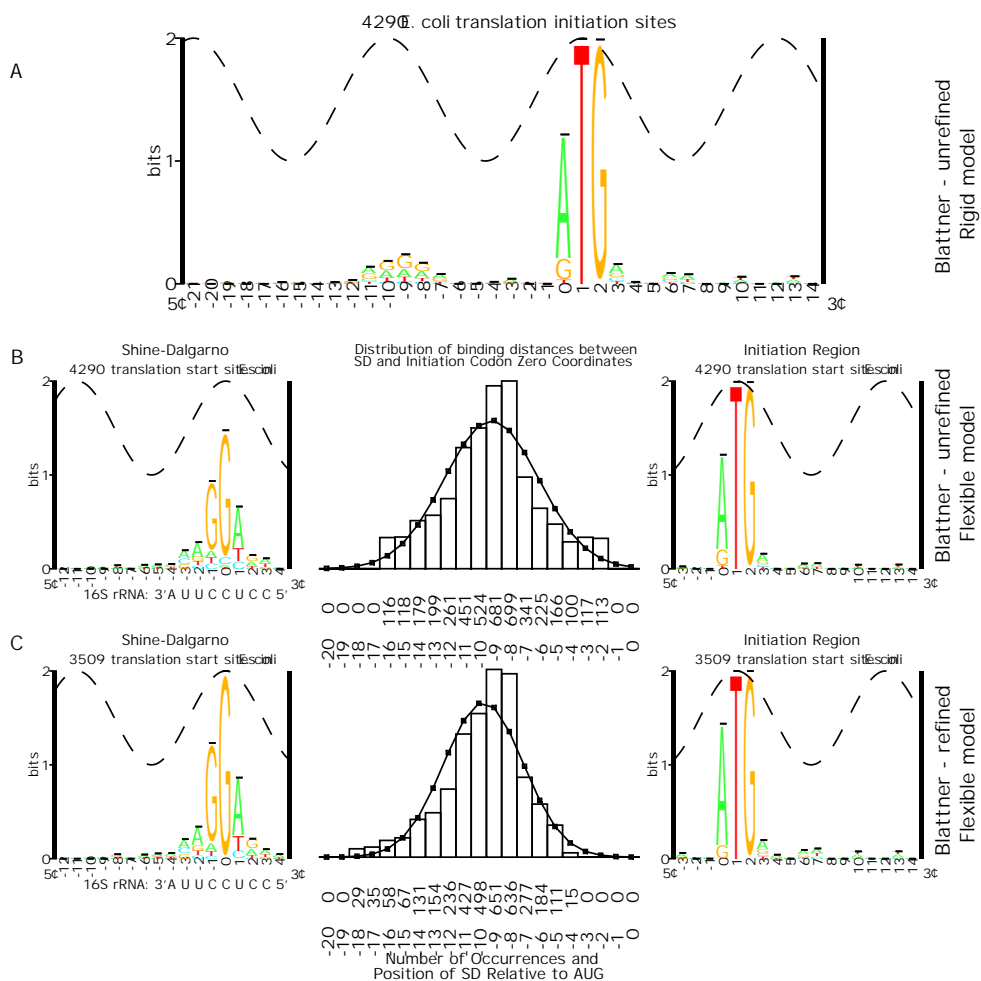


Figure 2: Representing Blattner's ribosome binding sites with sequence logos. Sequence logos were constructed and displayed as described in Fig. 1. Part A is the rigid model of Blattner's entire set <sup>28</sup> while parts B-C represent flexible ribosome models. B) is the Blattner unrefined flexible model. C) is the Blattner refined flexible model.

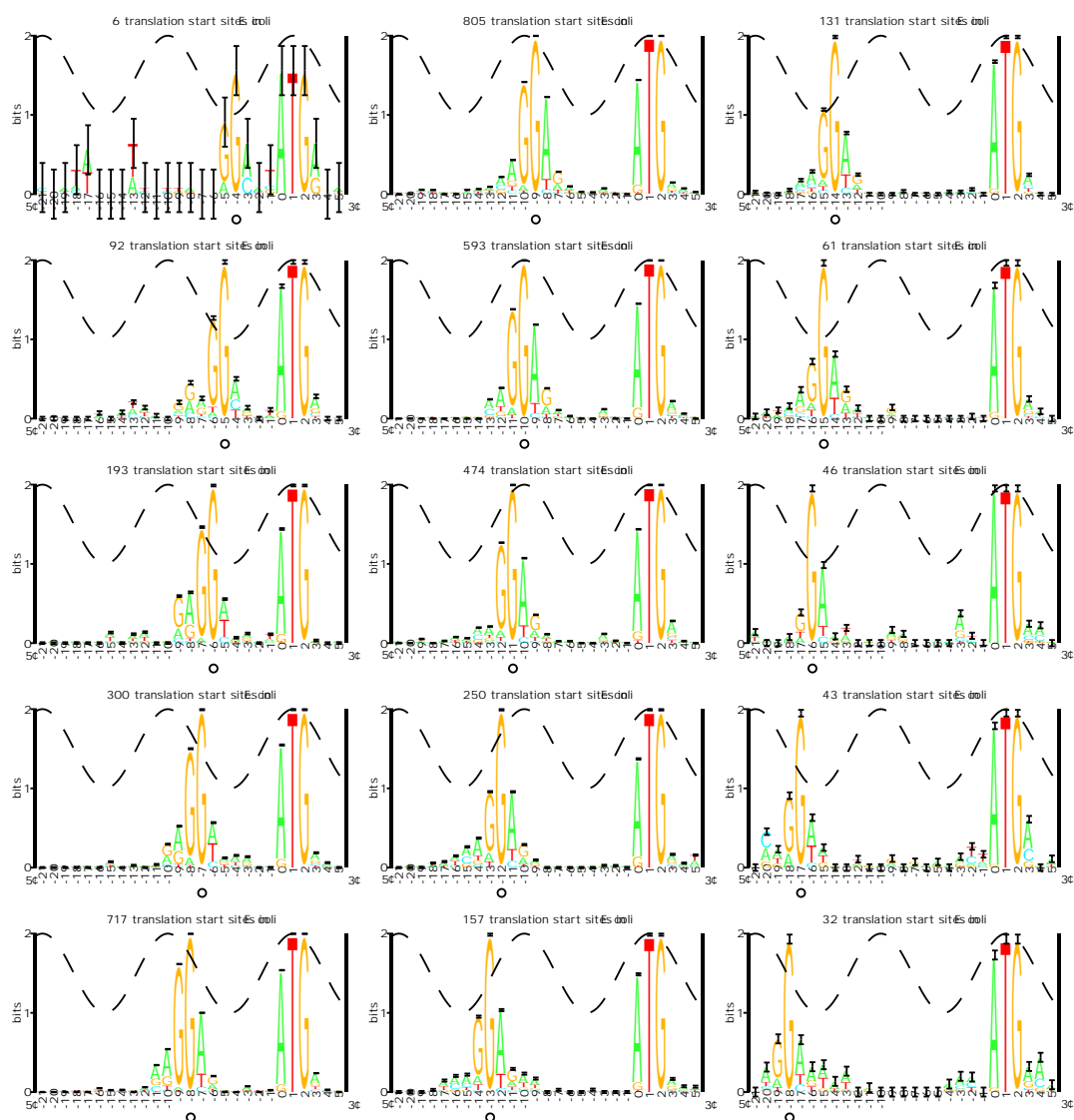


Figure 3: The Shine-Dalgarno as a function of spacing.

Sequence logos were constructed for all distances between the SD and IR zero coordinates observed in the EcoGene12 refined set. The black circle falls under the central G of the Shine-Dalgarno, which is the zero coordinate of the SD in the variable model.

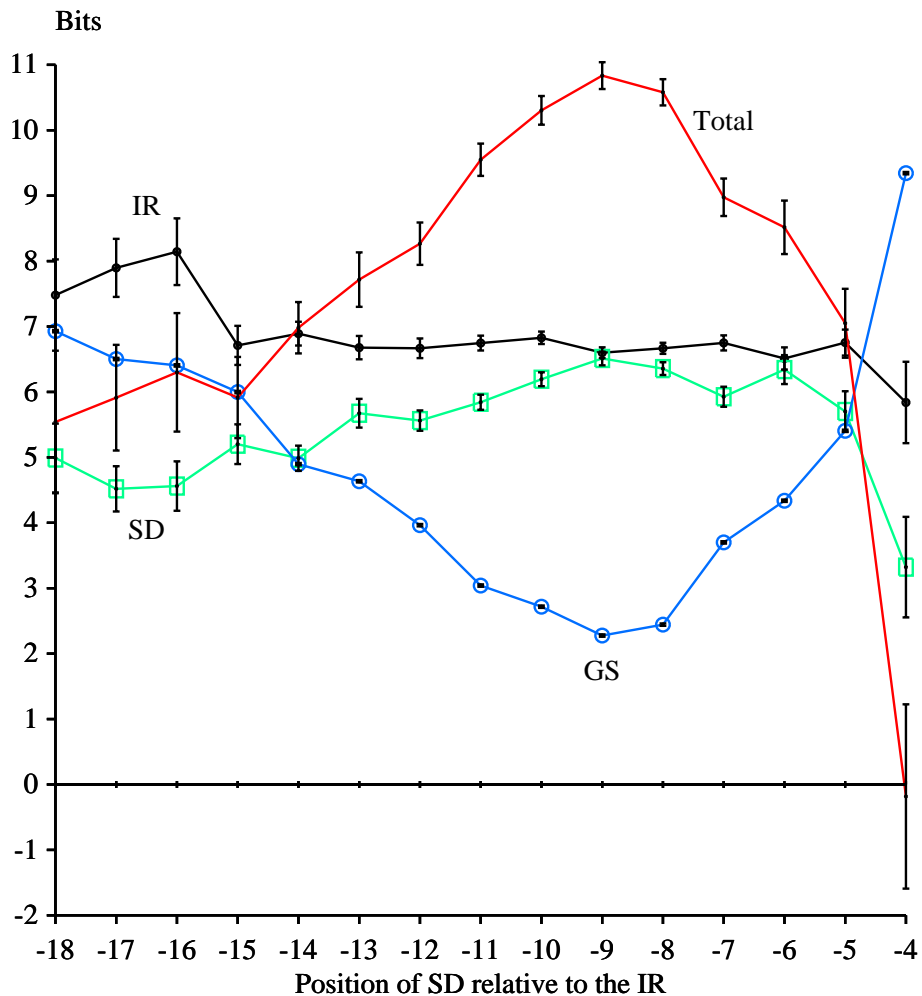


Figure 4: Information present in the Shine-Dalgarno subsets.

The information present in the Shine-Dalgarno regions of Fig. 3 (shown in green boxes) were plotted at their respective distances. The information content was measured over the region 12 bases prior to and 4 bases after the central G of the Shine-Dalgarno, except for the spacing of  $-4$ , whose information is measured over the range of  $-12$  to  $+3$  because of interference with the initiation region at position 0. The information present in the IR for the range of  $-3$  to  $+14$  for each of the SD regions is shown in black (with small filled circles). The gap surprisal GS computed by equation (4) from the distance distribution in Fig. 1C, is plotted in open blue circles. The red curve with no symbols shows the total flexible information at each spacing, as calculated by equation (1). Error bars are plotted with black “I” symbols (the error for GS is smaller than the circle).

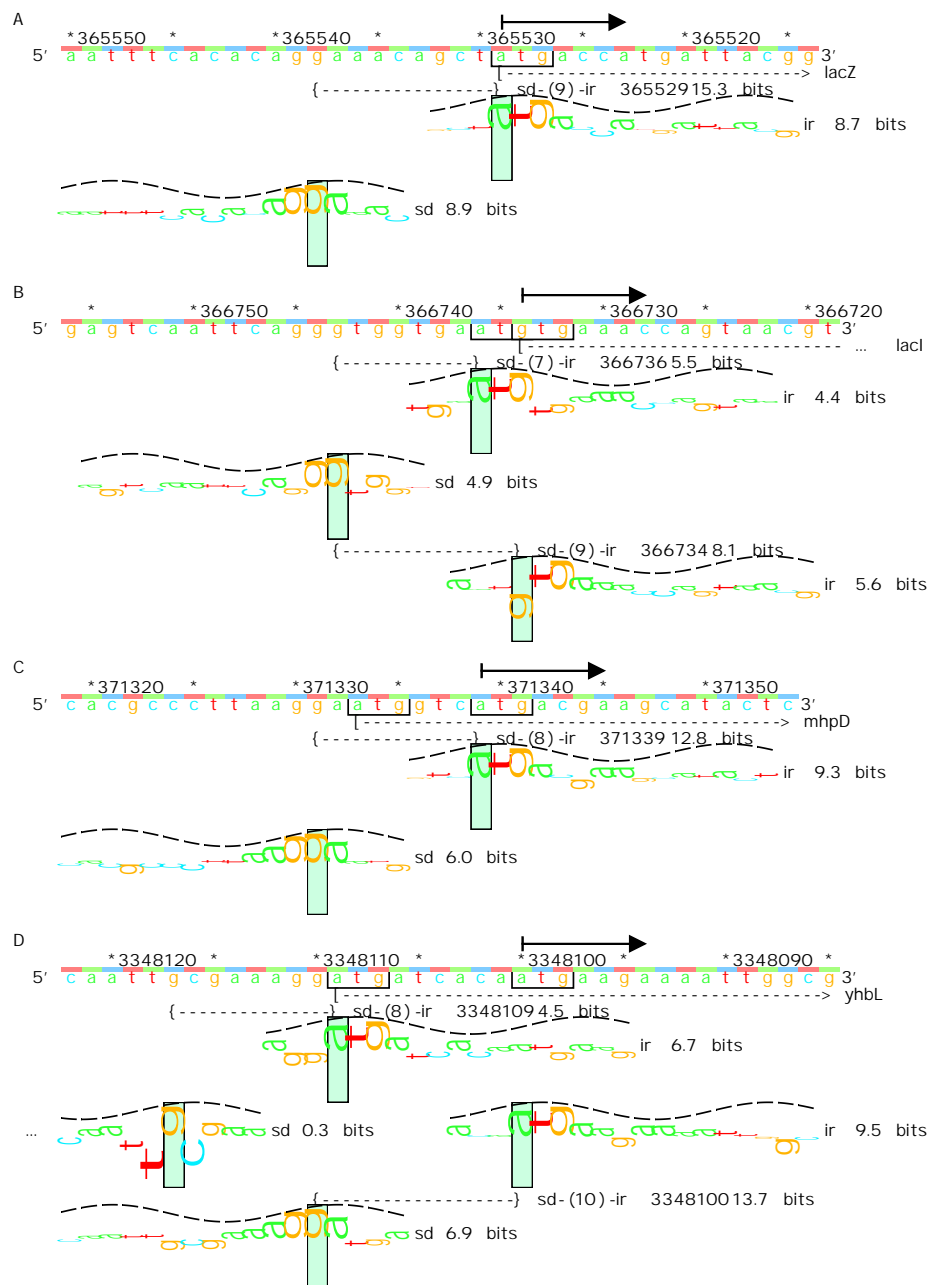


Figure 5: Lister maps with sequence walkers for four ribosome binding sites. Blattner’s sequence, GenBank accession number U00096, is annotated with 4290 gene starts<sup>28</sup>. Four of these start points are illustrated. All correct translation start points, based on experimental data, are identified by a black arrow starting at the initiation start point. The color bar above the sequence cycles through three colors to illustrate the reading frames. In cases C and D, it is obvious that the predicted (boxed) and corrected sites (boxed and arrow) fall in the same reading frame because the adenine bases lie under the same color. The vertical green rectangles indicate the zero coordinate of each sequence walker and provide a scale from  $-3$  to  $+2$  bits. The sine waves represent the 11 base twist of A-form RNA<sup>44</sup>. The asterisks and numbers above the sequence indicate the position of the site on the *Escherichia coli* genome<sup>28</sup>. The bracket “[ ]” followed by the dashed line shows Blattner’s predicted gene start. The braces “{ }” connected by a dashed line are used to link SD and IR walkers. This feature, created by the program **biscan**, also reports the distance of separation, the coordinate of the IR and the flexible site information value according to equation (1). The translational starts are: A: *lacZ*, B: *lacI*, C: *mhpD*, and D: *yhbL*.



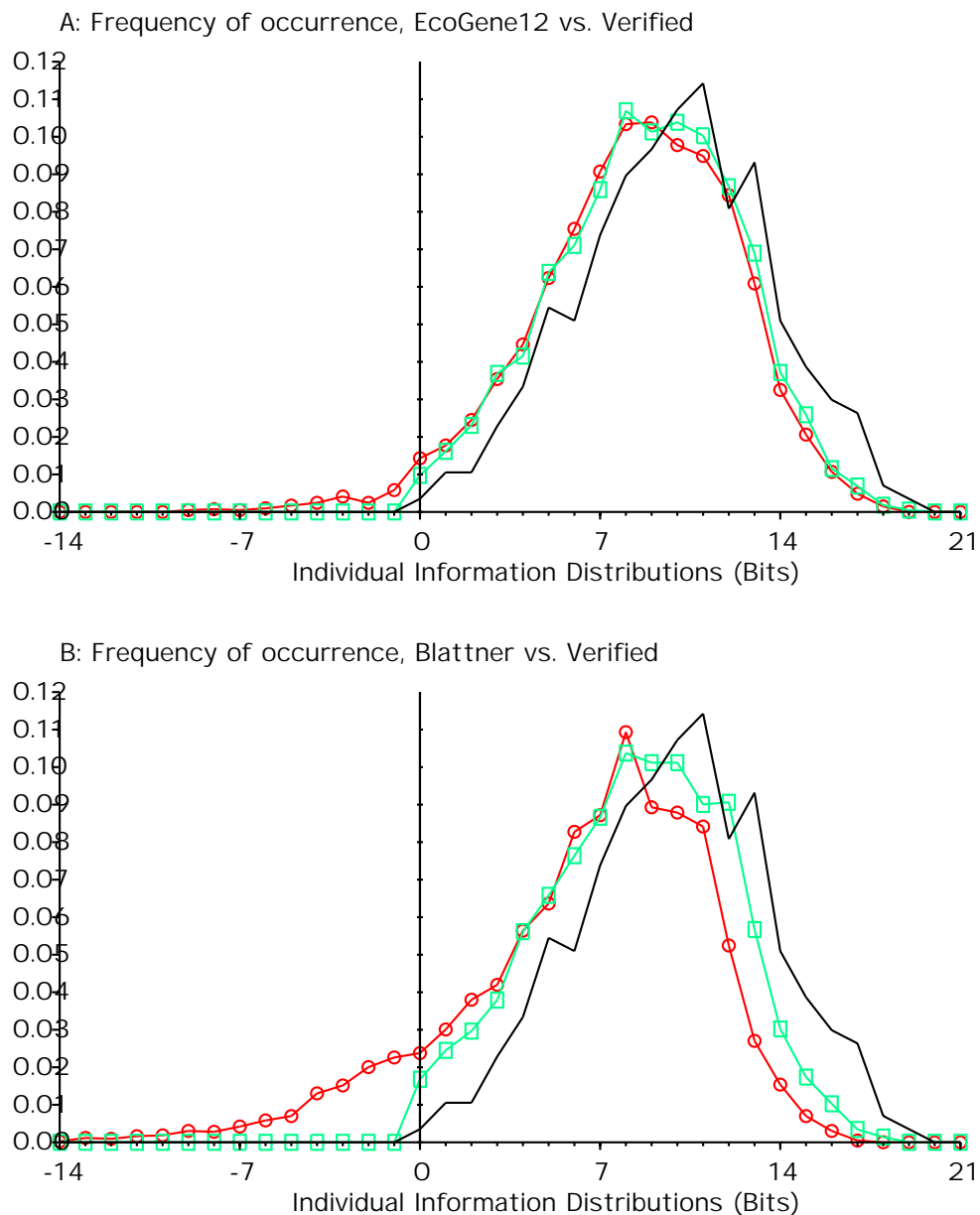


Figure 6: Individual information distributions for five ribosome binding site models. The ordinate is the individual information and the abscissa is the frequency of occurrence. Part A shows the information distributions for the EcoGene12 unrefined (red circles), the EcoGene12 refined (green boxes) and the Verified model (black, with no symbols). Part B shows the information distributions for the Blattner unrefined (red circles), the Blattner refined (green boxes) and the Verified model (black, with no symbols).

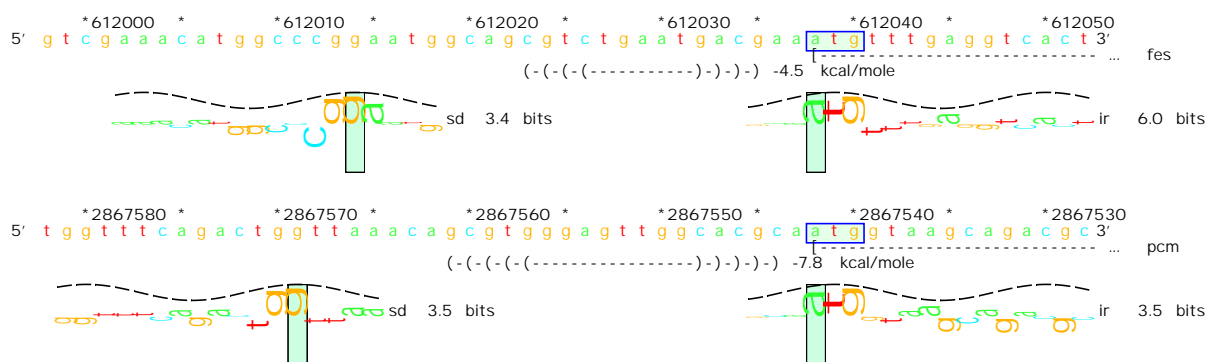


Figure 7: mRNA folding may rescue *fes* and *pcm* translation. Structure base pairings are indicated by parenthesis.