Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochemical and Biophysical Research Communications 335 (2005) 71–75
BBRC
www.elsevier.com/locate/ybbrc

Analysis of 5′ UTR composition and gene expression: Canonical versus non-canonical start codons

Anders Fuglsang *

Danish University of Pharmaceutical Sciences, 2 Universitetsparken, DK-2100 Copenhagen Ø, Denmark
Norwegian Medicines Agency, Sven Oftedals Vei 8, N-0950 Oslo, Norway

Received 1 July 2005
Available online 20 July 2005

Abstract

The overall composition upstream of start codons in *Escherichia coli* was evaluated and viewed in connection with global transcriptome data. Genes starting with AUG as initiation codon tended to be expressed at higher levels than the non-AUG genes, and the upstream region of the non-AUG genes showed negligible signs of Shine–Dalgarno sequences. The latter is in sharp contrast to the AUG genes. Viewing these findings in connection with the current literature, it is proposed that a distinct mechanism for initiation of translation might exist for non-AUG genes that are not preceded by a Shine–Dalgarno sequence. A survey covering a range of other eubacteria (Firmicutes, Proteoacteria, and Actinobacteria) reveals that it is mainly among the Proteobacteria that non-AUG genes do not display clear signs of Shine–Dalgarno regions.
© 2005 Elsevier Inc. All rights reserved.

Keywords: Expression; Shine–Dalgarno sequences; Translation initiation; Start codons; mRNA levels

Until the mid 1970s the mechanism of initiation of translation was unknown but was anticipated that some feature at the 5′-end of mRNA directed ribosomal binding. The studies undertaken by Shine and Dalgarno in the 1970s towards the understanding that in many bacteria highly expressed genes tend to have a consensus sequence that is complementary to a stretch of nucleotides at the 3′-end of 16S rRNA [1,2]. Subsequently it has been shown that initiation complex is formed when the 30S subunit ribosome is bound to mRNA and that this ribosomal binding involves an interaction between a stretch of nucleotides on the mRNA (Shine–Dalgarno sequence) and the 16S rRNA (for a review of the initiation of translation see [3]). However, although this mechanism has been firmly established in *Escherichia coli*, it is also a fact that this bacterium is very well capable of initiating translation from mRNA which is short-leadered or leaderless [4,5]. This means that there must be other factors governing this process. Start codons themselves seem to be implied in the translational efficiency with AUG generally being the more efficient codon [6,7], and likewise, the nucleotides at the 3′ side of the start codon also have strong influence [7–9]. It has been suggested that a 'downstream box' exists, being a sequence capable of base-pairing with the penultimate stem of 16S rRNA (working much in the same fashion as a Shine–Dalgarno sequence, but located 3′ to the start codon) [10,11], although the evidence currently does not seem overwhelmingly solid [12–14].

Thus for any given mRNA, the efficiency of initiation of translation therefore seems to be determined by the composition of the start codon and the nucleotides on the 5′ side of it. It naturally comes into mind that sequence alignment, such as the Smith–Waterman algorithm [15], could be considered for studies of the Shine–Dalgarno sequences, perhaps in conjunction with estimates free energy estimates of duplex formation

* Fax: +45 35306020.
E-mail address: anfu@dfuni.dk.

0006-291X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2005.07.047

between the mRNA and the 16S rRNA. This would allow for study of single genes, and has been used in the study by Ma et al. [16].

There are however some drawbacks of this approach. First and foremost, it has not been validated that sequence alignment can predict when a stretch of nucleotides in fact is a Shine–Dalgarno sequence. And second, dynamic programming algorithms are subjective in the sense that the alignment result is highly dependent on subjective and empirical parameters such as gap opening and gap extension values, mismatch scores, etc. And third, the precise length of the 3′-end of 16S rRNA is typically not precisely determined, so often the flatfiles only indicate an estimated length of the 3′-end of 16S rRNA (Siv. G. Andersson, Uppsala University, Sweden, personal communication).

This is also an argument why computation of the estimated drop in free energy of duplex formation may be less appropriate. In a study by Schurr et al. [17], the relationship between calculated energy of duplex formation and efficiency of the Shine–Dalgarno sequence did not reveal any clear relationship. In addition to this there is no good way of knowing how long the mRNA is just from the sequence files; this itself also has to be determined specifically (experimentally, not by bioinformatics) for each cistron.

For this study, an objective method (non-randomness analysis) was used in stead. The advantage of this method is that it does not rely on assumptions regarding length of 16S rRNA at the 3′-end, length of mRNA upstream of start codons, and dynamic programming empirism, but the drawback is that it can only be applied to a pool of genes and not to one single gene (i.e., it cannot tell if a given sequence harbours a potentially strong Shine–Dalgarno sequence, etc.).

Materials and methods

Genomes. This study is centered around analysis of the *E. coli* genome (NC_000913 [18]) downloaded from GenBank (ftp://ftp.ncbi.nih.gov/genomes). In addition to this, the fully sequenced genomes of other bacteria belonging to Firmicutes, Actinobacteria, and Proteobacteria (α, β, γ, δ, and ε subgroups) were downloaded. These three classes were chosen because they represent the majority of pathogenic bacteria and those that have major industrial importance.

Inclusion criterion for this study was that the genes have correct start and stop codons, no undetermined nucleotides, no internal stop codons, and an intact frame.

Non-randomness analysis. Non-randomness analysis was used, as implemented in the NORA software package [19]. In brief, with this analysis the entire genomic composition is first assessed by counting all adenines, thymines, guanines, and cytosines. Next, it is assumed that all nucleotides are dispersed randomly throughout the genome. This assumption is naturally invalid, and non-randomness analysis measures the degree by which this assumption is compromised. If the total genomic fraction of guanine is f_G and we pick N nucleotides for analysis then we would expect to find $C_{exp,G} = N \times f_G$ guanines in our sample following the assumption, and this may or may not be close to the actually observed number $N_{obs,G}$. A χ^2 value integrates the knowledge for all four nucleotides:

$$\chi^2 = \sum_{i=1}^{4} \frac{(C_{obs,i} - C_{exp,i})^2}{C_{exp,i}}. \quad (1)$$

The higher the χ^2 value the larger the deviation between observed and expected counts. This analysis is carried out for all nucleotides that correspond to position one, two, three, and so forth upstream of start codons. A plot of χ^2 values versus position upstream of start codons is constructed on this basis.

For species utilizing Shine–Dalgarno sequences (*E. coli* as an example) a clear peak in χ^2 values is observed in that region, i.e., peaking around 10 nt upstream of the start codon. This is accompanied by an overrepresentation of guanine and adenine. The latter is because the nucleotides that are involved in basepairing with the 3′-end of 16S rRNA are 5′-...ACCUCCU...-3′. This analysis was carried out separately for genes starting with the canonical AUG start codons and for those starting with a non-AUG start codon.

Expression data. To find out if AUG-starting genes generally have higher expression than non-AUG genes, mRNA expression data from Bernstein et al. [20] for *E. coli* grown in rich medium were included in this study. A separate parser for this set of expression data was programmed, and the relative mRNA expression levels were retrieved and split on AUG vs. non-AUG genes.

Results and discussion

Escherichia coli

Non-randomness plots for *E. coli* are shown in Fig. 1 (genes starting with an AUG codon) and Fig. 2 (genes starting with a non-AUG codon). The peak around position 10 upstream of the start codons appears more sharp in Fig. 1 than in Fig. 2. However, χ^2 values cannot be readily compared when the number of genes (N) in the two pools (AUG vs. non-AUG) is not equal. The workaround in this situation is to scale the χ^2 value by dividing with N (it follows from Eq. (1) that the χ^2 value is linearly proportional to the number of genes sampled). This way, a plot showing the average contribution

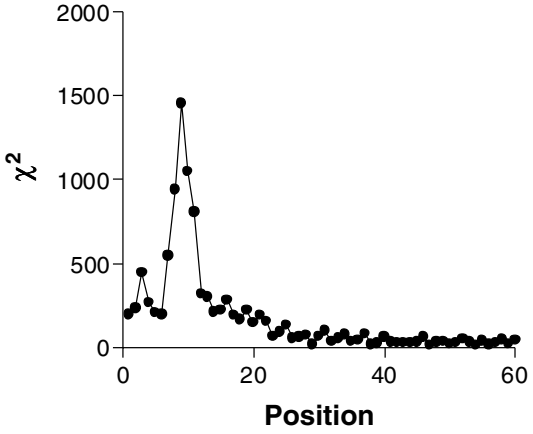

Fig. 1. Plot of non-randomness for the genes in *E. coli* having AUG as start codon ($N = 3543$). A clear peak in the χ^2 value is observed around 10 nt upstream of start codons.

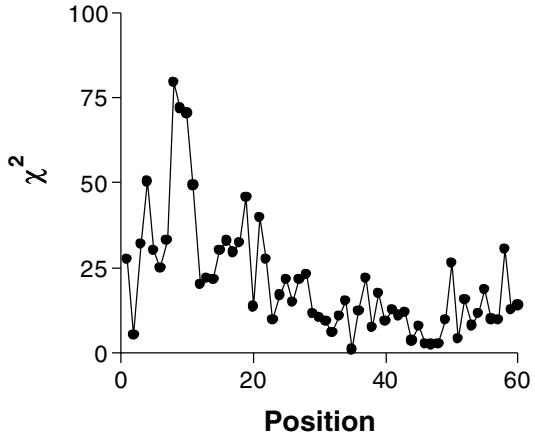

Fig. 2. Similar to Fig. 1, but for genes with non-AUG start codons ($N = 727$).

per gene to the χ^2 value can be made as a function of position upstream of start codons. This plot is shown in Fig. 3. And here the difference between the two datasets becomes very clear, in that the AUG-starting genes clearly have a much more restricted composition, with guanine and adenine being overrepresented, and cytosine being underrepresented (not shown), as expected for genes using Shine–Dalgarno regions. There is thus much stronger compositional indication of Shine–Dalgarno sequences in the AUG genes. As illustrated in Fig. 4, the median mRNA level in the AUG group is higher than in the non-AUG group.

The data suggest that in *E. coli* the prevalence of Shine–Dalgarno sequences is higher in the AUG group than in the non-AUG group, which accords well with the finding of Ma et al. [16], and this in turn suggests that an "alternative way" to initiate translation is employed more often with non-AUG genes. Currently, it

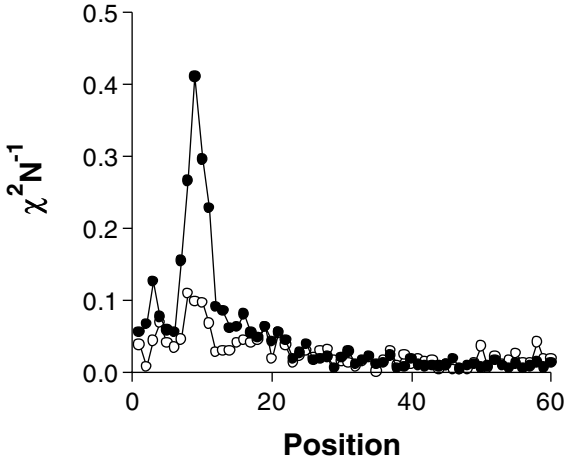

Fig. 3. The data from Figs. 1 (●) and 2 (○) scaled to show the average contribution to χ^2 for genes in the two datasets. Clearly, the composition upstream of start codons in AUG-starting genes is much more non-random than in genes having non-AUG start codons.

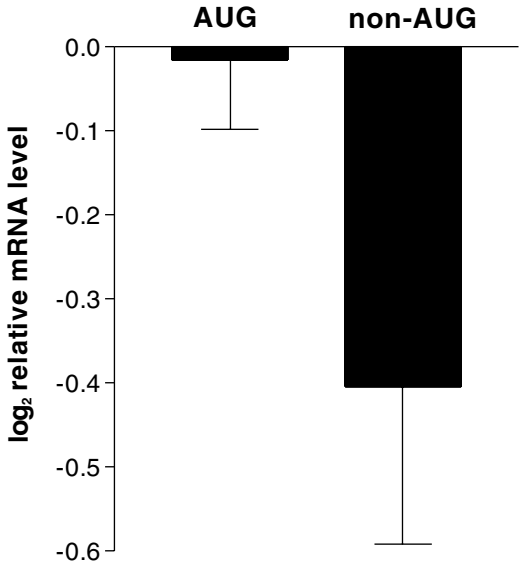

Fig. 4. mRNA levels of AUG-staring genes vs. non-AUG starting genes, using the microarray mRNA expression data from Bernstein et al. [20], for *E. coli* grown in rich medium (Luria–Bertani broth). The median mRNA level for the AUG genes is significantly higher than for non-AUG genes ($P < 0.001$).

has been established that the rpsA protein plays an important role in the translation of *E. coli*, as regards leadered mRNA [21]; rpsA is involved in the initiation process through an interaction with initiation factor III (IF3). Thus, *E. coli rpsA* null mutants are better capable of initiating translation from mRNA that is leaderless. When IF3 levels are low, the translation from leadered mRNA is prone to increased erroneous initiation of translation on upstream AUGs [22,23]. Thus, rpsA and IF3 seem to positively promote translation from leadered mRNA, and this is to some degree in contrast to IF2, which seems to have positive effect on translation of transcript that does not have leaders [24]. It is important to note that it has been observed that translation of leaderless mRNA is promoted by binding directly to 70S ribosome particles [25,26]. In effect the leaderless mRNA is highly depending on AUG start codons [27].

Summing it up in general terms, AUG is prominent either in leadered mRNA with Shine–Dalgarno sequences (traditional mechanism for initiation of translation), or in mRNA that does not have leaders (initiation by the 70S particle). Thus, there might exist a third mechanism of initiation for the genes having non-AUG start codons; these are highly likely to have leaders [27] and have less trace of Shine–Dalgarno sequences.

Other bacteria

A range of other bacteria of Firmicutes, Proteobacteria, and Actinobacteria were analysed the same way. Most, if not all, bacteria of Firmicutes and Actinobacte-

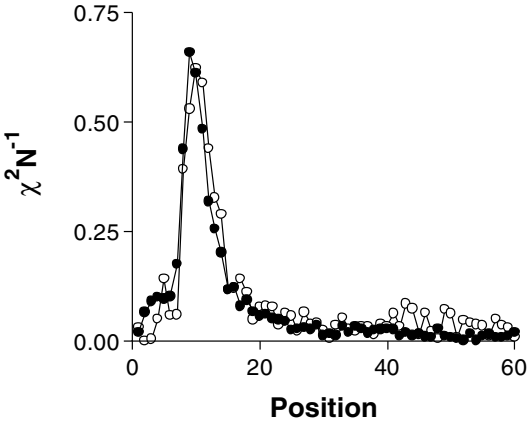

Fig. 5. Plots similar to Fig. 3 but for *S. pyogenes* ($N = 1535$ in the AUG group, $N = 162$ in the non-AUG group).

ria show more marked peaks in χ^2 values for the group of non-AUG genes than what was observed in *E. coli* (Fig. 3). Fig. 5 shows a plot generated with *Streptococcus pyogenes* (NC_002737); in this bacterium there is no visible difference between the AUG group and the non-AUG group. The very flat curve for non-AUG genes is in fact only seen with Proteobacteria, but there seems to be no rule about how closely related bacteria are (phylogenetically) and to what extent their non-randomness plots are similar. For example, with *Pseudomonas aeruginosa* (NC_002516), which is a member of the γ subdivision of the Proteobacteria, and thus a close relative to *E. coli*, the non-AUG peak is just slightly lower than the AUG peak, while with *Caulobacter crescentus* (AE_005673), which is a member of the α subdivision Proteobacteria and thus more distant to *E. coli* than *Ps. aeruginosa*, the plot appears very similar to that obtained with *E. coli*, as shown in Fig. 6.

On basis of Figs. 3 and 6, we can readily see that two species that are closely related (*E. coli* and *Ps. aeruginosa*) may give plots that appear more different from each

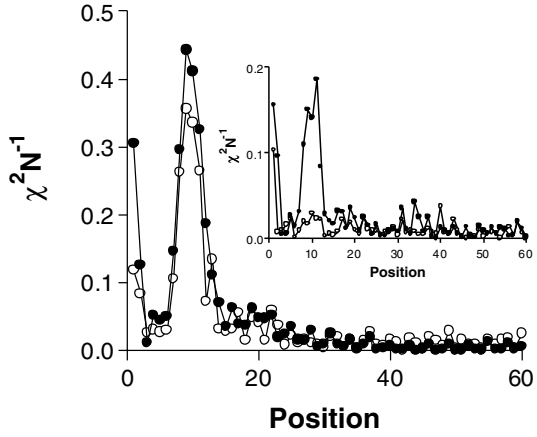

Fig. 6. Plots similar to Figs. 3 and 5, but for *Ps. aeruginosa*, which, like *E. coli*, belongs to the γ-Proteobacteria. The inset figure is a similar plot for *C. crescentus*, which belongs to the α-Proteobacteria.

other than plots generated for two more distantly related species (*E. coli* and *C. crescentus*). The phenomenon that non-AUG genes do not give rise a non-randomness peak is for this reason at present of limited phylogenetic value. It should be emphasized, though, that the very clear visible differences in scaled χ^2 values between AUG genes and non-AUG genes were observed only for Proteobacteria. This may be because the methodology needs further improvement or because there truly might be no general rule about the evolution of start codons and Shine–Dalgarno sequences.

Acknowledgment

I thank prof. Siv. G. Andersson, Uppsala University, Sweden, for useful information about sequencing of ribosomal rRNA.

References

[1] J. Shine, L. Dalgarno, The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites, Proc. Natl. Acad. Sci. USA 71 (1974) 1342–1346.
[2] J. Shine, L. Dalgarno, Determinant of cistron specificity in bacterial ribosomes, Nature 254 (1975) 34–38.
[3] C.O. Gualerzi, C.L. Pon, Initiation of mRNA translation in prokaryotes, Biochemistry 29 (1990) 5881–5889.
[4] C.J. Wu, G.R. Janssen, Expression of a streptomycete leaderless mRNA encoding chloramphenicol acetyltransferase in *Escherichia coli*, J.l Bacteriol. 179 (1997) 6824–6830.
[5] S. Grill, C.O. Gualerzi, P. Londei, U. Blasi, Selective stimulation of translation of leaderless mRNA by initiation factor 2: evolutionary implications for translation, EMBO J. 19 (2000) 4101–4110.
[6] S. Ringquist, S. Shinedling, D. Barrick, L. Green, J. Binkley, G.D. Stormo, L. Gold, Translation initiation in *Escherichia coli*: sequences within the ribosome-binding site, Mol. Microbiol. 6 (1992) 1219–1229.
[7] C.M. Stenstrom, E. Holmgren, L.A. Isaksson, Cooperative effects by the initiation codon and its flanking regions on translation initiation, Gene 273 (2001) 259–265.
[8] C.M. Stenstrom, L.A. Isaksson, Influences on translation initiation and early elongation by the messenger RNA region flanking the initiation codon at the 3′ side, Gene 288 (2002) 1–8.
[9] C.M. Stenstrom, H. Jin, L.L. Major, W.P. Tate, L.A. Isaksson, Codon bias at the 3′-side of the initiation codon is correlated with translation initiation efficiency in *Escherichia coli*, Gene 263 (2001) 273–284.
[10] M.L. Sprengart, E. Fuchs, A.G. Porter, The downstream box: an efficient and independent translation initiation signal in *Escherichia coli*, EMBO J. 15 (1996) 665–674.
[11] M.L. Sprengart, H.P. Fatscher, E. Fuchs, The initiation of translation in *E. coli*: apparent base pairing between the 16srRNA and downstream sequences of the mRNA, Nucleic Acids Res. 18 (1990) 1719–1723.
[12] E.P. Rocha, A. Danchin, A. Viari, The DB case: pattern matching evidence is not significant, Mol. Microbiol. 37 (2000) 216–218.
[13] I. Moll, M. Huber, S. Grill, P. Sairafi, F. Mueller, R. Brimacombe, P. Londei, U. Blasi, Evidence against an interaction

between the mRNA downstream box and 16S rRNA in translation initiation, J. Bacteriol. 183 (2001) 3499–3505.
[14] J.P. Etchegaray, M. Inouye, DB or not DB in translation? Mol. Microbiol. 33 (1999) 438–439.
[15] T.F. Smith, M.S. Waterman, Identification of common molecular subsequences, J. Mol. Biol. 147 (1981) 195–197.
[16] J. Ma, A. Campbell, S. Karlin, Correlations between Shine–Dalgarno sequences and gene features such as predicted expression levels and operon structures, J. Bacteriol. 184 (2002) 5733–5745.
[17] T. Schurr, E. Nadir, H. Margalit, Identification and characterization of *E. coli* ribosomal binding sites by free energy computation, Nucleic Acids Res. 21 (1993) 4019–4023.
[18] F.R. Blattner, G. Plunkett, C.A. Bloch, N.T. Perna, V. Burland, M. Riley, J. Collado-Vides, J.D. Glasner, C.K. Rode, G.F. Mayhew, J. Gregor, N.W. Davis, H.A. Kirkpatrick, M.A. Goeden, D.J. Rose, B. Mau, Y. Shao, The complete genome sequence of *Escherichia coli* K-12, Science 277 (1997) 1453–1474.
[19] A. Fuglsang, INTROnSPECTIVE and NORA: computerized compositional characterization of regions within introns and around start codons on basis of non-randomness analysis, J. Biochem. Biophys. Methods 62 (2005) 175–181.
[20] J.A. Bernstein, A.B. Khodursky, P.H. Lin, S. Lin-Chao, S.N. Cohen, Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays, Proc. Natl. Acad. Sci. USA 99 (2002) 9697–9702.
[21] K. Tedin, A. Resch, U. Blasi, Requirements for ribosomal protein S1 for translation initiation of mRNAs with and without a 5′ leader sequence, Mol. Microbiol. 25 (1997) 189–199.
[22] I. Moll, A. Resch, U. Blasi, Discrimination of 5′-terminal start codons by translation initiation factor 3 is mediated by ribosomal protein S1, FEBS Lett. 436 (1998) 213–217.
[23] K. Tedin, I. Moll, S. Grill, A. Resch, A. Graschopf, C.O. Gualerzi, U. Blasi, Translation initiation factor 3 antagonizes authentic start codon selection on leaderless mRNAs, Mol. Microbiol. 31 (1999) 67–77.
[24] S. Grill, C.O. Gualerzi, P. Londei, U. Blasi, Selective stimulation of translation of leaderless mRNA by initiation factor 2: evolutionary implications for translation, EMBO J. 19 (2000) 4101–4110.
[25] I. Moll, G. Hirokawa, M.C. Kiel, A. Kaji, U. Blasi, Translation initiation with 70S ribosomes: an alternative pathway for leaderless mRNAs, Nucleic Acids Res. 32 (2004) 3354–3363.
[26] S.M. O'Donnell, G.R. Janssen, Leaderless mRNAs bind 70S ribosomes more strongly than 30S ribosomal subunits in *Escherichia coli*, J. Bacteriol. 184 (2002) 6730–6733.
[27] W.J. Van Etten, G.R. Janssen, An AUG initiation codon, not codon–anticodon complementarity, is required for the translation of unleadered mRNA in *Escherichia coli*, Mol. Microbiol. 27 (1998) 987–1001.