The EMBO Journal vol.7 no.12 pp.3957–3962, 1988

Messenger RNA recognition in *Escherichia coli*: a possible second site of interaction with 16S ribosomal RNA

G.B.Petersen, P.A.Stockwell and D.F.Hill

Department of Biochemistry, University of Otago, Dunedin, New Zealand

Communicated by G.Brownlee

Examination of the nucleotides following the ATG or GTG initiation codons of a file of 251 genes from *Escherichia coli* has shown that 247 (98.4%) of them contain a sequence of at least three and 168 (66.9%) of them a sequence of at least four consecutive nucleotides that is complementary to some part of the 16 nt at the 5′ terminus of the bacterial 16S rRNA. It is proposed that this sequence, which falls within the first 24 nt coding for the genetic message, might be involved in mRNA recognition through a mechanism analogous to the well-established 'Shine–Dalgarno' interaction with the 3′ terminus of the 16S rRNA. Comparison of these data with data derived from a file of 117 'false' gene starts that have a Shine–Dalgarno-like sequence followed by a suitably spaced ATG or GTG triplet but which are believed not to lie at the beginnings of genetic messages shows the association that we have found to be statistically significant at the 99.9% level.

Key words: Escherichia coli/mRNA/prokaryotic genes/rRNA/translational initiation

Introduction

The availability of successful techniques for the rapid determination of nucleotide sequences in DNA molecules has led to an increased need for the development of reliable methods for the conversion of 'raw' DNA sequence data into genetic maps. In the absence of direct evidence from protein primary structure, it is often possible to make predictions about the genetic activity of an unknown DNA sequence from a study of the sequence itself (Staden, 1985). Steitz (1979) has reviewed the experiments and thinking that followed the observation by Shine and Dalgarno (1974, 1975) that the initiation codons of known genes in the *Escherichia coli*/coliphage system are preceded by a short sequence (SD sequence) of nucleotides that is directly complementary to some part of the -ACCUCCUUA-OH nucleotide sequence lying at the extreme 3′ terminus of *E.coli* 16S rRNA. The present view of the mechanism of initiation of mRNA translation in prokaryotic systems is that a preliminary interaction between the 16S rRNA sequence and the mRNA 'points' a 30S ribosomal subunit towards an AUG or GUG triplet lying 5–9 nt 'downstream' (i.e. 3′) in the mRNA sequence as the first codon in the genetic message and thus allows the formation of a properly oriented 30S subunit–mRNA–fmet tRNA complex. The initiation complex is completed by the addition of the 50S ribosomal unit to give an active, in-phase ribosome. Thus, in considering a new DNA sequence, it is a simple matter to translate the sequence in all three phases and to identify open reading frames (ORFs) that might code for viable proteins *in vivo*. This usually involves the selection of ORFs of plausible length (i.e. that are capable of coding for a peptide of greater than a specified minimum size) that contain an initiation codon that is preceded by a suitably spaced SD sequence. If the sequences of one or two genes in the system are already known, a comparison of the codon usage of an unknown nucleotide sequence can provide strong supporting evidence for the assignment of genetic function to a new sequence and a number of computer-assisted methods exploiting this and other approaches have been described (Staden, 1985).

In many cases, this approach gives convincing results, but the difficulties experienced with the interpretation of the sequence of the DNA of bacteriophage λ (Sanger *et al.*, 1982) highlighted the uncertainties involved. Application of the methods outlined above to the λ sequence allowed 46 ORFs to be identified, but in ~50% of these cases the initiating codon could not be identified with certainty, while a further 20 ORFs that do not meet the usually accepted criteria can be distinguished in the sequence. Steitz (1979) drew attention to the likely importance of mRNA secondary structure and of protein–RNA interactions (possibly involving nucleotides lying downstream of the initiation codon as well as upstream sequences) in enhancing the specificity of the interaction and in stabilizing the ribosome–mRNA–tRNA initiation complex.

In a critical and original review, Gold *et al.* (1981) discussed the evidence that an SD sequence, an initiation codon and a suitable spacing between the two are sufficient to specify the translational start of a coding sequence. The evidence is convincing that they are necessary, but equally convincing that they are not sufficient. Their own studies and those of others (discussed in Gold *et al.*, 1981) support the view that other upstream sequences, distinct from the SD sequence, are probably involved. More strikingly, downstream sequences (i.e. nucleotides lying within the amino acid coding sequence itself) are also implicated in the initial recognition of translational start codons by the 30S ribosomal subunit.

The evidence for the participation of downstream sequences, summarized by Gold *et al.* (1981), is based mainly upon statistical analyses of coding sequences. In an earlier study, Scherer *et al.* (1980) scored the occurrence of nucleotide triplets in *E.coli* ribosome binding sites and deduced a 'consensus' sequence that is optimal for the binding of ribosomes to mRNA in this system. This sequence included 25 nucleotides upstream of the initiating codon and ran some 20 nucleotides downstream, into the coding region itself. Gold and his colleagues (Gold *et al.*, 1981; Stormo *et al.*, 1982a) used a statistical analysis to obtain a measure of the significance of any one nucleotide occupying a position from −25 to +18, relative to the initiation codon (these workers called the first nucleotide of this codon position 0). They compared a series of known 'gene starts' from *E.coli*, coliphages and *Salmonella typhimurium* with a similar series of 'false' starts, constructed by selecting ATG or GTG codons known to be located within genes but fortuitously preceded by a suitably spaced nucleotide sequence of the SD type. Their analysis revealed several regions—three (including the SD sequence) lying upstream from the initiation codon and two lying downstream—in which the nucleotide sequence of the known gene starts differed significantly from those of the control selection of false sequences. One of the downstream sequences was centred around nucleotides 3–6 of the coding sequence, the second at nucleotides 10–13. These data were used to develop a 'perceptron' algorithm for the identification of likely gene starts in an unknown DNA sequence (Stormo *et al.*, 1982b).

There are two ways by which a sequence of nucleotides might be 'recognized' by a ribosome in setting up the mRNA–30S subunit–fmet tRNA initiation complex: by protein–RNA interaction or by RNA–RNA interaction. Although the first of these is likely to be an important mechanism in the overall process, we have in this study concentrated on the possibility that RNA–RNA interaction might also be involved in the recognition of some, at least, of the downstream sequences and have used this as a working hypothesis. If this hypothesis is correct, then there are only two RNA species—fmet tRNA and 16S rRNA—that can be considered as candidates for this interaction. Clear precedent for the involvement of the 16S rRNA is already provided by the Shine–Dalgarno hypothesis (Shine and Dalgarno, 1974, 1975) and we have therefore looked in detail at the possibility that the 16S rRNA is involved in a second interaction with nucleotides of the mRNA sequence that lie downstream from the initiating codon.

Results and discussion

Preliminary studies showed that an unexpectedly large number of nucleotide sequences lying in positions +3 to +18 of *E.coli* genes showed a direct complementarity to five or more consecutive nucleotides from some part of the 16S rRNA sequence. Four regions, comprising nucleotides 1–18, 49–77, 79–99 and 105–132 of the 16S RNA, stood out as candidates for investigation. Closer examination of short lists composed of gene starts known to be the first translated genes after a promoter (and therefore presumed to be efficiently used ribosome binding sites) focused our attention on nucleotides 1–18 (i.e. the extreme 5′ terminus) of the 16S RNA. On the basis of the examination of a file of 251 gene starts we now propose that, in addition to the upstream SD sequence (complementary to the 3′ terminus of the 16S RNA), *E.coli* genes usually contain a downstream sequence, lying in the region containing nucleotides +4 to +21, that is complementary to three or more consecutive nucleotides of the first 16 nucleotides from the 5′ terminus of the 16S rRNA. Stated more precisely: most *E.coli* gene starts contain at least three consecutive nucleotides from the string TCAAACTCTTCAATTT within the first 21 nucleotides of the coding sequence immediately following the initiation codon. A list of the contents of our database, arranged to emphasize these observations, is given in Appendix 1 and some representative examples of the evidence supporting this proposal are given in Tables I and II. Table I lists 10 examples of *E.coli* gene starts selected from list (a) of Appendix 1 that show homology of four or more consecutive nucleotides to this complementary sequence. The SD sequence and the new, downstream, sequence are shown in upper case. To make it easier to see the limits to the downstream sequences, the file has been arranged in increasing order of spacing from the initiating codon. Nucleotides that can extend the SD and downstream sequences if G:U base pairing is allowed are underlined. Table II lists a further 10 gene starts from the 79 of list (b) in Appendix 1 that show complementarity to only three consecutive nucleotides of the 16S RNA sequence. Many of the entries in list (b) of Appendix 1 have more than one such triplet and in the examples given in Table II only the first triplet occurring from nucleotide +4 onwards is marked. Again, many of the sequences in this file can be extended by invoking G:U base pairing and the additional nucleotides that will then take part in the overall interaction have been underlined. The four gene starts in our database that do not satisfy the criteria of either Table I or Table II are listed in part (c) of Appendix 1.

The boundaries of the downstream sequence

The formation of an initiation complex requires interaction between mRNA, fmet tRNA and 16S RNA. In general, the nucleotides involved in the SD interaction are separated from the initiation codon by at least five nucleotides, a spacing that is, perhaps, dictated by steric considerations and the need to provide 'room' for the fmet tRNA–initiation codon interaction. Instinctively, it might be thought that a downstream interaction with the 16S RNA would involve a similar spacing from the initiation codon. The data of Table I suggest that this is not necessarily the case. It is not possible to be definite about this since in a number of cases in Appendix 1 where a sequence of at least four or five complementary nucleotides starts immediately after the initiation codon another sequence of at least four or five consecutive nucleotides can be found further downstream. At this stage, however, there would seem to be no reason to exclude the early nucleotides in the gene from participating in the interaction and some of the examples in Tables I and II have been selected to display this. The question of how far the downstream sequence extends into the coding sequence must also be addressed. For all of the sequences listed in Appendix 1(b), the match of three nucleotides lies completely within the range +4 to +18. Gene *hisreg* (see Table I) has a match of eight consecutive nucleotides comprising nucleotides +11 to +18. Similarly, gene *rplJ* has a match of seven consecutive nucleotides terminating at the same position. Although in this case there are several possible triplets lying upstream, the interaction of these will be energetically much less favourable for helix formation and we would argue that the 7-nt interaction will be the one that is actually used. The sequence for gene *dnaB*, however, provides support for extending the likely region further into the coding sequence since a much more favourable interaction can be found with the sequence CTTCAA (ΔG^0_{37} = −6.4 kcal/mol) lying further downstream than with the sequence AAAC (ΔG^0_{37} = −2.6 kcal/mol) that we have highlighted in Table I. The last sequence in Table I extends to nucleotide +21. Although triplet interactions further upstream could be formed by this sequence, the pentanucleotide interaction that we have highlighted is

Table I. Examples of gene starts from *E.coli* that obey the new rule and have four or more downstream nucleotides complementary to some part of nucleotides 1–16 of the 16S rRNA

aacgagcataaacAGGAtcgccatc	atg	CAAAaagacgcgctgaataacgtac	*aroF*
aaaatcaggcacAGGcagaacaaca	atg	aTCAAggcgacggacagaaaactgg	*ftsA*
agattgctattttttGGAGtcata	atg	gatTCTTgtcataaaattgattatg	*hlyB*
aagcaacaaatttctGAGacttgta	atg	aacagAACTgacgaactccgtactg	*aroH*
tgaataaacattcacaGAGactttt	atg	acacgcgTTCAATTTaaacaccacc	*hisreg*
tggcaaacatccAGGAGcaaagcta	atg	gctttaaaTCTTCAAgacaaacaag	*rplJ*
ctttatctcGGTaactccattcact	atg	gcaggaaatAAACccttcaacaaac	*dnaB*
taaagtcctcgcgtacgAAGtgcgt	atg	ttgatgaccgaCAAACTgcaactgg	*envZ*
tactgaacgagAAGGcgggtgcgta	atg	accgataaaatccgtACTCTgcaag	*rpsQ*
agcaaacataagAAGGggtgttttt	atg	tcatccgatattaagaTCAAAgtgc	*mtlA*

The SD sequences and the downstream sequences are shown in upper case. Nucleotides that extend the SD and downstream sequences if G:U base pairing is allowed are underlined. The gene name used in the EMBL database is given in the right hand column.

Table II. Examples of gene starts from *E.coli* that obey the new rule but have only three downstream nucleotides complementary to some part of nucleotides 1–16 of 16S rRNA

gctatccttaaccAGGgagctgatt	atg	AAAaaagccacatgcttaactgacg	*ada*
gcctaaaagataaacGAGGaaacaa	atg	gCTCgtacaacacccatcgcacgct	*ef-g*
ttgttagctgagtcAGGAGatgcgg	atg	ttAAAgcgtgaaatgaacattgccg	*glyA*
gacttgcaatatAGGAtaacgaatc	atg	gcaCAAgtcattaataccaacagcc	*hag*
gacacattttAAGGggattttcgca	atg	cgtaTCAttctgcttggcgctccgg	*adk*
tttgcggtctggtgtGAGGTttacc	atg	agtggATTacgtccggcattatcaa	*kdpC*
gaggcgttagccacAGGAGGgatct	atg	tccgggTTTtatcataagcatttcc	*argI*
cacatagccagtaGAGtcaggactg	atg	aagacgttaTCTcccgctgtgatta	*pabB*
ttactgatatgaaaaGAGtttaaca	atg	cagcagttacagAACattattgaaa	*dapD*
acatattaaatagtAGGAGtgcata	gtg	gcccgtatagcaggcATTaacattc	*S13*

The SD sequences and the downstream sequences are shown in upper case. Nucleotides that extend the SD and downstream sequences if G:U base pairing is allowed are underlined. The gene name used in the EMBL database is given in the right hand column.

energetically much more favourable and we therefore propose the limits +4 to +21 for the downstream interaction. This range includes the nucleotides found by Stormo *et al.* (1982a) to be significant when a group of gene starts (largely differing from those in our file) was studied using the χ^2 test. Scherer *et al.* (1980) studied a file of 68 gene starts that, in common with the file used by Stormo *et al.* (1982a), and unlike ours, contained a large number of coliphage genes as well as sequences from *E.coli* genes. They deduced a downstream consensus sequence that appears to be preferred in this system. This sequence:

```
AUGAAAAAAAUUAAAAAACUCAA
   CUC              G
```

contains many elements that are complementary to the 5′-terminal sequence of *E.coli* 16S rRNA and an examination of the tables of gene starts on which their consensus is based strengthens our proposal that this preference for certain nucleotides reflects a requirement for a limited complementarity with the 5′-terminal sequence of the 16S RNA.

The boundaries of the 16S rRNA sequence

The 16-nt sequence from the 5′ end of the 16S RNA that we propose to be involved in the recognition of the downstream genetic sequence is longer than the 3′ sequence involved in the SD interaction. While the advantage of the greater flexibility that this longer sequence introduces through allowing a larger number of amino acid coding possibilities is readily appreciated, the tetramer -TCAA- appears twice in the sequence and an obvious question to ask is whether the putative binding sequence could be shortened (and the case for it perhaps strengthened) by choosing a more limited region of the 16S RNA sequence. An examination of the contents of our database and of the consensus sequence of Scherer *et al.* (1980) emphasizes the frequency with which runs of consecutive A residues appear in *E.coli* gene starts and it would seem likely that, if a mechanism of the sort we propose operates, it should be important that the isolated sequence -AAA- should be capable of being recognized. This could, of course, be achieved by a sequence terminating at nucleotide 14 of the 16S RNA. In 40 of the entries of Appendix 1(a), the -AAA- sequence is extended further at least to -AAAC-, with a consequent significant increase in the overall negative free energy change on formation of the helical interaction (Freier *et al.*, 1986). For 32 other sequences, however, the additional C residue is at the 5′ end of the AAA sequence. While the increase in stability of interaction will be fractionally smaller it is still significant and there seems to be no reason to exclude this possibility at this stage. The extension of the sequence to include the T residue that is complementary to the A at position 16 of the 16S RNA is justified at this stage by the observation that the stabilities of interaction of eight of the entries of Appendix 1(a) involving the sequence CAAA are significantly increased by the inclusion of this residue in the overall interaction, despite the consequent duplication of the TCAA tetramer. (A comparable advantage of the duplication of the -AGG- triplet in the SD nucleotide should be noted.)

The single exception in the *E.coli*/coliphage system to the requirement for an SD sequence is that of gene CI of bacteriophage λ. This gene can be transcribed from either of two promoters, depending upon the physiological conditions prevailing. When transcription is initiated from the proximal promoter the transcribed mRNA, which is translated in the infected cell, has the A of the ATG initiation codon as its 5′ terminus. This mRNA thus lacks an SD sequence (Walz *et al.*, 1976). The first seven codons of this gene (ATGAGCACAAAAAAGAAACCA-) include both -CAAA- and -AAAC- within the sequence range that we propose to be significant. Interestingly, although both of these sequences are absent from the corresponding sequence (ATGAGTATTTCTTCCAGGGTA-) from the closely related bacteriophage 434 (Kuziel and Tucker, 1987), the downstream sequence from this phage gene includes two other sequences, -ATTT- and -TCTTC-, that are complementary to that part of the 5′ region of the 16S RNA that we propose to be involved in this initiation interaction.

In an early paper, van Knippenberg (1975) noted some striking features of the ribosome binding sites of the A protein and coat protein genes of the RNA coliphages R17 and Qβ. He pointed out that these sequences had extensive regions both of direct homology and complementarity to the 5′-terminal sequence of *E.coli* 16S rRNA and proposed mechanisms by which these might be involved in translational initiation of the messages. The possible mechanisms that he proposed involved interactions with nucleotides lying on both sides of the initiation codon and include nucleotides that are believed to take part in the SD interaction. Such extensive interaction would not appear to be possible with *E.coli* genes in general, but the downstream interactions that he envisaged as taking part in the interaction with the 5′-terminal sequence of the 16S RNA are identical to those that we would now propose to have more universal importance.

Statistical considerations

The question whether the complementarity that we observe between *E.coli* gene starts and the 5′ end of 16S rRNA is the result of chance can be addressed in several ways. There are 4^4 tetramers that can be constructed from the four nucleotides A, G, C and T. The entries of Table I, however, contain only the 12 tetramers found in the sequence TCAAACTCTTCAATTT (remembering that the sequence -TCAA- appears twice). If, as we propose, the sequence with which the 16S RNA can interact involves the 21 nucleotides within the range +4 to +24 of the gene, then there are 18 possible nucleotides in each gene start at which a tetramer can be initiated. The probability of one of the 12 'allowed' tetramers occurring in such a sequence at random is thus $12/(18 \times 4^4) = 1{:}384$. Similarly, the probability of the random occurrence of five consecutive nucleotides complementary to this part of the 5′ end of the 16S RNA is $12/(17 \times 4^5) \approx 1{:}1450$. Of the 251 entries in our file, 168 (66.9% of the total file) contain a sequence of at least four consecutive nucleotides and 47 of those 168 (18.7% of the total file) have a sequence of five or more consecutive nucleotides that is complementary to the 5′-terminal sequence of *E.coli* 16S RNA. The calculations of probability of occurrence assume that the four nucleotides are randomly distributed in *E.coli* gene starts and ignores the fact that, for example, the sequences immediately following the initiation codon tend to be low in G residues (Stormo *et al.*, 1982a). Although it does not completely overcome this problem, another approach is to consider the distribution of tetramers in a selection of 'false' gene starts of the type used by Stormo *et al.* (1982a) in their statistical studies. We have analysed a file of 'false' gene starts, selected from some of the longer sequence entries in the EMBL database on the basis of the occurrence of a possible SD sequence followed by a suitably spaced ATG codon. Of the 117 entries in the file, 11 (9.4% of the file) have a match of five or more consecutive nucleotides; 46 (39.3%) have a match of four or more consecutive nucleotides; 35 (47.9%) have a match of no more than three, and 14 (18.8%) have no match of at least three nucleotides with the corresponding 16S RNA sequence.

These data were used to calculate χ^2 values to test the hypothesis that the occurrence downstream of sequences of consecutive nucleotides complementary to the 5′ terminus of the 16S rRNA is significant. The χ^2 values testing the hypothesis that the occurrence of three such consecutive nucleotides is significant ($\chi^2 = 35.3$) and the more refined hypothesis that a sequence of four such nucleotides is significant ($\chi^2 = 25.0$) are both significant at the 99.9% level. The facts that not all of the gene starts in our file obey the new rule and that sequences that obey both the SD hypothesis and the new rule but are not gene starts can be found make it clear that other determinants, as yet unknown, are involved in the setting up of the initiation complex. We have, of course, taken no account of the role of secondary structure, nor have we any clear conception at this stage of the importance of our proposed interaction in determining the 'strength' of ribosome binding or the efficiency of initiation.

It can be argued that interactions of this kind should involve associations that are easily made but easily broken and that a weak, but specific, interaction may be more likely to lead to enhanced translational efficiency than a stronger interaction. Munson *et al.* (1984) studied the effects on expression induced by 16 separate point mutations near the beginning of the *lac*Z gene of *E.coli* and Looman *et al.* (1987) have extended these observations by studying the effect on synthesis of the protein product of altering the three nucleotides that form the first codon following the initiating ATG codon. This gene (which we did not include in our file because the upstream sequence was not given in the version of the EMBL database from which we constructed our data set) has two possible ATG initiation codons, separated by one triplet. The sequence is given by Munson *et al.* (1984) as: TAACAATTTCACACAGGAAACAGCTATGACCATGATTACGGATTCACTGGTCG (the two ATG codons are underlined). Although the spacing between the second ATG and the SD sequence (13 nt) is towards the upper limit that is usually observed, ~20% of the protein molecules synthesized *in vivo* are initiated at the second of the two ATG codons (Munson *et al.*, 1984). This gene obeys our rule, since the tetramer TTCA is found within the proposed range of both of the two initiation codons. It is of particular interest to note that, in each case (CAA, TCA, TTC) where the new second codon introduced a four-base sequence complementary to the 5′-terminal sequence of 16S RNA immediately after the first ATG initiation codon, the expression of the gene was drastically reduced (Looman *et al.*, 1987). Interpretation of these results in the light of our hypothesis is complicated by the fact that secondary structure involving both initiation codons is believed to play an important role in the regulation of the translational initiation of this gene (Munson *et al.*, 1984) and that the effect of these mutations on this secondary structure may be more important than the introduction of a new downstream interaction of the type that we propose. Equally, however, these results might be taken to indicate that a strong downstream interaction with the 16S RNA that is close to the ATG initiation codon has the effect of reducing translational efficiency.

The effectiveness of this type of theoretical study depends heavily upon the accuracy with which the gene starts that are used in constructing the database have been identified by the various authors in the first place. In many cases the interpretation of the nucleotide sequences cited in our tables was supported by direct determination of the primary structure of at least the N terminus of the corresponding protein product, but in a large number of cases the assignment of the initiation codon depended solely upon the application purely of rules based on the SD hypothesis, sometimes with additional support from measurements of amino acid composition and/or the mol. wt of the protein product. It is thus likely that some of the entries in our database are themselves 'false' but it is most unlikely that the number of these will be so high as to affect the significance of the effect that we observe. The issue is further clouded by the knowledge that initiation at internal ATG codons can occur *in vivo* (see Steitz, 1979, for a detailed discussion). It is therefore likely that some of the 'false' gene starts that seem to obey our rule are, in fact, capable of participating in translational initiation.

The immunochemical studies of Mochalova *et al.* (1982) and, more recently, the cross-linking studies of Brimacombe *et al.* (1988) have mapped the 5′ and 3′ termini of the 16S rRNA to opposite faces of the 30S ribosomal particle, separated by a distance of ~130 Å. An interaction of the sort that we propose would therefore require the mRNA to be wrapped around the ribosome and the question arises whether the initial encounter between the mRNA and the 16S RNA is at the 3′ or the 5′ end of the latter. The SD hypothesis assumes that the initiation complex involves recognition of the mRNA in the 5′ to 3′ direction, but there would seem to be no reason to exclude the possibility of the first interaction being with the downstream sequence, with formation of the complex being completed by recognition of the SD sequence and relaxation of the downstream interaction. This proposal gains some plausibility when it is noted that a number of the gene starts in our file (Appendix 1) have ATG or GTG triplets between the SD sequence and the true initiation codon, while a significant number have sequences resembling our downstream sequence lying 5′ to the SD sequence.

Our hypothesis that the 5′ end of the 16S rRNA might interact with the mRNA during the formation of the initial translation complex is, of course, based solely upon analyses of published sequence data. We hope, however, that our presentation of it at this stage as a theoretical possibility will stimulate experimental investigation since, if the hypothesis is correct, it will be important in taking us one stage closer to a complete understanding of the multiple interactions that this process of translational initiation involves, as well as providing additional criteria for the interpretation of new sequence data. Our hypothesis could be tested by a systematic study of the effects of modifying the nucleotide sequence in the first few codons of a suitable mRNA but the interpretation of such experiments would be complicated by the uncertainties introduced by possible changes to the secondary structure of the mRNA. Perhaps more rewarding would be to use the approach that has been used so successfully to obtain supporting evidence for the SD sequence (Hui and de Boer, 1987; Jacob *et al.*, 1987) by determining the effect of modifying the 5′ terminus of the 16S rRNA on the efficiency of translation of genes in a model system.

Eukaryotic mRNAs do not seem not to have an SD sequence associated with translational starts and Both (1979) has suggested the possibility of base pairing of eukaryotic mRNA with a conserved, internal sequence of the 18S rRNA. He did not, however, limit his proposal solely to downstream sequences and we are not aware of any experimental evidence that has been adduced to support his proposal. It would be of interest to re-examine the evidence for this sequence, given the very large number of eukaryotic gene sequences now available.

The SD sequence of the 16S rRNA of *E.coli* is very strongly conserved amongst different bacterial species (Dams *et al.*, 1988). The sequence at the 5′ terminus that we are proposing to be involved in the recognition of downstream nucleotide sequences is strongly conserved for part of its length, but there is some variation in the last six nucleotides or so of the sequence. This variability involves mainly the gene sequences for which the downstream 'recognition' sequence includes the nucleotides -AATTT. We are not, of course, able at this stage to decide whether these nucleotides should properly be considered part of the downstream sequence that we propose to be important. We note that, in most cases, a triplet or a tetramer that will base pair with the conserved nucleotides can be found elsewhere within these gene starts, but the possibility cannot be excluded that the use of these poorly conserved nucleotides might represent a species-specific feature of some gene starts.

Materials and methods

We have followed the route taken by others by compiling lists of *E.coli* 'gene starts' from the literature and using these as a basis for study. We have confined our attention to those nucleotides that are believed to be covered by the ribosome during the formation of the initiation complex and shown by the statistical studies of Stormo *et al.* (1982a) to be significant. Nucleotides +4 to +28 (we number the first A or G of the initiation codon +1) of each entry were scanned for direct homology of pre-determined lengths with the complement of the *E.coli* 16S rRNA sequence (Brosius *et al.*, 1978), converted to a DNA sequence by the substitution of T for U. The possibility of G:U base pairing was not considered in this first stage of selection. Matches were scored against the 16S RNA sequence and those regions of the 16S RNA that showed high scores were then reinvestigated in greater detail. Approximate free-energy changes for helix formation were calculated when required using the parameters determined by Freier *et al.* (1986).

Acknowledgements

Some of the early work of this study was performed while G.B.P. was on study leave in the MRC Laboratory of Molecular Biology, Cambridge, UK. The manuscript was completed while he was on study leave in the Sir William Dunn School of Pathology, University of Oxford. He is grateful to his colleagues in those laboratories, especially to Fred Sanger, Rodger Staden and George Brownlee, for their helpful discussion and encouragement. The help of Rodger Staden with computing in the early stages of the work was particularly valuable. We thank Derek Holton for advice on the probability calculations. This research was supported through a Programme Grant from the Medical Research Council of New Zealand and through grants for computer equipment from the Medical Research Council of New Zealand, the New Zealand Lottery Board and the New Zealand University Grants Committee.

References

Both,G.W. (1979) *FEBS Lett.*, **101**, 220–224.

Brimacombe,R., Atmadja,J., Stiege,W. and Schüler,D. (1988) *J. Mol. Biol.*, **199**, 115–136.,

Brosius,J., Palmer,M.L., Poindexter,J.K. and Noller,H.F. (1978) *Proc. Natl. Acad. Sci. USA*, **75**, 4801–4805.

Dams,E., Hendriks,L., Van de Peer,Y., Neefs,J.-M., Smits,G., Vandenbempt,I. and De Wachter,R. (1988) *Nucleic Acids Res.*, **16** (Suppl.), r87–r173.

Freier,S.M., Kierzek,R., Jaeger,J.A., Sugimoto,N., Caruthers,M.H., Neilson,T. and Turner,D.H. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 9373–9377.

Gold,L., Pribnow,D., Schneider,T., Shinedling,S., Singer,B.S. and Stormo,G. (1981) *Annu. Rev. Microbiol.*, **35**, 365–403.

Hui,A. and de Boer,H.A. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 4762–4766.

Jacob,W.F., Santer,M. and Dahlberg,A.E. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 4757–4761.

Kuziel,W.A. and Tucker,P.W. (1987) *Nucleic Acids Res.*, **15**, 3181.

Looman,A.C., Bodlaender,J., Comstock,L.J., Eaton,D., Jhurani,P., de Boer,H.A. and van Knippenberg,P.H. (1987) *EMBO J.*, **6**, 2489–2492.

Mochalova,L.V., Shatsky,I.N., Bogdanov,A.A. and Vasiliev,V.D. (1982) *J. Mol. Biol.*, **159**, 637–650.

Munson,L.M., Stormo,G., Neice,G.D. and Reznikoff,W.S. (1984) *J. Mol. Biol.*, **177**, 663–683.

Sanger,F., Coulson,A.R., Hong,G.F., Hill,D.F. and Petersen,G.B. (1982) *J. Mol. Biol.*, **162**, 729–773.

Scherer,G.E., Walkinshaw,M.D., Arnott,S. and Morré,D.J. (1980) *Nucleic Acids Res.*, **8**, 3895–3907.

Shine,J. and Dalgarno,L. (1974) *Proc. Natl. Acad. Sci. USA*, **71**, 1342–1346.

Shine,J. and Dalgarno,L. (1975) *Nature*, **254**, 34–38.

Staden,R. (1985) In Setlow,J.K. and Hollaender,A. (eds.), *Genetic Engineering: Principles and Methods.* Plenum Press, New York. Vol. 7, pp. 67–114.

Steitz,J.. (1979) In Goldberger,R.F. (ed.), *Biological Regulation and Development.* Plenum Press, New York, pp. 349–399.

Stormo,G.D., Schneider,T.D. and Gold,L.M. (1982a) *Nucleic Acids Res.*, **10**, 2971–2996.

Stormo,G.D., Schneider,T.D., Gold,L. and Ehrenfeucht,A. (1982b) *Nucleic Acids Res.*, **10**, 2997–3011.

van Knippenberg,P.H. (1975) *Nucleic Acids Res.*, **2**, 79–85.

Walz,A., Pirrotta,V. and Ineichen,K. (1976) *Nature*, **262**, 665–669.

Received on June 20, 1988; revised on August 15, 1988

Appendix 1

Contents of the database analysed in this study. The entries are arranged in increasing order of spacing from the ATG or GTG initiation codon as is illustrated in Tables I and II. For each entry the name used in the EMBL database is given first, followed by the EMBL database identifier (ID).

(a) Gene starts from *E.coli* that obey the new rule and have four or more consecutive downstream nucleotides complementary to some part of nucleotides 1–16 of 16S rRNA:
aroF ECTYRA; *aroG* ECAROG; *aspA* ECASPA; *deoB* ECDEOAB; *dhfr* ECDHFR01; *dld* ECDLDH; *fumA* ECFUMA; *glysA* ECGLYS; *guaB* ECGUAB; *ilvN* ECILVBN; *lacI* ECLACI; *nusB* ECNUSB1; *phoA* ECPHOA; *rpmH* ECRPMH; *thyA* ECTHYA; *trpC*, *trpE* ECTRPX; *carB* ECCARB; *dnaN* ECDNAAN; *envA* ECFTSQAB; *ftsA* ECFTSQA; *polA* ECPOLA; *pyrE* ECDUTPYR; *rplM* ECRPSI; *rplV* ECRPOS1P; *rpmD* ECSPC; *rpsO* ECRPSO; *sdhA* ECSDHACD; *S11* ECRPA; *tyrB* ECTYRB; *uncA*, *uncB* ECUNC; *uncC* ECUNC; *asnA* ECORIASN; *deoR* ECDEOR; *dsdA* ECDSDA; *hlyD* ECHLY; *malG* DECMALG; *rplD* ECRPOS10; *ecoRV* ECECORV; *gshII* ECGSHII; *fimA* ECFIMA01; *metG* ECMETG; *nusA* ECNUSA; *pepN* ECPEPN5; *rpsB* ECRPSB; *rpsP* ECTRMD; *sdhB* ECSDHB; *uncE* , *uncF* , *uncI* ECUNC; *lspa* ECLSP; *tnaA* ECTNAA; *gyrB* ECGYRB; *hlyB* ECHLY; *metK* ECMETK; *pldA* ECPLDA1; *prlA* ECSPC; *rnh* ECRNHX; *rplA* ECROPBC; *rplE* ECSPC; *tsr* ECTSRX; *uvrB* ECUVRB; *argF* ECARGF; *hlyA* ECHLY; *lacY* ECLACY; *motA* ECMOTAB; *narG* ECNARPR; *rf1* ECRF1X; *rpmC* ECRPOS10; *rpoB* ECRPOJ; *rpoC* ECRPOBC; *sulA* ECSULA; *gap* ECGAP; *aac* ECAAC3IV; *aroH* ECAROH; *gidB* ECUNC; *rf2* ECRF2X; *rpsU* ECRPSU; *tsf* ECRPSB; *hlyC* ECHLYC; *lysC* ECLYSC; *motB* ECMOTAB; *argE* ECARGOP1; *trg* ECG; *cheZ* ECCHEY; *dacA* ECDACA; *kdpB* ECKDPABC; *metB* ECMETLB1; *ompR* ECOMPB; *ppcp* ECPPCP; *proA* ECPROAB; *rpsN* ECSPC; *rpsJ* ECRPOS10; *uvrD* ECUVRD; *rpsA* ECRPSA; *lamB* ECLAMBA; *melB* ECMELB; *npl* ECNPL; *glysB* ECGLYS; *uncG* ECUNC; *xylA* ECXYLA; *rpsD* ECRPA; *rpsS* ECRPOS10; *gltA* ECGLTA; *melA* ECMELOP; *rplP* ECRPOS10; *tyrA* ECTYRA; *ftsZ* ECFTSQA; *metF* ECMETF; *araB* ECARABM; *araE* ECARAE; *dapB* ECDAPB; *phoU* ECPHOWTU; *tonB* ECTONB; *hisreg* ECIHIS1; *crp* ECCRP; *eltA* ECELTA; *fur* ECFUR; *ompF* ECOMPF; *pheA* ECTYRA; *pyrD* ECPYRD; *rpsI* ECRPSI; *rplJ* ECRPOBC; *toxA* ECTOXA; *araC* ECARAC; *hisG* ECHIS1; *malP* ECMALP1; *malT* ECMALX; *pldB* ECPLDB; *dnaB* ECDNAB; *lpp* ECLPPX; *phoE* ECPHOE; *aceF* ECACEX; *aph4* ECAPH4; *pstB* ECPHOWTU; *cheY* ECCHEY; *rnc* ECRNC1; *recA* ECRECA; *hisB* ECHISCBH; *cdh* ECCDH; *trmD* ECTRMD; *envZ* ECOMPB; *aceE* ECACEX; *galK* ECGALK; *kdpA* ECKDPABC; *gdhA* ECGDHA; *pbpB* ECPBPB; *trpA* ECTRPX; *livK* ECLIVK; *dye* ECDYE; *fnr* ECFNR. *ssb* ECSSB; *nrdB* ECNRDA; *plsB* ECPLSB; *rplL* ECRPOBC; *tap* ECTARX; *leuA* ECLEUA; *malF* ECMALF; *pabA* ECPABA; *sdhC* ECSDHACD; *pstA* ECPHOWTU; *rplB* , *rpsQ* ECRPOS10; *ilvH* ECILVIH; *mtlA* ECMTLA; *rplS* ECTRMD; *glgA* ECGLGC.

(b) Gene starts from *E.coli* that obey the new rule but have only three consecutive downstream nucleotides complementary to some part of nucleotides 1–16 of *E.coli* 16S rRNA:
ada ECADA; *asd* ECASDX; *aspC* ECASPC; *deoC* ECDEOCA1; *dnaA* ECDNAAN; *dut* ECDUT; *ecoRVmet* ECECORV; *gidA* ECUNC; *ks71A* ECKS71AP; *lexA* ECLEXA; *mglB* ECMGLB; *molA* ECLAMBA; *nrdA* ECNRDA; *ompA* ECOMPA; *ompC* ECOMPC; *pfkA* ECPFKA; *phos* ECPHOS; *phr* ECPHRORF; *pin* ECPIN; *rplC* ECRPOS10; *rplF* ECSPC; *rplW* ECRPOS10; *sbp* ECPFKA; *tar* ECTARX; *usg* ECHIST1; *ef-g* ECSTR2; *fol* ECFOLX; *fus* ECFUSG; *glnL* ECGLNAL; *ilvB* ECILVBN; *proC* ECOC; *spc* ECSPCX; *trpB* ECTRPX; *glyA* ECGLYA; *gnd* ECGND; *rpsC* ECRPOS10; *argC* ECARGC; *dam* ECDAMX; *hag* ECHAGFLG; *lep* ECLEP; *metA* ECMFTA; *pyrB* ECPYRBIA; *rpln* ECSPC; *serB* ECSERB; *uncD* ECUNC; *rpmG* ECRPMG; *sucA* ECSUCA; *recF* ECRECF; *adk* ECADK; *gpt* ECGPT1; *ileS* ECRPST; *kdpC* ECKDPABC; *rpoA* ECRPA; *sdhD* ECSDHACD; *argI* ECARGI; *aroB* ECAROB; *birA* ECBIRA; *purF* ECPURF; *relB* ECRELB; *rplK* ECRPOBC; *rplX* ECSPC; *trpD* ECTRPX; *ams* ECAMS; *glnA* ECGLNA; *pfkB* ECPFKB; *rpmB* ECRPMB; *tpi* ECTPI; *dnaG* ECDNAG; *galR* ECGALR; *malK* ECMALK; *hisT* ECHIS1; *ilvI* ECILVIH; *ilvE* ECILVE; *pabB* ECPABB; *dapD* ECDAPD; *sucB* ECSUCA; *uvrA* ECUVRA; *cds* ECCDS; *S13* ECRPA.

(c) Gene starts from *E.coli* that have fewer than three consecutive downstream nucleotides complementary to nucleotides 1–16 of *E.coli* 16S RNA:
glnS ECGLNS; *proB* ECPROAB; *rplQ* ECRPOA; *uvrC* ECUVRC1.

A full listing of this database in the form of Tables I and II is available on application to the authors.