Volume 15 Number 1 1987 Nucleic Acids Research

Potential secondary structure at translation-initiation sites

M.C.Ganoza, E.C.Kofoid*, P.Marlière+ and B.G.Louis

Banting and Best Department of Medical Research, University of Toronto, 112 College Street, Toronto, Ontario M5G 1L6, Canada

Received May 21, 1986; Revised October 2, 1986; Accepted November 17, 1986

ABSTRACT

Since translational start codons also occur internally, more-complex features within mRNA must determine initiation. We compare the potential secondary structure of 123 prokaryotic mRNA start regions to that of regions coding for internal methionines. The latter display an unexpectedly-uniform, almost-periodic pattern of pairing potential. In contrast, sequences 5' to start codons have little self-pairing, and do not pair extensively with the proximal coding region. Pairing potential surrounding start codons was found to be less than half of that found near internal AUGs. In groups of random sequences where the distribution of nucleotides at each position, or of trinucleotides at each in-frame codon position, matched the observed natural distribution, there was no periodicity in the pairing potential of the internal sequences. Randomized internal sequences had less pairing: the ratio of pairing intensity between internals and starts was reduced from 2.0 to 1.6 by randomization.

We propose that the transition from the relatively-unstructured start domains to the highly-structured internal sequences may be an important determinant of translational start-site recognition.

INTRODUCTION

Start codons are generally homonyms of methionine, valine or leucine codons (for review see 1,2). Features other than the initiation triplet itself must therefore be necessary to determine a translation start site within mRNA. In principle, the specificity required to initiate translation selectively could be imparted by primary, secondary or tertiary structure of the molecule. In prokaryotes, a polypurine tract complementary to the 3' end of 16S rRNA (3) is important in start-site designation (4,5,6). Nevertheless, this is not always necessary (7,8,9,10,11,12) nor sufficient (13). Effects of nucleotides 5' (14,15) and 3' (16,17) to the start codon have been considered, as has the presence of UAA or UGA terminators (18,19). Statistically-significant base-frequency biases have been observed in domains neighboring start sites (20). Several authors have pointed out isolated instances where stem-and-loop structures, with the initiator codon buried in the stem or

exposed in the loop, can be postulated (1,21). It has been suggested that loop-exposed AUG or GUG correlates with efficiency of expression (22,23). Release of embedded Shine-Dalgarno or AUG signals from intramolecular association has been used to explain widely-varying translational efficiencies of the λ cro deletions (22), and of mutations in the lam B protein of E. coli (23).

We now ask whether there are secondary-structural features that can occur frequently in domains known to specify initiation of translation. We have analysed the potential local stem-and-loop formation of 123 prokaryotic start regions and compared these to regions coding for internal methionines.

METHODS

Initiation-site sequences were examined for common pairings essentially as described by Trifonov and Bolshoi (24). Sequences to be examined were compiled in a standard format in which the start codon was centered (as positions 35-37) in a 71-nucleotide window. First the 71-base mRNA sequences were compared. From any group of two or more sequences with 54 or more identical bases at corresponding positions, one sequence was arbitrarily chosen as group representative. This eliminated five of the 123 original sequences, leaving 118 sequences for further analysis. Self-pairing matrices (25) were formed for all of the class representatives with the aid of a computer program that finds, by exhaustive search, all possible foldings of a given sequence that exceed minimum criteria for length (4 or more) and stability (K_a at least 50,000).

Local potential mRNA secondary structure was deduced by an iterative procedure. Each nucleotide, n_i, was compared with the nucleotides, n_j ($j>i$), following it in the sequence. When a base-pairing match was found, the residues n_{i+1} and n_{j+1} were compared. The comparison continued until two successive mismatches were found; then the energy, length and position of the stem were stored provided that length and stability exceeded the given minima. G·U base pairs were tolerated internally, but were treated as mismatches when adjacent to two or more unambiguous mismatches. Approximate base-pairing energies were assigned: G·U=1, C·G=4, A·U=2. Mismatched bases were given a value of -2. Constraint produced by formation of small loops was taken into account using the calculations described by Ninio (26).

The triangular self-pairing matrices were then superimposed and summed. The average pairing potential was calculated as the sum of all pairings in the superposed matrix divided by the product of sequence length and number of

sequences. To reduce the effect of minor variation in the positions of common features (if any), the composite matrix was smoothed by averaging each cell with its eight nearest neighbours. The smoothed composite array was plotted as a density map (the algorithm used for this purpose allowed a resolution of one part in 50). The same array was then replotted showing only the cells with densities over half of the maximum (i.e., all other cell densities were set to 0), thus displaying prominent densities in isolation. Results from all of the analyses described in this and the next paragraph (except for the tRNA controls) were all plotted to a common scale to facilitate comparison of the overall densities.

Several controls were performed. A set of 51 tRNA sequences was examined. A set of 123 sequences surrounding non-initiator AUGs within the coding regions was created by taking one such sequence from each of the 123 genes making up the start-sequence library (in a few cases there were no internal AUGs, and the deficiency was made up by choosing one at random from another gene), and this was analysed in the same way. For both the start and the internal sequence libraries, the frequency of occurrence was then calculated for each base at each of the 71 positions in the sequence, and for each triplet at each of the 23 in-frame codon positions. These frequency distributions (Table 1) were used to generate four further libraries: one containing 2000 random sequences with base distribution at each position corresponding to that found in the corresponding position in the start sequences; one with 2000 random sequences where the bases were distributed as in the internal (coding) sequences; one with 2000 random sequences of length 69 with nucleotide triplets corresponding to the codon distributions at positions 2 to 68 of the start sequences; and the fourth with 2000 random sequences of length 69 matching the distributions of codons among the internal sequences. These too were analysed and mapped as described in the foregoing paragraph except that the sums were divided to match the density scales with those of the gene libraries.

Most initiator and phe-tRNA sequences in the control set were from Gauss and Sprinzl (27). Other Met initiator tRNA data (courtesy of Drs. Y. Kuchino and S. Nishimura) were for H. morrhuae, T. acidophilum and S. acidocaldarius; the sequence from S. faecalis was provided by Dr. J. Rabinowitz, and M. musculus mitochondrial initiator and phe-tRNAs were from Bibb et al. (28).

Translation start sequences and mRNA internal sequences were from both bacterial and bacteriophage genes. The compilation, including references, is available on request.

RESULTS

The secondary-structure-forming potential of random RNA molecules is expected to involve about 50% of the residues in base-paired structures (29). Before we tested the start regions for common or unique potential mRNA secondary structure, we checked that the programs accurately detect the cloverleaf structure of tRNA. Average density maps were produced as described in Methods from a set of 51 tRNAs, from which the initial screening yielded 21 representative sequences. Dense regions in such maps with their long axes perpendicular to the hypotenuse reflect potential to form paired structures involving similarly-located bases in a significant number of the sequences. Common or frequently-occurring complementary contacts thus cluster in characteristic positions, whereas other contacts disappear as a nearly-uniform background. The 21 representative tRNA sequences showed areas of density corresponding to the common aminoacyl, anticodon, D and T stems (Figure 1A), along with other regions of background density. The aminoacyl acceptor stem

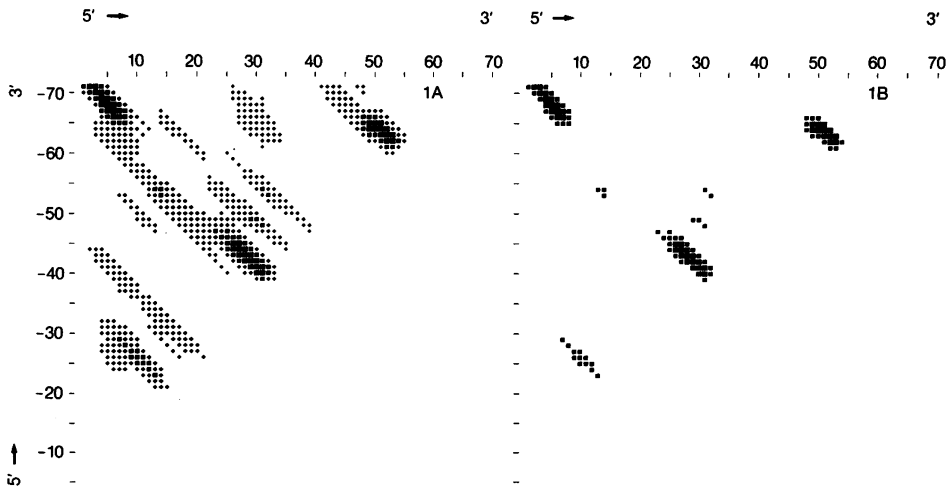

Figure 1. Self-Pairing of tRNA Sequences

Self-pairing matrices were formed for members of a library of tRNA sequences as described in Methods. The matrices were superimposed and summed and the composite matrix was smoothed. The results were plotted (A) such that each cell of the matrix is represented on the diagram as a 7x7-dot square, the blackness of which (on a scale from 0 to 49 black dots) is proportional to the number of sequences in the library that could form a pairing involving that cell.

Prominent densities were isolated by replotting the data (B) without cells having a pairing density beneath 50% of maximum (i.e., the densities of such cells were reset to 0 before plotting).

Table 1. Nucleotide Distributions for Start and Internal Sequences

NUCLEOTIDE DISTRIBUTIONS FOR START AND INTERNAL SEQUENCES

Position	Starts					Internals				
	A	C	U	G	Total	A	C	U	G	Total
1	40	25	33	25	123	34	35	25	29	123
2	49	24	27	23	123	20	35	24	44	123
3	29	26	37	31	123	35	29	37	22	123
4	42	20	33	28	123	41	26	25	31	123
5	40	32	29	22	123	22	33	30	38	123
6	37	31	31	24	123	36	24	40	23	123
7	34	28	34	27	123	36	36	22	29	123
8	35	32	31	25	123	13	35	30	45	123
9	46	26	29	22	123	35	21	48	19	123
10	37	26	32	28	123	43	29	24	27	123
11	35	29	33	26	123	23	23	34	43	123
12	36	30	30	27	123	37	30	38	18	123
13	37	32	31	23	123	35	36	24	28	123
14	32	24	38	29	123	14	23	37	49	123
15	25	27	56	15	123	27	28	43	25	123
16	31	26	50	16	123	44	32	20	27	123
17	39	35	29	20	123	33	28	21	41	123
18	40	20	39	24	123	34	33	40	16	123
19	31	31	47	14	123	32	35	21	35	123
20	40	22	37	24	123	18	26	30	49	123
21	32	27	42	22	123	20	47	33	23	123
22	25	17	49	32	123	49	24	30	20	123
23	22	10	45	46	123	22	25	32	44	123
24	11	11	37	64	123	35	27	41	20	123
25	8	8	42	65	123	38	33	27	25	123
26	12	8	41	62	123	21	30	25	47	123
27	22	11	43	47	123	35	33	28	27	123
28	30	16	47	30	123	41	33	24	25	123
29	42	14	40	27	123	25	28	25	45	123
30	41	15	46	21	123	34	28	38	23	123
31	36	18	50	19	123	37	41	22	23	123
32	26	28	51	18	123	24	28	29	42	123
33	47	25	36	15	123	38	31	34	20	123
34	31	29	44	19	123	34	24	31	34	123
35	1	0	116	6	123	1	0	116	6	123
36	123	0	0	0	123	123	0	0	0	123
37	0	0	0	123	123	0	0	0	123	123
38	24	13	46	40	123	16	36	28	43	123
39	15	55	35	18	123	40	35	22	26	123
40	48	22	38	15	123	43	22	22	36	123
41	10	24	61	28	123	18	27	36	42	123
42	23	21	56	23	123	32	27	41	23	123
43	41	23	36	23	123	45	25	24	29	123
44	36	19	49	19	123	23	27	33	40	123
45	49	29	35	10	123	33	30	36	24	123
46	39	22	47	15	123	43	31	18	31	123
47	13	26	55	29	123	25	31	33	34	123
48	35	24	52	12	123	44	28	35	16	123
49	38	29	36	20	123	38	32	22	31	123
50	21	26	49	27	123	19	27	32	45	123
51	39	33	34	17	123	31	25	52	15	123
52	41	29	26	27	123	41	24	29	29	123
53	23	30	34	36	123	24	26	29	44	123
54	35	33	35	20	123	28	29	43	23	123
55	33	37	29	24	123	45	37	19	22	123
56	22	31	33	37	123	23	18	32	50	123
57	36	24	42	21	123	26	27	48	22	123
58	40	34	27	22	123	37	40	23	23	123
59	16	30	47	30	123	18	25	33	47	123
60	38	24	38	23	123	34	28	37	24	123
61	37	36	26	24	123	36	35	24	28	123
62	19	35	26	43	123	23	27	30	43	123
63	39	33	34	17	123	37	27	35	24	123
64	28	31	29	35	123	43	29	21	30	123
65	27	24	26	46	123	18	35	33	37	123
66	38	35	37	13	123	39	31	28	25	123
67	30	36	25	32	123	48	33	20	22	123
68	26	31	29	37	123	23	28	33	39	123
69	38	38	30	17	123	35	26	43	19	123
70	39	35	22	27	123	41	35	20	27	123
71	18	29	28	48	123	23	22	27	51	123

Base sequences surrounding initiator codons at 123 known protein start sites were compiled such that each sequence comprised 34 nucleotides prior to the start site and 34 nucleotides following the start codon. The 5' nucleotide of the start codon was considered as occupying position 35 for ordinal reference. For each position from 1 through 71, the numbers of A, C, U and G nucleotides appear in the subcolumns so labelled. The column group labelled Internals contains data for a similarly-compiled library of 71-nucleotide sequences centered around non-initiator AUGs occurring in-frame within the coding regions of the genes used in the start-sequence library.

Table 2. Trinucleotide Distributions for Start Sequences

TRINUCLEOTIDE COMPOSITION: START SEQUENCES

pos=2	UUU: 9	UUC: 1	UUA: 1	UUG: 5	UCU: 3	UCC: 1	UCA: 2	UCG: 1
	UAU: 5	UAC: 2	UAA: 2	UAG: 3	UGU: 2	UGC: 4	UGA: 5	UGG: 3
	CUU: 2	CUC: 2	CUA: 1	CCU: 1	CCA: 2	CCG: 3	CAU: 2	CAC: 3
	CAA: 1	CAG: 2	CGU: 2	CGC: 2	CGG: 1	AUU: 3	AUA: 2	ACU: 1
	ACA: 3	ACG: 3	AAU: 3	AAC: 2	AAA: 4	AGU: 1	AGC: 1	AGA: 3
	AGG: 1	GUU: 1	GUA: 1	GUG: 1	GCU: 2	GCG: 4	GAU: 4	GAC: 1
	GAA: 3	GGU: 1	GGC: 1	GGA: 3	GGG: 1		TOTAL: 123	
pos=5	UUU: 5	UUC: 2	UUA: 3	UUG: 3	UCU: 1	UCC: 6	UCA: 4	UCG: 1
	UAU: 3	UAA: 4	UAG: 3	UGU: 2	UGC: 1	UGA: 2	CUU: 2	CUC: 2
	CUA: 2	CUG: 2	CCU: 3	CCA: 2	CCG: 2	CAU: 1	CAC: 4	CAA: 1
	CAG: 2	CGU: 4	CGC: 3	CGA: 2	AUU: 2	AUC: 1	AUA: 3	AUG: 3
	ACC: 3	ACA: 1	ACG: 3	AAU: 3	AAC: 2	AAA: 3	AGU: 3	AGC: 1
	AGA: 1	GUU: 3	GUA: 3	GUG: 1	GCU: 1	GCC: 1	GCG: 3	GAU: 1
	GAA: 3	GAG: 1	GGC: 2	GGG: 3			TOTAL: 123	
pos=8	UUU: 9	UUC: 4	UUA: 3	UCU: 2	UCA: 2	UCG: 3	UAU: 2	UAA: 2
	UAG: 3	UGU: 2	UGC: 3	CUU: 5	CUC: 3	CUA: 1	CUG: 4	CCC: 1
	CCG: 2	CAU: 2	CAC: 1	CAG: 5	CGU: 1	CGC: 4	CGA: 3	AUU: 2
	AUC: 1	AUA: 3	AUG: 1	ACU: 4	ACC: 3	ACA: 2	ACG: 2	AAU: 1
	AAC: 1	AAA: 5	AAG: 2	AGU: 1	AGA: 2	AGG: 1	GUU: 3	GUC: 1
	GUA: 5	GUG: 1	GCU: 1	GCA: 1	GCG: 3	GAU: 1	GAC: 2	GAA: 1
	GAG: 1	GGU: 1	GGC: 2	GGA: 2			TOTAL: 123	
pos=11	UUU: 8	UUC: 4	UUA: 4	UUG: 3	UCU: 2	UCA: 1	UAU: 1	UAC: 1
	UAA: 2	UAG: 2	UGU: 1	UGC: 3	UGA: 2	UGG: 1	CUG: 5	CCU: 5
	CCC: 3	CCA: 2	CCG: 1	CAU: 1	CAA: 2	CAG: 2	CGU: 3	CGA: 2
	CGG: 3	AUU: 2	AUC: 4	AUA: 1	AUG: 1	ACU: 2	ACC: 2	ACA: 1
	AAU: 6	AAC: 5	AAA: 4	AGU: 1	AGC: 3	AGG: 1	GUC: 1	GUA: 1
	GUG: 2	GCU: 3	GCC: 3	GCA: 4	GCG: 1	GAU: 1	GAA: 2	GAG: 1
	GGU: 1	GGC: 3	GGA: 3				TOTAL: 123	
pos=14	UUU: 5	UUC: 1	UUA: 1	UUG: 2	UCA: 5	UAU: 3	UAC: 1	UAA: 6
	UAG: 3	UGU: 1	UGC: 2	UGA: 2	CUU: 2	CUC: 2	CUA: 1	CUG: 1
	CCU: 1	CCC: 3	CCA: 1	CCG: 1	CAU: 1	CAC: 1	CAA: 5	CAG: 1
	CGU: 1	CGC: 2	CGA: 1	AUU: 3	AUA: 2	ACU: 3	ACC: 1	ACA: 2
	ACG: 3	AAU: 3	AAC: 8	AAA: 10	AAG: 1	AGC: 2	GUU: 1	GUA: 3
	GUG: 1	GCU: 2	GCC: 3	GCA: 1	GCG: 1	GAU: 4	GAA: 8	GAG: 1
	GGU: 1	GGA: 2	GGG: 1				TOTAL: 123	
pos=17	UUU: 6	UUA: 5	UUG: 1	UCC: 2	UCA: 3	UCG: 1	UAU: 2	UAC: 2
	UAA: 7	UGC: 8	UGA: 2	CUU: 7	CUA: 5	CUG: 2	CCU: 1	CCC: 2
	CCA: 2	CAU: 2	CAC: 1	CAA: 5	CAG: 3	CGC: 2	CGA: 2	CGG: 1
	AUU: 2	AUC: 4	AUA: 1	AUG: 1	ACU: 1	ACC: 1	ACA: 3	AAU: 4
	AAC: 3	AAA: 4	AGC: 2	AGA: 1	AGG: 2	GUC: 2	GUA: 2	GUG: 2
	GCU: 1	GCA: 2	GCG: 1	GAU: 4	GAA: 2	GGU: 1	GGC: 2	GGA: 1
							TOTAL: 123	
pos=20	UUU: 5	UUC: 1	UUA: 5	UUG: 4	UCU: 1	UCA: 5	UCG: 2	UAU: 2
	UAA: 4	UAG: 3	UGC: 2	UGA: 3	UGG: 3	CUU: 3	CUA: 3	CUG: 1
	CCU: 1	CCC: 2	CCA: 2	CAU: 1	CAA: 6	CAG: 1	CGC: 2	AUC: 1
	AUA: 1	AUG: 2	ACU: 2	ACC: 2	ACA: 5	ACG: 2	AAU: 3	AAC: 2
	AAA: 6	AAG: 5	AGU: 2	AGC: 1	AGG: 3	GUA: 4	GUG: 2	GCU: 1
	GCA: 2	GAU: 2	GAC: 4	GAA: 1	GAG: 2	GGU: 2	GGA: 2	GGG: 2
							TOTAL: 123	
pos=23	UUC: 1	UUA: 1	UUG: 1	UCA: 2	UAU: 3	UAA: 4	UAG: 3	UGU: 1
	UGC: 2	UGA: 2	UGG: 2	CUG: 2	CCG: 1	CGG: 7	AUA: 2	ACC: 1
	ACA: 2	ACG: 2	AAU: 1	AAA: 4	AAG: 7	AGA: 3	AGG: 23	GUU: 1
	GUC: 2	GUG: 1	GCU: 1	GCA: 1	GCG: 1	GAA: 3	GAG: 12	GGU: 1
	GGC: 2	GGA: 18	GGG: 3				TOTAL: 123	
pos=26	UUU: 1	UUA: 1	UCC: 1	UCA: 1	UCG: 1	UAC: 2	UAA: 2	UGA: 1
	UGG: 2	CUU: 2	CUA: 1	CCA: 1	CAU: 2	CAC: 1	CGA: 1	AUU: 3
	AUA: 4	AUG: 1	ACU: 2	ACA: 3	AAU: 4	AAC: 1	AAA: 3	AAG: 2
	AGU: 5	AGC: 3	AGA: 4	AGG: 6	GUU: 3	GUC: 2	GUA: 4	GCA: 1
	GCG: 1	GAU: 5	GAC: 6	GAA: 4	GAG: 11	GGU: 3	GGA: 16	GGG: 6
							TOTAL: 123	
pos=29	UUU: 8	UUC: 1	UUA: 6	UUG: 2	UCU: 1	UCC: 2	UCA: 3	UAU: 4
	UAA: 8	UAG: 1	UGC: 1	UGA: 3	UGG: 2	CUU: 3	CCU: 1	CCC: 1
	CCG: 1	CAU: 1	CAC: 2	CAA: 3	CAG: 1	CGG: 1	AUU: 4	AUC: 1
	AUA: 8	AUG: 2	ACA: 2	ACG: 1	AAU: 3	AAC: 5	AAA: 7	AAG: 2
	AGU: 2	AGA: 2	AGG: 1	GUU: 3	GUC: 1	GUA: 2	GCA: 1	GCG: 2
	GAU: 4	GAC: 3	GAA: 1	GAG: 1	GGU: 2	GGC: 1	GGA: 4	GGG: 2
							TOTAL: 123	
pos=32	UUU: 5	UUC: 3	UUA: 2	UCC: 3	UCA: 2	UCG: 1	UAU: 1	UAA: 4
	UGU: 1	UGC: 2	UGA: 1	UGG: 1	CUC: 1	CUA: 1	CUG: 9	CCU: 3
	CCC: 1	CCA: 1	CAU: 5	CAC: 2	CAA: 3	CGU: 1	CGC: 1	AUU: 9
	AUC: 4	AUA: 4	AUG: 1	ACU: 2	ACC: 3	ACA: 4	AAU: 1	AAC: 3
	AAA: 9	AAG: 5	AGU: 2	AGC: 1	AGA: 3	GUC: 2	GUA: 6	GCU: 1
	GCC: 1	GCA: 1	GCG: 2	GAC: 1	GAA: 2	GGC: 1	GGA: 1	
							TOTAL: 123	
pos=35	UUG: 1	AUG: 116	GUG: 6				TOTAL: 123	
pos=38	UUU: 3	UUC: 1	UCU: 10	UCA: 1	UCG: 2	UAU: 1	UAC: 5	UGU: 1
	CCU: 2	CCA: 2	CCG: 1	CAA: 2	CAG: 1	CGU: 1	CGA: 4	AUU: 1
	AUC: 3	AUA: 1	AUG: 3	ACU: 2	ACC: 1	ACA: 4	ACG: 3	AAU: 3
	AAC: 5	AAA: 10	AAG: 1	AGU: 4	AGC: 3	AGA: 1	AGG: 1	GUU: 1
	GUA: 2	GCU: 19	GCC: 3	GCA: 3	GCG: 2	GAC: 1	GAA: 5	GAG: 1
	GGA: 3						TOTAL: 123	

pos=41	UUU: 1	UUA: 1	UUG: 2	UCU: 3	UAU: 2	UAC: 1	CUC: 1	CUA: 1		
	CUG: 1	CCU: 1	CCA: 1	CAC: 1	CAA: 5	CAG: 3	CGU: 6	CGC: 2		
	CGA: 2	AUU: 5	AUC: 2	AUG: 3	ACU: 2	ACC: 2	ACA: 5	ACG: 4		
	AAU: 7	AAC: 5	AAA: 10	AAG: 8	AGU: 3	AGC: 2	AGA: 3	GUU: 3		
	GUA: 2	GUG: 1	GCU: 1	GCC: 1	GCA: 1	GAU: 4	GAC: 5	GAA: 4		
	GAG: 1	GGU: 3	GGC: 1	GGA: 1			TOTAL: 123			
pos=44	UUU: 5	UUC: 4	UUA: 8	UUG: 2	UCU: 5	UCA: 3	UCG: 1	UAU: 2		
	UAC: 4	UGG: 2	CUU: 4	CUC: 1	CUA: 1	CUG: 2	CCC: 1	CCA: 1		
	CAA: 4	CAG: 1	CGU: 1	CGA: 3	AUU: 8	AUC: 3	AUA: 1	AUG: 3		
	ACU: 6	ACC: 2	ACA: 6	AAU: 2	AAC: 4	AAA: 9	AAG: 2	AGC: 1		
	AGG: 2	GUU: 2	GUC: 1	GUA: 4	GCU: 3	GCA: 1	GAC: 1	GAA: 6		
	GGU: 1						TOTAL: 123			
pos=47	UUU: 4	UUC: 2	UUA: 2	UUG: 1	UCA: 1	UAU: 1	UAC: 2	CUU: 2		
	CUC: 1	CUA: 2	CUG: 5	CCA: 1	CCG: 1	CAU: 1	CAC: 1	CAA: 3		
	CAG: 3	CGU: 1	CGC: 3	CGA: 2	AUU: 8	AUC: 2	AUG: 1	ACU: 8		
	ACC: 3	ACA: 3	AAU: 7	AAC: 8	AAA: 10	AAG: 4	AGU: 1	GUC: 1		
	GUA: 1	GUG: 3	GCU: 2	GCA: 4	GCG: 1	GAU: 1	GAC: 4	GAA: 6		
	GAG: 1	GGU: 2	GGC: 2	GGA: 1			TOTAL: 123			
pos=50	UUU: 4	UUC: 3	UUG: 5	UCU: 2	UCA: 1	UAU: 3	UGC: 1	UGA: 1		
	UGG: 1	CUC: 1	CUA: 1	CUG: 4	CCU: 2	CCC: 3	CCA: 1	CCG: 2		
	CAA: 4	CAG: 3	CGU: 2	CGC: 2	CGA: 1	AUU: 5	AUC: 7	AUG: 2		
	ACU: 6	ACC: 2	ACA: 4	ACG: 2	AAU: 1	AAC: 4	AAA: 7	AAG: 6		
	AGU: 2	AGC: 1	GUU: 1	GUC: 2	GUA: 3	GUG: 1	GCU: 5	GCC: 1		
	GCA: 1	GCG: 1	GAU: 4	GAA: 2	GGU: 4	GGC: 2	TOTAL: 123			
pos=53	UUU: 3	UUC: 1	UUA: 2	UUG: 3	UCU: 5	UCC: 3	UCA: 1	UCG: 2		
	UAU: 1	UAA: 1	UGG: 1	CUU: 1	CUC: 7	CUG: 4	CCC: 2	CCA: 2		
	CAU: 1	CAA: 3	CAG: 2	CGU: 3	CGC: 5	AUU: 4	AUC: 1	AUA: 1		
	AUG: 3	ACU: 3	ACC: 2	ACA: 2	ACG: 1	AAU: 1	AAC: 6	AAA: 8		
	AAG: 1	AGA: 1	GUU: 2	GUC: 1	GUA: 1	GUG: 1	GCU: 3	GCC: 2		
	GCA: 2	GCG: 3	GAU: 2	GAC: 2	GAA: 4	GAG: 3	GGU: 4	GGC: 5		
	GGA: 1						TOTAL: 123			
pos=56	UUU: 2	UUC: 4	UUA: 3	UUG: 1	UCU: 3	UCA: 2	UAU: 1	UAC: 4		
	UGA: 1	UGG: 1	CUU: 2	CUC: 2	CUA: 1	CUG: 4	CCG: 2	CAU: 1		
	CAA: 5	CAG: 3	CGU: 5	CGC: 3	CGA: 2	CGG: 1	AUU: 4	AUC: 2		
	AUA: 1	AUG: 3	ACU: 2	ACC: 6	ACA: 1	AAU: 3	AAC: 4	AAA: 3		
	AAG: 4	GUU: 2	GUC: 1	GUA: 4	GCU: 2	GCC: 3	GCA: 2	GCG: 1		
	GAU: 8	GAC: 3	GAA: 1	GAG: 2	GGU: 5	GGC: 2	GGA: 1			
							TOTAL: 123			
pos=59	UUU: 5	UUC: 1	UUA: 4	UUG: 1	UCG: 2	UAU: 1	UGG: 2	CUU: 2		
	CUC: 1	CUA: 2	CUG: 4	CCU: 1	CCG: 1	CAU: 3	CAC: 2	CAA: 2		
	CAG: 2	CGU: 5	CGC: 5	AUU: 4	AUC: 4	AUA: 2	AUG: 1	ACU: 1		
	ACC: 2	ACA: 5	AAU: 3	AAC: 10	AAA: 7	AAG: 3	AGU: 1	AGC: 2		
	AGG: 2	GUU: 3	GUC: 2	GUA: 1	GUG: 1	GCU: 4	GCC: 4	GCA: 2		
	GCG: 2	GAU: 1	GAC: 1	GAA: 1	GAG: 2	GGU: 3	GGC: 2	GGG: 1		
							TOTAL: 123			
pos=62	UUC: 2	UUA: 2	UUG: 1	UCU: 4	UCC: 1	UCA: 3	UCG: 1	UAU: 1		
	UAC: 2	UGG: 2	CUU: 1	CUC: 3	CUA: 1	CUG: 9	CCU: 1	CCC: 1		
	CCA: 1	CCG: 1	CAA: 3	CAG: 6	CGU: 5	CGC: 1	CGA: 1	CGG: 1		
	AUU: 3	AUC: 4	AUG: 1	ACU: 1	ACC: 2	AAU: 2	AAC: 4	AAA: 5		
	AAG: 1	AGU: 1	AGC: 1	AGA: 1	GUU: 2	GUC: 5	GUA: 2	GUG: 3		
	GCU: 5	GCC: 2	GCA: 4	GCG: 6	GAC: 2	GAA: 6	GAG: 2	GGU: 2		
	GGC: 1	GGG: 1					TOTAL: 123			
pos=65	UUU: 1	UUC: 7	UUA: 1	UUG: 4	UCU: 5	UCC: 2	UCA: 1	UCG: 3		
	UAU: 1	UAC: 2	CUU: 1	CUC: 2	CUA: 3	CUG: 1	CCU: 1	CCA: 1		
	CCG: 4	CAU: 1	CAC: 1	CAA: 1	CAG: 3	CGU: 3	CGC: 2	AUU: 3		
	AUC: 1	AUG: 6	ACU: 1	ACC: 2	AAU: 1	AAC: 2	AAA: 6	AAG: 2		
	AGA: 2	GUU: 1	GUC: 3	GUA: 2	GUG: 2	GCU: 5	GCC: 4	GCA: 3		
	GCG: 3	GAU: 4	GAC: 6	GAA: 4	GAG: 3	GGU: 2	GGC: 2	GGA: 1		
	GGG: 1						TOTAL: 123			
pos=68	UUU: 2	UUC: 7	UUA: 3	UUG: 3	UCU: 5	UCC: 1	UCG: 2	UAC: 1		
	UGU: 2	CUU: 1	CUC: 1	CUA: 2	CUG: 1	CCU: 2	CCC: 1	CCA: 1		
	CCG: 3	CAU: 2	CAC: 2	CAA: 1	CAG: 7	CGU: 2	CGC: 4	CGA: 1		
	AUU: 3	AUC: 3	AUA: 1	ACU: 5	ACC: 1	ACG: 2	AAU: 4	AAC: 3		
	AAA: 1	AAG: 1	AGC: 2	AGA: 2	AGG: 1	GUU: 5	GUA: 3	GUG: 3		
	GCU: 4	GCC: 4	GCA: 5	GCG: 2	GAU: 1	GAC: 4	GAA: 2	GAG: 1		
	GGU: 1	GGC: 1	GGG: 1				TOTAL: 123			

Base sequences surrounding initiator codons at 123 known protein start sites were compiled such that each sequence comprised 34 nucleotides prior to the start site and 34 nucleotides following the start codon. The 5' nucleotide of the start codon was considered as occupying position 35 for ordinal reference. For each trinucleotide position in-frame with the start codon, beginning with position 2 and ending with position 68, the trinucleotides found are listed together with the number of each.

at positions 65-71 is seen paired with positions 1-8; similarly, the T stem (bases 48-53 with 61-66), the anticodon stem (25-33 with 39-47), and the D stem (8-14 with 23-29) are all visible as areas of high density. (Apparent

overlaps between the D and anticodon stems at positions 25-29, and between the aminoacyl and T stems at 65-66, are artefacts produced by summation of the individual pairing maps, among which differences in stem length are expressed as differences in stem positions relative to the anticodon. The important point here is the detection of a common structure; clearly, this technique indicates only the average position of any given feature.)

The use of a minimum threshold of 50% of the maximum observed density clarified the image as shown in Figure 1B. The variable loop of this set of tRNAs also remained visible when the threshold was applied.

Start domains of 71 nucleotides from 123 prokaryotic genes, for each of which the 36 bases 3' to the start codon were known to correspond to the first twelve N-terminal amino acids of the protein, were then investigated along with a collection of 123 internal sequences flanking internal in-frame AUGs from the coding regions of 123 different prokaryote mRNAs. The nucleotide distributions at each position along the 71-base sequences are given in Table 1, and the codon distributions in-frame with the start codon appear in Table 2. Of the 123 start sequences, 116 had AUG as the initiator codon; 6 had GUG and one started at UUG. The mononucleotide compositions for start sequences deviate from the random, as reported for another set of prokaryotic genes (20). As noted (20), the differences are even more apparent if di- and trinucleotides of internal and start sequences are compared.

A prominent difference apparent upon simple inspection of the composition data of Table 1 is the enrichment for G and A at positions -9 to -11 in the start library, as well as the higher proportion of A at positions -12 to -20. This position-specific purine enrichment defines the well-known Shine-Dalgarno domain of prokaryotic genes (3).

Figures 2A and 3A show the folding patterns of the 117 representative prokaryotic mRNA start regions, and figures 2B and 3B show the folding potential of the mRNA coding regions. The smoothed density maps of the gene start domains are, overall, less intense by a factor of 1.98 than those from regions surrounding internal AUGs, reflecting a corresponding difference in pairing potential.

Differences between the start and internal sequences become more apparent when prominent densities are isolated by application of a 50% threshold (Figures 3A and 3B). Less pairing is observed within the regions 5' to the AUG as compared with the 123 internal sequences. The latter exhibit a uniform, almost-periodic pattern of pairing potential with intensities

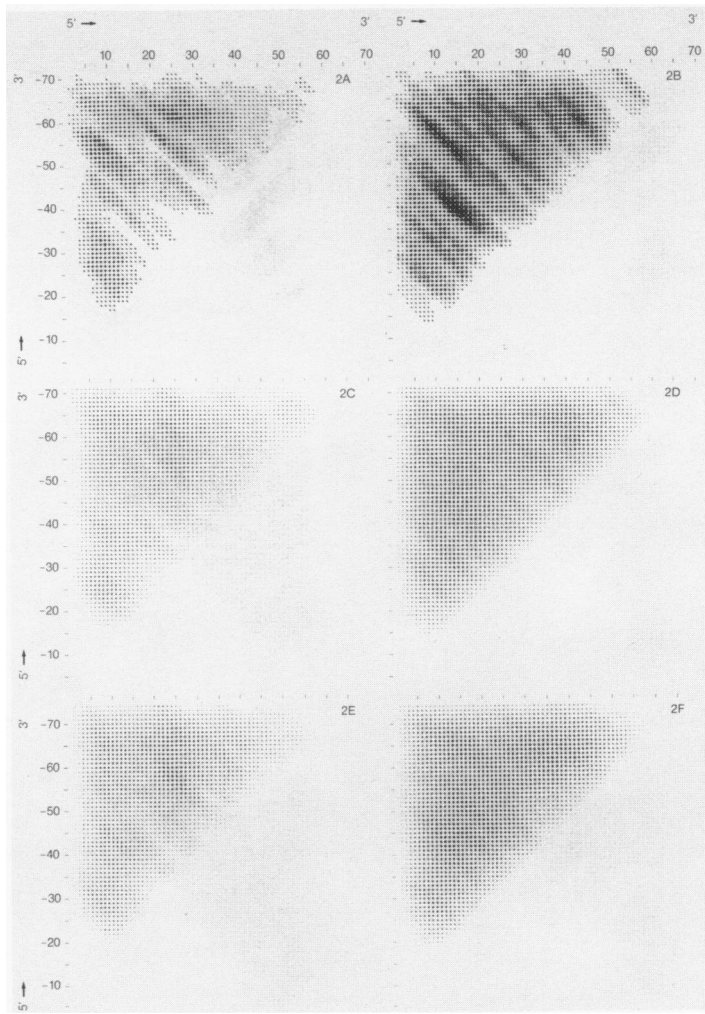

Figure 2. Self-Pairing of Natural and Simulated Start and Internal Sequences

Self-pairing matrices were formed for sequences in libraries of true protein start sequences (A), internal sequences (B), random-nucleotide simulated start sequences (C), random-nucleotide simulated internal sequences (D), random-trinucleotide simulated start sequences (E), and random-trinucleotide simulated internal sequences (F), as described in Methods. The simulated sequences had, at each position within the sequence, the same nucleotide or trinucleotide distribution as was found in the corresponding real sequence library. For each group, the matrices were superimposed and summed and the composite matrix was smoothed. The results were plotted as described in the legend to Figure 1, with the maximum density (49 black dots) corresponding to 15 sequences capable of pairing at a cell.

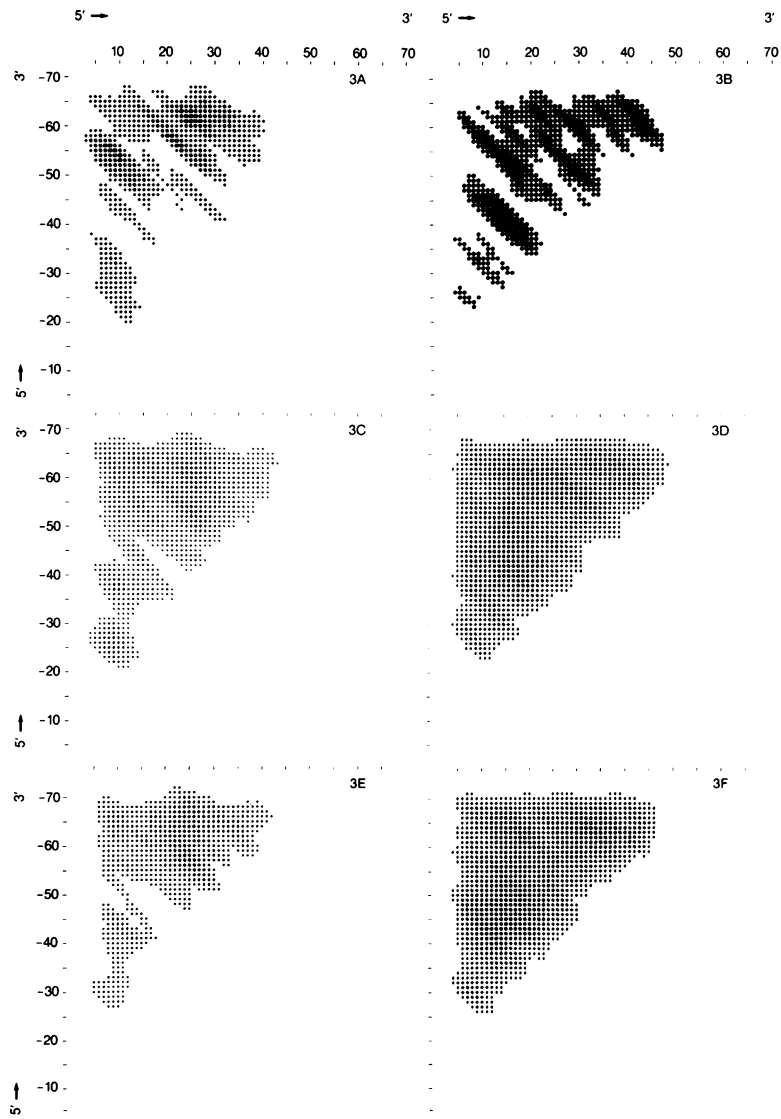

Figure 3. Regions of Elevated Pairing Potential within Natural and Simulated Start and Internal Sequences

Composite density matrices were generated as described in Figure 2 for protein start sequences (A), internal sequences (B), random-nucleotide simulated start sequences (C), random-nucleotide simulated internal sequences (D), random-trinucleotide simulated start sequences (E), and random-trinucleotide simulated internal sequences (F). All cells with densities beneath 50% of maximum for each matrix were then set to 0. The results were plotted as in Figure 1(B). This procedure has the effect of showing prominent densities (if any) in isolation.

separated by approximately six nucleotides without apparent interruption (Figures 2B and 3B).

Randomization of the start sequences (Figures 2C and 3C), and of the internal sequences (Figures 2D and 3D), where the distributions shown in Table 1 were preserved during the randomization process, altered both the overall densities and their spatial distributions.

The relative effect of this randomization was to increase density from 0.400 pairings per nucleotide to 0.432, an increase of 8%. The density of the internal sequences was 0.791, which is reduced 12% by randomization, to 0.698. A t test shows that the effect of randomization on the internal sequences is significant at the 0.01 level; however, the increase in density of the randomized starts is not significant.

Randomizations with preservation of the codon distributions were also performed. Figures 2E and 3E show the effect on the start sequences, where the distributions shown in Table 2 were preserved during the randomization process, and Figures 2F and 3F show what happened to the internal sequences when a similar process was applied.

Preserving the codon distributions led to greater changes in overall density than did preserving the distributions of individual nucleotides. The randomized starts had average density 0.490, an significant 22% increase, while a decrease to 0.637 (19%) was noted for the internal sequences.

The internal sequences were 98% more dense than the start sequences in the initial analysis. This difference in density between the start and internal sequences is reduced to 62% after randomization of nucleotides and to only 56% after randomization of codons. It should be obvious from these data that the different nucleotide compositions of the start and internal regions do account in part for the overall differences in density; otherwise randomization should obliterate these differences and not merely reduce them. Nevertheless, the fact that randomization has a significant effect clearly indicates that sequence does play a role as well.

Figures 2C and 2D show that the prominent pairings of the two sets of sequences are differently distributed. A significant effect of the nucleotide composition can be seen in the randomized start sequences, where a definite lack of pairing between the purine-rich Shine-Dalgarno region and the area immediately surrounding the start codon shows as a notch on the hypotenuse of the matrix. The bias this produces against structure in this area is also evident in the non-randomized starts (Figure 3A), where the start codon and proximal 3' region (approximately 15 bases) remain relatively unpaired.

In contrast, an almost-uniform, near-periodic pattern of increased pairing density, with a period of roughly six nucleotides, is clearly apparent among the internal sequences. This periodicity is obliterated by randomization.

DISCUSSION

Influence of mRNA secondary structure on translational initiation has been assumed for many years. Particularly clear are examples where the secondary structure masks initiation sites, *e.g.* AUG in the case of R17 or MS2 replicase, or Shine-Dalgarno signals as in the case of the *lam* B protein of *E. coli* (30,31,23). Various treatments that disrupt secondary structure tend to increase the ability of ribosomes to recognize correct (and a few incorrect) sites (32,33,34). Munson *et al.* (35) have observed that mutations that disrupt the secondary structure of the ribosome-binding site increase *lac* Z expression. The effects of many polar mutations on translation can be largely explained if the secondary structure of the mRNA masks the signals needed to initiate protein synthesis (30,32).

Casual inspection has failed to reveal a common secondary structure among the sequences (for review see 1). However, the density maps presented here reveal what appear to be some interesting common features.

The rules used in establishing local mRNA pairing potential (see Methods) were somewhat simplistic, but are adequate for enumeration, ranking and classification of local structures. Calculating more exact energetic data would only be justifiable in a recursive calculation of "best" secondary structure - a concept which is essentially meaningless when one considers local regions in isolation from the rest of the sequence. It is important to emphasize, too, that the measurement of "pairing potential" in a local region does not reflect the actual structure or structures to which that region is constrained *in situ*, but merely estimates roughly the degree to which the region is likely to be exposed or masked due to local interactions. When the pairing potential within a given region is high, the likelihood of one or more of the possible interactions actually participating in the biological structure is high also. Though the relationship is not linear as density increases, so that summing possible interactions carries a risk of overestimating the actual probability of interaction in areas of high potential, it should be noted that the present data yield a maximum of less than one pairing per nucleotide on average even in the densest map; the degree of overestimation should therefore be small.

The choice of a window of 71 nucleotides was justified by genetic and statistical evidence indicating that the most important 5' information lies within 35 nucleotides of the start codon (20; also see reviews 1,2). In biochemical experiments, it has been shown that the 35 AUG-proximal nucleotides are needed to initiate synthesis of MS2 replicase (31). Similarly, in experiments where the λ cro gene was placed under the control of the lac promoter, marked differences in λ cro expression were observed as a result of sequence modifications about 25 bases from the start AUG (13). Fourfold better expression of gal E was observed from transcripts that had 31 rather than 28 bases 5' to AUG (36). In contrast, binding of ribosomes to the start of MS2 replicase is not affected by deletion of bases 3' to AUG (31). Mutations downstream from AUG have limited effects on initiation of the T4 rIIB protein (1). One notable exception is known to us: a mutation in the start codon of T7 protein 0.3 is suppressed by an alteration 64 bases 3' to the mutated start (6, and Dunn, personal communication). But in the bulk of cases the 71-nucleotide window should be long enough to detect broad structural features recognized by ribosomes on either side of the initiation triplet.

As mentioned in Results, significant differences in distribution and overall pairing potential were found when start sites and regions surrounding internal AUGs were compared. It seems reasonable to conjecture that the denser structure associated with internal AUGs may play a part in preventing false starts, and conversely that ribosomes readily gain access to the more-open surroundings of the initiator codons. To confirm this, another study is needed in which the entire spectrum of translation-initiation efficiency (from untranslated internal AUGs through poor initiators to very efficient initiators) is examined for correlation with potential accessibility.

The results we obtained are likely to be due partly to the different base compositions of the start and coding regions (20) and partly to secondary or tertiary structural features of the different mRNAs. For the start sequences, randomization did reveal an interesting effect of base composition involving lack of pairing between the region of the Shine-Dalgarno domain and the bases 3'-proximal to the start codon. Except for this, randomization of both the start and the internal sequences sharply altered the density patterns, which would not be expected if the different statistical distributions of bases in starts and internals were solely responsible for producing the observed structures.

Analyses of individual sequences for potential base-pairing with the

updated pairing rules of Papanicolau et al. (37) confirmed that most start regions tend to have bases 5' to AUG that are free of strong secondary-structural constraints. Yet we observe that many of the start sequences have the Shine-Dalgarno region paired to the coding region 20 or more nucleotides 3' to the start codon. Pairing of these areas is not always deleterious to translational initiation, as has previously been documented for the coat and the replicase genes of MS2 (30,38). In addition, in the case of the lam B protein, mutations that altered the pairing of the Shine-Dalgarno region with bases 3' to AUG restored function and allowed maintenance of a weak hairpin in approximately the same location of the gene (23). Nevertheless, messengers that are well-expressed tend to have AUG and/or the Shine-Dalgarno region free of strong internal interactions (22,23,39,40,41,42). It may be significant that most of the pairings found are of low stability. Possibly an early event in translation may involve unfolding of weak secondary structure near the ribosome-binding site.

Genetic studies and statistical and biochemical data summarized here and elsewhere (1,2) indicate that the start codon and potential interaction of sequences 5' to it with 16S or 18S rRNA are probably insufficient to specify protein-synthesis start sites. In cases where the mRNA lacks an rRNA binding region, other signals may occur in the immediate vicinity of the initiation codon (43). The secondary-structure patterns we find could result in a spatial arrangement of these signals that is unique to start domains. Such features could suffice to specify the initial recognition by ribosomes of the message. Indeed it may be that the ribosome initially recognizes something as simple as the transition from the relaxed and unstructured region 5' to the start codon, to the more structured, flower-like 3' region. We note that self-complementarity is rather more likely in the coding sequences than in a random sequence of similar nucleotide distribution. This could itself explain in part the observed zone of transition.

Present addresses: *Department of Biology, University of Utah, Salt Lake City, UT 84112, USA and +Institut Pasteur, 25 et 28 rue du docteur Roux, F-75724 Paris, Cedex 15, France

REFERENCES

1. Gold,L., Pribnow,D., Schneider,T., Shinedling,S., Singer,B.S. and Stormo,G. (1981) Ann. Rev. Microbiol. 35, 365-403.
2. Steitz,J.A. (1979) In Goldberger,R.F. (ed), Biological Regulation and Development, Plenum Press, New York, Vol. I, pp. 349-399.
3. Shine,J. and Dalgarno,L. (1975) Nature (London) 254, 34-38.
4. Steitz,J.A. and Jakes,K. (1975) Proc. Natl. Acad. Sci. USA 72, 4734-4738.

5. Neilson,T., Kofoid,E.C. and Ganoza,M.C. (1980) Nucl. Acids Res. Symp. Ser. 7, 313-323.
6. Dunn,J.J., Buzash-Pollert,E. and Studier,F.W. (1978) Proc. Natl. Acad. Sci. USA 75, 2741-2745.
7. Ptashne,M., Backman,K., Humayun,M.Z., Jeffrey,A., Maurer,R., Meyer,B. and Sauer,R.T. (1976) Science 194, 156-161.
8. Beck,E., Sommer,R., Auerswald,E.A., Kurz,C., Zink,B., Osterburg,G., Shalder,H., Sugimoto,K., Sugisaki,H., Okamoto,T. and Takanami,M. (1978) Nucl. Acids Res. 5, 4495-4503.
9. Farabaugh,P.J. (1978) Nature 274, 765-769.
10. Godson,G.N., Barrell,B.G., Staden,R. and Fiddes,J.C. (1978) Nature 276, 236-247.
11. Pirrota,V. (1979) Nucl. Acids Res. 6, 1495-1508.
12. Singleton,C.K., Roeder,W.D., Bogosian,G., Somerville,R.L. and Weith,H.L. (1980) Nucl. Acids Res. 8, 1551-1560.
13. Roberts,T.M., Kacich,R. and Ptashne,M. (1979) Proc. Natl. Acad. Sci. USA 76, 760-764.
14. Ganoza,M.C., Fraser,A. and Neilson,T. (1978) Biochemistry 17, 2769-2775; Ganoza,M.C., Sullivan,P., Cunningham,C., Kofoid,E.C., Hader,P. and Neilson,T. (1982) J. Biol. Chem. 257, 8228-8232.
15. Eckhardt,H. and Luhrmann,R. (1981) Biochemistry 20, 2075-2080.
16. Taniguchi,T. and Weissmann,C. (1978) J. Mol. Biol. 118, 533-565.
17. Schmitt,M., Manderscheid,U., Kyriatsoulis,A., Brinkmann,U. and Gassen,H.G. (1980) Eur. J. Biochem. 109, 291-299.
18. Ganoza,M.C. (1977) Can. J. Biochem. 55, 257-281.
19. Atkins,J.F. (1979) Nucl. Acids Res. 7, 1035-1041.
20. Stormo,G., Schneider,T.D. and Gold,L. (1982) Nucl. Acids Res. 10, 2791-2996.
21. Bahramian,M.B. (1980) J. Theor. Biol. 84, 103-108.
22. Iserentant,D. and Fiers,W. (1980) Gene 9, 1-12.
23. Hall,M., Gabay,J., Debarbouille,M. and Schwartz,M. (1982) Nature 295, 616-618.
24. Trifonov,E.N. and Bolshoi,G. (1983) J. Mol. Biol. 169, 1-13.
25. Tinoco,I.,Jr., Borer,P.N., Dengler,B., Levine,M.D., Uhlenbeck,O.C., Crothers,D.H. and Gralla,J. (1973) Nature New Biol. 246, 40-41.
26. Ninio,J. (1979) Biochimie 61, 1133-1150.
27. Gauss,D.H. and Sprinzl,M. (1983) Nucl. Acids Res. 9, 41-53.
28. Bibb,M.J., Van Etton,R.A., Wright,C.T., Walberg,M.W. and Clayton,D.A. (1981) Cell 26, 167-180.
29. Gralla,J. and Delisi,C. (1974) Nature 276, 236-247.
30. Fiers,W., Contreras,R., Duerinck,F., Haegeman,G., Iserentant,D., Merregaert,J., Min Jou,W., Molemans,F., Raeymaekers,A., Van den Berghe,A., Volckaert,G. and Ysebaert,M. (1976) Nature (London) 260, 500-507.
31. Borisova,G.P., Volkova,T.M., Berzin,V., Rosenthal,G. and Gren,E.J. (1979) Nucl. Acids Res. 6, 1716-1774.
32. Lodish,H.F. (1970) J. Mol. Biol. 50, 689-702.
33. Kozak,M. and Nathans,D. (1972) Bact. Rev. 36, 109-134.
34. Wahba,A., Iwasaki,K., Miller,M.J., Sabol,S., Sillero,M.A.G. and Vasquez,C. (1969) Cold Spring Harbor Symp. Quant. Biol. 34, 291-299.
35. Munson,L.M., Stormo,G.D., Niece,R.L. and Reznokoff,W.S. (1984) J. Mol. Biol. 177, 663-683.
36. Queen,C.L. and Rosenberg,M. (1980) Fed. Proc. 39, 810.
37. Papanicolaou,C., Gouy,M. and Ninio,J. (1984) Nucl. Acids Res. 12, 31-44.
38. Kastelein,R.A., Berkhout,B., Overbeek,G.P. and Van Duin,J. (1983) Gene 23, 245-254.
39. Ray,P.N. and Pearson,M. (1975) Nature 253, 647-650.
40. Gheyson,D., Iserentant,D., Derom,C. and Fiers,W. (1982) Gene 17, 55-63.

41. Wood,C.R., Boss,M.A., Patel,T.P. and Emtage,J.S. (1984) Nucl. Acids Res. 12, 3937-3950.
42. Cone,K.C. and Steege,D.A. (1985) J. Mol. Biol. 186, 733-742.
43. Ganoza,M.C., Marlière,P., Kofoid,E.C. and Louis,B.G. (1985) Proc. Natl. Acad. Sci. USA 82, 4587-4591.