© 1994 Oxford University Press Nucleic Acids Research, 1994, Vol. 22, No. 7 1287–1295

Quantitative analysis of ribosome binding sites in *E.coli*

Doug Barrick+, Keith Villanueba, John Childs§, Rhonda Kalilϕ, Thomas D.Schneider¶, Charles E.Lawrence[1], Larry Gold and Gary D.Stormo*

Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, CO 80309-0347 and [1]New York State Department of Health, Wadsworth Center for Laboratories and Research, Empire State Plaza, Albany, NY 12201, USA

Received August 27, 1993; Revised and Accepted February 15, 1994

ABSTRACT

185 clones with randomized ribosome binding sites, from position −11 to 0 preceding the coding region of β-galactosidase, were selected and sequenced. The translational yield of each clone was determined; they varied by more than 3000-fold. Multiple linear regression analysis was used to determine the contribution to translation initiation activity of each base at each position. Features known to be important for translation initiation, such as the initiation codon, the Shine/Dalgarno sequence, the identity of the base at position −3 and the occurrence of alternative ATGs, are all found to be important quantitatively for activity. No other features are found to be of general significance, although the effects of secondary structure can be seen as outliers. A comparison to a large number of natural *E.coli* translation initiation sites shows the information profile to be qualitatively similar although differing quantitatively. This is probably due to the selection for good translation initiation sites in the natural set compared to the low average activity of the randomized set.

INTRODUCTION

The sequences of a large number of *E.coli* translation initiation sites have now been determined (1), and the important features of those sequences are fairly well established (for some reviews see 1–5). The initiation codon is usually an AUG, although GUG and UUG are also used in many cases, and there is nearly always a region 5′ to the initiation codon that can base-pair with the 3′ end of 16S rRNA, called the Shine/Dalgarno sequence (6). The importance of these features has been established by many studies, including statistical analyses (for example, 7), selection of functional sites (8) and a large number of mutations, both selected and directed (see the reviews for some of those references). The switching of sequences between the Shine/Dalgarno sequence on the mRNA and the 3′ end of 16S rRNA leads to mRNA-specific ribosomes (9). It is also well established that mRNA secondary structure can have large effects on translation initiation (10,11). In many cases an inhibitory RNA structure can essentially inactivate an otherwise good initiation site. This complicates the analysis of sequence effects on translation initiation because any change in structure is accompanied by a change in sequence, and any change in sequence has the potential to change the structure. Distinguishing effects due to sequence from those due to structure is not always easy.

In recent work (12) we synthesized a large number of synthetic ribosome binding sites and accurately measured the amount of protein made from each. We tested two different Shine/Dalgarno sequences, the three initiation codons AUG, GUG and UUG, a variety of spacings between those elements, and three different second codons, in a large number of combinations. Analysis of the data provided quantitative values for the relative contributions of each element to translation initiation efficiency and could be fit quite well to a simple kinetic model. In this paper we do a similar analysis on a different type of data. Rather than making a set of directed changes to defined features of ribosome binding sites, we have randomized a portion of that region and determined the translation initiation activity for a large number of sites. The results are compared to previous work on translation initiation sites, including the large number of natural sites from *E.coli* for which activities are not known.

MATERIALS AND METHODS

Bacteria and phage

The *E.coli* strain 79-02 (C600 *hsdR*, *hsdM*, *thr*, *leu*, *str*, Δ(*lacZY-pro*), F′ *traD36*, *lacI*^{q}, *lacZ*Δ*M15*, *pro*$^+$) was used in all experiments (13). The phage f1 (IR1) (14) was used for preparation of single stranded plasmid DNA for dideoxy sequencing.

Single-stranded DNA synthesis and 'notch' cloning

Single stranded DNA was synthesized on an Applied Biosystems model 380A DNA synthesizer. Detritylated, deprotected DNA

*To whom correspondence should be addressed

Present addresses: +Institute of Molecular Biology, University of Oregon, Eugene, OR 97403, §Energy Biosystems Corp., 4200 Research Forest Drive, The Woodlands, TX 77381, ϕMedical University of South Carolina, Department of Psychiatry, 171 Ashley Avenue, Charleston, SC 29425-0742 and ¶National Cancer Institute, Frederick Cancer Research and Development Center, Laboratory of Mathematical Biology, PO Box B, Frederick, MD 21702-1201, USA was gel-purified. Two randomized oligos were used for the experiments described in this paper. Oligo *des12* was synthesized as:

GATC AAA nnnnnnnnnnnn TGCA

and oligo *des19* as:

GATC AAA rrrrnnnnnnnn TGCA

where *n* means that a mixture of all four bases was used at that step in the synthesis and *r* means that a mixture of the purines (A and G) was used at that step in the synthesis. Both of these mixtures were provided by Applied Biosystems.

Plasmid pBC39 was the recipient of the synthetic oligos. It is derivative of pBC29 (13) from which the *pra2* promoter has been replaced by a *tacII* promoter (12,15). pBC39 also contains, between the *Bgl*II and *Pst*I sites a fragment from phage lambda, bases 35712 to 37005 (16), to facilitate purification of the doubly cut plasmid from singly cut plasmids. The complete plasmid, including the lambda fragment, is 8910bp, whereas the portion without the lambda fragment is only 7617bp. The sequence of pBC39 from the start of transcription to the site of the inserted, randomized ribosome binding sites is shown in (12). Cutting pBC39 with *Bgl*II and *Pst*I creates the following 'notch' into which the single-stranded oligos were cloned:

```
5'-...A A G C T T T A                          G G A T C C...
   ...T T C G A A A T C T A G          A C G T C C T A G G...-5'
      |         |   |          |     |         |         |
       Hind III       Bgl II          Pst I      BamH I
```

The TG that begin the fixed sequence at the 3′ end of the synthetic oligos (and which anneal to the 3′ overhang of the *Pst*I site) are the second and third bases of the initiation codon for *lacZ*. If they are preceded by an A as the last base of the randomized region of the oligos, that provides an ATG initiation codon, but all other bases are also possible at that position. We number that position 0 so the synthetic oligos are randomized at the twelve positions from −11 to 0.

8μg of pBC39 DNA was cut with 16 units of *Bgl*II (New England Biolabs) followed by removal of the 5′ terminal phosphates using Calf Intestinal Phosphatase (Boerhinger Mannheim Biochemicals) as described in (17). After ethanol precipitation the DNA was digested with 40 units of *Pst*I (New England Biolabs) followed by phenol extraction and ethanol precipitation. This method prevents religation of the parental pBC39 plasmid so that the only circular plasmid molecules formed are those with a random synthetic ribosome binding site preceding *lac*Z. The synthetic single stranded ribosome binding sites, *des12* and *des19*, had phosphates added at the 5′ end using T4 polynucleotide kinase (United States Biochemicals). 100 fmol of the synthetic ribosome binding sites were then mixed with 10 pmol of pBC39 which had been cut with *Bgl*II, treated with phosphatase, and cut with *Pst*I as described above, and ligated with 20 units of T4 DNA ligase (Bethesda Research Labs). 10 μl of the ligation mix was put into a 1.5 ml Eppendorf tube along with 100 μl of competent 79-02 cells. This mixture was then incubated for 15 minutes on ice, heated to 37 °C for 5 minutes, added to 2 ml of H-broth, and incubated at 37 °C shaking for 40 minutes. The entire 2 ml was then plated on EHA plates containing 50 μg/ml ampicillin with 40 μg of 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal, Boerhinger Mannheim Biochemicals), and incubated at 37 °C overnight. Single colonies were picked and restreaked on EHA plates containing ampicillin and X-gal, and were then used to inoculate H-broth containing 50 μg/ml ampicillin. These cultures were then made 20% in glycerol and were stored at −70 °C. Plasmid DNA was prepared from 5 ml of an overnight culture of each clone grown in H-broth containing 50 μg/ml ampicillin. 1 μl was digested with *Hind*III (New England Biolabs) and samples that showed a linear band at 7.6 kb were assumed to be the desired product.

DNA sequencing

DNA was sequenced using the Sanger dideoxy-chain termination method either with single stranded DNA obtained by f1 superinfection using methods described in (13) or by using double stranded plasmid DNA. Sequences were read from the *Bgl*II site through the beginning of the mRNA to be sure there were no other differences in that region. Sequences that were ambiguous or difficult to read were determined in both directions to make sure we had reliable sequence for each clone.

Assay of β-galactosidase

β-galactosidase activities were measured kinetically on 96-well microtiter plates. The procedure has been described briefly elsewhere (12,18), including a modified version that uses phage T4 to lyse the cells rather than chloroform (19). Four independent colonies were picked from plates for each clone to be assayed. Overnight cultures from these colonies were grown in 5ml H-broth plus 50 μg/ml ampicillin. 0.05ml of each overnight were added to 5ml of fresh H-broth with 50 μg/ml ampicillin and 1mM IPTG. These were grown at 37 °C to an OD_{600} of 0.3 to 0.4, after which the cultures were placed on ice for at least 10 minutes.

All reagents were the same as those listed in (20) except that the Z buffer was 100 mM β-mercaptoethanol. 0.4ml of each cell culture was added to 0.6ml of Z buffer and 30 μl of $CHCl_3$ in an Eppendorf tube. The tubes were vortexed immediately for 10 seconds and then left at room temperature for at least 10 minutes. Two blank tubes were made in the same manner with 0.4ml of the H-broth, ampicillin and IPTG mixture. The remainder of the cell cultures were left on ice for determining the cell densities later.

125 μl from the tubes were added to the corresponding wells of the 96-well microtiter plate. The blank solutions were added to the first column of the plate (wells A1 to H1). The solutions for the four independent cultures for each clone were added in adjacent rows of a particular column (i.e., wells Ax−Dx were from the same clone, as were Ex−Hx, where x = 2−12). After all the wells were filled, 50 μl of a solution of ONPG at 4mg/ml in Z buffer was added to each well using a 12-channel pipetman. OD measurements were made on a Titertek Multiskan MC plate reader, using a 414nm filter, until sufficient data were recorded to calculate an accurate activity (see description below). Typically, readings were taken every 5 to 20 minutes for 1 to 2 hours. Wells were emptied by aspiration when the OD exceeded 1.5 to avoid complications with saturation and because it was found that a high OD in one well can influence the reading in the next column of the same row. After all the OD_{414} measurements were taken, another 96-well plate was filled, in the corresponding wells, with 250 μl from each of the iced cultures (using the broth mixture for the blank wells). The Titertek filter was changed to 620nm and this plate was read to obtain the cell densities.

All of the Titertek data were transferred to an IBM-PC using a BASIC program, 'tk'. The program allows the user to label each well of the plate, then captures all of the OD measurements from the Titertek and stores the information on a floppy disk. Data were then analyzed by a Pascal program, 'Titer'. This program calculates the activity of each well as a least-squares best fit to the OD_{414} vs. time data. A specific activity for each well is determined as:

$$Specific\ Activity = 1200 \times \frac{\Delta OD_{414}\ /\ \Delta time(minutes)}{OD_{620} \cdot volume(ml)}$$

where 'volume' is the volume of cell lysate added to the tubes, 0.4ml in the above protocol. The factor of 1200 makes these units approximately equal to 'Miller' units (Barrick and Binkley, unpublished data). The 'Titer' program returns the specific activity for each well, and also calculates an average and standard deviation for the different cultures from the same clones.

Data storage and analysis

Sequence data and β-galactosidase activities were stored using the INGRES data base management system (copyright 1977 by the Regents of the University of California) on a Pyramid 90X computer running Unix. The sequence data were manipulated and analyzed using programs of the Delila system (21,22). Multiple regression and ANOVA analyses were performed using BMDP (BMDP Statistical Software, Inc., Los Angeles, CA) on a Vax computer at the New York State Department of Health, Albany, NY.

Regression analyses were performed essentially as described in (23,12). The sequences provide the independent variables and the natural logarithms of the activities are the dependent variables. The rationale for using logarithms of the activities is similar to that described in Ringquist *et al.* (12) except that we were then fitting the coefficients to the kinetic parameters in an initiation model. In this case we assume that each base in the region from -11 to 0 has an influence on the translational initiation activity and those influences are multiplied together to obtain the activity of any sequence. Using the logarithms of the activities as the dependent variables allows us to solve for parameters that are additive. These parameters may be thought of as the partial binding energies contributed by each base to the total binding energy of the site.

As described in (23), there are only three independent variables for each position. That is, we solve for the difference in activity for each base relative to one of the (arbitrarily chosen) bases. It is the differences between the coefficients at each position which are important, rather than their absolute values. (The absolute value of the intercept is important and is the activity of the 'reference sequence'). We have normalized the coefficients for each position with a procedure that maintains the differences between the coefficients:

$$e^{C'_{b,i}} = \frac{4e^{C_{b,i}}}{\Sigma_{b=A}^{T} e^{Cb,i}}$$

where $C_{b,i}$ is the coefficient for base b at position i determined from the multiple regression and $C'_{b,i}$ is the normalized coefficient. This normalization has the properties that positions with no influence on the activity get coefficients of 0 for all bases, and positions where one particular base is absolutely required for activity get coefficients of $C_{b,i} = \ln 4$ for the required base and $-\infty$ for each other base. The 'Information Content' of those conserved positions is therefore 2 bits, the maximum allowed for an absolute discrimination from four equiprobable possibilities (24,25). The coefficients determined in this way makes them proportional to 'Specific Free Energies' if the differences in activities are due to differences in equilibrium binding constants between the ribosome and the various initiation sites, and 'Information Content' can be calculated for each position from such 'Specific Binding Constants' as $I_{spec} = 1/4\ \Sigma_b\ [K_s(b) \log_2 K_s(b)]$ (25,26).

RESULTS AND DISCUSSION

Notch cloning of random ribosome binding sites

Colonies obtained from the transformations described in Materials and Methods varied from white to dark blue on X-gal plates. The white colonies were usually the parental plasmid pBC39; colonies of this type were seen relatively infrequently (less than 20%). Plasmids were isolated from the colonies that were at least light-blue. Those plasmids determined to be about 7600bp in size were assumed to contain a synthetic ribosome binding site insertion and were then sequenced. Of the 244 sequences determined, 203 were synthetic ribosome binding sites as designed; the remaining 41 sequences contained nucleotide insertions or deletions within the synthetic ribosome binding site or in the flanking cloning sites. Additionally, 18 synthetic ribosome binding sites were identical in sequence to another synthetic ribosome binding site. These repeat sequences probably arose by cell doubling in liquid culture incubation during transformation, since it is unlikely that they are the product of independent cloning events of two random oligonucleotides with the same sequence.

Sequences and β-galactosidase activities

β-galactosidase levels were measured for the 185 unique ribosome binding sites, as described in Materials and Methods. Table 1 lists the name, sequence, average activity, standard deviation and number of measurements for each of these clones, 122 from the *des12* cloning and 63 from the *des19* cloning. The total base composition for positions -11 to -8 (for which the *des19* synthesis inserted only purines) is 37% A, 15% C, 23% G and 24% T. None of the individual positions deviates significantly from this average composition. The total base composition for positions -7 to 0 is 28% A, 25% C, 14% G and 33% T. Position 0 deviates significantly from this average composition ($\chi^2 = 40$, $df = 3$), presumably due to selection for functional ribosome binding sites (*i.e.* we picked colonies that were at least light blue) and the large influence played by this position in that activity.

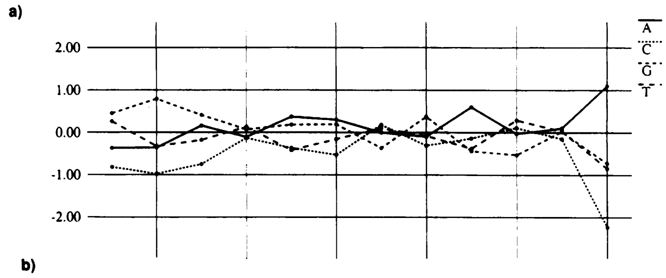
A total of 1012 independent specific activities were measured for the 185 cloned unique ribosome binding sites. The 79-02 host, without plasmid, gives -1.7 ± 0.5 units of activity (data not shown). These cells are deleted for the β-galactosidase gene. The negative specific activity means that cell debris inhibits the spontaneous breakdown of ONPG which occurs in the blank wells. Therefore 1.7 was added to each of the measured activities so that the numbers listed in Table 1 are relative to the host cell without any plasmid. There remain a few isolates with negative specific activities, although only one (1863) does not have 0 within the range of its standard deviation. The fact that each of these clones comes from colonies that are at least light blue on X-gal plates (while the 79-02 host is white) indicates that conditions on the plate are not equivalent to those in liquid broth, and can cause some synthesis to occur which is not seen in the

Table 1.

name	sequence	act.	s.d.	#	ln(act)	name	sequence	act.	s.d.	#	ln(act)	name	sequence	act.	s.d.	#	ln(act)
	des12 sequences:												des19 sequences:				
1217	CTGCCATTGTTA	88.7	14.9	4	4.48	1413	GCCGATCCCCCA	52.8	2.0	4	3.97	938	AAGAAAGAGTTG	51.4	4.0	4	3.94
1218	TATTTAAAGTCG	26.2	1.4	4	3.27	1420	TGACACCTATTG	3015.0	167.4	4	8.01	939	AGAACCCTAACA	597.4	75.4	7	6.39
1219	GGTTGAACAATG	88.9	0.7	4	4.49	1421	TGGTCGCCCCCT	532.0	51.1	4	6.28	942	AAAGTAGTATTG	39.9	1.9	8	3.69
1225	CATACCGAACAG	29.7	3.2	4	3.39	1423	TCTTGTATACAA	359.6	50.1	4	5.88	943	AGGGACAGGATA	601.4	22.4	4	6.40
1226	ACTTCGCAAGAA	95.7	5.2	4	4.56	1439	TTTTAAGTTATA	527.4	23.4	4	6.27	946	AAGGTGAATCCT	228.6	13.1	4	5.43
1227	TAAGTAATTTAA	187.9	26.9	4	5.24	1711	ATCCTTGCTGAT	-0.2	1.3	4	0.00	947	GGAATATTAGAT	491.1	23.0	4	6.20
1228	CATAACCGAAAG	49.8	7.1	10	3.91	1716	TCTCCTAAAACT	3.8	1.4	4	1.32	965	GGAGGTTCTCTG	500.5	27.0	4	6.22
1231	AAACAGTGAGTG	181.1	25.6	4	5.20	1730	TCCTTCACTTGA	17.3	0.5	4	2.85	1081	GGGACATTCAAT	251.4	12.2	4	5.53
1232	GGCACTAACTCT	112.1	15.7	4	4.72	1801	ATTTCTCGACCA	280.6	53.1	12	5.64	1082	AGGAGTGGACAT	908.4	29.9	4	6.81
1234	TTCGAAATCTTT	44.7	8.5	4	3.80	1802	ACAAACTTCCAT	15.0	3.1	4	2.71	1096	GAAATTCAATTG	105.2	6.4	4	4.66
1235	TTAAATATTTTA	685.1	109.0	8	6.53	1803	TGTTAATCTTTA	1695.8	393.8	8	7.44	1097	AGGGATACCTCT	60.6	4.3	4	4.10
1237	TCATAATCGACC	26.6	2.2	4	3.28	1804	CTCGCTTTTCGA	14.3	5.3	4	2.66	1099	AGAGTCCCCATG	28.7	1.1	4	3.36
1238	TCTTAACGTTTA	343.8	14.2	4	5.84	1805	TCTTACTAACCT	126.5	160.5	4	4.84	1113	GAAGGTTCCTCA	597.2	67.6	4	6.39
1242	GAAGCAAAGATA	543.2	21.7	4	6.30	1806	CCGTCATCATAT	6.2	3.2	4	1.82	1119	GAGACCTGGTTT	24.1	1.6	4	3.18
1243	GTTGATATCTCA	393.0	38.3	8	5.97	1808	ATCTCCTTTTGC	1.5	1.1	4	0.38	1120	AGGGTTTTCAAA	115.3	4.8	4	4.75
1247	TTGAACTATTCA	223.0	75.6	8	5.41	1809	CTCCCAACTCCT	6.0	5.2	4	1.79	1124	AAAAATAAATTG	36.2	2.7	8	3.59
1248	TGATAGCTTGCT	910.3	107.1	4	6.81	1810	GTACTGTTTTAA	48.3	11.4	4	3.88	1125	GAGGCGCTTCTC	26.1	8.9	4	3.26
1249	ATAAACCGAGCA	249.3	27.3	8	5.52	1813	TAATTCACCGCC	0.5	0.4	4	0.00	1126	AGAAGTTATTAG	29.2	10.4	4	3.37
1250	TTGATTTCGACA	215.4	19.1	4	5.37	1815	CCCAAGTCCTAC	0.6	2.0	12	0.00	1127	GAGATTTCAGTG	274.2	25.5	4	5.61
1253	ATTAAACTCTTA	613.8	94.3	12	6.42	1816	CATAGCCATTAC	2.1	1.0	4	0.73	1128	AGAACCCGACAA	412.3	24.3	8	6.02
1256	CTGAATTGGATT	68.8	3.0	4	4.23	1817	CCTACTTTGTAG	4.3	3.2	8	1.47	1129	AAGGAATGAGAC	123.4	3.9	4	4.82
1257	TGGGCCTCGGCT	126.9	13.4	4	4.84	1818	CCTATTTCACAC	1.0	1.6	12	0.03	1131	AAAGTCCCTAGA	145.9	3.7	4	4.98
1260	ATTAATCTCTTA	564.7	113.3	8	6.34	1820	GATTCCTTACTC	2.6	2.2	4	0.95	1132	AAAACGACCTGT	21.6	5.8	4	3.07
1262	TTCGGTTGCCGA	76.4	4.9	4	4.34	1821	CCCCACCACTAC	0.7	2.1	8	0.00	1143	AAGGATACCTCT	60.3	1.6	4	4.10
1265	TTTAAATTATTA	875.2	187.2	16	6.77	1822	CCTACCCGGCCT	1.8	0.9	4	0.58	1154	GAAAGATTGAAA	478.8	59.9	12	6.17
1266	AGGTAGGATAGG	145.0	44.9	12	4.98	1828	TTTGCCGTCGCT	-0.8	1.8	4	0.00	1167	GGGAGTCCACAA	1333.9	110.5	4	7.20
1267	TTAAACTTCTTA	691.7	95.6	4	6.54	1830	CACCCACTTCTT	2.2	0.3	4	0.80	1184	AAGAATAAACAT	32.4	2.6	4	3.48
1268	CTTTCAGTCGAA	92.7	46.6	8	4.53	1832	CTCTCCTAGAGA	37.8	15.7	12	3.63	1185	GAAAGTTCGAAT	26.5	2.3	4	3.28
1275	ATTAAACTATTA	667.1	43.1	4	6.50	1833	ACCGTCACCAGA	47.6	9.3	8	3.86	1186	GGGGCGCCTGAG	50.4	7.2	8	3.92
1276	ATTGAAATCTCA	293.8	144.9	12	5.68	1836	CGGAATGCAATG	28.3	4.2	8	3.34	1189	AAGAAAAAGCCT	24.0	2.3	4	3.18
1278	CCTAACTCATAG	19.9	3.4	4	2.99	1838	CACAACTTCCAA	156.5	22.9	4	5.05	1190	GAGAGGACCCAC	45.5	7.6	8	3.82
1283	AATTTCAGTGAA	74.9	2.4	4	4.32	1841	ACATTGTCCCCC	3.5	0.5	8	1.24	1193	AGGACAATTTTC	19.2	9.8	4	2.96
1284	AATATATAAAAA	348.1	27.4	4	5.85	1842	TGTCCTGGTCGT	159.1	20.1	8	5.07	1194	AGAGTATATGAT	34.3	2.2	4	3.53
1285	ATTTAGCCCCAA	118.5	7.9	4	4.77	1846	TGATATCCCTCT	182.2	27.4	4	5.21	1199	GGAGCCGGTTTT	32.7	2.0	4	3.49
1287	AGTTAGTTATTG	52.6	4.0	4	3.96	1853	CACTAACTATCT	26.8	4.7	4	3.29	1200	GGGGCATCCATG	48.8	3.1	4	3.89
1290	TCTCCAGAATGG	98.5	9.1	4	4.59	1862	AGATTTACCCTG	23.4	1.4	4	3.15	1201	AGAGTCTTCATG	71.1	24.8	12	4.26
1293	GTAACTTCATTA	825.7	111.8	12	6.72	1863	CCTACATTTAAC	-1.1	0.5	4	0.00	1203	GAGACTTAGTCT	51.8	15.1	12	3.95
1295	TTAAATCATTTA	628.4	28.2	4	6.44	1864	ACCGCCGTATTG	8.0	2.0	8	2.08	1208	AGGAATTTATTG	670.5	93.3	8	6.51
1296	GTTATACTATAA	340.4	26.5	4	5.83	1866	GAACCATCACCC	6.2	1.4	4	1.82	1209	GGAAGAGTACTG	509.5	201.5	4	6.23
1300	ATCGAATCATCG	44.7	2.0	4	3.80	1869	AAGCCCCTTCCC	-0.0	1.7	8	0.00	1426	GGAATAAATATG	43.7	6.7	8	3.78
1306	TTTAACCTTTAA	85.1	6.3	4	4.44	1870	TGTTTAAGTTAA	2122.3	353.1	8	7.66	1431	AAAAACAACTTG	23.2	7.9	8	3.15
1309	AGGGTGAAGCGG	94.8	10.0	4	4.55	1873	ATACAATTATGT	281.5	32.6	8	5.64	1434	GAAGAAACTTAG	103.9	34.5	4	4.64
1312	TGCCGTCCAGGA	111.5	14.0	4	4.71	1874	CTTTACTATTGT	4.5	3.9	8	1.51	1438	GAAATTTACCTG	69.6	10.9	8	4.24
1313	TCGAATCTCTCG	32.8	3.1	4	3.49	1875	ACAGACCCCTTG	27.9	68.2	12	3.33	1443	AGAGTCCGGTTA	353.5	41.4	4	5.87
1315	GTTTCTCTGAAA	316.0	171.8	8	5.76	1877	CCCATCCTTTCT	0.9	1.2	8	0.00	1450	GAAAATTATGTG	4.0	2.2	4	1.38
1316	TTTTCTATGGAA	35.8	4.3	4	3.58	1879	AAGGCGTCCCCT	8.7	0.6	4	2.17	1458	AAAGTCTCGCGG	0.5	0.6	4	0.00
1317	ACAGAATTCTCG	36.7	1.9	4	3.60	1881	CCCAAATGATTT	12.8	3.9	4	2.55	1462	GAAACTTCGTTT	30.8	18.2	4	3.43
1319	TAATTACATCGT	16.2	1.6	8	2.79	1884	AATACCGTCCAT	4.6	1.5	8	1.52	1464	AGAAACAAGGTG	9.2	1.5	8	2.22
1335	TATGTCTTTTAA	46.8	6.6	4	3.85	1885	TCTCCAAATCTG	10.9	5.0	8	2.39	1483	GAAACATGTCAT	1114.3	123.5	4	7.02
1336	GAGTTAATTTAG	124.3	16.3	4	4.82	1888	TCTAACGTTTCA	47.1	6.7	4	3.85	1485	AAAATAACGTTG	28.1	13.3	4	3.34
1340	TTGATATTCTCA	317.6	20.5	4	5.76	1892	ATACTTATCCCC	4.2	1.9	4	1.44	1487	AAGATTATCATG	10.9	0.4	4	2.39
1346	TGTCGCGACCAG	870.5	73.4	4	6.77	1894	TCCTCAATCACC	-1.7	2.8	4	0.00	1489	AAAATTTACGAT	8.4	1.8	4	2.12
1362	CGTTGAGGCTCA	44.0	4.5	4	3.78	1895	CCACACTTTCAC	2.4	0.6	4	0.86	1491	AAAGTTACTGTG	12.5	1.5	4	2.52
1363	AATTGACATCGG	76.3	3.5	4	4.34	1896	CGGAACCTCCCA	186.1	32.5	8	5.23	1493	GAAGATTTACAA	588.2	37.2	4	6.38
1368	CGATGACCTTTT	1064.7	108.8	4	6.97	1897	CTAATCCATCTT	4.0	2.0	4	1.39	1499	GAAACGGTGGTT	-0.5	1.0	4	0.00
1372	CTTGAAATATCA	744.4	87.7	8	6.61	1905	CATTGTCTTTCC	22.0	6.0	4	3.09	1502	GGAACAATAGTT	656.0	33.4	7	6.49
1378	CCAATACCATTG	48.5	19.5	12	3.88	1908	ACAGGAATTTCG	23.3	2.5	4	3.15	1509	GAAGACAACGAT	22.3	1.9	4	3.11
1381	TTAAATCTCTTA	585.7	100.0	8	6.37	1909	CGCGTTGCCCCT	1.8	0.5	4	0.58	1511	AAAATCGTATTT	36.4	2.1	4	3.60
1391	CCTTAGGCATAA	81.2	2.7	4	4.40	1910	TTACGCCGCAGT	4.2	1.1	4	1.43	1516	AAAATTATCGTG	6.9	0.5	4	1.93
1392	ATTAACTACTTA	667.3	29.7	4	6.50	1911	ATTACAACCCTT	5.2	1.9	4	1.65	c488	GAGGAGATACTG	1348.5	76.2	4	7.21
1396	TTTTTCAGAGAA	330.1	1.5	4	5.80	1912	CCCCTTATCCTT	-1.8	2.4	4	0.00	c493	AGAATTGTAATT	23.8	1.2	4	3.17
												c501	GGAGATACCCTG	3171.0	329.5	4	8.06
												c551	AAAATAATACTC	1.8	0.6	4	0.61

broth cultures. Jacques *et al.* have also noted differential specificity of ribosomes under different growth conditions (27). The activities range from these undetectable levels to nearly 3,200 units. All activities that are measured to be less than 1 are assigned a value of 1. Since the standard deviations of the low activity plasmids are almost always at least 1, the most that can be reliably said of these is that they have low activity, probably less than 1 unit. This approximation also makes easier the following analyses because the natural logarithms of the activities range from 0 to 8.1, without any negative values. The final column of Table 1 gives the natural logarithm of the average activity values for each sequence, using the approximation for low values described.

The 'pure error' of the sample is only 2% of the total variance. That is, the regression performed on the 185 isolate sequences as the independent variables versus the 1012 individual activity measurements, to determine how much of the total variance is due to the variability in activity measurements for individual clones, results in $r^2 = 0.98$. While different activity measurements for any clone do not usually give identical values, and some of the standard deviations are a fairly high fraction



b)

pos:	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0
A	-0.37	-0.36	0.16	-0.11	0.37	0.30	0.00	-0.10	0.60	-0.03	0.10	1.10
C	-0.82	-0.98	-0.75	-0.13	-0.37	-0.53	0.18	-0.31	-0.14	0.11	-0.15	-2.24
G	0.45	0.79	0.41	0.07	0.18	0.19	-0.37	0.37	-0.44	-0.53	0.02	-0.73
T	0.26	-0.33	-0.18	0.13	-0.42	-0.16	0.10	-0.08	-0.38	0.29	0.03	-0.85
												+4.80

Figure 1. a) Normalized values of base versus position for the mononucleotide regression analysis of the 185 ribosome binding sites. **b**) The same values in normal matrix form. Any sequence can be scored by the matrix by summing the base/position values corresponding to that sequence, as described in (28). A constant value of 4.80 is also added to each score; this is the value of the intercept from the regression analysis after the normalization procedure (23).

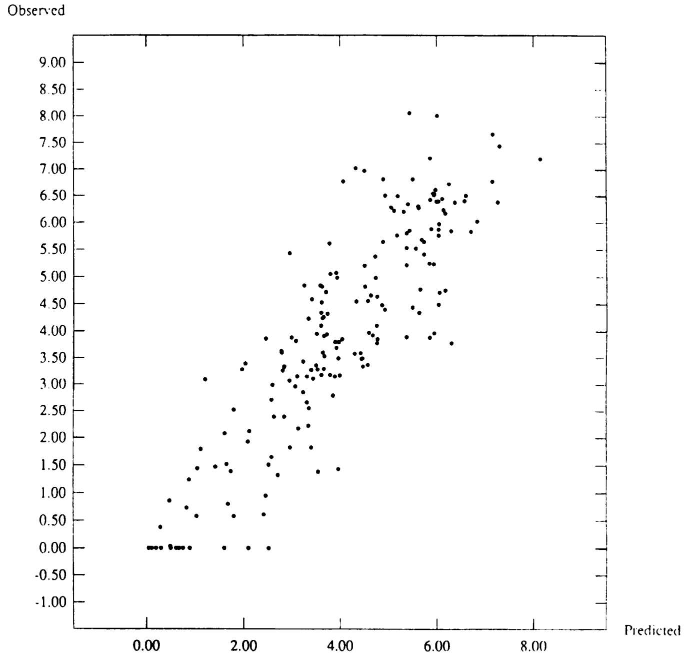


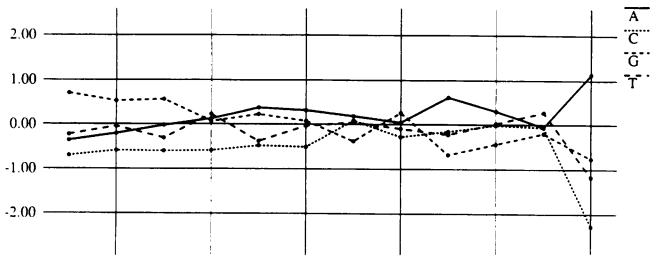
Figure 2. Plot of the predicted versus the observed *ln* (β-galactosidase activity) values for the 185 ribosome binding sites. The predicted values were determined by matrix evaluation of the corresponding sequences using the matrix in Figure 1b, as described.

of the activity, only 2% of the total variance in activity measurements is due to the variability of the measurements themselves, and the remaining 98% can be attributed to differences in the sequences. This also represents the upper limit of r^2 that we could get from the analysis based on the sequences of the isolates.

Mononucleotide analysis on all data

A regression analysis was performed [essentially as in Stormo *et al.*, (23)] on the entire data set of 1012 specific activities for the 185 sequences. Figure 1 shows the values of the base-position parameters obtained, both in normal matrix notation (23,28) and graphically. Figure 2 is a graph of the observed vs. expected values for each of the 185 sequences, using the matrix of Figure 1b for calculating the expected values. The correlation coefficient between the observed and predicted values is 0.89 (r^2=0.79; all r^2 are corrected for degrees of freedom). Many of the features displayed in Figure 1 are consistent with the known qualitative aspects of ribosome binding sites. For example, a G-rich sequence in the −6 to −11 region leading to high level translation is consistent with that being a Shine/Dalgarno sequence. The high activity of A at −3 is consistent with the statistics for that position (2), and some activity measurements (13). The fact that A at position 0 leads to the highest expression is also as expected, as it is that C at position 0 gives the lowest expression (or even works at all). However, some other features are surprising. For example, we expected G at 0 to be better than T (12). We also did not expect a T at −11 to lead to nearly as high expression as a G, and better than an A. These two features, in particular, can be explained by postulating that translation may begin at positions other than 0. In order for functional β-galactosidase to be produced, translation must begin in the same frame as 0, such as −3, −6, −9 or −12. In fact, position −12 is always an A, so whenever −11 is a T and −10 is a G (10 such cases, out of 42 Ts at −11), translation may be higher than expected due to in-frame initiation at this position. This could account for T at −11 leading to high levels of translation. Similarly, whenever there is a G at position 0 there is some probability that it is preceded by an AT (7 such cases, out of 51 Gs at position 0), which could lead to initiation in an alternative reading frame. Those cases would decrease the average contribution of a G at position 0 to the activity. Since the amino-termini of the β-galactosidase proteins were not sequenced we do not know the initiation codon that is actually used.

a)



b)

pos:	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0
A	-0.36	-0.21	-0.03	0.12	0.37	0.31	0.18	0.04	0.61	0.30	-0.06	1.14
C	-0.70	-0.59	-0.60	-0.59	-0.48	-0.51	0.08	-0.28	-0.16	-0.02	-0.07	-2.28
G	0.70	0.52	0.56	0.06	0.22	0.07	-0.38	0.26	-0.68	-0.44	-0.20	-0.76
T	-0.23	-0.05	-0.31	0.22	-0.38	-0.04	0.04	-0.10	-0.23	0.02	0.27	-1.18
												+4.64

Figure 3. a) Normalized values of base versus position for mononucleotide analysis of 158 ribosome binding sites without alternative ATGs. b) The same values in normal matrix form, as in Figure 1. The constant in this case is 4.64.

Mononucleotide analysis excluding alternative ATGs

Several analyses were performed to assess the effect of alternative initiation codons, and to determine the mononucleotide parameters in the absence of those effects. In the first analysis all of the sequences with ATG occurring in positions other than 0 were simply eliminated from the data set. This removed 27 sequences and a total of 140 activity measurements. The analysis described above was repeated on this smaller set of 872 activity measurements on 158 sequences. (In a separate analysis we also removed alternative GTG sequences, since they should have the second largest effect, but no additional improvement was seen; data not shown). Figure 3 shows the parameters obtained, and the filled squares in Figure 4 plot the observed vs. expected activities for this matrix and this set of data. The correlation coefficient is 0.92 (r^2=0.85), a substantial improvement over the previous analysis that demonstrates the effect of alternative ATGs upstream of the initiation codon. Wild-type initiation sites rarely have alternative ATGs in the vicinity of the initiation codon (1,2). The standard deviation of the difference between the predicted and observed values is only 0.82, indicating that most of the time the observed activities are within about 2-fold of those predicted by the matrix in Figure 3b.

Removing the alternative ATGs from the data did not significantly change most of the matrix values, but did make G much better than T at position −11. This analysis also separated G and T at position 0, with G being better, as expected. In fact, the values obtained for the relative activities for ATG, GTG and TTG initiation codons, 1, 0.15 and 0.10, respectively, are identical to the relative rate constants for the different initiation

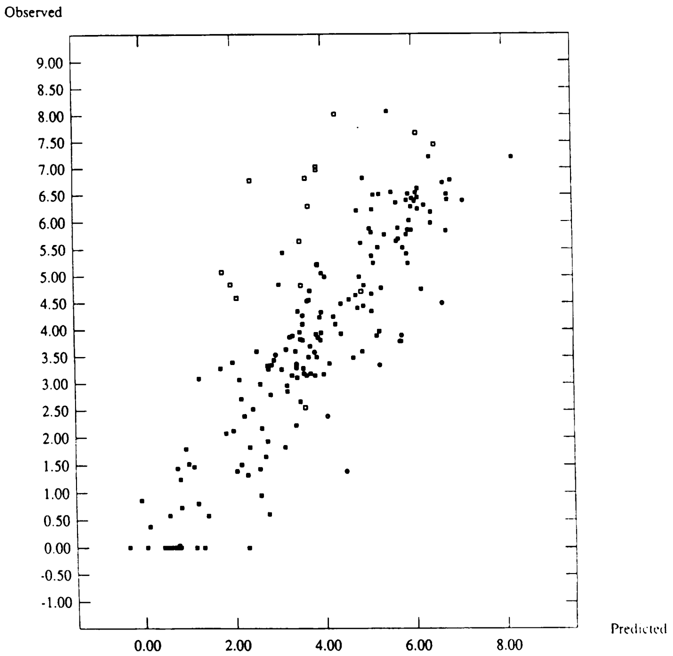


Figure 4. The filled symbols plot the predicted versus the observed *ln* (β-galactosidase activity) values for the 158 ribosome binding sites without alternative ATGs of Figure 3. The open symbols plot the values for those sequences with alternative ATGs. The predicted values of those sequences were also determined by the matrix of Figure 3. The open squares are sequences with in-frame ATGs and the open circles are those with out-of-frame ATGs.

codons determined by Ringquist *et al.* for a CAG second codon, as exists in these plasmids [see Table 4 of (12)]. The matrix in Figure 3b was then used to calculate the expected activities of the sequences that were removed from the data set (i.e., those sequences with ATGs at positions other than 0). Those values are plotted versus the observed values as the open symbols in Figure 4. In-frame ATGs generally increase the synthesis and out-of-frame ATGs generally decrease the synthesis relative to that expected from the matrix. This result is consistent with the hypothesis that these alternative ATGs can serve as initiation codons and justifies eliminating them from the analysis because alignment of the sequences by their initiation codons becomes ambiguous in those cases.

Two other analyses were performed on the entire data set, but with extra parameters included in the regression for sequences with alternative ATGs. In one analysis we added a parameter for each position at which an ATG could occur (−12 through −2), and in the other we added only two parameters, one for an ATG being in-frame and another for being out-of-frame. The results from the two analyses are essentially the same, both in terms of r^2 (=0.83) and the individual base-position parameters. Both are better than the first analysis but neither is as good as eliminating alternative ATGs from the data. This is due to non-additive effects of the alternative ATGs. If the initiation codon at position 0 is an ATG then the alternative ATGs have much smaller effects than if position 0 is not an A. In Figure 4, all of the open symbols that appear as outliers are from clones that do not have an A at position 0. On average, an in-frame ATG increases the activity about 12-fold over expected if position 0 is not an A, but only about 2-fold over expected if it is an A.

a)

pos	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2
A	368	363	351	333	367	373	346	360	367	428	283	303	971	0	0
C	137	108	83	78	103	129	163	207	218	195	298	309	0	0	0
G	350	476	511	540	443	365	259	224	198	257	152	173	71	0	1054
T	200	108	110	104	142	188	287	264	272	175	322	270	13	1055	1

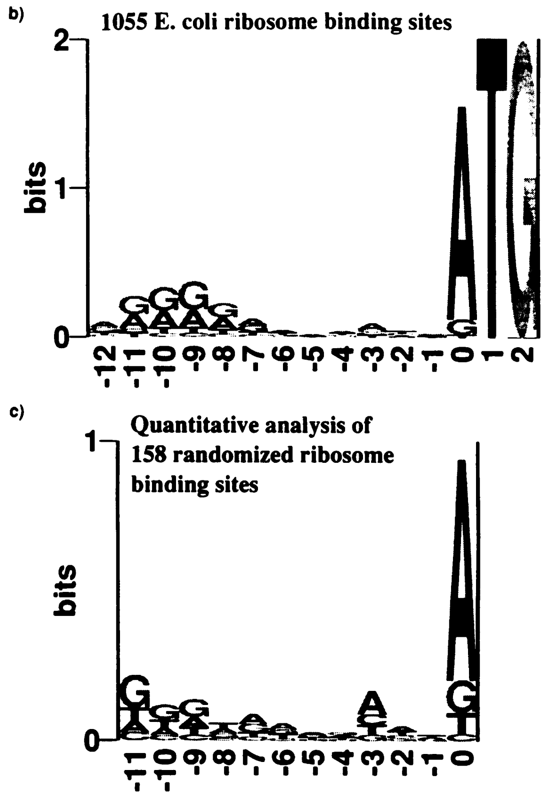


Figure 5. a) The number of times each base occurs at each position, from −12 to +2, in the 1055 natural initiation sites collected by Rudd and Schneider (1). **b)** The 'logo' for the data in part a), showing the information content at each position and the relative frequencies of each base (30). **c)** The 'logo' based on the information content of the matrix from Figure 3b (25,26).

An out-of-frame ATG decreases activity about 3.5-fold if position 0 is not an A, but only by about 10% if it is an A.

Higher order analyses

Despite the fact that the matrix in Figure 3 is fairly good at predicting the activities of initiation sites without alternative ATGs, there are still some outliers obvious in the filled symbols of Figure 4. The three farthest outliers below the diagonal, for which the observed activity is much less than expected from the sequence, are clones 1499, c551 and 1362. It is known that RNA structure that includes the initiation site can inhibit initiation, so we looked for structures that might explain the activities of these clones. Both 1499 and 1362 have long complementarities with their synthetic ribosome binding sites. In 1499, the 10-long sequence from positions −11 to −2 is complementary (with three G−U pairs) with the coding region of *lac*Z just 3′ of the *Bam*HI cloning site. In 1362, the 8-long sequence from positions −9 to −2 is complementary (with three G−U pairs) to the leader region overlapping the *Hind*III and *Bgl*II sites. The clone c551, with a very A-rich ribosome binding site, has many potential interactions with the leader region, but none of them are particularly long or stable. We checked for structures in the first 200 bases of the mRNA sequence using the Zuker suboptimal structure program (29) but found nothing significant. It is possible there are longer range structures that inhibit initiation in clone c551, but we have currently no good explanation for its low activity. It does have a CTG initiation codon which contributes to the low activity, but other sequences with CTG are much better and the matrix of Figure 3b includes the discrimination against CTG initiation codons.

The three farthest outliers above the diagonal, for which the observed activity is much higher than expected from the sequence, are c501, 946 and 1082. Each of these has a good Shine/Dalgarno sequence at an appropriate distance 5′ of the initiation codon. c501 has AAGGAG from positions -13 to -8 and 1082 has the same sequence from -12 to -7. 946 has AAGG from -11 to -8, or alternatively AGGTGA from -10 to -5. These suggest that the mononucleotide matrix does not adequately account for good Shine/Dalgarno sequences. This is not too surprising since the free energy of base-pairing comes from stacked dinucleotides and is not additive over single base interactions. Yet the matrix of Figure 3 does fairly well at predicting activities except for a few cases with especially good Shine/Dalgarno interactions. It appears that the parameters in the region of -8 to -11 are a compromise whereby sites without Shine/Dalgarno sequences are somewhat over predicted and sites with Shine/Dalgarno sequences are somewhat under predicted, but on average all are predicted within an accuracy of about 2-fold except for those with exceptionally good Shine/Dalgarno sequences. Practically, this means that in addition to calculating predictions using the matrix, it is also necessary to increase the prediction for those sequences with good Shine/Dalgarno sequences with appropriate spacing.

In a further analysis we tested whether taking Shine/Dalgarno sequences into account specifically could improve the fit. We chose AGG, GAG and GGA to be sequences we counted as Shine/Dalgarno-like, and added a parameter for each possible position (-12 to -2) where they could occur. This regression resulted in an improved fit with $r^2 = 0.89$. The coefficients obtained indicate that sequences with a Shine/Dalgarno-like sequence, with a spacing of 5 to 9 bases between it and the initiation codon, results in an average of 2.4-fold increased activity over the individual base contributions. Shine/Dalgarno-like sequences with smaller spacings had essentially no effect on the activity. While these results are consistent with the model of Shine/Dalgarno sequences contributing in excess of the sums of their mononucleotide contributions, they must be taken with caution. By using trinucleotides as potential Shine/Dalgarno sequences we end up with only a few examples per parameter which limits the reliability of those values. Even with this large data set we do not have a sufficient sample size to do an extensive analysis of higher order contributions to the activity. Those types of analyses are most reliably done by directed studies of particular features, both for the effects of sequences (12) and of structures (11).

Comparisons of the quantitative predictions

In order to test the accuracy of the quantitative predictions on a collection of data which were not included in the regression analysis, we compared the predictions from the matrix of Figure 3b to the measured activities for the set of oligos used in Ringquist *et al.* (12) that were in the identical context for the rest of the sequence. The set of clones sd2−sd8, for each of the initiation codons ATG, GTG and TTG, are identical in sequence to those reported in this paper except for the region from -11 to 0. That is a total of 21 sequences; all had the Shine/Dalgarno sequence AAGGA with spacings to the initiation codon varying from 2 to 8 bases. Using the matrix of Figure 3b for the predictions, all of the sequences with spacings from 5 to 8, regardless of the initiation codon, had measured activity within 2-fold of the prediction. For the shorter spacings, from 2 to 4, the predictions were consistently too high. In this region our data set had very few examples of good Shine/Dalgarno sequences and so does not adequately account for their inhibitory effect (12). In our analysis with parameters for Shine/Dalgarno-like sequences at all possible spacings, described above, we found that such sequences with spacings between 2 and 4 had no effect on activity, either positive or negative. This analysis of the Ringquist *et al.* data suggests that short spacings may actually reduce activity over what would be observed with no Shine/Dalgarno sequence. In other work (Barrick, unpublished) it appeared that a Shine/Dalgarno sequence upstream of the initiation codon too far to be effective for translation initiation could, in fact, inhibit translation if it was near enough that ribosomes bound to it would block access to an appropriately spaced Shine/Dalgarno sequence. It may be possible in general for ribosomes to be repressors of translation by binding to Shine/Dalgarno sequences in such a way that they are ineffective at translation initiation but can inhibit binding at other locations that would be effective. This possibility could also contribute to the paucity of Gs in the vicinity of natural ribosome binding sites (1−4).

We also calculated the predicted activities from a large collection of natural ribosome binding sites of *E.coli* (1) using only the region of -11 to 0. Of course these are not expected to be accurate predictions of the translational yields of those genes because the sequences differ at most other positions as well, and many of those will contribute to the initiation rates. However, it is interesting to compare the differences in sequences between our randomized set of ribosome binding sites and those selected to serve as the translational initiation sites in *E.coli*. The highest predicted activity from the natural set is about 16,500 units, nearly as high as the 24,000 units predicted for the best possible sequence from the matrix in Figure 3b. The highest activity in our set of sequences was just under 3,200 units. The mean predicted activity of the natural sites was about 500 units, much higher than the mean of 45 units from our set. The minimum activity predicted for the natural set was about 4 units, significantly better than the less than one unit of the randomized set.

Figure 5a gives the counts for each base at each position in the 1055 *E.coli* translation initiation sites tabulated by Rudd and Schneider (1), for the region -12 to 2. Figure 5b shows the information content 'logo' for that data (30), in which the height at each position is the total information content of the position, and the height of each base is written in proportion to its occurrence at the position, with the most common base on top. The total information content of the natural sites is about 8.7 bits, of which about 3.2 bits comes from the region of -11 to 0. This is based only on the sequences without taking their different activities into account. If the same measurement is made on the sequences in Table 1 (which are only weakly selected) they contain less than 1.1 bits. Information content can also be calculated directly from the matrix which takes into account the different activities of different sequences (25,26). By such an analysis the matrix of Figure 3b has an information content of about 1.9 bits, and the 'logo' for it is shown in Figure 5c.

Qualitatively the graphs are very similar but they differ quantitatively, most importantly as seen in the differences in the vertical scales. The small amount of information seen in both sets is consistent with most mutation studies, where single base changes in this region, unless they create or eliminate secondary structures, usually have only moderate affects on translational yields. Conversely, changes to the TG nucleotides of the initiation codon have much more dramatic affects. The sequence in this upstream region can be used to set translational activities over a wide range, over 3000-fold in our data and even more in the natural sites, in small increments.

One can also make a predictive matrix from example sequences, as is typically done with 'weight matrix' approaches for representing sites (28). An appropriate method for constructing a weight matrix from example sites uses the logarithms of occurrences (31). If the logarithms are taken of the numbers in Figure 5a, over the region of -11 to 0 (with the value of 0 for C at position 0 replaced by 1), the correlation coefficient between those values and the elements of the matrix in Figure 3b is $r = 0.88$, showing they contain mostly the same information. In an earlier study we used a neural network called a 'perceptron' to find a weight matrix that could discriminate between translation initiation sites and other sites in the *E.coli* genome (32). That perceptron was trained with only 124 examples sites, most of them from phage. Nonetheless, the weight matrix with the best discriminatory ability, W101, has a correlation coefficient of $r=0.80$ (over the region -11 to 0) with the matrix of Figure 3b. These results show that three different methods, a simple weight matrix based on example sites, a weight matrix found by a neural network to provide discrimination between sites and non-sites, and a weight matrix based on a best-fit to quantitative activity data, all converge to very similar representations of the important features in those sites. This happens even though each analysis was performed on almost completely independent sets of data; the only overlap was 41 initiation sites that were common to the 124 sites used in the perceptron work and the 1055 natural sites.

While the distributions of the predicted activities of the natural sites and the activities in our data overlap, the natural sites are selected to be very active compared to the sites picked up in our screen of random sequences. The mean of the predicted activities of the natural sites is more than 10-fold higher than for the random sites, and 75% of the natural sites have predicted activities higher than 75% of the random sites; that is, the 25th percentile of the natural sites corresponds to the 75th percentile of the random sites. The sites picked up in our screen are on a high copy number plasmid and driven by a strong promoter under inducing conditions. Those which have low, or even moderate, β-galactosidase activity under those conditions are not likely to be active enough for most real genes. The selection of the real genes for higher activity, compared to the whole range of possible activities, explains the high information content of them as a group. However, it seems puzzling that CTG has never been observed as a natural initiation codon. According to our results it can function at about 3% the level of an ATG, well within the range of observed variation in natural initiation sites. Over 10% of the random sites contain a CTG initiation codon and the highest of them (clone 1190) has about 45 units of activity (clone 1129 has an in-frame ATG which probably contributes to its high activity), putting it at about the 10th percentile for the predictions of the natural sites. Therefore one might expect there to be genes that initiate with CTG codons, but in over 1000 examples none has been identified (1). One possibility is that some genes on the list really do begin with CTG and the initiation codon has been mis-identified because CTG is not expected, or at some time CTG initiation codons will be identified. Another possibility is that CTG is really worse for translation than our measurements suggest, at least under some growth conditions, so that selection prevents them from being utilized. For example, both ATA and ACG can function as initiation codons in *E.coli*, but both are also temperature sensitive so that they provide much less synthesis at 37 °C than at 21 °C (33), and neither is seen as a natural initiation codon (1).

CONCLUSIONS

Sequence and structure are both important for the activity of translation initiation sites. This analysis of randomized initiation sites has revealed features of initiation sites that were already known, verifying their importance. It also indicates there are unlikely to be other signals in the region of -11 to 0 that are used to facilitate initiation, or that such other signals are improbable enough to not be seen in our limited randomized collection. Translational enhancer sequences have been identified but they exist outside of this region (34–36). The quantitative results of this analysis are not as accurate as more directed studies, such as those in Ringquist *et al.* (12), for assessing the contributions of particular features, such as Shine/Dalgarno sequences and their spacings from the initiation codons. However, overall it gives a fairly reliable picture of the relative importance of different features, and has the advantage of an unbiased search for possibly important features. This general method, of randomization followed by quantitative activity measurements and multiple regression analysis, could be used to identify signals before enough was known to do more directed experiments. It should be even more effective in determining the quantitative importance of sequence features for DNA binding activities where complications of secondary structure will not interfere with the analysis.

ACKNOWLEDGEMENTS

We thank Ken Rudd for supplying us with the *E.coli* translation initiation sites electronically and before publication, and Vishnumohan Jejjala for technical assistance. This work was supported by NIH grants to Gary Stormo (GM28755) and Larry Gold (GM28685).

REFERENCES

1. Rudd, K.E. and Schneider, T.D. (1992) In A Short Course in Bacterial Genetics: A Laboratory Manual and Handbook for *Escherichia coli* and related bacteria. (Miller, J., Ed.) Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. pp. 17.19–17.45.
2. Gold, L., Pribnow, D., Schneider, T. D., Shinedling, S., Singer, B.S. and Stormo, G. D. (1981). Ann. Rev. Microbiol. 35, 362–403.
3. Stormo, G. D. (1986). In Maximizing Gene Expression Reznikoff, W. and Gold, L. eds., pp. 195–224, Butterworths, Stonesham, MA.
4. Gold, L. (1988) Ann. Rev. Biochem. 57, 199–233.
5. McCarthy, J. and Gualerzi, C. (1990) em Trends Genet. 6, 78–85.
6. Shine, J. and Dalgarno, L. (1974). Proc. Natl. Acad. Sci USA 71, 1342–1346.
7. Stormo, G. D., Schneider, T. D. and Gold, L. M. (1982). Nucl. Acid Res. 10, 2971–2995.
8. Dreyfus, M. (1988) J. Mol. Biol. 204, 79–94.
9. Hui, A. and de Boer, H. (1987) Proc. Natl. Acad. Sci. USA 84, 4762–4766.
10. Looman, A.C., Bodllaender, J., de Gruyter, M., Vogelaar, A. and van Knippenberg, P.H. (1986) Nucl. Acids Res. 14, 5481–5497.

11. de Smit, M.H. and van Duin, J. (1990) Proc. Natl. Acad. Sci. USA 87, 7668–7672.
12. Ringquist, S., Shinedling, S., Barrick, D., Green, L., Binkley, J.,Stormo, G.D. and Gold, L. (1992) Molec. Microbiol. 6, 1219–1229.
13. Childs, J., Villanueba, K., Barrick, D., Schneider, T. D., Stormo, G. D., Gold, L., Leitner, M. and Caruthers, M. (1985). In Sequence Specificity in Transcription and Translation, UCLA Symposia on Molecular and Cellular Biology New Series, Volume 30, (Calender, R., and Gold, L., eds.) pp 341–350, Alan R. Liss, Inc., New York.
14. Dotto, G. P., Enea, V. and Zinder, N. D. (1981). Virology 20, 1645–1655.
15. de Boer, H.A., Comstock, L.J. and Vasser, M. (1983) Proc. Natl. Acad. Sci. USA 80, 21–25.
16. Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F. and Peterson, G.B. (1982) J. Mol. Biol. 162, 729–773.
17. Maniatis, T., Fritsch, E. F. and Sambrook, J. (1982). Molecular Cloning, A Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
18. Gardella, T., Moyle, H. and Susskind, M.M. (1989) J. Mol. Biol. 206, 579–590.
19. Arvidson, D.N., Youderian, P., Schneider, T.D. and Stormo, G.D. (1991) BioTechniques 11, 733–738.
20. Miller, J. (1972). Experiments in Molecular Genetics, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
21. Schneider, T. D., Stormo, G. D., Haemer, J. S. and Gold, L. (1982). Nucl. Acid Res. 10, 3013–3024.
22. Schneider, T. D., Stormo. G. D., Yarus, M. A. and Gold, L. (1984). Nucl. Acid Res. 12, 129–140.
23. Stormo, G. D., Schneider, T. D. and Gold, L (1986). Nucl. Acid Res. 14, 6661–6679.
24. Schneider, T. D., Stormo, G. D., Gold, L. and Ehrenfeucht, A. (1986) J. Mol. Biol. 188, 415–431.
25. Stormo, G.D. and Yoshioka, M. (1991) Proc. Natl. Acad. Sci. USA 88 5699–5703.
26. Stormo, G.D. (1991) In Protein–DNA Interactions, Methods in Enzymology, Vol. 208 (Sauer, R.T., Ed.) Academic Press, Inc. pp.458–468.
27. Jacques, N., Guillerez, J. and Dreyfus, M. (1992) J. Mol. Biol. 226, 597–608.
28. Stormo, G.D. (1988) Ann. Rev. Biophys. Biophys. Chem. 17, 241–263.
29. Zuker, M. (1989) Science 244, 48–52.
30. Schneider, T.D. and Stephens, R.M. (1990) Nucl. Acids Res. 18, 6097–6100.
31. Stormo, G.D. (1990) In Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences, Methods in Enzymology, Vol. 183 (Doolittle, R.F., Ed.), Academic Press, Inc. pp. 211–221.
32. Stormo, G.D., Schneider, T.D., Gold, L. and Ehrenfeucht, A. (1982) Nucl. Acids Res. 10, 2997–3012.
33. Shinedling, S., Gayle, M., Pribnow, D. and Gold, L. (1987) Mol. Gen. Genet. 207, 224–232.
34. Boni, I.V., Isaeva, D.M., Musychenko, M.L. and Tzareva, N.V. (1991) Nucl. Acids Res. 19, 155–162.
35. Zhang, J. and Deutscher, M.P. (1992) Proc. Natl. Acad. Sci. USA 89, 2605–2609.
36. Olins, P. and Rangwala, S. (1989) J. Biol. Chem. 264, 16973–16976.