

24 NOVEL GENES WITH NON-CANONICAL START CODONS FOUND IN *ESCHERICHIA COLI* K12

Anton Forsberg[†], Arne Elofsson[†], Leif Isaksson^{*}

[†]*Stockholm Bioinformatics Center, Stockholm University*

^{*}*Department of Microbiology, Stockholm University, S-106 91 Stockholm, Sweden*

ABSTRACT

Escherichia coli (*E.coli*) K12 is one of the most studied organisms there is. The complete genome was sequenced for about five years ago (Blattner, et al., 1997). The genome consists of 4639221 nucleotides with 4279 annotated proteins. Almost all genes found to date have AUG, UUG, or GUG as start codon. Recent findings suggest that with the “right” sequences upstream and downstream of the start codon either the first or third base can be shifted (AUN/NUG) and the codon can still function as initiation codon (Jin, 2002).

Using different criteria concerning the upstream Shine-Dalgarno sequence (SD) and the downstream region (DR) relative to the start codon, we have been able to find 94 new sequences with a potential initiation site. These putative genes are either located in regions not yet recognized as expressed, in regions annotated as open reading framed (ORFs) or in genes with known proteins. From these 94 sequences we focused on 29 that seemed the most promising to be expressed. Among these 29 sequences, 24 sequences reside in regions with no known corresponding protein. These 24 sequences is considered new putative genes. In this report we present 3 sequences that will serve as examples for each of these categories mentioned above.

The sequences that have been found have SD and DR sequences that strongly suggests that they would be translated if mRNA is produced. If they are not translated it would be equally interesting to find out what in these sequences that suppress expression. Our conclusion is nonetheless that these sequences have a great potential to be expressed, but this remains to be confirmed.

Aside from this main problem we also analyzed the downstream region of 727 genes with the weaker (relative to AUG) start codons UUG, GUG or AUU. We found a bias for AAA, AUG, and AAU in the first five codons (considered as the DR).

Keywords: Translation, initiation, non-canonical start codons, *Escherichia coli*

INTRODUCTION

The complete genome of *Escherichia Coli* K-12 (*E.coli*) was first sequenced in 1997. (Blattner, et al., 1997). The genome consists of 4,639,221 bases and had in 1997, 4288 annotated genes, 38% of these genes had no clear function. These findings are based on the common knowledge that *E.coli*, together with other prokaryotes, uses the initiation codons AUG, UUG and GUG, (also AUU in two genes).

It has for long been known that the regions around the initiation codon can influence the initiation

process and the expression level of genes. The influence of the regions downstream and upstream can compensate for a weaker start codon (compared with AUG). Taken this into consideration could it not be possible that other start codons than the known ones could, with the “right“ sequences around it, function as start sites. If this is true it would mean that many genes have so far been undiscovered! But not for long...

I will now guide you through the mysteries of translation in prokaryotes and argue that the analysis of the *E.coli* genome that we have conducted was worth trying.

Figure 1

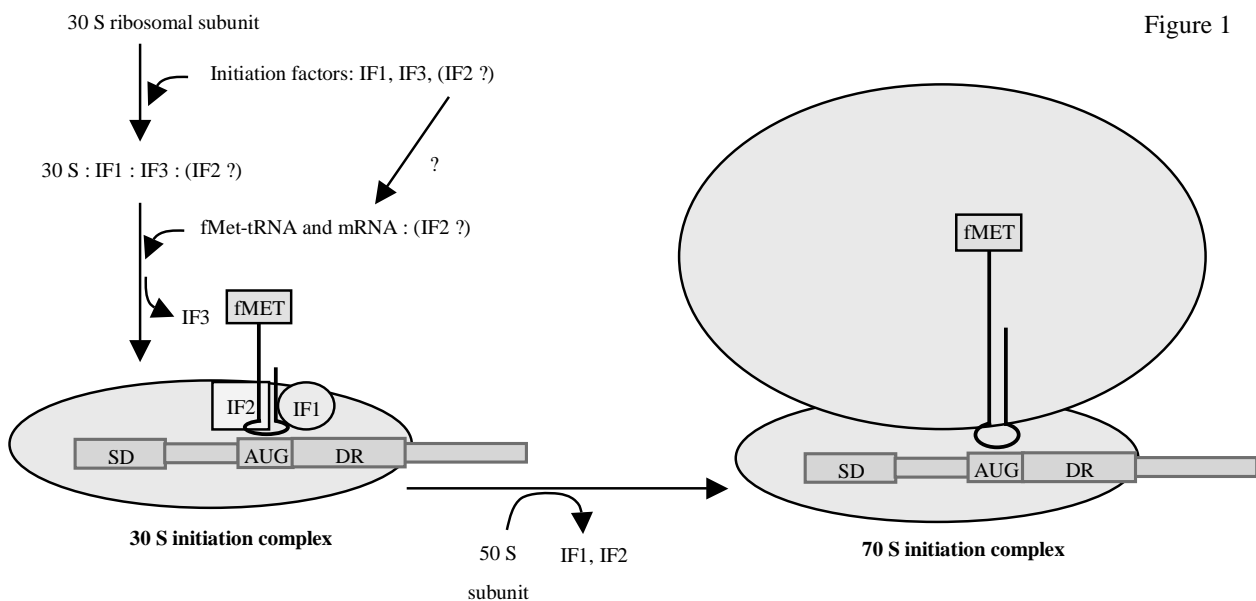


Figure 1. The initiation process starts with the formation of the 30 S initiation complex. First the three initiation factors, IF1 and IF3, forms a complex with 30S ribosomal subunit (Kozak, 1999). Whether the GTP binding protein IF2 binds the 30S before the joining of fMet-tRNA or accompanies it to the ribosome is not clear. Recent studies support the latter theory. When fMET-tRNA joins the 30S ribosomal subunit IF3 leaves the complex. The mRNA binds to the 30S subunit either before or after that fMet-tRNA has joined. Later studies suggest that the 30S ribosomal subunit uses pre-bound fMet-tRNA to select the correct mRNA start site. The complex is now referred to as the 30S initiation complex. As IF3 is gone the 50S ribosomal subunit can join the complex. When this happens the GTP bound to IF2 is hydrolyzed and IF2 leaves the complex together with IF1. This finishes the initiation process as the 70 S initiation complex has been formed

TRANSLATION

Initiation

As first step of translation a complex between the small ribosomal subunit (30S) and 3 initiation factors (IF1, IF2, IF3) is formed. This complex is then joined by initiator tRNA, and mRNA.

Initiation factor 1 (IF1) resides in the A site (see figure 2) and modulates subunit association (Dahlquist, Puglis, 2000). It might also influence tRNA selection in the P site (see figure 2) by conformational changes in the 16S rRNA. IF2 modulates subunit association by promoting dissociation of the 70S ribosome and thereby maintain a free pool of free 30S subunits (Kozak, 1999)

Initiation factor 3 (IF3) ensures that the initiation of translation starts at the right position (Meinzel, et al., 1999; O'Connor, et al., 2001). The

inspection involves both the codon-anticodon interaction, the unique features of the initiation fMet-tRNA (O'Connor, et al., 2001). IF3 also suppresses initiation at other potential start sites in the mRNA.

The initiation complex is formed at a specific site, of the mRNA, called the initiation region. The initiation region consists of at least three important parts: The Shine-Dalgarno sequence (SD), the start codon, and the downstream region (DR). All these parts will be explained later in this introduction.

The third and fourth step of the process is the joining of the mRNA and the fMet-tRNA (Kozak, 1999). The order of which these parts bind to the complex is not clear. The initiation tRNA (fMet-tRNA) is positioned at the start codon by help of IF2 and GTP. Recent studies suggests that this binding of fMet-tRNA precedes the joining of mRNA and that the pre-bound assists in the selection of a start site. IF1 and IF3 play important

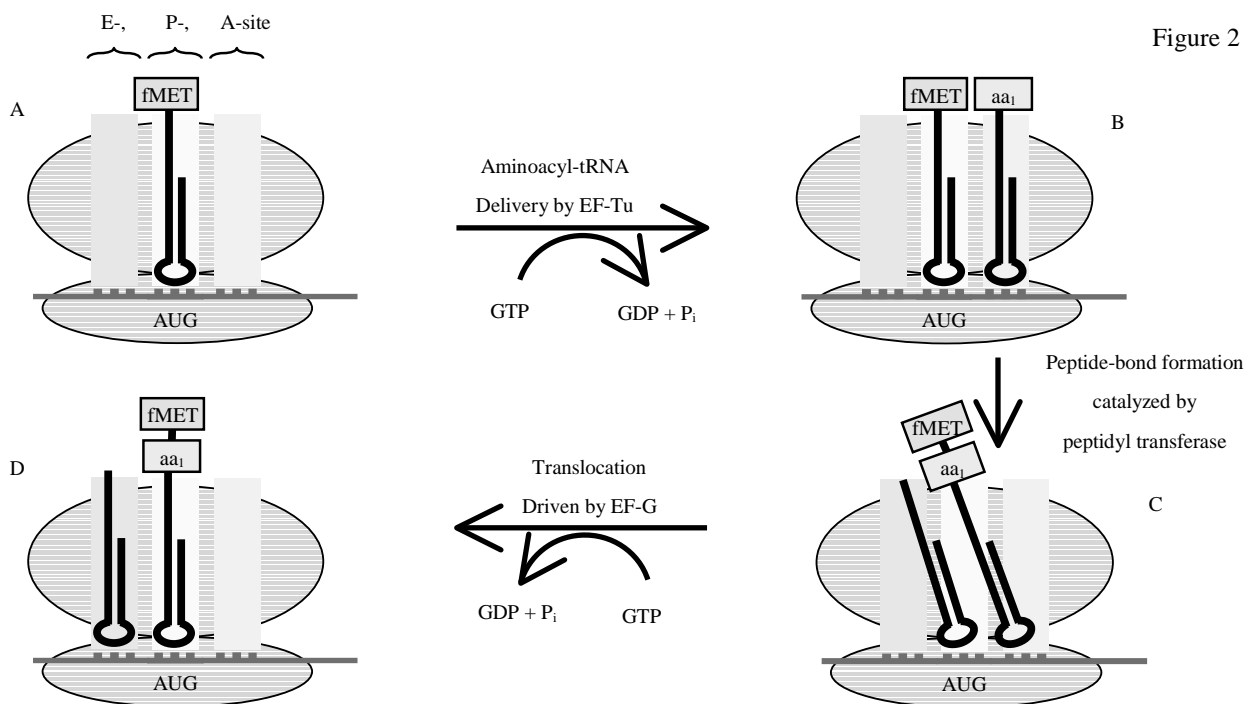


Figure 2

Stryer L., 4th ed., 1996

Figure 2. At the beginning of the elongation process the fMet-tRNA is in the P-site of the ribosome (Fig. 2A). The first step is that a new aminoacyl-tRNA binds to the mRNA in the free A-site. The aminoacyl-tRNA is delivered to the ribosome by the elongation factor Tu (EF-Tu) (Fig. 2B). After delivery the GTP bound to EF-Tu is hydrolyzed and EF-Tu dissociates from the ribosome. EF-Tu is reused by help from EF-Ts.

It is important that the “right” aminoacyl-tRNA is positioned in the A-site before peptide bond formation. This is managed by the time it takes for EF-Tu to leave the A-site. This takes longer than the time a non-complementary aminoacyl-tRNA is bound to the mRNA. Therefore, no peptide bond is formed unless it is the correct aminoacyl-tRNA. The codon-anticodon interaction is also altered when EF-Tu hydrolyzes GTP. The correct aminoacyl-tRNA will still bind strongly but not incorrect.

The next step is the peptide bond formation. The reaction is catalyzed by peptidyl transferase in the 50 S subunit of the ribosome. The activated formylmethionine unit of the fMet-tRNA_f is transferred to the amino group of the aminoacyl-tRNA. Then a dipeptidyl-tRNA is formed (Fig. 2C). The peptide bond formation alters the interactions of the both tRNAs with the 50 S subunit but not the 30 S. By this change the both tRNAs will be positioned as seen in fig. 2C. After the peptide bond formation the Translocation is the next step in the elongation cycle.

roles in the selection of the right start site on the mRNA (Meinzel, et al., 1999; O'Connor, et al., 2001).

The last step of the initiation process is the joining of the large ribosomal subunit (50S). The total complex is now called 70 S, and is now ready for the next step in the translation process, elongation.

Elongation

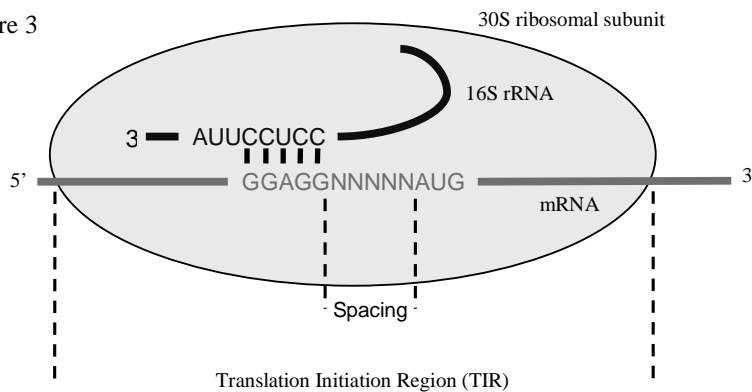
After the ribosome complex 70S has been formed, another aminoacyl-tRNA enters the A-site of the

ribosome (See figure 2). The tRNA with amino acid 1 (aa₁) is delivered to the A site by the elongation factor EF-Tu. The tRNA wanders through different states in the ribosome and simultaneously with the mRNA moves through the ribosome. During this motion new aminoacyl-tRNAs enter the ribosome.

Termination and recycling

A stop codon (UAA, UAG, and UGA) is recognized by either of two release factors (RF1,

Figure 3



Ma, J., et al., 2002

Figure 3. The core of the anti-SD region of the 16S rRNA binds to the Shine-Dalgarno sequence (SD) on the mRNA. A strong binding increases the rate and efficiency of translation. (Ma, et al., 2002).

RF2). The RF triggers the peptidyl transferase in the 23S rRNA, which hydrolyzes and releases the peptide. The complex is then dissociated by the help of ribosome recycling factor (RRF) and GTP. After this the ribosome is once again ready for a new cycle of protein production.

THE PARTS OF THE INITIATION REGION

Shine-Dalgarno sequence (SD)

The SD sequence is conserved, purine rich region 5-9 base pairs upstream of the start codon in the mRNA (Shine & Dalgarno, 1975). In the 16S rRNA there is a region (5'-ACCUCCUUA-3') that binds complementary to the SD region of the mRNA. This binding positions the mRNA in the ribosome and together with the start codon initiates the translation. The interaction between mRNA and rRNA has been confirmed in structural analysis (Yusupova, et al., 2001). The length of the SD region has been shown to influence the translation rate (longer = better) (Kozak, 1999). The core part of the SD region GGAGG is the most abundant sequence for SD regions in highly expressed genes in *E.coli* (Ma, et al., 2002). This does not entirely correspond to the consensus sequence of SD if you look at the entire genome (Fuglsang, et al., 2003).

The length of the SD is of importance for the strength of binding to the ribosome and is correlated to gene expression levels (Ma, et al., 2002). It was shown that genes characterized as highly expressed genes had a higher percentage of

SD sequences than genes expressed at a lower level.

Table 1

Strong Shine-Dalgarno sequences

- | |
|---------------------------|
| • UAAGGAGG (-12 kcal/mol) |
| • AGGAGG (-9,8 kcal/mol) |
| • AAGGAG (? kcal/mol) |
| • GGAGG (? kcal/mol) |
| • GAGGU (-6,6 kcal/mol) |
| • AGGAG (-6,5 kcal/mol) |
| • AAGGA (-5,3 kcal/mol) |
| • UAAGG (-4,2 kcal/mol) |
| • GGAG (-4,4 kcal/mol) |
| • GAGG (-4,4 kcal/mol) |
| • AGGA (-4,4 kcal/mol) |

Table 1. Strong Shine-Dalgarno sequences (SD) and their free energy (ΔG_{SD}) for duplex with anti-SD sequence AUCACCUCCUUU (Ma, J., et al., 2002).

This implies that the SD could influence the genes to be expressed at a higher level. In the same study it was found that genes with UUG or GUG as initiation codon had almost never been characterized as highly expressed and a less percentage of the genes had a SD sequence.

Also the spacing between the SD sequence and start codon is of importance for the influence the SD will have on expression rate (Schultzaberger, et al., 2001). Most SD sequences reside at an aligned

spacing of 5-13 bases from the start codon (Ma, et al., 2002).

It has been shown that translation initiation can occur even without a SD region (Moll, et al., 2002). This implies that there are other parts of the initiation region that can compensate for the loss of interaction between SD and rRNA. What other determinants for gene expression there are will be discussed below.

Start codon

The standard start codon is AUG, which forms a stable interaction with the anticodon of fMet-tRNA. Also UUG and GUG are used as start codons in ~10% of the genes (Blattner, et al., 1997). The start codon AUU is only used in two genes, *InfC* (IF3) and *pcnB* (Binns, Masters, 2002). It is interesting that initiation factor 3 (IF3, gene: *InfC*) has a “weak” start codon. IF3 is in part responsible for the discrimination between “right” and “wrong” start codons. IF3 will then have an own negative feedback of the translation of the gene *InfC*. A lot of IF3 will reduce the translation of *InfC* and decrease the production of IF3, thereby keeping an appropriate level of IF3.

Characterization of genes regarding their expression levels revealed that the highly expressed genes almost always relied on AUG as a start codon (Ma, et al., 2002). And genes with UUG and GUG were either just ORFs or found expressed at a lower level. This confirms the idea that the start codon is, in part responsible for determining the expression level of a gene.

When weaker start codons (not AUG) are used to initiate translation, the sequences around is of great importance (Stenström, et al., 2001). A truncated protein of polymerase II transcription factor E subunit α , is produced with AUC as start codon and a SD sequence upstream of AUC (Chalut, Egly, 1995). With appropriate sequences around the start codon even other non-canonical start codons (AUN, NUG) can function as initiator codons (Jin, 2002). Also the -1 position can influence the binding of mRNA to initiator tRNA and add specificity for non-standard initiator codons (Esposito, et al., 2003).

There is also results implying that initiation factor 3 (IF3) inhibits initiation at non-canonical start codons (not: AUG, GUG, UUG) (Meinzel, et al., 1999; O'Connor, et al., 2001). IF3 is thought to recognize complementary between start codon and

initiator tRNA. If no codon-anticodon complementarity is found the initiation process is stopped.

Downstream box (DB) or Downstream region (DR)

The first evidence pointing at the importance of the region downstream of the start codon came from analysis of leaderless mRNA (lacking SD sequence) (Kozak, 1999; Martin-Farmer, Janssen, 1999; Etchegaray, Inouye, 1999; Moll, et al., 2002). It was proposed that a region, downstream box (DB), complementary to a part of the 16 S rRNA (anti-DB) (Etchegaray, Inouye, 1999). Interaction between these could compensate for lack of SD or together with SD increase the level of expression. This interaction has been questioned and substantial evidence supporting the DB – anti-DB interaction is currently lacking (Moll, et al., 2001; Sato, et al., 2001). But the results that the region downstream of the start codon, in some way, influences the expression rate remain unquestioned.

Other tests has focused on the 5 codons downstream of the start codon, not as a downstream box, but more as a downstream region (DR).

Evidence for the importance of the DR region was found when different DRs were analyzed in cooperative with or without a strong SD (Stenström, C.M., et al., 2001). The results showed that a gene with a weak SD but with the “good” DR could be as efficient as a gene with a strong SD but a “bad” DR. It was also shown that the weaker initiation codons UUG and GUG (compared to AUG) could, with a good DR, give as high expression levels as AUG with another DR. It was determined that the +2-codon strongly influenced the “strength” of the DR, but did not by it self entirely determine the effect of the DR in question.

Expression levels of genes, without a strong SD, showed great differences when the +2-codon was changed (Stenström, et al., 2001). The results are, in part, showed in table 1. It clearly shows the strong influence of the +2-codon. The conclusion in the article is that that the effect seen should be associated with the decoding tRNA. They base this conclusion on the facts that the expression levels are quite similar between codons encoded by the same tRNA, and tRNA, pools, mRNA stability and

mRNA secondary structure did not seem to be what effected the expression.

In another study it was also shown that changes in the +2-codon affected the gene expression rate, both *in vitro* and *in vivo* (Sato, et al., 2001). It was in this study, contradictory to Stenström et al., not found that codons encoded by the same tRNA yielded the same expression levels. Although this study is not as complete as that of Stenström et al., the results so far are not conclusive.

When analyzing natural occurring DR regions no correlation whit known expression levels of the corresponding genes could be found (Stenström, Isaksson, 2002). Changes in the sequence of these DRs greatly influenced the expression levels. The differences in expression were shown by changing the iso-codon sequence and maintaining the produced amino-acid composition. This implies that it is the mRNA and not the amino-acids that influence the expression.

Table 2

Codon	tRNA	Expression (~)
UUU	Phe	0,32
UUC	Phe	0,18
UUA	Leu5	0,27
UUG	Leu4,5	0,13
CUU	Leu2	0,08
CUC	Leu2	0,09
CUA	Leu3	0,18
CUG	Leu1,3	0,10
AUU	Ile1	0,62
AUC	Ile1B	0,51
AUA	Ile2	0,50
AUG	Metm	0,62
GUU	Val1,2A,B	0,14
GUC	Val2A, B	0,29
GUA	Val1B	0,12
GUG	Val1	0,11
UCU	Ser1,5	0,19
UCC	Ser5	0,09
UCA	Ser1	0,29
UCG	Ser1,2	0,18
CCU	Pro2,3	0,19
CCC	Pro2	0,11
CCA	Pro3	0,05
CCG	Pro1,3	0,07
ACU	Thr1,3,4	0,07
ACC	Thr1,3	0,32
ACA	Thr4	0,58
ACG	Thr2,4	0,17
GCU	Ala1,2	0,14

GCC	Ala2	0,11
GCA	Ala1B	0,20
GCG	Ala1	0,15
UAU	Tyr1,2	0,33
UAC	Tyr1,2	0,47
UAA	Term	0,01
UAG	Term	0,02
CAU	His	0,42
CAC	His	0,34
CAA	Gln1	0,56
CAG	Gln1,2	0,24
AAU	Asn	0,74
AAC	Asn	0,61
AAA	Lys	1,0
AAG	Lys	0,47
GAU	Asp	0,23
GAC	Asp	0,21
GAA	Glu2	0,07
GAG	Glu2	0,08
UGU	Cys	0,20
UGC	Cys	0,16
UGA	Term	0,04
UGG	Trp	0,13
CGU	Arg2	0,32
CGC	Arg2	0,22
CGA	Arg2	0,27
CGG	Arg3	0,07
AGU	Ser3	0,44
AGC	Ser3	0,47
AGA	Arg4	0,67
AGG	Arg4,5	0,11
GGU	Gly3	0,14
GGC	Gly3	0,17
GGA	Gly2	0,07
GGG	Gly1,2	0,08

Table 2. The effect of different +2-codons on gene expression relative to the expression level with AAA as +2-codon (Stenström, et al., 2001).

AIM WITH OUR STUDY

The expression of a gene is tightly regulated by a number of mechanisms both prior to, and after transcription. One of the mechanisms is the interaction between mRNA and the ribosome. As previously described there are three parts of the mRNA contributing to the initiation of the translation. These regions have also been shown to influence the rate of translation. I will now describe some recent results that directly resulted in the aim of my study.

That sequences around the start codon strongly influence the expression rate has clearly been shown (Stenström, Isaksson, 2002; Stenström, et al., 2001, Stenström). The area that has yielded

much interest is the Shine-Dalgarno sequence 5-8 bases upstream of the start codon. The SD sequence strongly influences the expression rate of a gene (Kozak, 1999). Genes known to be expressed at high level usually have a stronger SD (Ma, et al., 2002). More debate has risen around the question whether the downstream region of the start codon influences the translation, or not. If the DR influences the translation, in what way does it do it? There is no doubt that the DR influences the translation rate. But the results suggesting that the influences came by complementary binding to rRNA has recently been contradicted (Moll, et al., 2001). Nevertheless a strong SD and DR are together capable of efficiently initiate translation. Genes with a weaker start codon, UUG, are usually expressed at lower levels and have a significant lower presence of strong SD regions (Ma, et al., 2002). Nevertheless, the expression levels of genes with UUG as start codon, and a “good” +2-codon was comparable to the rates of genes with AUG as start codon. With the “right” SD and DR sequences translation can even be initiated in genes with the non-canonical start codons AUN or NUG (Jin, 2002).

With this in mind it does not seem impossible that there would be genes in the E.coli genome, not yet discovered, using non-canonical start codons and with strong SD and DR compensating for the lack of a strong start codon.

This argument brings us to the aim of our study. We want to find out if there are putative genes in the E.coli genome, not yet annotated as genes, that has strong SD and DR. These sequences would then be considered to contain the signals needed to initiate translation.

MATERIAL & METHODS

PROGRAMS AND DATA BASES

The programs used for finding new genes in the genome of *Escherichia coli* K12, and analyzing codon distribution in DR, was engineered by Anton Forsberg*, in collaboration with Arne Elofsson. A simplified flowchart of the program is seen in figure 4.

* Questions regarding programs are answered by Anton Forsberg, Phone: +468-745 59 25, +4673-6976650. E-mail: antfo370@home.se.

Figure 4.

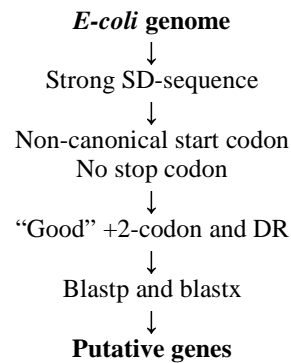


Figure 4. Flowchart of program structure. SD = Shine-Dalgarno sequence. DR = Downstream Region.

Allot of ideas for the programs were found in the book *Beginning Perl for Bioinformatics* (Tisdall, 2001).

Further studies of these putative genes were conducted using “Standard protein-protein BLAST” (blastp, ver. 2.2.2, Jan-08-2002) (Altschul, et al., 1997) and “Nucleotide query - Protein db” (blastx, ver. 2.2.2, Jan-08-2002) (Altschul, et al., 1997).

The complete DNA sequence, for *Escherichia Coli* K12, in fasta-format was downloaded from NCBI homepage. The protein databases were also downloaded from the NCBI homepage.

FINDING PUTATIVE GENES

A program was made, able to find putative genes that correspond to the criteria previously described. Below there is a short description of the functions of the constructed program (See figure 4).

First of all, the program tried to find strong Shine-Dalgarno sequences (SD) in the genome of *Escherichia coli* K12 (See table 1). When the program found a SD it took out a segment of 150 bases downstream of the SD. The program stored the, 150 base pair long, sequences with a “good” SD.

The next step in the program was to see if there was a possible start codon 4-9 bases downstream of each SD. The start codon should not be the already known start codons AUG, GUG or UUG. If a possible start codon was found, the program looked for a stop codon within the sequence. If it found a stop codon the sequence was discarded.

The sequences left all had “good” SDs, possible start codons and no stop codons. As previously argued putative genes should have a “good” +2-codon and all together favorable downstream region (DR). The program first excluded sequences with extremely bad codons (CGG, AGG, UGG, GGG). As cutoff value for the +2-codon the program used 0.5 (See table 2) (Stenström, et al., 2001).

The results were stored in a file now containing putative genes with “good” SD, +2-codon and DR.

CODON USAGE DOWNSTREAM OF START CODON

The analysis of the downstream regions of genes with GUG, AUU or UUG included 727 genes. The sequences of the genes were extracted from a database downloaded from NCBI-homepage. Analysis of the genes was conducted with a program that extracted the codons from the +2-codon to the +16 codon of all 727 genes. The frequency of the different codons were then analyzed and graphs were made using OpenOffice.org 1.0.

RESULTS

NON-CANONICAL START CODONS

Our initial search for strong Shine Dalgarno sequences (AGGAGG, AAGGAG, AAGGA, AGGAG, GGAGG) rendered in enough positive hits to stay at this fairly stringent criteria for the SD sequence. The specific length of the sequences (150 bases) was chosen because this is a sufficient length to give functional proteins. Looking at longer sequences from the beginning could have sorted away putative genes.

By neglecting genes with extremely bad codons (CGG, AGG, UGG, GGG) in the DR, false positive sequences that most likely would not be translated in the cell were discriminated (Stenström, et al., 2001).

Further studies of the DR sorted out genes with a “good” downstream region. The cutoff values 0,5 for the +2-codon and 1,5 for the entire 5 codon long DR were used to further narrow down the search (See table 2 for details about the values for each codon). From a huge number of sequences with strong SD sequences, 54 sequences in the

forward frame and 40 sequences in the reverse/complement frame fulfilled the search criteria.

The sequences were further studied using blastp and blastx. This analysis showed if the sequences corresponded to already annotated genes and/or proteins. Based on these results, 17 sequences in the forward frame and 12 in the reverse frame were chosen for further studies. Of these 29 sequences 7 are located in areas categorized as not expressed. 17 sequences are positioned in regions that are annotated as open reading frames (ORFs). Then there are 5 sequences that lie entirely or partly inside genes with annotated proteins.

Three sequences are presented as examples of the discoveries. These sequences have different localization in the genome, and are good representatives for the other sequences found. The first sequence described clearly lies outside any annotated gene, then one sequence that lie inside an ORF. Finally we will present a sequence that is located in a gene with an annotated protein.

ID: 4435849

This sequence starts with a strong SD sequence (AGGAGG) with an aligned spacing of 7 bases (middle of SD to start codon) (See table 3). After the start codon, AUU, there is a perfect +2-codon (AAA). The DR region overall has a total value of 1,53. The total length of the sequence until the stop codon is 363 bases. The region downstream of the start codon does not show any evident complementarity to the SD sequence that would imply that the SD sequence would become hidden from the ribosome.

ID: 5724

Start codon in this sequence is AUC spaced 9 bases from the core of the SD sequence GGAGG (See table 3). The total length of the sequence until a stop codon is found is 210 bases. The overall value of the DR is 1,69 with AAA as the +2-codon. Downstream of the start codon the sequence CCTTC is found. This region might bind to and hide the SD region. The translation initiation might then be influenced negatively.

ID: 2041329

Last of the chosen sequences is a putative gene of 147 nucleotides, located within the annotated gene *dapA* (See table 3). The SD sequence is AAGG,

with an aligned spacing of 7 bases to the start codon AUA. The DR has a high overall value of 3,24 with perfect AAA codons in position +2 and +3. There is nothing downstream of the start codon that implies a complementary binding and hiding of the SD region.

CODON DISTRIBUTION

The results from the analysis of the downstream region of the start codon are shown in figure 5. The genes analyzed are 727 genes with UUG, GUG or AUU as start codon. The graphs illustrate the number of genes with a specific codon in each position. The frequency of all codons has been measured. 16 codons out of 64 are shown in figure 5. These include the ones that vary between the positions and codons that don not vary for comparison.

There is a clear overrepresentation of the codons AAA, AAT, ATG, and ATT in the first five codons of the genes (See figure 5).

DISCUSSION

PUTATIVE GENES

The number of interesting genes found exceeded our expectations and shows that there is a great chance that there are a number of genes that have been overlooked with the criteria used today to find open reading framed and new genes. That non-canonical start codons can be used by the ribosome to initiate translation has clearly been shown (Jin, 2002). With our approach we have found 94 sequences that fulfill the criteria, strong SD sequence, favorable +2-codon and overall “good” DR.

Bearing in mind the results from studies, of the great influence on expression rates from SD and DR, it would be a mystery why sequences that appear to fulfill all criteria would not be expressed.

Possible expression in new areas

7 sequences has been found in the genome of *E.coli* K12 that fulfill our stringent criteria of strong SD sequence, “good” +2-codon and DR. All the SD sequences we have used in our search represent strong SD’s (Ma, et al., 2002).

In the example we have chosen (ID: 4435849, See table 3) the sequence has the possible start codon

AUU. This is not a novel start codon but has previously only been described to act as initiator codon for two genes, *infC* and *pcnB*.

In our sequence the +2-codon is AAA, which has been proven to be the most efficient codon to increase the expression rate (Sato, et al., 2001; Stenström, et al., 2001). The overall value of the DR shows a normal distribution of codons. The influence of the codons in positions +3 to +5 is not fully understood. Some codons influence the expression rate in the same way (although to less extent further away from the start codon) regardless the position (Stenström, Isaksson, 2002). But some codons can change their effect on the expression when the position of the codon is changed. It is therefore hard to clearly evaluate the complete DR in this case.

The SD sequence, AAGGAGG, is considered a very strong SD sequence (see table 1) that clearly would favor the binding of the ribosomal subunit 30S (Kozak, 1999; Ma, et al., 2002). The presence of such a strong SD sequence should alone suggest that an initiation complex could form around this sequence.

Taken all these facts together it looks like this sequence could be expressed in *E.coli*. There are also other sequences (data not shown) that seem as promising as the one presented in this report.

Nevertheless further studies need to be done to see that nothing in the sequence or around it could render in problems for transcription and/or translation. The problems could be unfavorable secondary structure, or downstream SD sequences that could inhibit ribosome binding (Jin, 2002). As a final step, to see if this putative gene is expressed or not, mRNA-probes could be used to see if there is any mRNA matching this sequence in *E.coli*.

New genes within known ORF’s

17 of the sequences found to be promising resided in parts of the *E.coli* genome annotated as open reading frames (ORF’s). There has not been any protein found that matches these regions. There is simply just a start codon and a long enough sequence without a stop codon to imply that it might be a gene. This on the other hand does not mean that the annotation might be wrong. There for it is just as interesting to find new putative start sites within an ORF then outside any annotated region.

The sequence we show as example (ID: : 5724) has AUC as start codon. AUC has been shown to function as start codon with the “right” sequences around it (Jin, H., 2002; Chalut, C., Egly, 1995). The SD region, GGAGG, is thought to be a strong one that would bind to the 16S ribosomal subunit in the initiation complex (Ma, et al., 2002). The DR region appears to be favorable for translation with AAA in the +2 position. These facts led to the conclusion that this sequence should be expressed. As for the previously described sequence, ID: 4435849, there are questions regarding sequences around and secondary structure of the mRNA that need to be addressed. The sequence CCTTC inside the downstream region could bind to the SD and inhibit ribosome binding. Probes against the mRNA can tell if the putative gene is transcribed or, not.

It is not unlikely that what is now considered an ORF might still come out to be expressed but not with the start codon described in current databases but with a non canonical start codon like AUC in our presented sequence.

Expression of truncated proteins

The last type of position that our proposed initiation regions have is inside genes with known protein products. If these sequences were found to be expressed they would produce a truncated protein. These proteins might have a completely different tertiary structure than the whole-length protein but might also be a copy of a part of the larger protein in question.

The sequence we use as an example (ID: 2041329) has AUA as suggested start codon, a quite strong SD (see table 1), and a very good DR with AAA codons in both the +2 and +3 position. There is strong evidence supporting that this sequence might be expressed.

SELECTION OF START CODON

Which parts of a mRNA that will be translated is determined by the possibility of a specific part of the mRNA to bind to the ribosome. This binding must also be able to start the binding of the factors used in translation. One of the processes that discriminate “wrong” sequences from “right” is the binding of the initiation fMet-tRNA (O’Connor, et

al., 2001; Meinnel, et al., 1999). This discrimination involves base pairing between the fMet-tRNA and the start codon with assistance of initiation factor 3 (IF3). The initiation factor 1 (IF1) may play an additional role in this discrimination (Dahlquist, Puglis, 2000). The sequences that we have found all show favorable SD and DR regions (See table 3 for examples). The discrimination machinery described is not fully understood, but could be involved stopping translation initiation of our sequences. If the gene for IF3 is lacking, genes with non-canonical start codons can be expressed (Meinnel, et al., 1999). But also with the gene for IF3 intact non-canonical codons can be used to initiate translation if the SD and DR are “good” enough (Jin, 2002). Why is GUG used as start codon and not CUG? This question has not been answered it has been shown that also CUG can function as start codon with strong SD and DR (Jin, H., 2002). Maybe there is no answer to that question, and there are genes with CUG, and other non-canonical start codon, as start codon. Our results clearly point in that direction.

Based on the results from our study and compared with sequences found to be translated *in vitro* I can not find anything that would predict that initiation of translation would not occur on mRNA containing our sequences (see table 3 for examples of our results and Jin, for comparison).

Translation initiation can be obstructed by unfavorable secondary structure or discrimination against the non-canonical start codon. These processes are difficult to predict and only further studies in the lab can give the final answers.

CODON DISTRIBUTION

The result that the codons AAA, AAT and ATT are over represented in the +2-codon (See figure 4) is in concord with previous results (Stenström, et al., 2001). This overrepresentation continues for a few codons, and then declines to normal levels. It seems that evolution has favored codons that increase the rate, and efficiency of expression. If the presence of “good” codons in DR helps the genes with weaker start codons to be expressed at sufficient levels is harder to determine.

Table 3

A

Identification(ID): 4435849

aggaggtggaattaaagaatgcgatggatgggtatatttacgaaaaataattgatgataatttgccatagtcataagaatctttgggc
 ttgcgcttaaaatctatggaggggatgaacgttttcgtaatggttcctctgttgttttggga

Gene position(P): 4435869..4436222, **SD-Start-spacing(S):** 4, **Count(C):** 13, **DR-value(V):** 1.53, **Gene length(L):** 363

Blastx:	gi 1786858 gb AAC73740.1	orf, hyp prot	[E.coli...	25	1.3	Frame(F): +2
Blastp:	gi 1786858 gb AAC73740.1	orf, hyp prot	[E.coli...	25	1.2	
Upstream:	4435285..4435785	+ 167	16132037	ytfI	b4215	orf, hyp prot
Downstream:	4436285..4436839	- 184	16132038	ytfJ	b4216 R COG3054	orf, hyp prot

B

ID: 5724

ggaggaatcttcatcaaagaagtaaccttcgctattaaaaccagtcagttgctctgggttgggtcagccgatcttcaataatgaaacga
 ctcatcagaccggtgctttcttagcgtagaagctgatgatcttaaatttggcggttctctc

P: 5736..5946, **S:** 7, **C:** 15, **V:** 1.69, **L:** 210

Blastx:	gi 1786187 gb AAC73117.1	orf, hyp prot	[E.coli...	103	2e-24	F: -1
Blastp:	No hits found					
Inside:	5683..6459	- 258	16128000	yaaA	b0006 S COG3022	orf, hyp prot
Upstream:	5234..5530	+ 98	16127999	-	b0005 - -	orf, hyp prot
Downstream:	6529..7959	- 476	16128001	yaaJ	b0007 E COG1115	inner membr transp prot

C

ID: 2041329

aaggaaagcataaaaaaaaaacatgcatacaacaatcagaacggttctgtctgcttggcttttaatggccataccaaacgtaccattgagac
 acttgtttgacagaggatggcccatgttcacggaagtattgtcgcgattgttactccgat

P: 2597736..2597883, **S:** 4, **C:** 12, **V:** 3.24, **L:** 147

Blastx:	gi 1788823 gb AAC75531.1	dihydrodip synth	[E.coli...	27	0.26	F: +2
Blastp:	gi 1788937 gb AAC75636.1	orf, hyp prot	[E.coli...	22	5.7	
Inside:	2596902..2597780	- 292	16130403	dapA	b2478 E COG0329	dihydrodipicoli synthase
Upstream:	2595851..2596888	- 345	16130402	nlpB	b2477 M COG3317	lipoprot-34
Downstream:	2597860..2598498	+ 212	16130404	gcvR	b2479 E COG2716	transcr regul gcv operon

Table 3A-C. In 3A-C 3 examples of new possible genes in *E.coli* K12 are shown. Start codons are marked in red and all AUG, GUG and UUG are marked in green regardless of frame. The sequence in 3A, is in forward frame, and lies outside any known gene or open reading frame (ORF). The total length of the sequence until stop codon is 363 bases. The SD sequence is AGGAGG with an aligned spacing of 7 bases to the start codon AUU. +2 codon is the favorable AAA-codon.

In 3B the sequence, is in forward frame, and is poisoned inside an ORF in the *E.coli* K12 genome. The SD is GGAGG with an aligned spacing of 9 bases to the start codon AUC and the total length from start to stop is 210 bases. As +2 codon is also here AAA.

The sequence in 3C is, in reverse/complement frame, and positioned inside an annotated gene. The SD sequence is AAGGA, with an aligned spacing of 7 nucleotides to the start codon AUA. The length of the sequence is 147 bases until stop codon. The DR is really strong with the optimal start codon AAA in both +2 and +3 position.

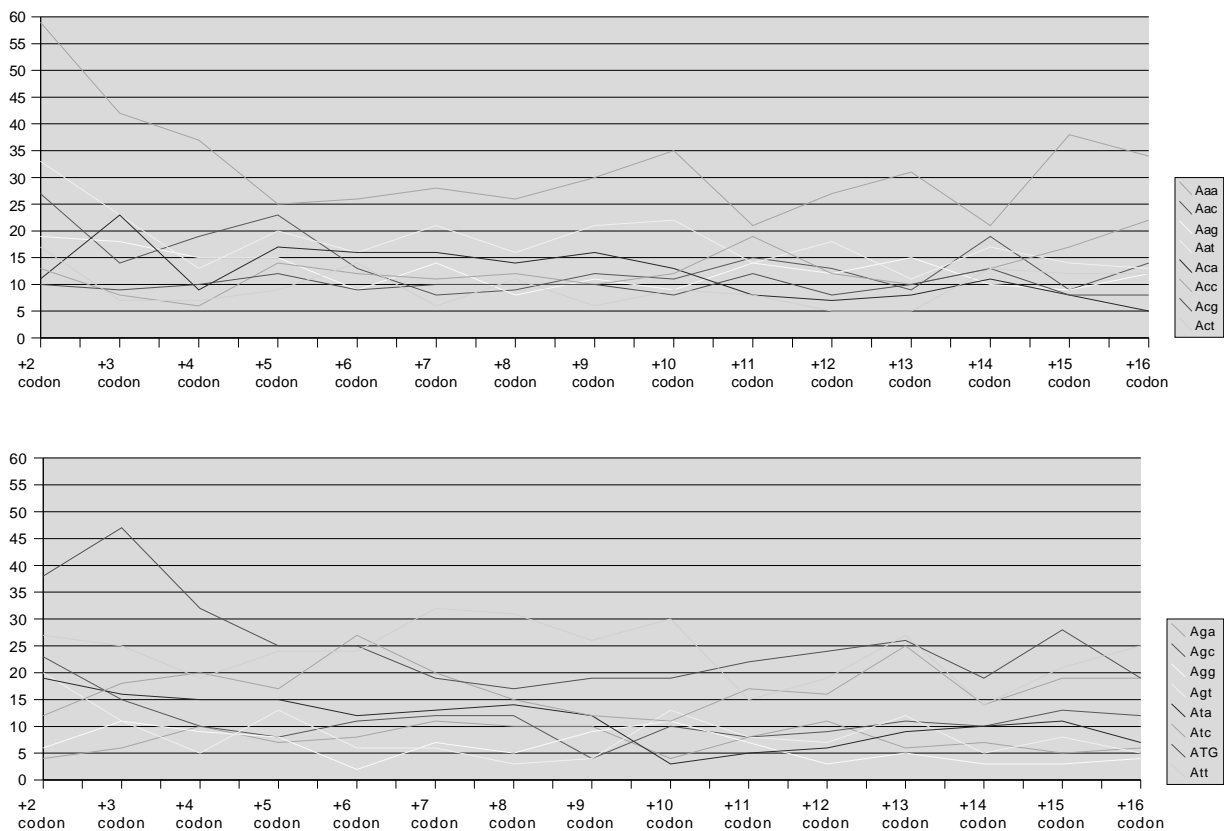


Figure 5. Results from analysis of the region downstream of the start codon in 727 genes with UUG, GUG, or AUU as start codon. The y-axis represent the number of genes with the particular codon in the position showed on the x-axis. There is an over representation of the codons AAA, AAT, ATG and ATT in the positions +2 to +5, in relation to the codons further downstream.

Genes with GUG, UUG, AUU, as start codon, are to less extent, then genes with AUG as start codon, categorized as highly expressed genes (Ma, et al., 2002). The SD sequences, in genes that in vivo are expressed at lower levels, are not as strong as in highly expressed genes. These facts taken together, sheds light on the effects that the regions around the start codons have. Nevertheless it is hard to say whether these genes are expressed at a lower level because of the weaker start codon, shorter SD sequence or lower frequency of favorable codons in DR.

An interesting remark is that there are about 40 genes out of 727 that have AUG in the +2 position and about 45 that have AUG in the +3 codon. These regions clearly could be used to initiate translation. If these genes are wrongly annotated is not for me to answer but since the efficiency to

initiate ribosome binding is much higher for AUG the question is justified (O'Donnel, Janssen, et al., 2000). It might be so that it is really the non-AUG codon that is used. If that it is the case it would be interesting to investigate what makes the ribosome choose a weaker start codon instead of AUG.

CONCLUSION

The common view of start codons in *E.coli* has been that AUG is the major initiator codon and that GUG and UUG is used in about 10% of the genes (Blattner, et al., 1997). It is also known that AUU is used in two genes, *infC* and *pcnB*.

With our approach to focus our attention to the sequences around the possible start codon we have found 94 sequences that could attract the initiation complex. Of these we looked closer at 29 extra

promising sequences. All knowledge until today implies that these sequences should be expressed, and if they, for some undiscovered reason, are not expressed it would be most interesting to find out why. Further studies of these putative genes need to be done. And finally mRNA-probes can be used against the most promising ones to find out if they truly are novel genes.

If the sequences we have found produce protein *in vivo* or *in vitro*, it would mean that many genes are yet to be discovered. Are the sequences not translated it would be most interesting to find out what obstructs the translation. Nevertheless our findings greatly influence our view of translation initiation.

ACKNOWLEDGMENTS

This work has been done in collaboration with Arne Elofsson at Stockholm Bioinformatics center (SBC), Stockholm University (SU) and Leif Isaksson at the Department of Microbiology, SU. The project has been done within the Stockholm Graduate School of Molecular Life Sciences.

REFERENCES

Altschul, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*, 1:25:17:3389-402, 1997

Binns, N., Masters, M., Expression of the *Escherichia coli* *pcnB* gene is translationally limited using an inefficient start codon: a second chromosomal example of translation initiated at AUU, *Mol Microbiol*, 44:5:1287-1298, 2002

Chalut, C., Egly, J-M., AUC is used as start codon in *Escherichia coli*, *Gene*, 156:43-45, 1995

Esposito, D., Fey, J.P., Eberhard, S., Hicks, A.J., Stern, D.B., *In vivo* evidence for the prokaryotic model of extended codon-anticodon interaction in translation initiation, *The EMBO Journal*, 22:3:651-656, 2003

Fuglsang A., Engberg, J., Non-randomness in Shine-Dalgarno regions: links to gene characteristics, *Biochemical and Biophysical Research Communications*, 302:296-301, 2003

Jin, H., Functional Studies of mRNA in Translation Initiation and Termination in *E.coli*, *Dissertation thesis*, Stockholm University, Ackademityrck, Sweden, 2002

Ma, J., Campbell, A., Karlin, S., Correlation between Shine-Dalgarno sequences and gene features such as expression levels and operon structures, *Journal of Bacteriology*, 184:20:5733-5745, 2002

Marin-Farmer, J., Janssen, G.R., A downstream CA repeat sequence increases translation from leadered and unleadered mRNA in *Escherichia coli*, *Molecular Microbiology*, 31:4:1025-1038, 1999

Meinzel, T., Sacerdot, C., Graffe, M., Blanquet, S., Springer, M., Discrimination by *Escherichia coli* initiation factor IF3 against initiation on non-canonical codons relies on complementarity rules, *J. Mol. Biol.*, 290:825-837, 1999

Moll, I., Grill, S., Gualerzi, C.O., Bläsi, U., Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control, *Molecular Microbiology*, 43:1:239-246, 2002

Moll, I., Huber, M., Grill, S., Pooneh, S., Mueller, F., Brimacombe, R., Londei, P., Bläsi, U., Evidence against an interaction between the mRNA downstream box and 16S rRNA in translation initiation, *Journal of Bacteriology*, 183:11:3499-3510, 2001

O'Connor, M., Gregory, S.T., Rajbhandary, U.L., Dahlberg, A.E., Altered discrimination of start codons and initiator tRNA by mutant initiation factor 3, *RNA*, 7:969-978, 2001

Sato, T., Terabe, M., Hidemi, W., Gojobori, T., Hori-Takemoto, C., Miura, K., Codon and base biases after the initiation codon of the open reading frames in the *Escherichia coli* genome and their influence on the translation efficiency, *J. Biochem.*, 129:851-860, 2001

Shultzaberger, R.K., Bucheimer, R.E., Rudd, K.E., Schneider, T.D., Anatomy of *Escherichia coli* ribosome binding sites, *J. Mol. Biol.*, 313:215-228, 2001

Stenström, C.M., Holgren, E., Isaksson, L.A., Cooperative effects by the initiation codon and its flanking regions on translation initiation, *Gene*, 273:259-265, 2001

Stenström, C.M., Isaksson, L.A., Influences on translation initiation and early elongation by the messenger RNA region flanking the initiation codon at the 3' side, *Gene*, 288:1-8, 2002

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

Tisdall, J., *Beginning Perl for Bioinformatics*, O'Reilly, UK, 2001

Yusupova, G.Zh., Yusupov, M.M., Cate, J.H.D., Noller, H.F., The path of messenger RNA through the ribosome, *Cell*, 106:233-241, 2001