Volume 8 Number 17 1980 Nucleic Acids Research

The ribosome binding sites recognized by E. coli ribosomes have regions with signal character in both the leader and protein coding segments

Günther F.E.Scherer[1], Malcolm D.Walkinshaw[2], Struther Arnott and D.James Morré

Purdue University, Department of Biological Sciences, West Lafayette, IN 47907, USA

Received 24 July 1980

SUMMARY

Oligonucleotide analysis, by a novel computerized procedure, was first applied to determine the sequence of an ideal E. *coli* promoter (Scherer *et al.*, *Nucl. Acids Res.* 1978, 5:3759-3773) and has now been used to obtain the sequence of nucleotides that should be present in a messenger RNA for optimum binding to the E. *coli* ribosome. This sequence is:

UU·UUAAAAAUUAAGGAGGUAUAUUAUGAAAAAAAUUAAAAAACUCAA
AA U A AUA A CUC G

Comparison of this sequence with each of the 68 ribosome binding site sequences used to generate it shows a preference rather than an absolute requirement for a specific base in any given position. The preference for certain bases persists along the whole length of the RNA within the ribosome binding domain even though nearly half of that length includes translated codons. Thus messages without leader sequences (like λCI mRNA) can still have some affinity for the ribosome. Part of the model sequence is complementary to the 3'end of 16S rRNA.

INTRODUCTION

A start codon AUG (or GUG) is a necessary but not sufficient requirement for translation by E. *coli* ribosomes. The start codon needs also to be embedded in an appropriate sequence (1) which (we think) is 46-48 nucleotides long (Fig. 1). The 5' ends of all known sequences but one (2) possess a leader sequence preceding the initiator codon. The 3' end of 16S rRNA (3) (Fig. 1) is thought to form a base-paired complex with the leader region (4,5). Two of these complexes have been isolated and characterized (6,7). It is clear, however, that more than base-pairing with the leader sequence is involved because B. *stearothermophilus* ribosomes have the same sequence as E. *coli* at the 3' end of their 16S rRNA (7) but interact only weakly with some E. *coli* ribosome

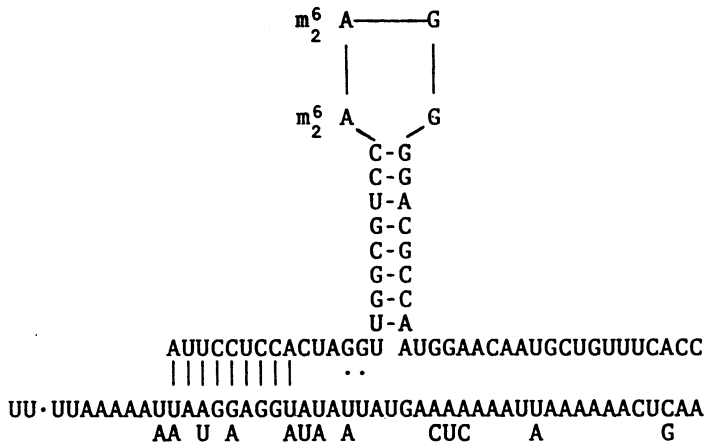

Fig. 1 - Model sequence for the E. *coli* ribosome binding site and its possible relationship with the 3' end of the 16S rRNA.

binding sites (8-11) and with two non-initiator binding sites for E. *coli* ribosomes that are found on Qβ RNA (8).

Our study of E. *coli* promoters (12) showed that the nucelotide sequences involved were longer than previously supposed and were characterized by a *preference* for certain bases but *no requirement* for any base in a particular site. This study was based on no more than 17 base sequences. We report here an analogous study using 68 non-identical sequences (13-49) of E. *coli* ribosome binding sites. It provides an objective comparison of all these sequences and enables us to define (but not necessarily to explain) additional, subtler signal elements not immediately obvious from casual inspection.

METHOD

The procedure used was the same as in our study of E. *coli* promoter sequences (12). The mRNA sequences were lined first as shown in Table 1. Homology was evaluated by comparing them pairwise. If one of the 64 possible triplet combinations of bases occurred at an identical position in two sequences, a score of 1 was given to each of the three bases in both sequences. If a triplet occurred in both sequences but at positions deviating by

±1 or ±2 then scores of 1/3 and 1/5 respectively were given to each base in both triplets. By comparing all possible pairs of sequences and checking each position for the occurrence of identical triplets, a total score was obtained for each base in all sequences. For each position the scores for each base type (A,C,G or U) were summed to give four overall scores (Σ). These amounts of Σ (12) depend on the number n of sequences used for the analysis. They were transformed into values independent from n so that analyses of differing numbers of sequences can be compared conveniently on the same scale. The rationale is as follows: The average value for Σ of an analysis of n sequences is $\Sigma = 0.0242\ (n)\ (n - 1)$ where n is the number of sequences at a given position and $(n - 1)/2$ is the number of pairs that can be formed among n sequences. The constant value of twice 0.0242 is the probability of a base occurring in a pair of trimers; $2 \cdot 3 \cdot (1/4)(1/64) \cdot 1 + (1/64) \cdot (2/3) + (1/64) \cdot (2/5)$. Division of Σ by n X $(n - 1)$ yields transformed Σ values which in an analysis of random sequences are always close to 0.0242 but may vary characteristically in an analysis of related sequences. Obviously, with higher values of n better approximations are achieved.

Using the AUG codon to define a common origin gives artificially high scores to the bases X and Y in the pentamer XAUGY. The AUG codon was therefore excluded and, in effect, separate analyses performed on the preceding leader sequence and on the succeeding protein coding sequence.

A second analysis of the leader sequences was conducted using a signal complementary to the 3' end of the 16S rRNA to define the origin in each mRNA sequence. Similarly, we conducted a third analysis using a CUC at the 3' end of the ribosome binding sequences as a model for a hypothetical signal element which is only weakly indicated in the first analysis due to its variable distance from the AUG codon.

The level above which transformed Σ values should be considered significant was determined empirically by applying an identical analysis to an equal number of nonsense RNA produced using a random number generator. The results suggest a value of 0.05 for transformed Σ to be the lower limit for significance. This limit may be slightly lower for bases adjacent to either the A or G of

Table 1 - Sequences of 74 ribosome binding sites of E. coli lined up as for the analysis in Fig. 2. λ CI was used as transcribed from the λ_{pre} promoter (2), the λ N RNA as a best fit (25). On top of the sequences bars indicate the portions of the sequences thought to be homologous to the AAGGAGGU octamer in the leader region and homologous to the hypothetical CUC signal element at the 3' end of the model sequence. Bases agreeing with the model are underlined. a) Sequences were not included into the analysis because they are identical with related phage sequences. b) Sequences were not included into the analysis of the 3' end.

SOURCE (REF.)	SEQUENCE
R17 A prot. (13) [a,b]	GAUUCCUAGGAGGUUUGACCUAUGCGAGCUUUUAGUG
R17 coat(13) [a,b]	CCUCAACCGGGGUUUGAAGCAUGGCUUCUAACUUU
R17 repl.(13) [a,b]	AAACAUGAGGAUUACCCAUGUCGAAGACAACAAAG
Q beta A prot.(14) [b]	UCACUGAGUAUAAGAGGACAUAUGCCUAAAUUACCGCGU
Q beta coat(15) [b]	GAAACTTTGGGUCAAUUUGATCAUGGCAAAAUUAGAGACUAGACUGU
Q beta repl.(16,17) [b]	AGUAACUAAGGAUGAAAUGCAUGUCUAAGACAGC
fd VIII(18,19)	UUACGUAUUUUACCCGUUUAAUGGAAACUUCCUCAUGAAAAAGUCUUUAGUCCUCAAAGCCUCCGUAGCCG
fd V(18,19)	GUUCUUAAAAUCGCAUAAGGUAAUUCAAAAUGAUUAAAGUUGAAAUUAAACCGUCUCAAGCGCAAU
fd IV(18,19)	UUAUGUACUGUUUCAAUUAAAAAAGGUAAUUCAAAUGAAAUUGUUAAAUGUAAUUAAUUUUGUUUUCUUGA
fd VII(19)	AUUGACCGUCUGCGCCUCGUUCCGGCUAAGUAACAUGGAGCAGGUCGCGGAUUUCGACACAAUUUAUCAGG
fd III(19)	CCUUUUGGAGCCUUUUUUUUGGAGAUUUUCAACGUGAAAAAAUUAUUAUUCGCAAUUCCUUUAGUUGUUC
fd VI(19)	UUGCUAACAUACUGCGUAAUAAGGAGUCUUAAUCAUGCCAGUUCUUUUGGGUAUUCCGUUAUUAUUGCGUU
fd I(19)	AAAUCGUUUCUUAUUUGGAUUGGGAUAAAUAAAUAUGGCUGUUUAUUUUGUAACUGGCAAAUUAGGCUCUG
fd II(19)	UUUGGGGCUUUUCUGAUUAUCAACCGGGGUACAUAUGAUUGACAUGCUAGUUUUACGAUUACCGUUCAUCG
f2 coat(20) [a,b]	CCUCAACCGAGGUUUGAAGCAUGGCUUCCAACUUUACUCAG
MS2 A prot.(21)	GUAGCCGGAAUUCCAUUCCUAGGAGGUUUGACCUGUGCGAGCUUUUAGUACUCUCGAUAGGGAGAACGAGA
MS2 coat(22)	UCUCUAGAUAGAGCCCUCAACCGGAGUUUGAAGCAUGGCUUCUAACUUUACUCAGUUCGUUCUCGUCGACA
MS2 repl.(23)	AUAGACGCCGGCCAUUCAAACAUGAGGAUUACCCAUGUCGAAGACAACAAAGAAGUUCAACUCUUUAUGUA
Phi X A(24)	AAAUCUUGGAGGCUUUUUUAUGGUUCGUUCUUAUUACCCUUCUGAAUGUCACGCUG
Phi X 3(24)	UGGACCUUGCUGCUAAAGGUCUAGGAGCUAAAGAAUGGAACAACUCACUAAAAACCAAGCUGUCGCUACUU
Phi X D(24)	GAUGCUGUUCAACCACUAAUAGGUAAGAAAUCAUGAGUCAAGUUACUGAACAAUCCGUACGUUUCCAGA
Phi X E(24)	UCUCGUGCUCGUCGCUGCGUUGAGGCUUGCGUUUAUGGUACGCUGGACUUUGUGGGAUACCCUCGCUUUCC
Phi X J(24)	UGCGUCAAAAAUUACGUGCGGAAGGAGUGAUGUAAUGUCUAAAGGUAAAAAACGUUCUGGCGCUCGCCCUG
Phi X F(24)	GCAGGGGCUUCGGCCCCUUACUUGAGGAUAAAUUAUGUCUAAUAUUCAAACUGGCGCCGAGCGUAUGCCGC
Phi X K(24)	AUGACGCAGAAGUUAACACUUUCGGAUAUUUCUGAUGAGUCGAAAAAUUAUCUUGAUAAAGCAGGAAUUAC
Phi X G(24)	AAGCGCGGUAGGUUUUCUGCUUAGGAGUUUAAUCAUGUUUCAGACUUUUAUUUCUCGCCAUAAUUCAAACU
Phi X H(24)	UUAUUUGUCUCCAGCCACUUAAGUGAGGUGAUUUAUGUUUGGUGCUAUUGCUGGCGGUAUUGCUUCUGCUC
Lambda N(25)	GAAGGGCAGGATTCAAAGCAGAAGGCUUUGGGGUGUGUGAUACGAAACGAAGCAUUGGCCGUAAGUGCGAU
Lambda CI(2)	AGAUAUUUAUCCCUUGCGGUGAUAGAUUUAACGUAUGAGCACAAAAAAGAAACCAUUAACACAAGAGCAGC
Lambda cro(26,27)	AUGUACUAAGGAGGUUGUAUGGAACAACGCAUAACCCUGAAAGAUUAUGCAAUGC
Lambda CII(27)	UUCAAUCAAUUGUUAUCUAAGGAAAUACUUACAUAUGGUUCGUGCAAACAAACGCAACGAGGCUCUACGAA
Lambda O(27)	GAGGUCAUUACUGGAUCUAUCAACAGGAGUCAUUAUGACAAAUACAGCAAAAAUACUCAACUUCGGCAGAG
G4 A (28)	CUCUAAUAUGCCUCCCAUCAAACGGAGGCUUUUCAUGUUUAAAGUACAUUCUGACUACUUCAGCAAACCUA
G4 A* (28)	AGAUGAGUGCUCAAAAAUCUUGGAGGAGUCAACUAUGAAGUCUCGACGUGGCUUUACCGUUCAAAGAUUAA

G4 B (28)	UUGACAUGGCCGUAAAAGGCCUAGGGAAUAAAGAAUGGAACAAUUCACUCAAAACCAAAAUCAACCUCAUA
G4 K (28)	AUGACCACGAAAUUAACAAGUACGGAUAUUUCUGATGAAACCAAAAACUACGUUGCUUCUGCAGGAAUUGC
G4 C (28) [a,b]	UACGAAUUAAAUCGAAGUGGACUGCUGGUGGAAAAUGAGGAAAUUCAAUCUCAACUUAAAAAACUCGAGGU
G4 D (28)	UCGUCAACUGACACAAACCACAAAGGAAACUGAAAUGUCUAAAUCAAACGAAUCUGCUGUAGCCUUUCAAA
G4 E (28) [a,b]	UCACGCGCUCGCCGUGCUAUCGAGGCUUGCGUAUAUGGAACACUGGACUUUGUCGGGUAUCCUCGCUUUCC
G4 J (28)	GAUUUUAUCGUCUUCACUUUUAAGGAGUUAUGUAAUGAAAAAAUCAAUUCGCCGCUCUGGUGGCAAAUCUA
G4 F (28)	GGACCGCGGUCCCACUCUAUUUAAGGAUACAAAAAUGUCUAACGUUCAAACAUCUGCGGACCGUGUACCUC
G4 G (28)	GCCGUCCCUACUGCAAAGCCAAAAGGACUAACAUAUGUUCCAGAAAUUCAUUUCUAAGCACAAUGCUCCAA
G4 H (28)	ACCGUCCUUCAACCUCUGAAAUAAGGAUUAUCCUAUGUUUGGCUCUAUCGCUGGCGGUAUCGCCUCCGCAC
Lambda 434 CI(29,30)	UGUGAAGAUUGGGGGUAAAUAACAGAGGUGGCUUAUGAGUAUUUCUUCCAGGGUAAAAGCAAAAGAAUU
Lambda 434 cro(29,30)	AGUUUGUUGAUGGAGGCGAUAUGCAAACUCUUUCUGAACGCCUCAAGAAGAGGCGAA
T7(T3) in vitro(31)	AACAUGAGGUAACACCAAAUGAUUUUCACUAAAGAGCCGCGAACG
T7a (32)	GAUAUUCACUAAUAACUGCACGAGGUAACACAAGAUGGCUAUGUCUAACAUGACUUACAAUAACGUUUUCG
T4 rII B (33) [b]	CCUAAUAAGGAAAAUUAUGUACAAUAUUAAAUG
trp leader (34)	AAGUUCACGTAAAAAGGGUAUCGACAAUGAAAGCAAUUUUCGUACUGAAAGGUUGGU
trp E (34)	CGGGCUUUUUUUUGAACAAAAUUAGAGAAUAACAAUGCAAACACAAAAACCGACUCUCGAACUGCU
trp A (35)	UUCACGAUAUUUUGAAAGCACGAGGAGAAAUCUGAUGGAACGCUACGAAUCUCUGUUUGCC
trp B (36)	UUUUCAGACACUGCGCGCAUAUUAAGGAAAAAACAUGACAACACUUCUCAACCCCUACUUUGGUGAAUUUC
lac Z (37)	GUGAGCGGAUAACAAUUUCACACAGGAAACAGCUAUGACCAUGAUUACGGAUUCACUGG
lac I (38)	GGAAGAGAGUCAAUUCAGGGUGGUGAAUGUGAAACCAGUAACGUUAUACGAUGUCGCAGAGUAUG
gal T (39)	CCAUCCACAGGGAUAUCCCGAUUAAGGAACGACCAUGACGCAAUUUAAUCCCGUUGAUCAUCCACAUCGCC
gal E (40)	ATACCATAAGCCUAAUGGAGCGAAUUAUGAGAGUUCUGGUUACCGGUGGUAGCGGUUACAUUG
ara BAD (41)	ACCCGUUUUUUUUGGAUGGAGUGAAACGAUGGCGAUUGCAAUUGGCCUCGAUUUUGCAGUGAUUC
lipoprotein (42)	CAUGGAGAUUAACUCAAUCUAAGAGGUUAUUAAUAAUGAAAGCUACUAAACUGGUACUGGGCGCGGUAAUCC
L 11 prot. (43)	CAGAGGCGUUAUUACCCAACUUGAGGAAUUUAUAAUGGCUAAGAAAGUACAAGCCUAUGUCAAGCUGCAGG
L 1 prot. (43)	ACGUUCCAUGGGCCUGGUAGUGGAGGACUAAGAAAUGGCUAAACUGACCAAGCUGCAUGCGUGUUAUCCGCG
L 7 prot. (43)	GAUUUUCUCUGGCAAACAUCCAGGAGCAAAGCUAAUGGCUUUAAAUCUUCAAGACAAACAAGCGAUGUUG
L 12 prot. (43)	UAUAAACUUAUUCUGAUAUUCAGGAACAAUUUAAAUGUCUAUCACUAAAGAUCAAAUCAUUGAAGCAGUUG
rpo B (43)	GGGUCGUCGACUUGUCAGCGAGCUGAGGAACCCUAUGGUUUACUCCUAUACCGAGAAAAAACGUAUUCGUA
amp gene (44)	AUAAAUGCUUCAAUAAUAUUGAAAAAGGAAGAGUAUGAGUAUUCAACAUUUCCGUGUCGCCCUUAUUCCCU
phe leader (45)	AAGUCACUUAAGGAAACAAACAUGAAACACAUACCGUUUUUCUUCGCAUUCUUUUUUA
phe A (45)	GGGCCUUUUUUAUUGAUAACAAAAAGGCAACACUAUGACAUCGGAAAACCCGUUACUGGCGC
his leader (46)	GGUAUCAAAUGAAUAAGCAUUCAUCGGAAUUUUUAUGACACGCGUUCAAUUUAAACACCACCAUCAUCACC
his G (46)	AGACCGGUUCAGACAGGAUAAAGAGGAACGCAGAAUGUUAGACAACACCCGCUUACGCAUAGCUAUUCAGA
bio A(47)	ACCUAAAUCUUUUCAAUUUGGUUUACAAGUCGAUUAUGACAACGGACGAUCUUGCCUUUGACCAACGCCAU
bio B (47)	CGAAUUAACAACAAAAAACACGUUUUGGAAGCCCCAUGGCUCACCGCCCACGCUGGACAUUGUCGCAAGUC
threo leader (48)	ACAGAUAAAAAUUACAGAGUACACAACAUCCAUGAAACGCAUUAGCACCACCAUUACCACCACCAUCA
threo A_1 (48)	UUUUUUUUCGACCAAAGGUAACGAGGUAACAACCAUGCGAGUGUUGAAGUUCGGCGGUACAUCA
spc (49)	GUCUCAGUAGUAGUUGACAUUAGCGGAGCCUAAAAUGAUCCAAGAACAGACUAUGCUGAACGUCGCCGACA
str (49)	UUACGAAGCAAAAGCUAAAACCAGGAGCUAUUUAAUGGCAACAGUUAACCAGCUGGUACGCAAACCACGUG

the AUG codon so that A at position -2 and G at position +2 may become significant, but, at present, this cannot be clearly assessed.

RESULTS

The results of the trimer analysis using the initiator codon as the common reference oligomer are shown in Fig. 2. The model sequence so derived exhibits considerable ambiguity in positions -14 through -6 where experiments have suggested that base-pairing

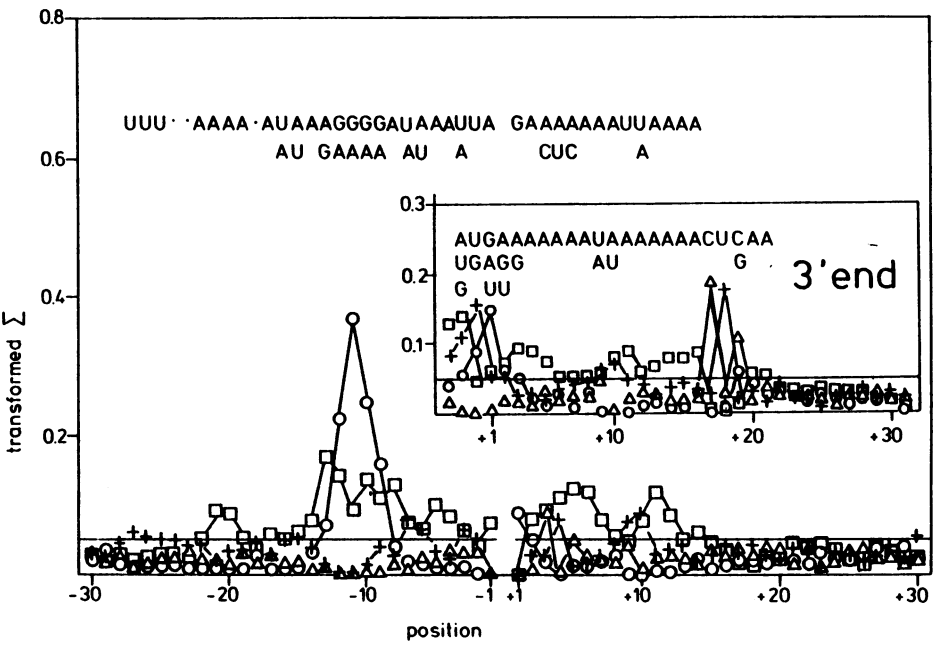

Fig. 2 - Trimer analysis of 68 ribosome binding sites with AUG as the common origin. 68 sequences were aligned as in Table 1 and the trimers analyzed in two separate runs excluding the starting codons. The untranslated leader regions carry negative position numbers, the protein coding regions carry positive position numbers. The sequences of the three R17 and the f2 coat ribosome binding sites were not included in the analysis because they are almost exact duplicates of the three MS2 ribosome binding sites. The inset shows the analysis using the hypothetical CUC$_G$ signal or the segments assumed to correspond to it in each sequence as indicated in Table 1 by overlining. Sequences too short to identify this CUC$_G$ or duplicate sequences were not included in this analysis. Bases with Σ values exceeding baseline are summarized. Symbols: Σ(A): □; Σ(G): ○; Σ(C): △; Σ(U): +.

with the 16S rRNA would be likely. Rather than accept the possibility that the ribosome recognizes a great variety of sequences equally well we considered the possibility that the segment of leader sequence base-pairing with the 16S rRNA might not be at a single fixed distance from the initiator codon (1).

The sequence AAGGAGGU, which is one of the possibilities for positions -14 through -7, is precisely complementary to an octanucleotide sequence on 16S rRNA. Therefore the mRNA binding sequences were realigned to make this sequence - or their best approximation to it - the common origin and a new trimer analysis of the leader region was performed. The results shown in Fig. 3 provide a much simpler picture of the model leader sequences but now the AUG start codon has a variable locus 6-9 positions upfield from the AAGGAGGU octamer.

Inspection of the sequences showed that CUC$_{G}$ seemed to be the

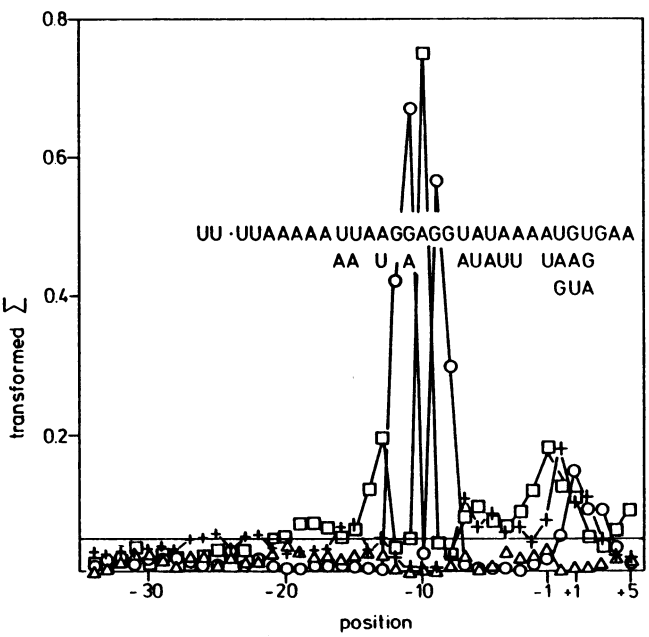

Fig. 3 - Trimer analysis of 68 ribosome binding sites using as the origins the AAGGAGGU segment or the segments assumed to correspond to it in each leader sequence. These segments are those overlined in Table 1. Duplicate sequences were omitted. Symbols: Σ(A): □; Σ(G): ○; Σ(C): △; Σ(U): +.

3' end of the model sequence. Again, due to the variable distance from the AUG codon this is only very weakly recognized in the first analysis (Fig. 2). A realignment of the sequences and subsequent analysis (inset in Fig. 2) showed a contiguous and almost identical model sequence between the AUG codon and the assumed CUC/G signal but, in contrast, a complete random behaviour beyond the CUC/G was found.

Therefore, we conclude that the complete model sequence for the mRNA binding site should consist of a part of the sequence in Fig. 3 viz.

UU·UUAAAAAUUAAGGAGGUAU
AA U A AUA

followed by a portion of the sequence surrounding the AUG codon in Fig. 2 viz.

AUUAUGAAAAAAAUU
A CUC

and completed at the 3' end as in the inset in Fig. 2 with

AAAAAACUCAA
G

DISCUSSION

The model sequence (Fig. 1) and the individual sequences (Table 1) have few *common* elements but many *similar* elements. Evidently the ribosome binding sites are coded in some degenerate way analogous to that found for promoter sequences. It is quite possible that secondary structures available to individual sequences have a role (1,23,38,45,46,48,50) but such effects would be almost invisible by our method which focusses on primary structures.

The fact that the signal with strong (base-pairing) complementarity to 16S rRNA as well as the distant signal portions at the 3' end appear not to have fixed distances from the initiator codon not only complicates the analyses but also limits further discussion of the coarser features of the model sequence.

It is noteworthy that there is a significant preference for certain bases *throughout* the mRNA region protected from digestion

by being ribosome-bound. This region includes the codons for the first seven N-terminal peptides (13,18,20,31,32,37,38). We think it unlikely that the significant Σ scores are the trivial consequence of all *E. coli* proteins having similar N-termini despite the observation that GCX seems to be preferred as the second codon in bacterial sequences. Consequently, the nucleotide sequences observed must have been selected for the double function of ribosome-binding and peptide-coding. In this context the sequence of λCI RNA is important. This mRNA can be transcribed either from the promoter λ_{pRE}, which has a leader sequence preceding the AUG codon (51), or from the λ_{pRM} promoter which has no leader segment at all (2). The coding portion of this mRNA (Table 1) is in very good agreement with our model sequence. Thus, despite the missing leader segment, there still should be some affinity to the ribosome binding site in addition to the signal represented by the AUG codon.

Fig. 3 shows that the most prominent feature in the leader region of the model sequence is a segment that would allow base-pair formation with the 3' end of the 16S rRNA in *E. coli*. Since base pairing has been demonstrated experimentally (6,7), our method is clearly suitable for highlighting *functional* bases within homologous RNA or DNA sequences despite the difficulties introduced by signal elements with variable distances from one another. The base pairing between the model sequence and the 3'end of the 16S rRNA is not complete, being interrupted on the 3' side of the AAGGAGGU signal. The same limited complementarity is seen with individual sequences (Table 1). The φXH and λ*cro* leader regions are notable exceptions to this rule. The preference for bases in the model sequence outside the base-pairing region suggests that RNA-protein interactions as well as RNA-RNA interactions are important in the binding of the leader region (1).

It is not clear whether the oligo (U) segment at the 5' end is indeed optional as its absence from some binding site fragments suggests (Table 1). It is quite possible that its expected lability would result in its being lost in many digestion experiments. (In Table 1 contrast the results of ref. 13 with ref. 21-23, of ref. 18 with ref. 19, and of ref. 37 with 50.)

Mutations in regulatory sequences provide direct clues to the

function of specific bases at specific positions. Only a few mutants involving ribosome binding sites have been characterized and sequenced. Two of these mutants suggest that a U immediately to the left of the AUG codon and an A immediately to the right might be important for translation (33,52). Possibly these bases participate in additional base-pairing with the $tRNA^{fmet}$. Functional significance for C at position +3 is suggested by a mutant which cannot be explained in terms of base-pairing (53). Other mutants would decrease the strength of base-pair formation between the AAGGAGGU signal and the 16S rRNA (54,55). An interesting set of mutants at the extreme 5' end of the *lac Z* ribosome binding site has been characterized (50). Apparently, some of these would enhance the stability of a loop and stem secondary structure in the *lac* RNA and therefore, influence translation negatively by decreasing the accessibility of the oligo (U) stretch present in the *lac Z*. The restart sites described for the *lac I* gene can be regarded as a special class of ribosome binding sites (56). Their similarity with the model sequence is rather distant (38) which may make it understandable that their efficiency is below 10% of that of wild type *lac I*.

Several proteins have been identified as participating in initiation (1). The roles of IF3 and of S1 in the selection of the correct AUG codon have been investigated but, it is still impossible to pinpoint exactly how these two proteins fullfill their functions. Only the AAGGAGGU and the codon-containing UAUGA signals are clearly involved in base pairing (with the 16S rRNA or $tRNA^{fmet}$ respectively).

It may be that any or all of the remaining segments of our model sequence are involved in binding with proteins. One peculiar feature of these prospective zones of protein-mRNA interaction is that they do not contain guanine. Guanine appears in the model sequence almost exclusively where we presume it is used for base pairing. The significance of this finding remains obscure but it is interesting that all known RNA sequences that bind to the S1 protein are also devoid of guanine (57-61).

In future studies it might become more useful to analyse groups of sequences having similar functional properties rather than looking for a common model for all ribosome binding sequences.

Obviously analysis of B. *stearothermophilus* ribosome binding sites is needed to reveal the subtler differences from those of E. *coli* (8-11). Certainly analyses analogous to the kind presented here are needed for all signal sequences where recognition is not based on unique nucleotide sequences.

ACKNOWLEDGEMENTS

This research was supported in part by a U.S. Public Health Service Grant (GM 17371 to SA).

Present addresses: [1]Botanisches Institut, Universität Bonn, D 5300 Bonn, Venusbergweg 22, GFR, and [2]Department of Chemistry, University of Edinburgh, King's Buildings, West Mains Road, Edinburgh, UK.

REFERENCES

1. Steitz, J.A. (1978) In Biological Regulation and Control (Plenum Publishing Co., R. Goldberger, Ed.), Vol. 1, in press
2. Ptashne, M., Backman, K., Humayun, M.Z., Jeffrey, A., Maurer, R., Meyer, B. and Sauer, R.T. (1976) Science 194,156-161
3. Baan, R.A., Hilbers, C.W., van Charldorp, E., van Leerdam, E., van Knippenberg, P.H. and Bosch, L. (1977) Proc. Natl. Acad. Sci. USA 74,1028-1031
4. Shine, J. and Dalgarno, L. (1974) Proc. Natl. Acad. Sci. USA 71,1342-1346
5. Steitz, J.A. and Stakes, K. (1975) Proc. Natl. Acad. Sci. USA 72,4734-4738
6. Steitz, J.A. and Steege, D.A. (1977) J. Mol. Biol. 114,545-558
7. Sprague, K.U., Steitz, J.A., Greenley, R.M. and Stocking, C.E. (1977) Nature 267,462-465
8. Steitz, J.A. (1973) J. Mol. Biol. 73,1-16
9. Goldberg, M.L. and Steitz, J.A. (1974) Biochemistry 13,2123-2129
10. Held, W.A., Gette, W.R. and Nomura, M. (1974) Biochemistry 13, 2115-2122
11. Isomo, K. and Isomo, S. (1976) Proc. Natl. Acad. Sci. USA 73, 767-770
12. Scherer, G.F.E., Walkinshaw, M.W. and Arnott, S. (1978) Nucleic Acids Res. 5,3759-3773
13. Steitz, J.A. (1969) Nature 224,957-964
14. Staples, D.H., Hindley, J., Billeter, M.A. and Weissman, C. (1971) Nature New Biol. 234,202-204
15. Hindley, J. and Staples, D.H. (1969) Nature 224,964-967
16. Staples, D.H. and Hindley, J. (1971) Nature New Biol. 234,211-212
17. Steitz, J.A. (1972) Nature New Biol. 236,71-75
18. Pieczenik, G., Model, P. and Robertson, H.D. (1974) J. Mol. Biol. 90,191-214

19. Beck, E., Sommer, R., Auerswald, E.A., Kurz, Ch., Zink, B., Osterburg, G., Schaller, H., Sugimoto, K., Sugisaki, K., Okamoto, T. and Takanami, M. (1978) Nucleic Acid Res. 5,4495-4503
20. Gupta, S.L., Chen, J., Schaefer, L., Lengyel, P. and Weissman, S.M. (1970) Biochem. Biophys. Res. Commun. 39,883-888
21. Fiers, W., Contreras, R., Duerink, F., Haegeman, G., Merregaert, J., Min Jou, W., Raeymaekers, A., Volckaert, G., Ysebaert, M., van de Kerckhove, J., Nolf, F. and van Montagu, M. (1975) Nature 256,273-278
22. Min Jou, W., Haegeman, G., Ysebaert, M. and Fiers, W. (1972) Nature 237,82-88
23. Fiers, W., Contreras, R., Duerink, F., Haegeman, G., Yseterentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., van den Berghe, A., Volckaert, G. and Ysebaert, M. (1976) Nature 260,500-507
24. Sanger, F., Coulson, A.R., Friedmann, T., Air, G.M., Barrell, B.G., Brown, N.L., Fiddes, J.C., Hutchison III, C.A., Slocombe, P.M. and Smith, M. (1978) J. Mol. Biol. 125,225-246
25. Dahlberg, J.E. and Blattner, F.R. (1975) Nucleic Acids Res. 2,1441-1448
26. Steege, D.A. (1977) J. Mol. Biol. 144,559-568
27. Schwarz, E., Scherer, G., Hobom, G. and Kössel, H. (1978) Nature 272,410-414
28. Godson, G.N., Barrell, B.G., Staden, R. and Fiddes, J.C. (1978) Nature 276,236-247
29. Pirrotta, V. (1979) Nucleic Acids Res. 6,1495-1508
30. Bayev, A.A., Zakharyev, V.M., Krayev, A.S., Skryabin, K.G., Monastryskaya, G.S., Sverdlov, E.D. and Ovchinnikov, Y.A. (1978) Bioorganic Chimia 4,1563-1565
31. Arrand, J.R. and Hindley, J. (1973) Nature New Biol. 244,10-13
32. Steitz, J.A. and Bryan, R.A. (1977) J. Mol. Biol. 114,527-543
33. Belin, D., Hedgpeth, J., Selzer, G.B. and Epstein, R.H. (1979) Proc. Natl. Acad. Sci. USA 76,700-704
34. Squires, C., Lee, F., Bertrand, K., Squires, C.L., Bronson, M.J. and Yanofsky, C. (1976) J. Mol. Biol. 103-351-381
35. Platt, T. and Yanofsky, C. (1975) Proc. Natl. Acad. Sci. USA 72,2399-2403
36. Selker, E. and Yanofsky, C. (1979) J. Mol. Biol. 130,135-143
37. Maizels, N. (1974) Nature 249,647-649
38. Steege, D.A. (1977) Proc. Natl. Acad. Sci. USA 74,4163-4167
39. Grindley, N.D.F. (1978) Cell 13,419-426
40. Musso, R.E., de Crombugghe, B., Pastan, I., Sklar, J., Yot, P. and Weissman, S. (1974) Proc. Natl. Acad. Sci. USA 71,4940-4944
41. Lee, N. and Carbon, J. (1977) Proc. Natl. Acad. Sci. USA 74, 49-53
42. Pirtle, R.M., Pirtle, I.L. and Inoue, M. (1978) Proc. Natl. Acad. Sci. USA 75,2190-2194
43. Post, L.E., Strycharz, G.D., Nomura, M., Lewis, H. and Dennis, P.P. (1979) Proc. Natl. Acad. Sci. USA 76,1697-1701
44. Sutcliffe, J.G. (1978) Proc. Natl. Acad. Sci. USA 75,3737-3741
45. Zurawski, J.G., Brown, K., Killingly, D. and Yanofsky, C. (1978) Proc. Natl. Acad. Sci. USA 75,4271-4275
46. Barnes, W.M. (1978) Proc. Natl. Acad. Sci. 75,4281-4285
47. Otsuka, A. and Abelson, J. (1978) Nature 276,689-694
48. Gardner, J.F. (1979) Proc. Natl. Acad. Sci. USA 76,1706-1710

49. Post, L.E., Arfsten, A.E., Reusser, F. and Nomura, M. (1978) Cell 15,215-229
50. Cannistraro, V.J., Kennell, D. (1978) Nature 277,407-409
51. Spiegelman, W.G., Reichardt, L.F., Yaniv, M. and Heinemann, S.F. (1972) Proc. Natl. Acad. Sci. USA 69,3156-3160
52. Tanaguchi, T. and Weissmann, C. (1978) J. Mol. Biol. 118,533-565
53. Atkins, J.F., Steitz, J.A., Anderson, C.W. and Model, P. (1979) Cell 18,247-256
54. Dunn, J.J., Buzash-Pollert, E. and Studier, F.W. (1978) Proc. Natl. Acad. Sci. USA 75,2741-2745
55. Roberts, T.M., Kacich, R. and Ptashne, M. (1979) Proc. Natl. Acad. Sci. USA 76,760-764
56. Files, J.G., Weber, K. and Miller, J.H. (1974) Proc. Natl. Acad. Sci. USA 71,667-670
57. Dahlberg, A.E. and Dahlberg, J.E. (1975) Proc. Natl. Acad. Sci. USA 72,2940-2944
58. Weber, H., Billeter, M.A., Kahane, S., Weissmann, C., Hindley, J. and Porter, A. (1972) Nature New Biol. 237,166-170
59. Senear, A.W. and Steitz, J.A. (1976) J. Biol. Chem. 251,1902-1912
60. Draper, D.E. and van Hippel, P.H. (1978) J. Mol. Biol. 122, 339-359
61. Krol, A., Branlant, C., Ebel, J.-P. and Visentin, L.P. (1977) FEBS Letters 80,255-260