Volume 10 Number 9 1982 Nucleic Acids Research

Characterization of translational initiation sites in *E. coli*

Gary D.Stormo, Thomas D.Schneider and Larry M.Gold

Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, CO 80309, USA

Received 26 October 1981; Revised and Accepted 5 April 1982

ABSTRACT

We characterize the Shine and Dalgarno sequence of 124 known gene beginnings. This information is used to make "rules" which help distinguish gene beginning from other sites in a library of over 78,000 bases of mRNA. Gene beginnings are found to have information besides the initiation codon and Shine and Dalgarno sequence which can be used to make better "rules".

INTRODUCTION

We have set out to study translational initiation by investigating the features of an mRNA that allow recognition by the ribosome. We wish to know, qualitatively, the requirements for a ribosome binding site on an mRNA; secondly, we wish to know quantitatively what features of an mRNA dictate the rate of translation of one mRNA versus another. We present here the first attempt, using a nucleic acid sequence library and assorted programs, at defining nucleotides that may play a role in the selection of initiation codons by the ribosomes of Escherichia coli.

The analyses performed do not fall into a conceptual vacuum. When Shine and Dalgarno sequenced the 3' end of 16S rRNA, they found that it contained nucleotides complementary to each of the then known translational initiation regions, just 5' to the initiation codon (1). They hypothesized that base pairing between the mRNA and rRNA might be an important step in the ribosomal selection of initiation sites. There is now considerable evidence, both biochemical and genetic, that supports that hypothesis (2, 3, 4).

It is clear, however, that there is not a simple relationship between the translational efficiency of a particular gene and its complementarity to 16S rRNA. Inhibitary effects of mRNA secondary structure have been proposed (5, 6) to account for the results in some cases. Scherer et al. (7) also showed that nucleotides from the entire region from -20 to +15, relative to the initiation codon, occur nonrandomly. They did not, however, try to use that information to distinguish the initiation regions from other sites with

0305-1048/82/1009-2971$2.00/0

similar sequences. We have used a rather large library of E. coli, S. typhimurium, and coliphage nucleic acid sequences to study the non-randomness of nucleotides surrounding initiation codons. We have found that the non-randomness is limited to the nucleotides protected by the ribosome against RNase when an mRNA is placed into an initiation complex. Furthermore, when other sequences from the library are selected because they contain an initiation codon and an appropriately-placed polypurine domain, those sequences are random everywhere else around the initiation codon. We infer that some, or all, of the positions of non-random nucleotides may be involved in translational initiation.

METHODS

All sequences are stored in our nucleic acid sequence data bank, as described in Schneider et al. (8). At the time of these analyses, this library contained 95,979 bases from E. coli, coliphage and S. typhimurium. (The S. typhimurium genes included have been shown to be translated in E. coli.) Information in the literature was used to determine where the transcribed and translated regions are in these sequences.

mRNA library

Since we wanted to compare real to potential ribosome binding sites, the entire library was pared down to only those sequences which occur as mRNA. This new data base, called the mRNA library, contained 78,612 nucleotides. It was constructed so that all regions which are transcribed are included, but only once. If a region of DNA is included on two distinct transcripts, only the longest one was included in this library.

Gene beginnings

All gene start sites which are known to us are recorded in this library. A Delila instruction set was made which named each gene and then requested: "get from gene beginning -X to gene beginning +Y;". The current library contains 124 known gene beginnings, and the set from -25 to +20 is shown in Table 1. For most of the analyses the region of -60 to +40 was used.

Nongenes

We have called "nongenes" those sites which resemble ribosome binding sites, by various criteria, but for which there is no evidence that they function as such. A search program was used to find sequences which are "gene beginning-like". One output of that program is a set of instructions of the form: "get from (some site) -X to (that site) +Y;". Such a search would

TABLE 1.

TRANSLATIONAL INITIATION REGIONS: -25 TO 20

```
                ----------------                    ++++++++++
                2222221111111111--------- +++++++++11111111112
                5432109876543210987654321012345678901234567890
                ..........................................
```

ORGANISM MS2		
A	1	ATTCCATTCCTAGGAGGTTTGACCTGTGCGAGCTTTTAGTACTCTC
COAT	2	AGAGCCCTCAACCGGAGTTTGAAGCATGGCTTCTAACTTTACTCAG
LYSIS	3	AAGGCAATGCAAGGTCTCCTAAAAGATGGAAACCCGATTCCCTCAG
REPLICASE	4	GGCCATTCAAACATGAGGATTACCCATGTCGAAGACAACAAAGAAG
ORGANISM QBETA		
A	5	TACTTCACTGAGTATAAGAGGACATATGCCTAAATTACCGCGTGGT
COAT	6	GTTGAAACTTTGGGTCAATTTGATCATGGCAAAATTAGAGACTGTT
REPLICASE	7	TGCTTAGTAACTAAGGATGAAATGCATGTCTAAGACAGCATCTTCG
ORGANISM PHIX174		
A	8	AAATCTTGGAGGCTTTTTTATGGTTCGTTCTTATTACCCT
A*	9	TTCAAGATTGCTGGAGGCCTCCACTATGAAATCGCGTAGAGGCTTT
B	10	CTGCTAAAGGTCTAGGAGCTAAAGAATGGAACAACTCACTAAAAAC
K	11	AAGTTAACACTTTCGGATATTTCTGATGAGTCGAAAAATTATCTTG
C	12	AATCGAAGTGGACTGCTGGCGGAAAATGAGAAAATTCGACCTATCC
D	13	TTCAACCACTAATAGGTAAGAAATCATGAGTCAAGTTACTGAACAA
E	14	CGTCGCTGCGTTGAGGCTTGCGTTTATGGTACGCTGGACTTTGTGG
J	15	AATTACGTGCGGAAGGAGTGATGTAATGTCTAAAGGTAAAAAACGT
F	16	TCGGCCCCTTACTTGAGGATAAATTATGTCTAATATTCAAACTGGC
G	17	AGGTTTTCTGCTTAGGAGTTTAATCATGTTTCAGACTTTTATTTCT
H	18	TCCAGCCACTTAAGTGAGGTGATTTATGTTTGGTGCTATTGCTGGC
ORGANISM G4		
A	19	ATCAAACGGAGGCTTTTCATGTTTAAAGTACATTCTGAC
A*	20	CTCAAAAATCTTGGAGGAGTCAACTATGAAGTCTCGACGTGGCTTT
B	21	CCGTAAAAGGCCTAGGGAATAAAGAATGGAACAATTCACTCAAAAC
K	22	AAATTAACAAGTACGGATATTTCTGATGAAACCAAAAACTACGTTG
C	23	AATCGAAGTGGACTGCTGGTGGAAAATGAGGAAATTCAATCTCAAC
D	24	GACACAAACCACAAAGGAAACTGAAATGTCTAAATCAAACGAATCT
E	25	CGCCGTGCTATCGAGGCTTGCGTATATGGAACACTGGACTTTGTCG
J	26	GTCTTCACTTTTAAGGAGTTATGTAATGAAAAAATCAATTCGCCGC
F	27	TCCCACTCTATTTAAGGATACAAAAATGTCTAACGTTCAAACATCT
G	28	ACTGCAAAGCCAAAAGGACTAACATATGTTCCAGAAATTCATTTCT
H	29	CAACCTCTGAAATAAGGATTATCCTATGTTTGGCTCTATCGCTGGC
ORGANISM FD		
II	30	TTTCTGATTATCAACCGGGGTACATATGATTGACATGCTAGTTTTA
X	31	GTTTAAAGCATTTGAGGGGGATTCAATGAATATTTATGACGATTCC
V	32	TTAAAATCGCATAAGGTAATTCAAAATGATTAAAGTTGAAATTAAA
VII	33	CTGCGCCTCGTTCCGGCTAAGTAACATGGAGCAGGTCGCGGATTTC
IX	34	TTGGTATAATCGCTGGGGGTCAAAGATGAGTGTTTTAGTGTATTCT
VIII	35	TTACCCGTTTAATGGAAACTTCCTCATGAAAAAGTCTTTAGTCCTC
III	36	GCCTTTTTTTTTGGAGATTTTCAACGTGAAAAAATTATTATTCGCA
VI	37	TACTGCGTAATAAGGAGTCTTAATCATGCCAGTTCTTTTGGGTATT
I	38	CTTATTTGGATTGGGATAAATAAATATGGCTGTTTATTTTGTAACT
IV	39	GTTTCAATTAAAAAAGGTAATTCAAATGAAATTGTTAAATGTAATT

TABLE 1. CONTINUED

```
                    ----------------                    ++++++++++
                    2222221111111111--------- +++++++++11111111112
                    5432109876543210987654321012345678901234567890
                    ........................................

ORGANISM LAMBDA
XIS            40 TGCGTGTAATTGCGGAGACTTTGCGATGTACTTGACACTTCAGGAG
INT            41 GGTGGCCTTTTGAAGAGGATCAGAAATGGGAAGAAGGCGAAGTCAT
N              42 AAAGCTAACTGACAGGAGAATCCAGATGGATGCACAAACACGCCGC
CI             43 TCCCTTGCGGTGATAGATTTAACGTATGAGCACAAAAAAGAAACCA
CRO            44        ATGTACTAAGGAGGTTGTATGGAACAACGCATAACCCTG
CII            45 TTGTTATCTAAGGAAATACTTACATATGGTTCGTGCAAACAAACGC
O              46 ACTGGATCTATCAACAGGAGTCATTATGACAAATACAGCAAAAATA

ORGANISM T4
RIIB           47 CCCTTGCGGCCTAATAAGGAAAATTATGTACAATATTAAATGCCTG
G23            48        TTTTAAAGGTTAACACAAATGACTATCAAAACT
G32            49 AAATTAAATTAAAAAGGAAATAAAAATGTTTAAACGTAAATCTACT
G36            50 TACTATTAAAATAAAGGGGCATACAATGGCTGATTTAAAAGTAGGT
G37            51 ATTTCGGCTATTATTAAGAGGACTTATGGCTACTTTAAAACAAATA
G38            52 GAGAGGGGCTTCGGCCCTTCTAAATATGAAAATATATCATTATTAT

ORGANISM T7
0.3            53 TAATAACTGCACGAGGTAACACAAGATGGCTATGTCTAACATGACT
0.4            54 AGGAGTACGAGGAGGATGAAGAGTAATGTCTACTACCAACGTGCAA
0.5            55 TTATCACTTTACTTATGAGGGAGTAATGTATATGCTTACTATCGGT
0.7            56 TAACGAACATAAAGGACACAATGCAATGAACATTACCGACATCATG
1              57 ATTTACTAACTGGAAGAGGCACTAAATGAACACGATTAACATCGCT
1.1            58 GAATTACTAAGAGAGGACTTTAAGTATGCGTAACTTCGAAAAGATG
1.2            59 AAGCGTAGCTGGGAGGGTCAGTAAGATGGGACGTTTATATAGTGGT
1.3            60 ATTTAACCAATAGGAGATAAACATTATGATGAACATTAAGACTAAC
1.7            61 TCAAACGAATCAAGGAGGTGTTCTGATGGGACTGTTAGATGGTGAA
2              62 TTTGGAAATCGAGAGGTCAATGACTATGTCAAACGTAAATACAGGT
2.5            63 ACGAAACCTAAAGGAGATTAACATTATGGCTAAGAAGATTTTCACC
3              64 TGTGATATACGCAAAGGGAGGCGACATGGCAGGTTACGGCGCTAAA
3.5            65 TAAAAAGGAAAGGAGGAAAGAAATAATGGCTCGTGTACAGTTTAAA
4A             66 ATTTATAGAACTAGGAGGGAATTGCATGGACAATTCGCACGATTCC
4B             67 ACGGAAACCCTCAGGAGGTAAACCAATGACTTACAACGTGTGGAAC
9              68 GTTCAACTTTAAGGAGACAATAATAATGGCTGAATCTAATGCAGAC
10             69 TTAACTTTAAGAAGGAGATATACATATGGCTAGCATGACTGGTGGA
17             70 CTTAGATTTACTTTAAGGAGGTCAAATGGCTAACGTAATTAAAACC

ORGANISM ECOLI
THRLEADER      71 AAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACC
THRA           72 GACCAAAGGTAACGAGGTAACAACCATGCGAGTGTTGAAGTTCGGC
FOLA           73   ATTTTTTTTATCGGGAAATCTCAATGATCAGTCTGATTGCGGCG
ARAC           74 TCTGAATGGCGGGAGTATGAAAAGTATGGCTGAAGCGCAAAATGAT
ARAB           75 CCGTTTTTTTGGATGGAGTGAAACGATGGCGATTGCAATTGGAATC
LACY           76 CGTATTTCGCGTAAGGAAATCCATTATGTACTATTTAAAAAACACA
LACI           77 AGAGAGTCAATTCAGGGTGGTGAATGTGAAACCAGTAACGTTATAC
LACZ           78 TAACAATTTCACACAGGAAACAGCTATGACCATGATTACGGATTCA
GALE           79 TACCATAAGCCTAATGGAGCGAATTATGAGAGTTCTGGTTACCGG
GALT           80 GGGATATCCCGATTAAGGAACGACCATGACGCAATTTAATCCCGTT
OMPA           81 TTTGGATGATAACGAGGCGCAAAAAATGAAAAAGACAGCTATCGCG
TRPLEADER      82 AGTTCACGTAAAAAGGGTATCGACAATGAAAGCAATTTTCGTACTG
TRPE           83 TTTTGAACAAAATTAGAGAATAACAATGCAAACACAAAAACCGACT
TRPD           84 CGCATCATGCACAGGAGACTTTCTGATGGCTGACATTCTGCTGCTC
```

TABLE 1. CONTINUED

```
                    ----------------                    ++++++++++
                    2222221111111111--------- +++++++++111111111112
                    5432109876543210987654321012345678901234567890
                    ........................................

TRPC          85          GCGGCACGAGGGTAAATGATGCAAACCGTTTTAGCGAAA
TRPB          86 TGCGCGCATATTAAGGAAAGGAACAATGACAACATTACTTAACCCC
TRPA          87 TTTTGAAAGCACGAGGGGAAATCTGATGGAACGCTACGAATCTCTG
LIPO          88 TAACTCAATCTAGAGGGTATTAATAATGAAAGCTACTAAACTGGTA
HISLEADER     89 TGAATAAACATTCACAGAGACTTTTATGACACGCGTTCAATTTAAA
HISG          90 GTTCAGACAGGTTTAAGAGAATACAATG
PHELEADER     91      AAGTCACTTAAGGAAACAAACATGAAACACATACCGTTTTTC
PHEA          92 TTATTGATAACAAAAAGGCAACACTATGACATCGGAAAACCCGTTA
RECA          93 ATTACCCGGCATGACAGGAGTAAAAATGGCTATCGACGAAAACAAA
RPOA          94 AGCTTAGTACCAAAGAGAGGACACAATGCAGGGTTCTGTGACAGAG
S11           95 AAGAAATAATCGGGGTGATTGAATAATGGCAAAGGCACCAATTCGT
S13           96 ACATATTAAATAGTAGGAGTGCATAGTGGCCCGTATAGCAGGCATT
L14           97 GTAGTTGACATTAGCGGAGCCTAAAATGATCCAAGAACAGACTATG
EFG           98 GCCTAAAAGATAAACGAGGAAACAAATGGCTCGTACAACACCCATC
S12           99 AAAAGCTAAAACCAGGAGCTATTTAATGGCAACAGTTAACCAGCTG
S7           100 CCTGAATTAACAACGGAGTATTTCCATGCCACGTCGTCGCGTCATT
ASN          101 GCAACAGGACGCAGGAGTATAAAAAATGAAAACCGCTTACATTGCC
ILVLEADER    102 CGAACAAGATGCAAGAAAAGACAAAATGACAGCCCTTCTACGAGTG
ILVG         103 CAAGATTCAGGACGGGGAACTAACTATGAATGGCGCACAGTGGGTG
ILVE         104 GCCTGAGCGCAAAAGGAATATAAAAATGACCACGAAGAAAGCTGAT
L11          105 TATTACCCAACTTGAGGAATTTATAATGGCTAAGAAAGTACAAGCC
L1           106 GGGCCTGGTAGTGGAGGACTAAGAAATGGCTAAACTGACCAAGCGC
L10          107 TGGCAAACATCCAGGAGCAAAGCTAATGGCTTTAAATCTTCAAGAC
L7/L12       108 ATTCTGATATTCAGGAACAATTTAAATGTCTATCACTAAAGATCAA
BETA         109 ACTTGTCAGCGAGCTGAGGAACCCTATGGTTTACTCCTATACCGAG
LAMB         110 GAAAAGCAATGACTCAGGAGATAGAATGATGATTACTCTGCGCAAA
TRPR         111 CCCCCGCTAACAATGGCGACATATTATGGCCCAACAATCACCCTAT
CLOIMM       112 TAAAAAATATCTTTAAGAGGTAAATATGGGGCTTAAATTACATATT
TNPA         113 AAAGCCGTCAGGCAGGGAGGCCGATATGCCCGTTGATTTTTTGACC
TNPR         114 GATATTCGATTTAAGGTACATTTTTATGCGAATTTTTGGTTATGCG
BLA          115 TCAATAATATTGAAAAAGGAAGAGTATGAGTATTCAACATTTCCGT
CAM          116 ATTTTCAGGAGCTAAGGAAGCTAAAATGGAGAAAAAAATCACTGGA
ST1          117 GTGAAACAACATGACGGGAGGTAACATGAAAAAGCTAATGTTGGCA
ELTB         118 TATATAACAGAATTCGGGATGAATTATGAATAAAGTAAAATGTTAT

ORGANISM STYPH
HISLEADER    119 TGAATAAGCATTCATCGGAATTTTTATGACACGCGTTCAATTTAAA
HISG         120 CAGACAGGATAAAGAGGAACGCAGAATGTTAGACAACACCCGCTTA
HISD         121 GGAACCCTGTGATGGAGTAAAGACCATGAGCTTCAATACCCTGATT
HISC         122 GCGTAAACGCCCTCAAGGAGCAAGCATGAGCACTGAAAACACTCTC
HIN          123 CTTATCTGATGTAAAGGAGAAAATCATGGCTACTATTGGGTATATT
H2           124 AAAAACCTTCCAAAAGGAAAATTTTATGGCACAAGTAATCAACACT
                  ........................................
```

REFERENCES (BY SEQUENCE NUMBER): 1-3(6);4(10);5-6(11);7(12);
8-18(13);19-29(14);30-39(15);40-41(16);42(17);43-46(18,19);
47(20);48(YOUNG,PERSONAL COMMUNICATION);49(21);50-52(22);
53-67(23);68-70(24);71-72(25);73(26);74-75(27,28);76(29);
77(30);78(31);79(32);80(33);81(34);82-87(35-38);88(39);
89-90(40);91-92(41);93(42);94-100(43-46);101(47);102-103(48);
104(49);105-109(50);110(51);111(52);112(53);113-115(54);
116(55);117(56);118(57);119-120(58);
121-122(BARNES,PERSONAL COMMUNICATION);123-124(59).

return sites which belong to both the gene and nongene classes. Another program was used which separated out the gene sites from that set so that one is left with a set of instructions to get a class of sequences which resemble gene beginnings, but are not known to be so. Dunn and Studier (23) identified in T7, besides the beginnings of known genes, eight sites which they called potential genes because of their Shine and Dalgarno sequence preceding an AUG and followed by a long open reading frame. Since this work is attempting to define what is involved in determining ribosome binding, these sites were not included in either the gene or nongene sets.

Programs and data flow

The programs which were used in the analyses are listed in Table 2 and will be described in the sections where the results from them are discussed. All programs (except the secondary structure predicter) are written in Pascal and are part of the Delila system (9).

Figure 1 diagrams the overall flow of data described in this paper. As in Schneider et al. (8) boxes represent data and dots represent programs with the associated arrows showing the flow of information. Starting with the Master Library (box #1) the instructions to pull out only the mRNA sequences

TABLE 2
List of Programs Used

Name	Version	Purpose
Achsq	2.01	Aligned chi-squared analysis. See also Alist.
Alist	3.00	Produces an aligned listing of a library using Delila instructions (Table 1)
Comp	4.20	Composition of a library
Count	2.30	Counts the number of bases in a library
Delila	1.20	Librarian: extracts libraries based on Delila instructions
Helix	1.10	Find all possible helices between two libraries
Hist	3.01	Histogram columns of aligned libraries. See also Alist
Search	2.01	Search for patterns by rules, produces Delila instructions
Sepa	1.07	Separates Delila instruction sets
Struct	none	Find best RNA structure by dynamic programming algorithm (designed by Eugene Myers and not part of Delila system)

(box #2) are written, and the mRNA Library (box #3) is generated. Then the instructions to get the regions around translational initiation sites (called "gene beginnings", box #7) are written. Now only the "rule" input to the search program (box #5) is specified by hand. All other manipulations of the data are automated and require the user only to initiate whatever analysis is wanted.

RESULTS AND DISCUSSION

(a) mRNA Library Composition

Many of the analyses that follow are based on the deviation of nucleotides from expected frequencies. We counted (box #4, Figure 1) the number of each mono-, di- and tri-nucleotide in the transcript library (Table 3). While the mononucleotides are not equally frequent, varying from 23.1% for C to 26.2% for A, the di-nucleotides are even more nonrandom. There are 1.7 times as many AAs as CCs, rather than 1.3 as expected from the mono-nucleotide composition. AAA is 3.2 times as abundant as CCC, rather than the

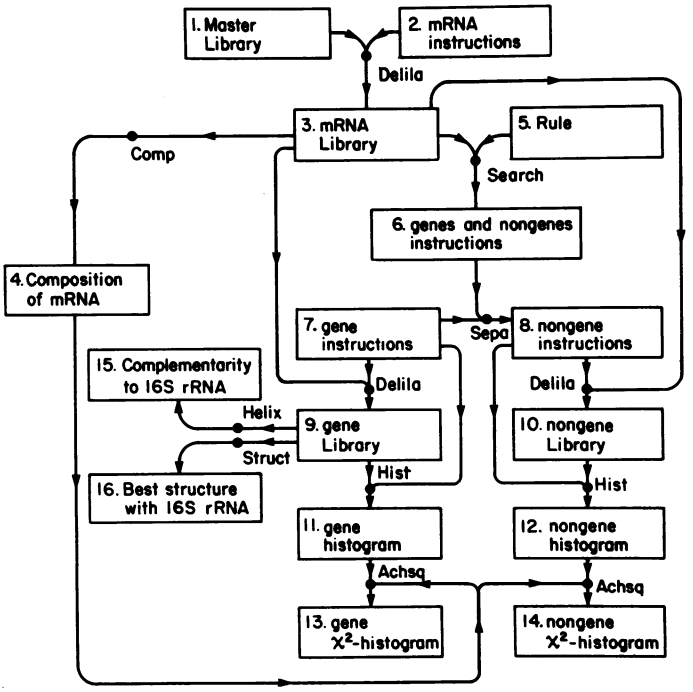

Figure 1. Flow diagram. Boxes are data files (described in the text) and dots are programs (described in Table 2). The arrows indicate the direction of information flow.

TABLE 3

COMPOSITION UP TO TRIPLETS OF MRNA LIBRARY

A	20597	C	18175	G	19271	T	20569
AA	6673	AC	4501	AG	4130	AT	5285
CA	4627	CC	3816	CG	4754	CT	4962
GA	4959	GC	5269	GG	4563	GT	4462
TA	4315	TC	4583	TG	5801	TT	5854
AAA	2228	AAC	1460	AAG	1476	AAT	1508
ACA	1145	ACC	1088	ACG	1112	ACT	1155
AGA	1105	AGC	1110	AGG	1060	AGT	851
ATA	1029	ATC	1241	ATG	1547	ATT	1467
CAA	1460	CAC	906	CAG	1209	CAT	1049
CCA	970	CCC	688	CCG	1175	CCT	977
CGA	1094	CGC	1385	CGG	1018	CGT	1257
CTA	880	CTC	1055	CTG	1802	CTT	1223
GAA	1580	GAC	1109	GAG	915	GAT	1352
GCA	1260	GCC	1108	GCG	1358	GCT	1538
GGA	1037	GGC	1358	GGG	811	GGT	1347
GTA	997	GTC	877	GTG	1125	GTT	1460
TAA	1394	TAC	1024	TAG	529	TAT	1367
TCA	1252	TCC	930	TCG	1105	TCT	1292
TGA	1718	TGC	1410	TGG	1665	TGT	1004
TTA	1409	TTC	1409	TTG	1325	TTT	1701

1.5 times expected. Different di-nucleotides with the same mono-nucleotide composition can have very different frequencies; there are 5801 occurrences of TG, but only 4462 occurrences of GT (see also 60). The disparity between frequencies is most noticeable in the tri-nucleotides. The most abundant triplet, AAA, occurs more than 4 times as often as the least abundant, TAG. Furthermore, TAG is less than one-third as frequent as TGA, a triplet with the same mono-nucleotide composition and the same function (as a translational terminator) when encountered by a ribosome in the proper frame. This clearly indicates that there are oligo-nucleotide biases in mRNA sequences.

(b) An Analysis of the Shine and Dalgarno Hypothesis

By the Shine and Dalgarno hypothesis, an initiation complex (containing initiation factors, GTP, initiator tRNA, mRNA, and a 30S and 50S ribosomal particle) uses mRNA-rRNA annealing to place the initiation codon into the ribosomal P site so that codon-anticodon scanning may occur. Successful codon-anticodon recognition will be followed by entry of the next charged amino-acyl tRNA into the ribosomal A site. The set of rRNA-mRNA-tRNA interactions that lead to translational initiation are shown in Figure 2.

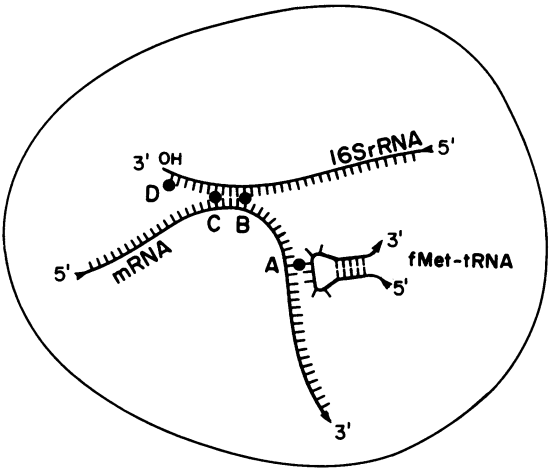

Figure 2. Base pairing interactions between rRNA, mRNA, and tRNA are shown as they occur during initiation of protein synthesis. Positions of particular interest are: A - base pair between the 5' nucleotide of the initiation codon and the 3' nucleotide of the anticodon; B and C - end points of the rRNA-mRNA annealing; D - 3' end of the 16S RNA.

To characterize the rRNA-mRNA interaction, we used a program (Helix) which finds complementary regions between two nucleotide sequences (box #15 of Figure 1). We compared the last 13 nucleotides of 16S rRNA with the 15 nucleotides preceding the initiation codon of each of the 124 genes; we did not allow G-U pairs, and we required at least three contiguous base pairs. Only one sequence, the trpR gene (#111 of Table 1), does not yield three contiguous base pairs. However, several of the mRNAs do not have a unique best complementarity with 16S RNA. For instance, sequence #16, the F gene of ϕX174, has the sequence GAGGA which can anneal to either UCCU or CUCC, but no contiguous five base pairs are possible. There are 29 cases of ambiguous "Shine and Dalgarno" alignments in the set of 124 genes. Rather than include multiple alignments for some genes in the following analysis, we used only those genes for which a single best complementarity between the rRNA and mRNA is obvious.

For the 94 mRNA sequences with unambiguous, three base pair or longer complementarity to the 3' end of 16S RNA, we analyzed the predicted interactions to get a measure of typical Shine and Dalgarno elements. Figure 3 shows the number of times each base, on both the rRNA and mRNA, is involved in the predicted annealing. The interactions center about the nucleotides

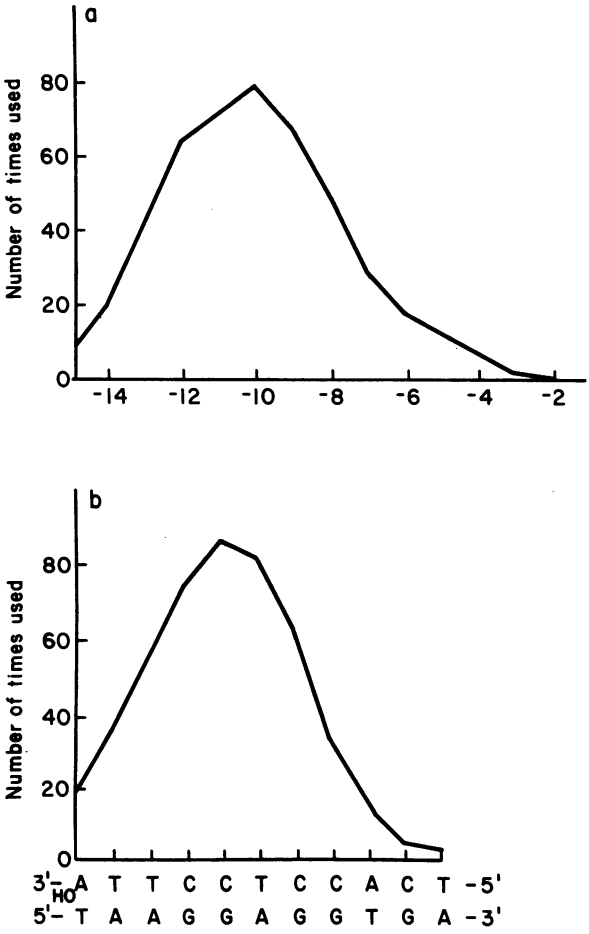

Figure 3. Nucleotides involved in rRNA-mRNA annealing are shown for the 94 sequences with unambiguous Shine and Dalgarno regions.

a) mRNA nucleotides that may anneal to 16S RNA; the positions are relative to the 5' base of the initiation codon

b) rRNA nucleotides that may anneal to the mRNAs; the top line of the ordinate represents the 3' nucleotides of 16S RNA (written as DNA), and the bottom line represents the mRNA nucleotides that would anneal to each rRNA nucleotide

5'- UCC -3' of the 16S rRNA and about position -10 of the mRNA.

We then measured the distances (in nucleotides) between points A, B, C, and D of Figure 2 for the 94 mRNAs that have unambiguous predicted alignments with the 3' end of 16S RNA. Each distance is given in Table 4, although only three distances are independent. For these 94 mRNAs an average picture of an initiation complex is obtained.

The Shine and Dalgarno hypothesis explicitly proposes annealing between the mRNA and rRNA, and one can ask about binding energies rather than (visual) nucleotide complementarity. We assume that any allowed RNA-RNA interaction is also legitimate during the initiation of protein synthesis. We calculated the free energies of all 124 genes (as a most stable predicted structure: G:U pairs and bulges are allowed) using published rules (61) (box #16 of Figure 1), with the convention that no energy contribution for a hairpin was added. The distribution of ΔG's is shown in Figure 4.

In most cases, the lowest energy structure is identical to the one predicted from nucleotide complementarity. However, there are exceptions. The most stable predicted structure for the Qβ A cistron (#5) is:

```
                C
               / \
3' -- A U U C U C C — 5' (16S RNA)
5' — U A A G A G G -- 3' (mRNA)
```

We know of no data which addresses whether such bulged structures are allowed to occur on 16S rRNA. The most stable predicted structure for trpE (#83) is:

```
3' -- U C C U — 5' (16S RNA)
5' — A G G A -- 3' (mRNA)
       \ /
        A
```

This structure is nearly 2 Kcal more stable than one involving only the GAG (paired to CUC). Perhaps the mRNA is allowed to bulge in these initiation intermediates, but, again, no experimental data are available.

(c) Placing the Initiation Codon into the P Site

TABLE 4

Distances Between Points in Figure 2

(measured in nucleotides)

Points	Mean Distance	Standard Deviation
A - B	6.9	2.0
A - C	11.9	2.1
A - D	15.1	2.2
B - C	5.0	1.2
B - D	8.1	1.5
C - D	3.1	1.6

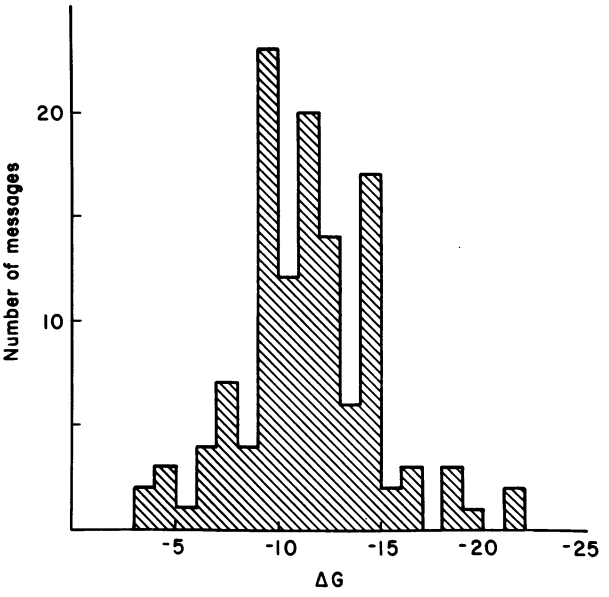

Figure 4. Energy of rRNA-mRNA annealing was calculated using a secondary structure predicting program. We used the 3' terminal 9 nulceotides of 16S RNA and the nucleotides from -18 to -4 (relative to the initiation codon) for each mRNA. Published rules for secondary structure calculations were used. Since these data are for an intermolecular annealing, no energy was added for loop closure. The data presented are for the most stable predicted structures.

The precise biochemical role of the Shine and Dalgarno element on an mRNA has not been defined (see 2). One plausible role for this element might be to allow the initiation codon [nearly always an AUG (see Table 1)] to interact favorably with the fMet.tRNA$_f^{met}$ that is (or will arrive) in the ribosomal P site. Clearly the E. coli initiator tRNA is different from other prokaryotic tRNAs (62), and has a special affinity toward the P site (63). Interestingly, annealing of a short oligonucleotide to the 3' end of 16S RNA stimlulates initiator tRNA binding to the ribosome even when the bound oligonucleotide contains no AUG (64).

When initiation complexes are treated with RNases, the 3' domains of the bound mRNA are cleaved independently of their A to B or A to D distances (see Figure 2). The published data for "bind-and-chew" experiments with mRNAs #47 (T4 rIIB), #53 (T7 0.3), and #87 (trpA) (65, 66, 67) are consistent with alignment of the 3' domains in an initiation complex by the initiation codon, not the Shine and Dalgarno element or the space between it and the initiation codon. Similar data have been reported for the f2 coat cistron (68). One

might imagine that mRNAs are also aligned by the Shine and Dalgarno sequence, and that the 5' boundaries of protected mRNAs will extend a fixed distance from point C or the projection of point D on to those mRNAs (Figure 2). The 5' domains of protected mRNAs tend to be more ragged than the 3' sides, and we are unable to draw a conclusion for this end. The "bind-and-chew" data for the T4 rIIB mRNA suggest that a hairpin 5' to the Shine and Dalgarno sequence can be protected by the ribosome against RNAase attack (65). This observation, plus our thought about sequence #52 (below), suggests that the

TABLE 5

Ribosome Binding Site "Rules"

Rule Number	Description[1] (S/D[2] - space - ATG)			Genes (%) Found	Nongenes Found
0			A/GTG	124 (100)	2548
0.5			ATG	120 (97)	1419
1	AGG GGA GAG	4E5N	ATG	105 (85)	201
2	AGG GGA GAG	3E6N	ATG	103 (83)	167
3	AGG GGA GAG	2E7N 9N 1E6N	ATG	91 (73)	95
4	AAGG AGGA GGAG GAGG AGGT	6E4N	ATG	95 (77)	111
5	AAGG AGGA GGAG GAGG	4E5N	ATG	90 (73)	68
6	AGGA GGAG GAGG	4E5N	ATG	83 (67)	44
7	AAGGA AGGAG GGAGG	4E5N	ATG	48 (39)	21

1. N is any base, E is any base or no base. 4E5N is the search command for "five to nine unspecified bases".
2. S/D is the searched for Shine and Dalgarno sequence.

ribosome is "roomier" 5' to the initiation codon than 3' to it. This is, of course, the domain of an mRNA upon which the ribosome, during elongation, no longer acts.

The A to B distance is rarely less than 5 nucleotides. The constraint reflects, probably, the role of the A to B distance on the rate at which the initiation codon can contact the ribosomal P site. The A to B distance may, under some circumstances, be rather large. The thr leader (sequence #71) has an A to B value of 12 nucleotides, suggesting that the ribosome can accummodate some extra nucleotides between the Shine and Dalgarno region and the initiation codon. An extreme example of this hypothesis is the situation prediced for the T4 gene 38 mRNA (#52). In this case, a Shine and Dalgarno region is separated from the initiation codon by 23 nucleotides; by hypothesis an intramolecular hairpin reduces the A to B distance to 5 nucleotides (see 2) and facilitates the entry of the initiation codon into the P site.

(d) Rules for Translational Initiation Sites

We wrote a search program which allows the user to find strings of specified bases. We used this program to define different "Shine/Dalgarno-space-initiation codon" sequences (called "rules" box #5, Figure 1) and then searched the mRNA library for those strings. As an example, consider A/GTG as a rule (rule 0 in Table 5). From the composition (Table 3), we see that there are 2672 ATGs and GTGs in the mRNA library. This rule, of course, finds all 124 genes, but also finds 2548 other sites which we call nongenes. Probably some of the nongenes function as ribosome binding sites, but not all 2548 of them could; rule 0 is not good at defining a translational initiation region. If we limit ourselves to ATGs (rule 0.5), we find 120 of the 124 genes, and now find only 1419 nongenes [excluding 8 sites in T7 that Dunn and Studier (23) call potential genes, but that could be nongenes; see Methods]. This seems an improved rule, but still far from meeting the implicit goal: can one define a rule of the type "Shine/Dalgarno- space-initiation codon" which is nearly perfect, i.e., which identifies nearly all the genes and almost no other sites? Table 5 shows the results of several rule attempts. The choices for the allowed Shine and Dalgarno sequences and the spacing come from our characterization of those features (Figure 3 and Table 4). Other choices do worse as predictors of gene beginnings. Rule 3 (of Table 5) is an attempt to use different spacings for different Shine and Dalgarnos. It is clear that we can diminish the number of nongenes only by finding fewer genes. In fact, if we require a five long Shine and Dalgarno, as in rule 7, the nongenes get quite low, but 60% of the genes are not found. Since we know the gene sites do function during translational initiation, while we suspect that some

nongenes also function as ribosome binding sites, the more important criterion in judging a rule should be that we identify a large portion of the genes. We arbitrarily require that any rule should find at least 80% of the genes. That makes only rules 0, 0.5, 1 and 2 acceptable, and rule 2 wins by virtue of fewest nongenes found.

Are nongenes, identified by rule 2 (or any other rule in Table 5), actually sites of translational initiation? If they are, the peptides made from them are unknown for lack of either biochemical or genetic evidence. We expect that some examples of this occur, but we believe that the majority of gene beginnings in our library are known, since most of these sequences have been studied intensely. Therefore, most of the 167 nongenes are probably sites that don't function as ribosome binding regions.

For those sites there must be other information in the sequence which allows the ribosome to distinguish genes from nongenes. The information could

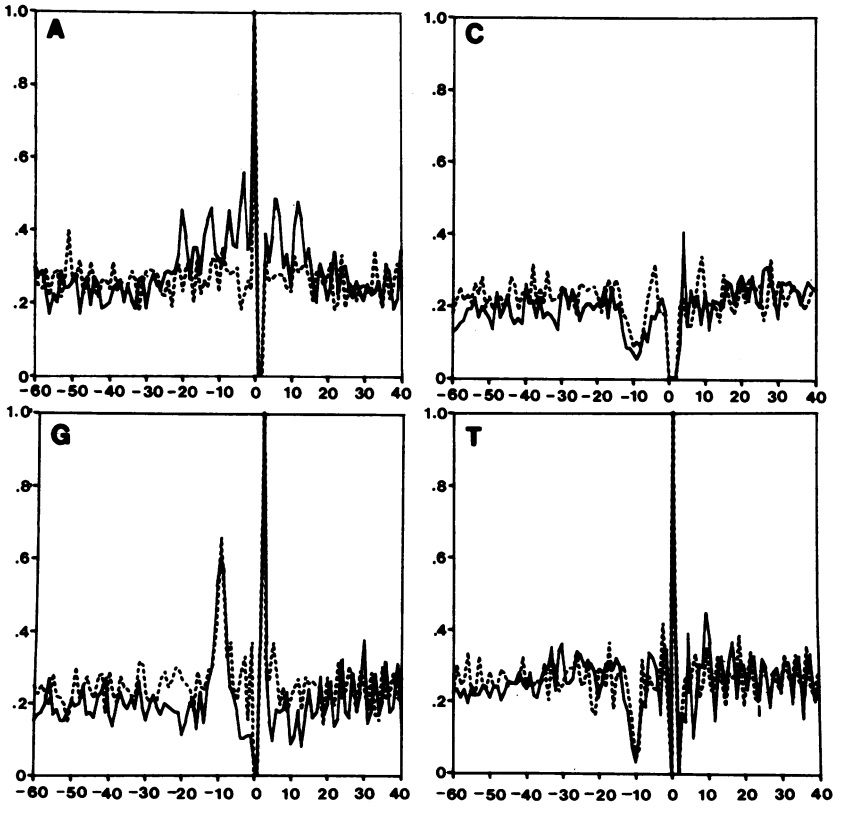

Figure 5. Mono-nucleotide histograms are shown for the region -60 to +40 around the initiation codon, for both genes (———) and nongenes (....).

be either in the primary sequence or in the 3-dimensional folding of the mRNA. Scherer et al. (7) have shown evidence that the primary sequence of ribosome binding sites in the region -20 to +15 has information besides the Shine and Dalgarno sequence and the AUG. Iserentant and Fiers (5) have made a consistent model using secondary structure predictions to account for different translational efficiencies of a set of constructs involving identical linear sequences at the initiation regions.

(e) Do Translational Intiation Regions have Information Beyond "Shine/Dalgarno-Space-Initiation Codon"?

i) Histogram data (boxes #11 and #12 of Figure 1)

The failure of the rules in Table 5 pushed us to explore other differences between genes and nongenes. We asked if translational initiation regions contained, at locations other than the Shine and Dalgarno and initiation codon domains, biased compositions. To do this, we wrote a program (Hist) which counts the number of bases or short oligonucleotides at each position relative to some fixed base, and over a variable distance from that base. In this case, we aligned the gene and nongene sequences by the initiation codons, just as in Table 1. We used the region -60 to +40 because we anticipate that this will include the entire area scanned by the ribosome in an initiation complex. Then, with the sequences so aligned, the computer simply goes down each column and counts the occurrence of each base (and also the number of each short oligo-nucleotide which begins at that position). Figure 5 shows the plot of the histograms for each of the mono-nucleotides over this range. Figure 5 also includes the histograms of sequences selected to have an ATG preceded by a Shine and Dalgarno but which don't function as translational initiation sites. These "nongenes" (from rule 2, above) serve to indicate what fluctuations are likely in sequences that may be functionally random. The gene plots appear generally random outside the nucleotides from -30 to +20, as one would expect if this is beyond the region inspected by the ribosome. Within that region, however, there are nonrandom positions besides the initiation codon and the Shine and Dalgarno domain.

The histograms for G are very different in the gene and nongene set. One simple idea, implicit in Figure 2, proposes selection against the inclusion of G's around the Shine and Dalgarno region or the initiation codon. Annealing to the 3' end of 16S RNA is a G-driven reaction, since the important 16S nucleotides include CCUCC; similarly, the initiator tRNA of E. coli can recognize XUG (including reinitiation codons, 69), again reflecting a partially G-driven binding. The gene set is low in G's from -30 to +10

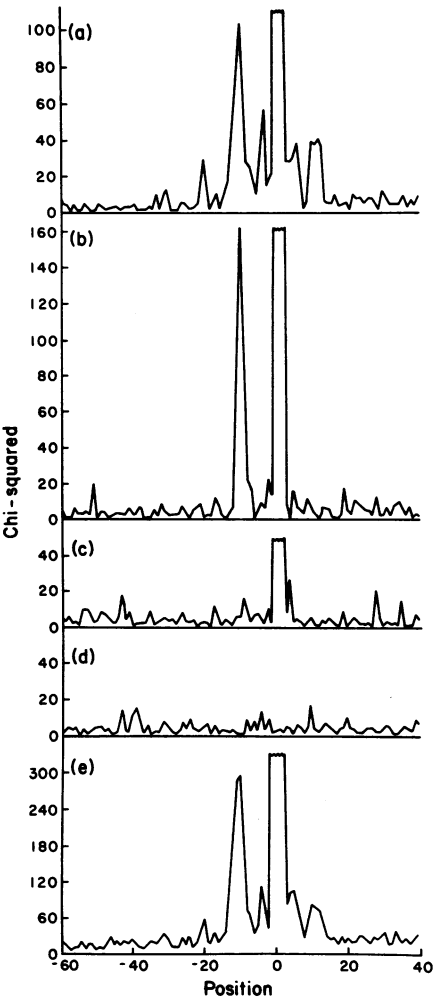

Figure 6. χ^2 test of the histograms. The observed data for each data set is a histogram, and the expected numbers come from the composition of the mRNA library (Table 3). The data sets are:

a) the 124 gene beginnings
b) the 167 nongenes found by rule 2
c) every tenth ATG in the mRNA library (so the sample size would be 157, in the range of the other sets)
d) 160 arbitrarily chosen phage sequences
e) the gene set (as in a), using di-nucleotide composition and histogram data.

(excluding the S/D), as though true initiation regions have evolved to exclude confusion in the alignment depicted in Figure 2. Some genetic data (2) weakly support the idea of G's as a negative element if they are close to the Shine and Dalgarno domain. The paucity of G's in the gene set results in UG also

TABLE 6

PARTIAL CHI-SQUAREDS FROM PEAKS

POSITION	TOTAL CHI-SQUARED	BASE * A	C	G	T
-20	30.2	19.8(+)	0.2(-)	9.0(-)	1.2(-)
-4	32.5	18.5(+)	2.6(-)	11.1(-)	0.2(+)
-3	58.5	41.0(+)	0.8(-)	11.1(-)	5.6(-)
-2	15.7	2.8(+)	0.1(-)	10.0(-)	2.8(+)
-1	20.8	6.5(+)	1.6(-)	10.0(-)	2.8(+)
3	29.4	7.7(+)	10.7(-)	6.4(+)	4.6(-)
4	29.4	1.0(+)	16.4(+)	4.9(-)	7.2(-)
5	31.3	6.8(+)	4.6(-)	12.2(-)	7.8(+)
6	40.2	23.9(+)	0.1(+)	2.2(-)	13.9(-)
10	30.0	0.3(-)	0.0(+)	13.5(-)	16.2(+)
11	29.3	9.8(+)	5.4(-)	9.8(-)	4.3(+)
12	32.0	22.2(+)	0.4(-)	2.2(-)	7.2(-)
13	27.6	12.1(+)	0.2(-)	14.8(-)	0.5(+)

* (+) MEANS OBSERVED GREATER THAN EXPECTED
(-) MEANS OBSERVED LESS THAN EXPECTED

being low. We assert that P site filling should occur with a minimum of choices among nearby codons.

ii) Chi-squared Histograms (boxes #13 and #14 of Figure 1)

Given the histogram data, one can ask where the nucleotides deviate from expected behavior. We used a chi-squared test (program Achsq). Here the histogram data serve as the observed numbers, and the expected numbers come from the mRNA composition (Table 3). For example, there are 34 A's at position -15 in the gene set, and, from the composition, one would expect that out of 124 sequences there should be 32 A's. This gives a chi-squared for A's at position -15 of 0.1. The same operation is done for each nucleotide at each position and the resulting values are plotted versus position (Figure 6a). Three controls were performed. The nongene set (Figure 6b), every tenth ATG in the mRNA library (Figure 6c) and arbitrary sequences (Figure 6d) were similarly analyzed. When compared one sees that genes have a region of higher than expected chi-squared values between positions -20 and +13. Figure 6e shows the χ^2 data for the dinucleotide histogram, using the dinucleotide

composition (Table 3) for the expected numbers. The χ^2 values are roughly 3 times higher because there are 3 times as many degrees of freedom in the data (9 vs 3). The peaks are all still there, though the one at -20 is diminished close to background. That peak is probably due only to mononucleotide bias. The peak at -3 is also diminished, probably for the same reason, but it is still highly significant.

The high chi-squared values for the genes fall into six distinct peaks. The peak at the initiation codon is not surprising since every gene starts with either ATG or GTG; the peak centered around -10 is expected because that is the region of the Shine and Dalgarno polypurine sequences. However, the peaks at -20, -3, +3 to +6 and +10 to +13 are not predicted by current models of initiation. The size of the region is very similar to the length and location of mRNA protected by the ribosome in bind-and-chew experiments (see 2). Those very peaks are included in the envelope of information found by Scherer et al. (7). Table 6 shows the partial chi-squareds for each nucleotide at each position within the unexpected peaks, so that one can determine which nucleotides are aberrant within each peak.

From -20 to -1 the high χ^2 values always reflect excess A's and diminshed G's. As noted above, a paucity of G's might diminish misalignment of either the polypurine tract or the initiation codon itself. One might also imagine ribosomal recognition of a resultant adenine, especially A's at position -3, in mRNAs.

The χ^2 peak at positions +3 to +6 is due to the abundance of the two codons GCT and AAA in positions 3 to 5 (see Table 1) and the preference for A and discrimination against T at position 6. Whether this is a codon-specific effect (having to do either with the preference for an amino acid or tRNA at that position) or nucleotide recognition by the ribosome is not known. Since the peak is four positions wide rather than two (the first two nucleotides of a codon generally sufficing to select most tRNAs), we might opt for actual sequence recognition by the ribosome. However, GCTA and AAAA might be recognized by tRNAs in the ribosomal acceptor site by virtue of a four base codon-anticodon interaction, since all tRNAs have a U 5' to the anticodon (71). The tRNAs which read these two codons are among the most abundant in the cell (72) so that perhaps they are selected for in order that the transition from initiation to elongation be enhanced. Another possiblity is that the enzyme which cleaves the fMet from the protein prefers alanine and/or lysine. However, some data suggest that alanine is good for the amino peptidase and lysine is not (73). Furthermore, the amino peptidase activity could be regulated with many more second codon selections than seem to be

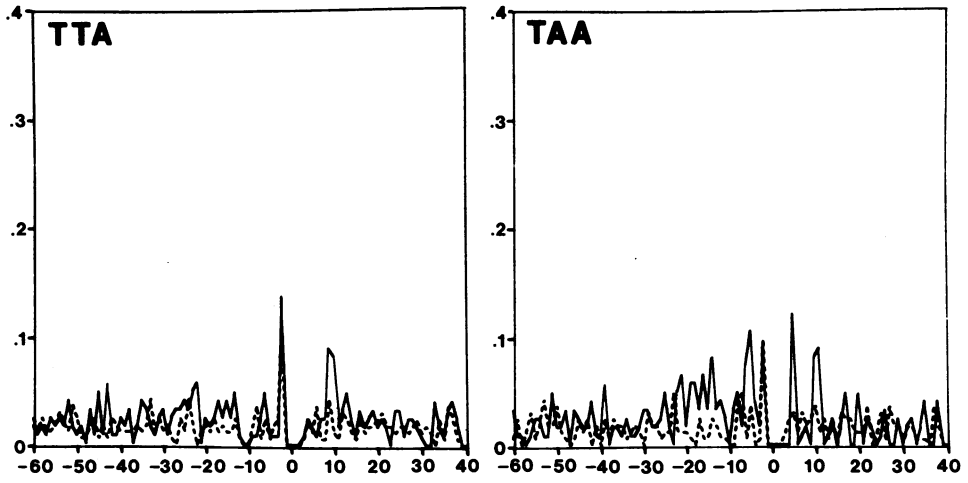

Figure 7. Histograms for the tri-nucleotides TTA and TAA are shown (full scale is 0.4).

utilized.

The peak at +10 to +13 appears the most complicated. In terms of mononucleotides, the high χ^2 is due to all of the following: A is high at positions 11 to 13; C is low at position 11; G is low at positions 10, 11 and 13; and T is high at positions 10 and 11 and low at position 12. Positions +10 to +13 are 70% A and T. The triplet histograms of this region show that both TTA and TAA are abundant here (Figure 7). Each of these triplets is frequent at two adjacent positions, suggesting that we are not revealing codon preference. We believe ribosomes might scan for T's and A's in this region and perhaps even for the sequence TTAA, which is quite common here (Table 1).

We note here again the congruence of the 3' domain χ^2 peaks and the ribosome-protected piece of mRNA. This coincidence leads us away from disclaimers of a general form that rely on a purported non-randomness of amino-terminal peptide sequences so as to facilitate enzyme activity. We know of no data that point toward the first five amino acids for any activity, even membrane attachment and/or transport. We scanned the carboxy terminal peptides in our library; they are random. Thus, we tend to think that at least some of the differences in the χ^2 values for genes and nongenes (Figures 6a and 6b) reflect elements used for translational initiation.

The preceeding analyses utilize data sets aligned by the initiation codons. The biochemical data showing preinitiation complex formation between mRNAs and 30S particles are quite striking (2); these complexes certainly

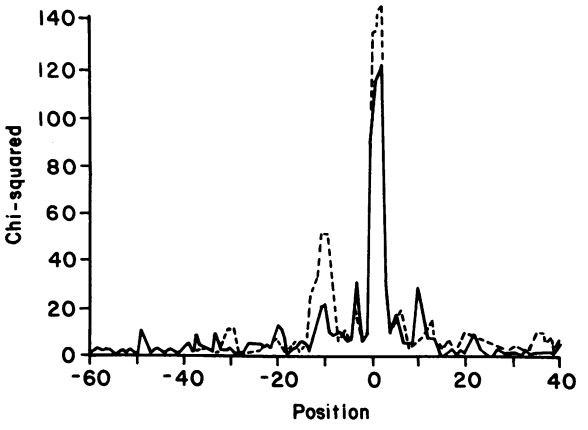

Figure 8. χ^2 test of histograms of two gene subsets. The observed values are from the histograms of the genes found by rule 7 (----) and the genes not found by rule 6 (———). The expected values are from the mRNA library composition (Table 3).

involve recognition of the Shine and Dalgarno elements and no P site codon-anticodon interaction. We aligned the genes by their Shine and Dalgarno sequences and did the same χ^2 analysis as we did for the ATG alignments. This was done for both choices of the Shine and Dalgarno element described earlier: the 94 sequences with unambiguous complementarity, and all genes using the alignment predicted by RNA annealing rules. Neither of these χ^2 data reveal information 5' to the Shine and Dalgarno domain. Peaks on the 3' side are not as significant and are more spread out than when the sequences were aligned by their initiation codons. Perhaps preinitiation complex formation uses annealing to the 3' end of 16S RNA and rearrangment of the complex (as implicit in Figure 2) to allow alignment of the initiation codon and information further 3'. This notion is consistent with the observation that preinitiation complexes dissociate much more slowly than simple complexes between 30S particles and an oligonucleotide complementary to the 3' end of 16S RNA (2).

iii) Rule Extensions

There appears to be information around gene beginnings, besides the AUG and Shine and Dalgarno sequence, which distinguishes the genes from other sites. Most sites with particularly good polypurine domains are genes (rules 6 & 7). Perhaps the other information somehow helps the ribosome bind to those genes with poor Shine and Dalgarno elements. If so, those genes should have higher χ^2 at the other peaks than the genes with strong Shine and Dalgarno elements. Figure 8 shows the χ^2 plots for the genes found by rule 7

(good Shine and Dalgarno) and for the genes not found by rule 6 (poorer Shine and Dalgarno). The peak at +3 to +6 is not more significant but is shifted somewhat. However, the peaks at -3 and +10 follow the prediction. In fact, for the genes not found by rule 6, those peaks are more significant than their Shine and Dalgarno peak.

Thus we can construct better rules than any of those previously used by including information besides ATG and S/D. Rule 6 gives us 83 genes and 44 nongenes, a reasonable ratio but too few of the genes are found. Adding in the new sites found by rule 2 (sites found by rule 6 are a subset of those found by rule 2) increases the genes by 20 (to 103) and the nongenes by 123 (to 167). By adding requirements to rule 2, we can eliminate most of the additional nongenes without losing too many of the additional genes. As an example, we can define a new rule (rule 8) which picks out all the sites found by rule 6 and additional sites found by rule 2 only if they meet stricter criteria:

$$\text{rule 8} \equiv \begin{array}{c}(\text{rule 6}) \\ \underline{\text{or}} \\ (\text{rule 2}) \quad \underline{\text{and}} \begin{pmatrix}\text{not G} \\ \text{at } -3\end{pmatrix} \quad \underline{\text{and}} \begin{pmatrix}<2 \text{ G's} \\ -7 \text{ to } -1\end{pmatrix} \quad \underline{\text{and}} \begin{pmatrix}\text{A or T} \\ \text{at } +5\end{pmatrix} \quad \underline{\text{and}} \begin{pmatrix}\text{A or T} \\ \text{at } +10\end{pmatrix}.\end{array}$$

This rule finds 99 of the genes (80%) but only 52 other sites! Although the number of alternative extended rules is very large, we believe this exercise demonstrates that the additional information uncovered by the χ^2 tests is relevant to ribosome initiation.

The extension of the rules gives about a two to one ratio of genes to nongenes. We of course cannot decide (without precise experimental data for each nongene sequence) if the identified sequence serves to initiate translation *in vivo*. We have been driven, in this analysis, by the assumption that these heavily studied genes from *E. coli* and its phages are unlikely to contain a vast number of undiscovered gene products translated from mRNAs that overlap their more well-known siblings. The assumption is explicitly the opposite of the well-worn assertion that an initiation codon preceded by a Shine and Dalgarno element *is* a ribosome binding site. We also note that the best rule (now rule 8) continues to miss 20% of the known genes, an oversight that would be lethal to a cell. In the subsequent paper (74), we describe an approach to this problem that is capable of identifying every gene, including those with exceptionally weak Shine and Dalgarno elements.

CONCLUSIONS

Information has been identified in a set of sequences known to allow the

initiation of protein synthesis in E. coli. That information is more extensive than had been previously suspected. Biochemical experiments and rigorous manipulation of a set of similar potential ribosome binding sites are required if we are to test the deduction that this information participates in the process of translational initiation. This information could reflect gratuitous non-random distribution that plays a role in other cellular processes.

The analysis suffers from two defects. Firstly, the role of higher order RNA structures (beyond linear representations of bases) has not been included, even though we are quite certain that some secondary structures can occlude essential elements of initiation from ribosomal inspection (75, 76) and even though we have deduced that in some cases secondary structures of an mRNA may well facilitate initiation (77 and above). We have excluded higher order structures of the mRNA because the available data do not permit reliable prediction of even secondary structures of lengthy mRNAs (78). Furthermore, simple linear tests have found significant information. Clearly the analysis must ultimately cope with higher order structures. Secondly, our analysis has only been concerned with the qualitative aspects of a ribosome binding site. We have not included data (nor does much data exist) for relative translational efficiencies in our χ^2 histograms; in principle, were the data available, we might have weighted sequences from the gene set in proportion to the relative translational efficiency of the particular sequences. Since we have not attempted to include such data, our histograms represent parameters related to a consensus translational initiation region. Consensus sequences probably do not contain the proper mixture of recognition elements required for, in this case, high level translation.

In spite of these two defects, we remind the reader again that the information uncovered falls within the so-called "bind-and-chew" site protected by ribosomes when the mRNA is in an initiation complex. This information must be scanned by the ribosome during some early partial reaction of protein synthesis.

ACKNOWLEDGEMENTS

We wish to thank Dr. Jeffrey Haemer for thoughtful consultation and Dr. Eugene Myers for designing and writing the secondary structure predicting program we used. This work was supported by NIH grant #GM28755.

REFERENCES

1. Shine, J. and Dalgarno, L. (1974) Proc. Natl. Acad. Sci. USA 71, 1342-1346.
2. Gold, L., Pribnow, D., Schneider, T., Shinedling, S., Singer, B.S. and Stormo, G. (1981) Ann. Rev. Microbiol. 35, 365-403.
3. Grunberg-Manago, M. (1980) in Ribosomes, Chambliss, G., Craven, G.R., Davies, J., Davis, K., Kahan, L. and Nomura, M. Eds., pp. 445-477, University Park Press.
4. Steitz, J.A. (1979) in Ribosomes, Chambliss, G., Craven, G.R., Davies, J., Davis, K., Kahan, L. and Nomura, M. Eds., pp. 479-495, University Park Press.
5. Iserentant, D. and Fiers, W. (1980) Gene 9, 1-12.
6. Fiers, W. (1979) Comprehensive Virology 13, 69-204.
7. Scherer, G.F.E., Walkinshaw, M.D., Arnott, S. and Morre, D.J. (1980) Nucl Acid Res. 8, 3895-3907.
8. Schneider, T.D., Stormo, G.D., Haemer, J.S. and Gold, L. (1982) Nucl. Acid Res.
10. Model, P., Webster, R.E. and Zinder, N.D. (1979) Cell 18, 235-246.
11. Steitz, J.A. (1972) Nature New Biol. 236, 71-75.
12. Staples, D.H., Hindley, J., Billeter, M.A. and Weissmann, C. (1971) Nature New Biol. 235, 202-204.
13. Sanger, F., Coulson, A.R., Friedmann, T., Air, G.M., Barrell, B.G., Brown, N.L., Fiddes, J.C., Hutchison III, C.A., Shocombe, P.M. and Smith, M. (1978). J. Mol. Biol. 125, 225-246.
14. Godson, G.N., Barell, B.G., Staden, R. and Fiddes, J.C. (1978) Nature 276, 236-247.
15. Beck, E., Sommer, R., Auerswald, E.A., Kurz, C.H., Fink, B., Osterburg, G. and Schaller, H. (1978) Nucl. Acid Res. 5, 4495-4503.
16. Hoess, R.H., Foeller, C., Bidwell, K. and Landy, A. (1980) Proc. Natl. Acad. Sci. USA 77, 2482-2486.
17. Franklin, N.C. and Bennett, G.N. (1979) Gene 8, 107-119.
18. Humayan, Z., Jeffrey, A. and Ptashne, M. (1977) J. Mol. Biol. 112, 265-277.
19. Schwarz, E., Scherer, G., Hobom, G. and Kossel, H. (1978) Nature 272, 410-414.
20. Pribnow, D., Sigurdson, D.C., Gold, L., Singer, B.S., Napoli, C., Brosius, J., Dull, T.J. and Noller, H.F. (1981) J. Mol. Biol. 149, 337-376.
21. Krisch, H.M., Duvoisin, R.M., Allet, B. and Epstein, R.H. (1980) ICN-UCLA Symposium on Molecular and Cellular Biology, 29, 517-526.
22. Oliver, D.B. and Crowther, R.A. (1981) J. Mol. Biol. 153, 545-568.
23. Dunn, J.J. and Studier, F.W. (1981) J. Mol. Biol. 148, 303-330.
24. Rosa, M.D. (1981) J. Mol. Biol. 146, 55-72.
25. Gardner, J.F. (1979) Proc. Natl. Acad. Sci. USA 76, 1706-1710.
26. Smith, D.R. and Calvo, J.M. (1980) Nucl. Acids Res. 8, 2255-2274.
27. Smith, B.R. and Schleif, R. (1978) J. Biol. Chem. 253, 6931-6933.
28. Cass, L.G., Horwitz, A.H., Miyada, C.G., Greenfield, L. and Wilcox, G (1980) Molec. gen. Genet. 180, 219-226.
29. Buchel, D.E., Gronenborn, B. and Muller-Hill, B. (1980) Nature 283, 541-545.
30. Farabauth, P.J. (1978) Nature 274, 765-769.
31. Maizels, N. (1974) Nature 249, 647-649.
32. Musso, R., DiLauro, R., Rosenberg, M. and deCrombrugghe, B. (1977) Proc. Natl. Acad. Sci. USA 74, 106-110.
33. Grindley, N.D.F. (1978) Cell 13, 419-426.
34. Movva, N.R., Nakamura, K. and Inouye, M. (1980) Proc. Natl. Acad. Sci. USA 77, 3845-3849.

35. Lee, F., Bertrand, K., Bennett, G. and Yanofsky, C. (1978) J. Mol. Biol. 121, 193-217.
36. Nichols, B.P. Miozzari, G.F., vanCleempat, M., Bennett, G.N. and Yanofsky, C. (1980) J. Mol. Biol. 142, 503-517.
37. Crawford, I.P., Nichols, B.P. and Yanofsky, C. (1980) J. Mol. Biol. 142, 489-502.
38. Nichols, B.P. and Yanofsky, C. (1979) Proc. Natl. Acad. Sci. USA 76, 5244-5248.
39. Nakamura, K. and Inouye, M. (1979) Cell 18, 1109.
40. DiNocera, P.P., Blasi, F., DiLauro, R., Franzio, R., Bruni, C.B. (1978) Proc. Natl. Acad. Sci. USA 75, 4276-4280.
41. Zurawski, G., Brown, K., Killingly, D. and Yanofsky, C. (1978) Proc. Natl. Acad. Sci. USA 75, 4271-4275.
42. Horri, T., Ogawa, T. and Ogawa, H. (1980) Proc. Natl. Acad. Sci. USA 77, 313-317.
43. Post, L.E. and Nomura, M. (1979) J. Biol. Chem. 254, 1064-10606.
44. Post, L.E. and Nomura, M. (1980) J. Biol. Chem. 255, 4660-4666.
45. Post, L.E., Arfstein, A.E., Reusser, F. and Nomura, M. (1978) Cell 15, 215-229.
46. Post, L.E., Arfsten, A.E., Davis, G.R. and Nomura, M. (1980) J. Biol. Chem. 255, 4653-4659.
47. Hirota, Y., Yasuda, S., Yamada, M., Nishimura, A., Sugimoto, K., Sugisaki, H., Oka, A. and Takanami, M. (1978) Cold Spring Harbor Symposium on Quantitative Biology, 42, 129-138.
48. Lawther, R.P. and Hatfield, R.P. (1980) Proc. Natl. Acad. Sci. USA 77, 1862-1866.
49. Lawther, R.P., Nichols, B., Zurawski, G. and Hatfield, G.W. (1979) Nucl. Acids Res. 2289-2301.
50. Post, L.E., Strycharz, G.D., Nomura, M., Lewis, H. and Dennis, P.P. (1979) Proc. Natl. Acad. Sci. USA 76, 1697-1701.
51. Hedgpeth, J., Clement, J., Marchal, C., Perrin, D. and Hofnung, M. (1980) Proc. Natl. Acad. Sci. USA 77, 2621-2625.
52. Singleton, C.K., Roeder, W.D., Bogsian, G., Somerville, R.L. and Weith, H.L. (1980) Nucl. Acids Res. 8, 1551-1560.
53. van den Elzen, P.J.M., Gaastra, W., Spelt, C.E., de Graaf, F.K., Veltkamp, E. and Nijkamp, H.J.J. (1980) Nucl. Acids Res. 8,4349-4363.
54. Heffron, F., McCarthy, B.J., Ohtsabo, H. and Ohtsubo, E. (1979) Cell 18, 1153-1163.
55. Kirby, N.K. and Vapnek, D. (1979) Nature 282, 864-869.
56. So, M. and McCarthy, B. (1980) Proc. Natl. Acad. Sci. USA 77, 4011-4015.
57. Dallas, W.S. and Falkow, S. (1980) Nature 288, 499-501.
58. Barnes, W.M. (1978) Proc. Natl. Acad. Sci. USA 75, 4281-4285.
59. Zieg, J. and Simon, M. (1980) Proc. Natl. Acad. Sci. USA 77, 4196-4200.
60. Nussinov, R. (1980) Nucl. Acid Res. 8, 4545-4562.
61. Tinoco, I., Borer, P.N., Dengler, B., Levine, M.D., Uhlenbeck, O.L., Crothers, D.M. and Gralla, J. (1973) Nature New Biol. 246, 40-41.
62. Woo, N.H., Roe, B.A. and Rich, A. (1980) Nature 286, 346-351.
63. van der Laken, K., Bakker-Steeneveld, H., Berkhout, B. and van Knippenberg, P.H. (1980) Eur. J. Biochem. 104, 19-23.
64. Jay, E., Seth, A.K. and Jay, G. (1980) J. Biol. Chem. 255, 3809-3812.
65. Belin, D., Hedgpeth ,J., Selzer, G.B. and Epstein, R.H. (1979) Proc. Natl. Acad. Sci. USA 76, 700-704.
66. Steitz, J.A. and Bryan, R.A. (1977) J. Mol. Biol. 114, 527-543.
67. Platt, T. and Yanofsky, C. (1975) Proc. Natl. Acad. Sci. USA 72, 2399-2403.
68. Gupta, S.L., Waterston, J., Sopori, M.L., Weissman, S.M. and Lengyel, P. (1971) Biochem. 10, 4410-4412.
69. Napoli, C., Gold, L. and Singer, B.S. (1981) J. Mol. Biol. 149, 433-450.

70. Taniguchi, T. and Weissman, C. (1978) Nature 275, 770-772.
71. Sprinzl, M., Grueter, F., Spelzhaus, A. and Gauss, D.H. (1980). Nucl. Acids Res. 8, r1-r22.
72. Ikemura, T. (1981) J. Mol. Biol. 146, 1-21.
73. Sherman, F., Stewart, J.W. (1981) Molecular Biology of the Yeast Saccharomyces, Cold Spring Harbor Laboratory, in press.
74. Stormo, G.D., Schneider, T.D., Gold, L. and Ehrenfeucht, A. (1982) Nucl. Acid Res.
75. Schwartz, M., Roa, M. and Debarbouille, M. (1981) Proc. Natl. Acad. Sci. USA 78, 2937-2941.
76. Hall, M.N., Gabay, J., Debarbouille, M. and Schwartz, M. (1982) Nature 295, 616-618.
77. Singer, B.S., Gold, L., Shinedling, S.T., Hunter, L.R., Pribnow, D. and Nelson, M.A. (1981) J. Mol. Biol. 149, 405-432.
78. Ninio, J. (1979) Biochimie 61, 1133-1150.