Volume 7 Number 7 1979 Nucleic Acids Research

Do eukaryotic mRNA 5' noncoding sequences base-pair with the 18 S ribosomal RNA 3' terminus?

Rupert De Wachter

Departement Celbiologie, Universiteit Antwerpen, Universiteitsplein 1, B-2610 Wilrijk, Belgium

Received 28 August 1979

ABSTRACT

Protein synthesis initiation on prokaryotic mRNAs involves base-pairing of a site preceding the initiation codon with the 3' terminal sequence of 16 S rRNA. It has been suggested that a similar situation may prevail in eukaryotic mRNAs. This suggestion is not based on experiments, but on observation of complementarities between mRNA 5' noncoding sequences and a conserved sequence near the 18 S rRNA 3' terminus. The hypothesis can be evaluated by comparing the number of potential binding sites found in the 5' noncoding sequences with the number of such sites expected to occur by chance. A method for computing this number is presented. The 5' noncoding sequences contain more binding sites than expected for a random RNA chain, but the same is true for 3' noncoding sequences. The effect can be traced to a clustering of purines and pyrimidines, common to noncoding sequences. In conclusion, a close inspection of the available mRNA sequences does not reveal any indication of a specific base-pairing ability between their 5' noncoding segments and the 18 S rRNA 3' terminus.

INTRODUCTION

There are currently two different hypotheses on the path by which the eukaryotic ribosome reaches the initiation codon in mRNAs. According to one proposal[1], ribosome binding may involve base pairing between the 5' noncoding sequence of the mRNA and the highly conserved sequence UGCGGAAGGAU near the 3' terminus of 18 S rRNA. Most 5' noncoding sequences contain sites, sometimes of considerable length[2], that could base-pair with the ribosomal sequence. On the other hand, the distance between the initiation codons and the putative binding sites is much more variable than in prokaryotic mRNAs. In the latter case, the Shine-Dalgarno hypothesis[3] on base-pairing between the mRNA and the 16 S rRNA 3' terminus has been experimentally proven[4] by ribosome binding studies, but no such proof is available yet for eukaryotes. An alternative proposal[5] for ribosome binding holds that the 40 S subunit binds at the cap site, scans the leader sequence until it encounters the initiation codon, which is usually the first AUG, and is then joined by the

60 S subunit.

If we want to make full use of the available sequence information in judging the relative merits of the "base pairing" and the "scanning" hypotheses, the following question demands an answer : do the known mRNA leader sequences contain appreciably more or longer sites complementary to the 18 S RNA 3' terminus than expected to occur by chance ? If the answer is no, then there is no reason to assume a special role for this particular ribosomal RNA sequence in the first place.

METHODS

Number of "ribosome binding sites" expected in a random RNA.

The number of "binding sites" expected to occur by chance in a random RNA sequence of given base composition can be computed as follows. The complement, allowing for G·U pairing, of the 18 S 3' terminal binding site postulated by Hagenbüchle et al.[1] is :

$$\begin{matrix} A \\ G \end{matrix} U \begin{matrix} U \\ C \end{matrix} \begin{matrix} U \\ C \end{matrix} U U \begin{matrix} U \\ C \end{matrix} \begin{matrix} U \\ C \end{matrix} G \begin{matrix} U \\ C \end{matrix} \begin{matrix} A \\ G \end{matrix} \qquad (1)$$

Let us call B_2 the set of 11 different dinucleotides distinguishable in this composite sequence :

AU, GU, UU, UC, CU, CC, UG, CG, GC, UA, CA

In this way we can consider 11 sets, B_1 to B_{11}, B_n being the set of all different oligonucleotides of length n that occur in sequence (1). On the other hand we consider a random RNA sequence with chain length N and fractional base composition p_U, p_C, p_A, p_G, such that

$$p_U + p_C + p_A + p_G = 1$$

This RNA can be regarded as a set, R_n, of N-n+1 overlapping oligonucleotides of length n. Again we can consider sets R_1 to R_{11}. The random RNA is said to show a "binding site" of length n for each oligonucleotide of set R_n that also belongs to set B_n, and is not part of a longer oligonucleotide R_m belonging to set B_m, where $m > n$. The mean number of "binding sites" of length n expected in the random RNA, μ_n, is given by the expression :

$$\mu_n = (N-n+1) \sum_{i=1}^{b} q_o \cdot q_{n+1} \cdot \prod_{j=1}^{n} p_j \qquad (2)$$

where the following symbols are used :

p_j the probability that the j^{th} base of an oligonucleotide of set R_n coincides with the j^{th} base of the i^{th} member of set B_n.

$\prod_{j=1}^{n} p_j$ the product of probabilities p_j for j=1 to j=n.

q_o the probability that the base preceding an oligonucleotide of set R_n in the random RNA differs from the base preceding the i^{th} member of set B_n in sequence (1).

q_{n+1} the probability that the base following an oligonucleotide of set R_n in the random RNA differs from the base following the i^{th} member of set B_n in sequence (1).

b the number of oligonucleotides in set B_n.

The derivation of equation (2) can be clarified by the following example. Consider the sequence UCUU, which belongs to B_4, the set of 37 different tetranucleotides found in sequence (1). The probability of finding a U at a given position of the random RNA is equal to p_U, the fraction of U's it contains. The probability of finding a C next to it is equal to p_C, etc.. The probability that a tetranucleotide from the random RNA, i.e. belonging to set R_4, happens to be UCUU is then

$$p_U \cdot p_C \cdot p_U \cdot p_U$$

The probability that the tetranucleotide is UCUU, and that the identity with sequence (1) does not extend further to the left or to the right, is

$$p_C \cdot p_U \cdot p_C \cdot p_U \cdot p_U \cdot (p_A + p_G) \qquad (3)$$

The first and the last factor arise as follows : UCUU appears in sequence (1) at positions 2-5 and 3-6. If a member of R_4 is UCUU, but is preceded by A, G, or U, or followed by U or C, then a "binding site" of at least length 5 will be scored. Only if C precedes and A or G follow is the binding site limited to 4 bases. The probability that a particular tetranucleotide from set R_4 coincides with any member of set B_4 is obtained by summing expressions such as (3) for all 37 members of B_4. This sum is then multiplied by N-3, the number of tetranucleotides in set R_4, to obtain μ_4, the mean number of binding sites of length 4 in a random RNA :

$$\mu_4 = (N-3) \sum_{i=1}^{37} q_o \cdot p_1 \cdot p_2 \cdot p_3 \cdot p_4 \cdot q_5$$

which is equation (2) for the case n=4. The particular values p_U, p_C etc. used in expression (3) for the example UCUU have been substituted by the gen-

eral symbols p_j for base identity, and q_j for base difference, at position j of the oligonucleotide.

Counting the binding sites in actual RNA sequences.

The values of μ_n can now be compared to the number of sites, i.e. oligonucleotides belonging to sequence (1), actually found in putative ribosome binding regions. How the sites are scored is demonstrated below with the sequence comprised between the cap structure $m^7GpppGm$ and the initiation codon AUG in reovirus messenger m52 [5] :

2 2 0 4 5 0 0 0 3 0 0 1 4 0 3 0 2 0 2 6 0 0 0 0 0 1 1 1 (4)
C U A A U C U G C U G A C C G U U A C U C U G C A A A G

Binding sites of length 4 or longer are underlined. A complete list of sites appears above the sequence, each figure indicating the length of the site starting at this position. A zero above a base means that it is not the first base of any binding site, although it may belong to one. As an example, position 8 from the right is scored 0 although the pentanucleotide CUGCA starting there appears in sequence (1). This is because the preceding base U is the first of the hexanucleotide UCUGCA found in sequence (1), so the C is only the second base of a site. The advantage of scoring binding sites at all positions, even if they have "length zero", is that the total number of sites equals the RNA chain length. This facilitates the comparison of the site length distribution found in an actual RNA with the one predicted by equation (2) for a random RNA of the same base composition. As for the number of sites of length 0 expected in the random RNA, this is obtained by subtracting the expected number of sites of length 1 to 11 (equation 2) from the chain length :

$$\mu_o = N - \sum_{i=1}^{11} \mu_i \qquad (5)$$

The fact that sites of less than 4 bases, and even some of the longer ones, cannot form stable base-paired structures with sequence (1) does not matter here. The only point of the operation is to compare the number of sites of each length found with the value forecast by equation (2).

Sequences examined for number of binding sites.

Binding sites of length 1 to 11, belonging to sequence (1), were scored in the 24 eukaryotic mRNA 5' noncoding sequences listed in Table 1. The total length is 1028 nucleotides, and the base composition is $U_{287}\ C_{258}\ A_{299}\ G_{184}$. As a control, the same operation was performed on a

computer-generated random sequence of identical length and base composition.

Table 1 also lists 43 prokaryotic ribosome binding regions with a total length of 860 nucleotides and base composition $U_{239}\ C_{140}\ A_{288}\ G_{193}$. In this case, binding sites were scored belonging to the composite sequence

$$U \begin{matrix} A \\ G \end{matrix} \begin{matrix} A \\ G \end{matrix} G\ G \begin{matrix} A \\ G \end{matrix} G\ G\ U \qquad (6)$$

which is the complement of the E. coli 16 S rRNA 3' terminus. The calculation of the expected number of sites in a random RNA is analogous to the case discussed for eukaryotic sequences.

For both eukaryotes and prokaryotes, as many data as possible were included, with two restrictions. When sequences are known for corresponding genes in closely related species, such as the mammals, phages MS2 and R17, ϕX174 and G4, only data from one species were used in order not to bias the result by including an appreciable fraction of partly homologous sequences. Eukaryotic sequences determined on the DNA level, for which the mRNA splicing pattern is not or incompletely known, were not considered.

RESULTS AND DISCUSSION

The comparison between the number of binding sites scored in the examined sequences and the number expected for a random RNA of the same length and composition is made in Table 2. In the eukaryotic mRNA 5' noncoding sequences, there is an excess of binding sites from length 5 onward, a deficit of sites of length 2 to 4, and again an excess of sites of length 1. A chi square test on the fit of the site length distributions found and expected yields a value of 33.25. Only one in about 50.000 random sequences of corresponding length and composition would yield such a high χ^2 value, which means that the fit is very bad. On the contrary, the site length distribution for a computer-generated random RNA sequence fits very well with the expected one, which means that the calculations adequately predict the random distribution. The prokaryotic ribosome binding regions give about the same excess-deficit pattern of binding site length distribution, but with more extreme deviations, as the eukaryotic sequences.

The binding site length distribution observed in eukaryotic mRNA 5' noncoding sequences could at first sight be explained by an evolution favouring the appearance of binding sites with a minimum length of 5 bases. Due to the composition of sequence (1), formation of long binding sites means appearance of long pyrimidine clusters. This also means that single pyrimidines and short pyrimidine clusters must be in deficit, and purine clusters in excess,

Table 1. Sequences analyzed for complementarity with the 3' terminus of the small ribosomal subunit RNA.

A. Eukaryotic mRNA 5' noncoding sequences.

Genome	mRNA	length of sequence examined	ref.
Chicken	ovalbumin	63	7
Man	α-globin	36	8
	β-globin	49	8
	γ-globin	52	11
Reovirus	s54,s45,s46,m52,m44,m30	30,26,17,28,17,12	5
Vesicular stomatitis virus	N,NS,L,M,G	12,9,9,20,14	12
Rous sarcoma virus	genome	80	13
Simian virus 40	early	61	14 †
	VP1	243	15
Brome mosaic virus	RNA4	8	16
Alfalfa mosaic virus	RNA4	35	17
Tobacco mosaic virus	genome	67	18
Satellite tobacco necrosis virus	genome	29	19
Turnip yellow mosaic virus	RNA4, genome	18,93	20

B. Prokaryotic mRNA ribosome binding regions.

Genome	gene	length of sequence examined	ref.
Escherichia coli	*lac* I	20	21
	lac Z	20	22
	gal E	20	23
	ara B, *ara* C	20 each	24
	trp E	20	25
	trp A	20	26
	phe A, phe leader peptide	20 each	27
	his G	20	28
	bio A, *bio* B	20 each	29
	rpl N, *rps* L	20 each	30
Plasmid pBR322	amp^r	20	31
Bacteriophage φX174	A,B,K,C,D,E,J,F,G,H	20 each	32
" MS2	A, coat, replicase	20 each	33
" Qβ	A, coat, replicase	20 each	33
" fd	I,II,III,IV,V,VI,VII,VIII	20 each	34
" λ	cI	20	35
	cro,cII,O	20 each	36

In the 24 eukaryotic mRNAs, the sequence comprised between the cap structure m^7GpppR and the initiating AUG was examined. The first purine R is a methylated base in some messengers and was omitted in all cases for uniformity. All 5' noncoding sequences are complete except VSV mRNAs M and G, for which only the AUG-adjacent end is known. In the 43 prokaryotic mRNA ribosome binding sites, the 20 bases preceding the initiation codon were examined.

† The SV40 early mRNA leader sequence extends from position 18 to 79 of the early region[14] (W. Fiers, personal communication)

Table 2. Binding sites expected and found in eukaryotic and prokaryotic mRNA initiation regions.

length of site n	eukaryotic mRNA 5' noncoding sequences		random sequence with composition of eukaryotic 5' noncoding sequences		prokaryotic mRNA initiation regions	
	expected	found	expected	found	expected	found
0	455.82	439	437.54	422	474.50	452
1	141.88	187	141.88	149	101.83	164
2	237.41	223	242.85	240	175.99	129
3	99.66	90	104.34	119	77.95	62
4	59.03	42	63.34	64	20.84	29
5	23.28	26	25.58	25	6.11	15
6	7.25	12	8.17	7	2.05	6
7	2.27	6	2.63	2	0.61	2
8	1.04	2	1.23	0	0.10	0
9	0.26	1	0.32	0	0.02	1
10	0.07	0	0.08	0		
11	0.03	0	0.04	0		
	$\chi^2 = 33.25$ ($\chi^2_{0.95} = 12.6$)		$\chi^2 = 5.35$ ($\chi^2_{0.95} = 12.6$)		$\chi^2 = 84.95$ ($\chi^2_{0.95} = 12.6$)	

Sites complementary with the 18 S rRNA 3' terminus were scored in 24 eukaryotic mRNA 5' noncoding sequences and a random sequence, and sites complementary with the 16 S rRNA 3' terminus in 43 prokaryotic mRNA initiation regions, all listed in Table 1. The random sequence was computer-generated and had the same length and base composition as the combined eukaryotic mRNA 5' ends. Hence the probability of finding a binding site of length n in any position is the same in both cases but the mean number of sites expected (equations 2 and 5) is slightly different because there are less overlapping oligonucleotides in 24 separate sequences than in a single sequence of the same overall length. χ^2 was computed after combining terms corresponding to length 6 and larger. The distribution then has 6 degrees of freedom, and in 95 % of tests on random sequences the χ^2 value will not exceed 12.6.

relative to a random sequence of the same base composition. This in turn results in a depletion of short binding sites of length 2 to 4. Further, as demonstrated in example (4), purine clusters are scored as binding sites of length 1, which hence must be in excess. An analogous mechanism may be invoked for the prokaryotic initiation regions, since sequence (6) is a purine

cluster.

A different light was thrown on the problem when sites belonging to sequence (1) were scored in three places where they are not expected in excess : the A protein cistron of phage MS2 [6] (1182 bases), the ovalbumin mRNA coding sequence[7] (1161 bases) and the combined 3' noncoding sequences of ovalbumin and human α-, β-, and γ-globin mRNAs[7, 8, 9] (962 bases). In the coding sequences, the site length distribution matched the predicted random distribution satisfactorily. However, the eukaryotic 3' noncoding sequences showed the same excess-deficit pattern as the 5' noncoding sequences. In Fig. 1, the latter distribution is compared with those discussed previously.

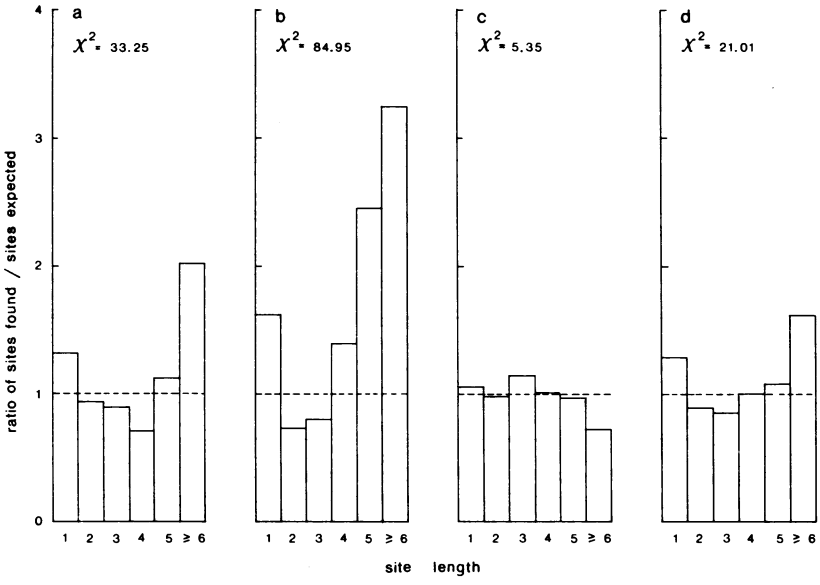

Fig. 1. Ratio of found to expected number of sites in several sequences.
The expected number of sites was computed from equation (2) and sites were scored in the sequences as explained in the text. The ratio of these two results is plotted as a function of site length.
(a) 24 eukaryotic mRNA 5' noncoding sequences, total length 1028 bases.
(b) 43 prokaryotic mRNA ribosome binding regions, total length 860 bases.
(c) random sequence with the same length and base composition as the combined sequences from (a).
(d) 3' noncoding sequences of 4 eukaryotic mRNAs, chicken ovalbumin and human α-, β-, and γ-globin[7,8,9], total length 962 bases.
18 S rRNA-complementary sites are plotted in (a), (c), and (d), 16 S rRNA-complementary sites in (b). The figures used in plots (a), (b) and (c) appear in Table 2, but binding sites of length 0 are not plotted. The χ^2 value for 6 degrees of freedom that will not be exceeded by 95 % of random sequences is 12.6.

It has been observed[7], and quantitatively confirmed[10], that clusters of identical bases tend to occur in eukaryotic mRNA 3' noncoding sequences. This phenomenon, as well as purine and pyrimidine clustering, is also found in the 5' noncoding sequences. It automatically results in an excess of all long pyrimidine sequences, among which are the sets belonging to sequence (1). Thus, although the eukaryotic mRNA 5' noncoding sequences do contain a definite excess of sites complementary to the 18 S RNA 3' terminus, this probably does not reflect a functional relation between the two sequences. Indeed, all pyrimidine sequences, not only those complementary to the 18 S RNA 3' end, are in excess. It could be argued that the occurence of pyrimidine-purine clustering, although not limited to 5' noncoding sequences, still may favour interaction of these sequences with the 18 S RNA 3' terminus. However, such an interaction would then be favoured with any purine or pyrimidine block, not just the particular purine block present near the 18 S RNA 3' terminus.

In conclusion, the analysis of sequence data presented here makes it very doubtful that the conserved sequence UGCGGAAGGAU near the 18 S rRNA 3' terminus would play a role in mRNA binding. The pyrimidine-purine clustering, found in both 5' and 3' noncoding sequences of eukaryotic mRNAs, may very well have a functional or evolutionary significance entirely foreign to ribosome binding.

ACKNOWLEDGMENT

I thank Dr. J. Haezendonck for discussions on the mathematical aspects of the problem.

REFERENCES

1. Hagenbüchle, O., Santer, M., Steitz, J.A., Mans, R.J. (1978) Cell 13, 551-563.
2. Ziff, E.B., Evans, R.M. (1978) Cell 15, 1463-1475.
3. Shine, J., Dalgarno, L. (1975) Nature 254, 34-38.
4. Steitz, J.A., Jakes, K. (1975) Proc. Nat. Acad. Sci. U.S. 72, 4734-4738.
5. Kozak, M. (1978) Cell 15, 1109-1123.
6. Fiers, W., Contreras, R., Dueringck, F., Haegeman, G., Merregaert, J., Min Jou, W., Raeymakers, A., Volckaert, G., Ysebaert, M., Van de Kerckhove, J., Nolf, E., Van Montagu, M. (1975) Nature 256, 273-278.
7. McReynolds, L., O'Malley, B.W., Nisbet, A.D., Fothergill, J.E., Givol, D., Fields, S., Robertson, M., Brownlee, G.G. (1978) Nature 273, 723-728.
8. Baralle, F.E. (1977) Cell 12, 1085-1095.
9. Poon, R., Kan, Y.W., Boyer H.W. (1978) Nucl. Acids Res. 5, 4625-4630.
10. De Wachter R. (1979) Arch. Int. Physiol. Biochim., 87, 403-404.

11. Chang, J.C., Poon, R., Neumann, K.H., Kan, Y.W. (1978) Nucl. Acids Res. 5, 3515-3522.
12. Rose, J.K. (1978) Cell 14, 345-353.
13. Haseltine, W.A., Maxam, A.M., Gilbert, W. (1977) Proc. Nat. Acad. Sci. U.S. 74, 989-993.
14. Fiers, W., Contreras, R., Haegeman, G., Rogiers, R., Van de Voorde, A., Van Heuverswyn, H., Van Herreweghe, J., Volckaert, G., Ysebaert, M. (1978) Nature 273, 113-120.
15. Ghosh, P.K., Reddy, V.B., Swinscoe, J., Choudary, P.V., Lebowitz, P., Weissman, S. (1978) J. Biol. Chem. 253, 3643-3647.
16. Dasgupta, R., Shih, D.S., Saris, C., Kaesberg, P. (1975) Nature 256, 624-628.
17. Koper-Zwarthoff, E.C., Lockard, R.E., Alzner-de Weerd, B., Raj Bhandary, U.L., Bol, J.F. (1977) Proc. Nat. Acad. Sci. U.S. 74, 5504-5508.
18. Richards K., Guilley, H., Jonard, G., Hirth, L. (1978) Eur. J. Biochem. 84, 513-519.
19. Leung, D.W., Browning, K.S., Heckmann, J.E., Raj Bhandary, U.L., Clark, J.M. (1979) Biochemistry 18, 1361-1366.
20. Briand, J.-P., Keith G., Guilley, H. (1978) Proc. Nat. Acad. Sci. U.S. 75, 3168-3172.
21. Farabaugh, P.J. (1978) Nature 274, 765-769.
22. Maizels, N. (1974) Nature 249, 647-649.
23. Musso, R.E., de Crombrugghe, B., Pastan, I., Sklar, J., Yot, P., Weissman, S. (1974) Proc. Nat. Acad. Sci. U.S. 71, 4940-4944.
24. Smith, B.R., Schleif, R. (1978) J. Biol. Chem. 253, 6931-6933.
25. Bennett, G.N., Schweingruber, M.E., Brown, K.D., Squires, C., Yanofsky, C. (1978) J. Mol. Biol. 121, 113-137.
26. Platt, T., Yanofsky, C. (1975) Proc. Nat. Acad. Sci. U.S. 72, 2399-2403.
27. Zurawski G., Brown, K., Killingly, D., Yanofsky, C. (1978) Proc. Nat. Acad. Sci. U.S. 75, 4271-4275.
28. Barnes, W.M. (1978) Proc. Nat. Acad. Sci. U.S. 75, 4281-4285.
29. Otsuka, A., Abelson, J. (1978) Nature 276, 689-694.
30. Post L.E., Arfsten, A.E., Reusser, F., Nomura, M. (1978) Cell 15, 215-229.
31. Sutcliffe, J.G. (1978) Proc. Nat. Acad. Sci. U.S. 75, 3737-3741.
32. Sanger, F., Coulson, A.R., Friedmann, T., Air, G.M., Barrell, B.G., Brown, N.L., Fiddes, J.C., Hutchinson, C.A., Slocombe, P.M., Smith, M. (1978) J. Mol. Biol. 125, 225-246.
33. Barrel, B.G., Clark, B.F.C. (1974) Handbook of nucleic acid sequences, pp. 75-93, Joynson-Bruyvers, Oxford.
34. Beck, E., Sommer, R., Auerswald, E.A., Zink, B., Osterberg, G., Schaller, H., Sugimoto, K., Sugisaki, H., Okamoto, T., Takanami, M. (1978) Nucl. Acids Res. 5, 4495-4503.
35. Humayun, Z., Jeffrey, A., Ptashne, M. (1977) J. Mol. Biol. 112, 265-277.
36. Schwartz, E., Scherer, G., Hobom, G., Kössel, H. (1978) Nature 272, 410-414.