The template RNAs of RNA polymerases can have compact secondary structure, formed by long double helices with partial violations of the complementarity

V.V. Solovyov, A.A. Zharkikh, N.A. Kolchanov and V.A. Ratner

Institute of Cytology and Genetics of the USSR Academy of Sciences, Siberian Department, Novosibirsk 630090, USSR

Received 23 September 1983

A new method of contextual analysis was used to search the long non-random inverted repeats and the complementary palindromes in the genes of *E. coli* and T7 RNA polymerases. These genes were found to contain from 25% to 50% of all the nucleotides involved in such helices. The 5′- and 3′-ends of mRNA can be protected by neighbouring double helices from the nuclease attack. Some double helices are competing and very similar to the attenuator of *E. coli* trp-operon.

RNA polymerase *Template RNA* *Complementary palindrome* *mRNA secondary structure*

1. INTRODUCTION

There are several different physical methods for the calculation of RNA secondary structures, based on the search of free energy global minimum for isolated RNA in solution, using the energy parameters [1–5]. But the problem of discovering the functionally important forms of the RNA secondary structure has not yet been solved. First, the parameters used contain some errors, and as the total error increases with the length of molecules they become decisive for RNA molecules as long as thousands of nucleotides. Furthermore, the functional forms of RNA secondary and tertiary structures, being stabilized through interaction with the proteins, might not correspond to a free energy global minimum of RNA [6]; for long RNA molecules (more than 1000–2000 bases) this minimum could not be realized because of the enormous number of possible conformation states [3]. Finally, the correct search of free energy global minimum for long RNAs leads to such enormously long computer time, e.g., thousands of hours for modern computers [7].

Therefore we have developed a completely new approach, based on the method of contextual analysis of genetic texts [8] – the search of non-random inverted repeats and complementary palindromes, corresponding to double-helical regions of RNA. These methods estimate combinatoric probabilities of such structures.

2. METHOD

A definite nucleotide sequence of length N has inverted repeats and complementary palindromes of size l with k violations (fig.1). Their real number $n(l,k)$ found by the elaborated procedure of contextual analysis is compared with expected values $\bar{n}(l,k)$ for a random sequence of length N with the same frequencies of nucleotides q_A, q_U, q_G, q_C. For inverted repeats our estimation of the expected value is

$$\bar{n}(l,k) = \phi(N,l)\, C_l^k p^{(l-k)}(1-p)^k = \phi(N,l)P(l,k) \quad (1)$$

where $\phi = (N-2l+1)(N-2l+2)/2$ is the number of possible dispositions of two segments of length l into a sequence of length N; $P(l,k)$ is the probability of two randomly chosen segments of

Published by Elsevier Science Publishers B.V.
00145793/84/$3.00

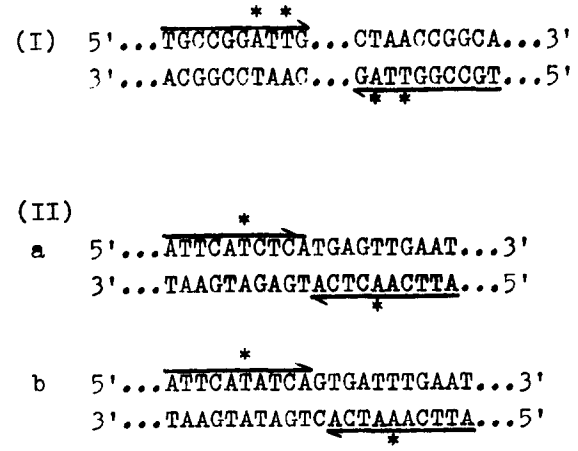

Fig.1. The inverted repeat and two types of complementary palindromes in DNA. (I) Inverted repeat of 10 base pairs with 2 violations of complementarity. Locations of repeated segments are indicated by arrows, the non-complementary pairs by asterisks. (II) Two types of complementary palindromes (of even and uneven length) containing 20 and 21 nucleotide pairs, respectively. The repeated segments are indicated by arrows.

length l to be complementary by $l-k$ positions and non-complementary by k positions; $p = 2(q_A q_U + q_G q_C)$ is the probability of complementation of two nucleotides; C_l^k is the number of possible dispositions of k violations of complementarity into the inverted repeats of length l.

If the number of inverted repeats actually found is $n(l,k) > \bar{n}(l,k)$, it is necessary to estimate the significance of this difference. Based on a binomial distribution, the probability of finding n inverted repeats of length l with k violations of complementarity was developed as

$$Q(n) = C_\phi^n P^n(l,k)(1-P(l,k))^{\phi-n}$$

This formula was used for estimation of the 95% confidence limit $n_o(l,k)$. Then, if $n(l,k) \geqslant n_o(l,k)$, the difference was significant and the corresponding inverted repeats were designated as non-random [8]. The same estimations were found for complementary palindromes.

It was very important to select combinations of l and k such that the expected value $\bar{n}(l,k)$ was less than or about 1, because only then could the estimations of $Q(n)$, based on the binomial distribution, be correct.

So, the proposed approach permits us to determine the non-random RNA secondary structure, composed of non-perfect double helices, based on the inverted repeats and complementary palindromes. The principal differences of this approach from those proposed earlier [1–4] are:

(i) it does not use the energy parameters;
(ii) it does not suppose correspondence between the functional forms of the secondary structure and the free energy global minimum;
(iii) the programs of the computer counts require a short computer time, and, as a consequence;
(iv) it is possible to determine the secondary structure in long nucleotide sequences (up to 10^5 bases).

3. RESULTS AND DISCUSSION

This method was applied to genes of *E. coli* RNA polymerase (β-, β'-, σ-subunits) [9–11] and of T7 RNA polymerase [12]. We describe in detail the procedure for manifestation of the non-random inverted repeats in mRNA of the β-subunit.

Firstly, using eq.1, the matrix of $\bar{n}(l,k)$ (expected numbers of inverted repeats) was constructed. The lengths of repeats were chosen as $l = 9,\ldots,50$ with violations of complementarity $k = 0,\ldots,22$. Table 1 shows the part of this matrix with $l = 9,\ldots,25$ and $k = 0,\ldots,8$. In accordance with the described procedure, the groups of inverted repeats with l and k were tested, corresponding to $\bar{n}(l,k)$ which is less than or about 1. All these values are shown below the solid line in table 1.

Eight of all tested groups of inverted repeats were shown to have $n(l,k) \geqslant n_o(l,k)$. Table 2 lists the characteristics of these 8 groups of non-random inverted repeats. It should be emphasized that some short non-random inverted repeats can be a part of longer ones. Fig.2 shows that a non-random inverted repeat from the gene of the β-subunit of *E. coli* RNA polymerase, of length 23 with 6 violations of complementarity, is a part of a longer repeat of length 25 with 7 violations. The short non-random inverted repeats evidently do not express additional information about the secondary structure of mRNA as compared to longer ones which include them; these short repeats were excluded from consideration.

The above 8 groups contain 15 inverted repeats

Table 1

The fragment of the matrix of expected numbers of inverted repeats of length l and k violations of complementarity for the RNA polymerase β-subunit gene in *E. coli*

	k								
l	0	1	2	3	4	5	6	7	8
9	37.71	–	–	–	–	–	–	–	–
10	9.63	–	–	–	–	–	–	–	–
11	2.46	44.33	–	–	–	–	–	–	–
12	0.63	13.22	–	–	–	–	–	–	–
13	0.16	3.86	40.53	–	–	–	–	–	–
14	0.04	1.11	13.32	93.21	–	–	–	–	–
15	0.01	0.32	4.25	34.03	–	–	–	–	–
16	*	0.09	1.33	11.96	71.74	–	–	–	–
17	*	0.02	0.41	4.07	27.50	–	–	–	–
18	*	0.01	0.12	1.35	10.15	54.81	–	–	–
19	*	*	0.04	0.44	3.63	21.79	98.04	–	–
20	*	*	0.01	0.14	1.26	8.35	41.76	–	–
21	*	*	*	0.04	0.43	3.10	17.07	73.17	–
22	*	*	*	0.01	0.14	1.12	6.74	31.79	–
23	*	*	*	*	0.05	0.40	2.60	13.30	54.83
24	*	*	*	*	0.01	0.14	0.97	5.38	24.20
25	*	*	*	*	*	0.05	0.35	2.11	10.31

–, Values > 100; *, values < 0.01

Table 2

The properties of groups of non-random repeats for the gene of the β-subunit of *E. coli* RNA polymerase

Number	Length of repeat	Number of non-complementary pairs	Number of repeats		
			Expected	95% border of confidence interval	Real
1	16	2	1.33	4	4
2	23	6	2.6	7	6
3	25	7	2.1	10	6
4	27	8	1.7	6	5
5	28	9	3.3	8	8
6	30	10	2.6	8	7
7	37	14	2.2	7	6
8	41	16	1.3	6	4

of maximum length (table 3). For genes of β'- and σ-subunits of *E. coli* RNA polymerase and T7 RNA polymerase, we found 12, 21 and 8 non-random inverted repeats of maximum length. Localization of these repeats in the gene of T7 RNA polymerase is shown in table 3.

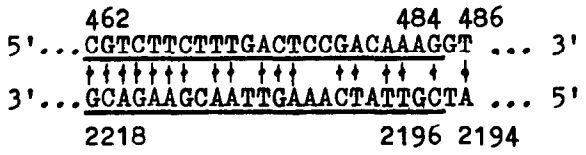

Fig.2. Inverted repeat in the gene of the β-subunit of *E. coli* RNA polymerase, of 25 base pairs with 7 non-complementary pairs, including the shorter inverted repeat of 23 base pairs with 6 non-complementary pairs.

The search of non-random complementary palindromes can be done in the same way. The maximum number of non-random complementary palindromes was found in the genes of β'- and σ-subunits of *E. coli* RNA polymerase (13 and 12, respectively; table 4); the gene of T7 RNA polymerase contains only 2 complementary palindromes and the gene of the *E. coli* β-subunit has none.

Table 3

The characteristics of the non-random inverted repeats of maximum length for genes of β-subunit of *E. coli* RNA polymerase and of phage T7 RNA polymerase

Gene	No.	Length of repeat	Number of violations of complementarity	Localization
β	1	16	2	2323–2338 3734–3749
	2	16	2	1298–1313 2774–2789
	3	16	2	999–1014 3796–3811
	4	23	6	1148–1170 2845–2867
	5	25	7	2352–2376 3291–3315
	6	25	7	1524–1548 3022–3046
	7	28	9	645– 672 718– 745
	8	28	9	615– 642 3622–3649
	9	28	9	462– 489 2191–2218
	10	37	14	28– 64 1717–1753
	11	37	14	1975–2011 2223–2259
	12	41	16	102– 142 1009–1049
	13	41	16	2170–2210 3257–3297
	14	42	16	601– 642 3806–3847
	15	42	16	1009–1050 2536–2577
T7	1	26	8	1298–1323 1896–1921
	2	26	8	2039–2064 2637–2662
	3	29	10	139– 167 326– 354
	4	30	10	1491–1520 1612–1641
	5	34	13	329– 362 2565–2598
	6	66	34	449– 514 2019–2084
	7	75	38	495– 569 1664–1738
	8	75	38	2140–2214 2441–2515

Thus, all the investigated genes were saturated by non-random inverted repeats and (or) complementary palindromes. The mRNAs, coded by them, can form a developed secondary structure by non-random double helices. Fig.3,4 presents such structures for mRNAs of the *E. coli* β-subunit and of T7 RNA polymerase. There are some common properties in mRNA secondary structures:

(i) They are formed by the long non-random helices (up to 80–100 base pairs) with partial violations of complementarity. The necessity of such non-random helices is motivated by their ability to provide the regular formation of a specific set of functionally important secondary structures. The absence of non-random double helices would lead every time to formation of different secondary structures, similar to a stochastic ball, instead of one functionally determined and important form.

(ii) The secondary structures are compact and highly helical (25–50% of all the nucleotides are involved in the non-random double helices). The compactness is the result of non-random double helices formed by RNA segments, located at remote parts of the nucleotide chain.

(iii) The 5'- and 3'-ends of mRNAs are blocked by these secondary structures, i.e., the neighbouring segments are involved in double helices (fig.3,4). This property seems to protect mRNAs from nucleases. However, the violations of complementarity lower the stability of the double helices and make possible their uncoiling and reconstruction in the course of translation.

(iv) Some groups of non-random helices are competing (fig.3,4). They are supposed to form the alternative secondary structures of functional importance. Such groups of competing helices at the 5'-end of RNAs are of special interest. For example, the helix (28–59, 1722–1753) in mRNA of the β-subunit of *E. coli* RNA polymerase blocking the 5'-end is competing with the alternative helix (1722–1730, 1733–1741), making the 5'-end free and available for translation.

Table 4

The properties of complementary palindromes for genes of β'- and σ-subunits of *E. coli* RNA polymerase and of phage T7 RNA polymerase

Gene	Number	Length of palindrome	Number of non-complementary pairs	Number of repeats			Localization
				Expected	95% border of confidence interval	Real	
β'	1	27	2	0.02	1	1	553– 579
	2	25	2	0.05	1	1	383– 407
	3	26	2	0.02	1	1	3789–3814
	4	40	7	0.06	1	1	3701–3740
	5	68	14	0.006	1	1	2603–2670
	6	67	15	0.04	1	1	4143–4209
	7	82	20	0.03	1	1	1067–1148
	8	99	24	0.008	1	1	337– 435
	9	96	26	0.07	1	1	379– 474
	10	107	28	0.03	1	1	3701–3803
	11	144	39	0.001	1	1	3775–3918
	12	152	42	0.001	1	1	630– 781
	13	158	45	0.006	1	1	1821–1978
σ	1	18	2	0.35	3	4	293– 310
	2						837– 854
	3						65– 82
	4						930– 947
	5	25	2	0.02	1	1	752– 776
	6	22	3	0.22	2	2	1331–1352
	7						1688–1709
	8	33	5	0.04	1	1	429– 461
	9	61	14	0.03	1	1	62– 122
	10	82	20	0.007	1	1	116– 197
	11	88	21	0.001	1	1	71– 158
	12	91	25	0.002	1	1	185– 275
T7	1	43	17	0.05	1	1	738– 780
	2	57	27	0.08	1	1	221– 277

Let us consider in more detail the group of such competing helices, localized at 11% of its length from the 5′-end of mRNA of the *E. coli* RNA polymerase β'-subunit: helix I (553–563, 569–579), corresponding to a non-random palindrome (table 2), and helix II (472–501, 553–582). This structure (fig.5) is similar to the attenuator of the *E. coli* trp-operon [13] and has several stable states. It is interesting that the segment (431–454), out of the main translation frame, can code the short peptide, being analogous with the leader peptide of the trp-operon [13]. The attenuator-like structures were also found in other mRNAs. At the gene level they can be used for transcription regulation.

The functional role of the compact secondary mRNA structures, with long double helices and partial distortions of complementarity, with blocking of 5′- and 3′-ends and with groups of competing helices, has not yet been fully elucidated. Perhaps, they are necessary either for mRNA protection from the action of nucleases, or for formation of mRNA sites, recognizing the regulatory proteins or ribosomes in the course of translation.

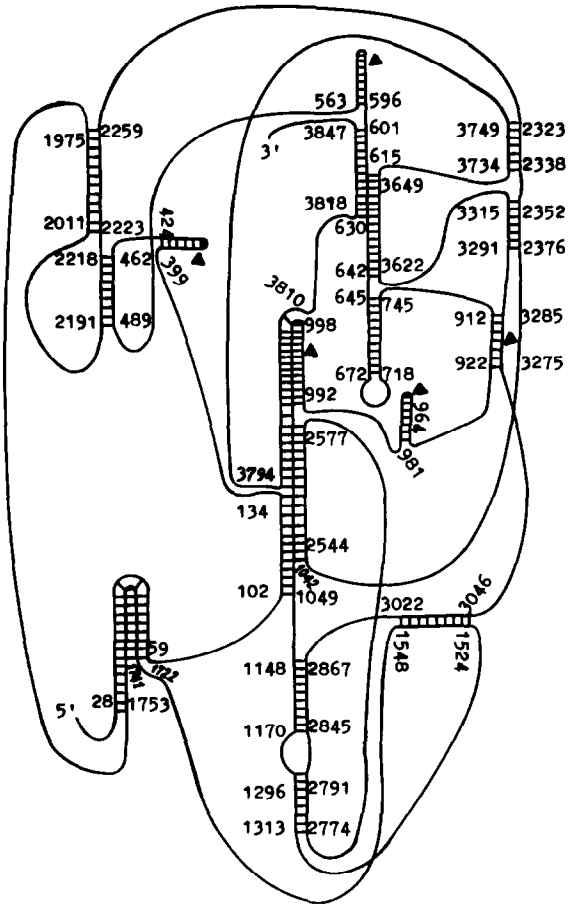

Fig.3. Schematic expression of the secondary structure of mRNA of *E. coli* RNA polymerase β-subunit, formed by double helices corresponding to non-random inverted repeats and complementary palindromes. The non-helical segments are depicted by single lines, the helices by double lines, with shading. The numbers indicate the helices' localization. (▲) Inverted repeats or complementary palindromes which are not listed in tables 3 and 4, because they are random, but take part in formation of the secondary structure. Some non-random helices are omitted for simplification of the secondary structure plate picture. The ends of some helices are lengthened by addition of G-U pairs as compared to the data of tables 3 and 4.

It is important to determine such structures in the course of planning of experiments for directed mutagenesis by complementary addressed modification.

ACKNOWLEDGEMENTS

We wish to express our gratitude to D.G. Knorre, V.V. Schamin, L.V. Omelyanchuk and I.N. Schindyalov for constructive criticism.

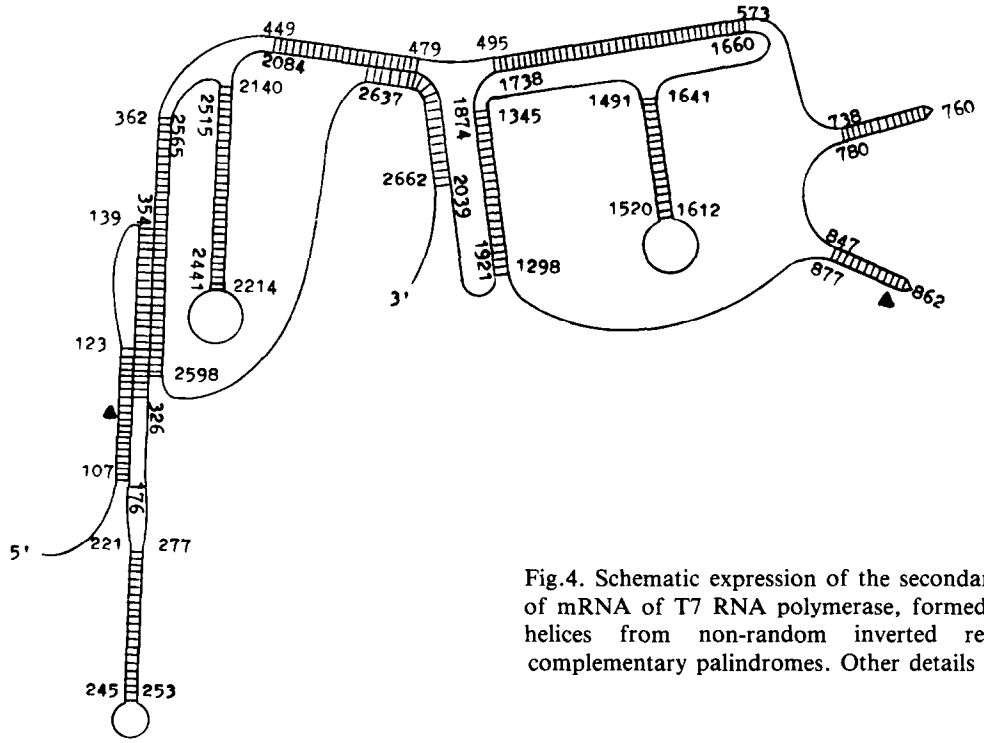

Fig.4. Schematic expression of the secondary structure of mRNA of T7 RNA polymerase, formed by double helices from non-random inverted repeats and complementary palindromes. Other details as in fig.3.

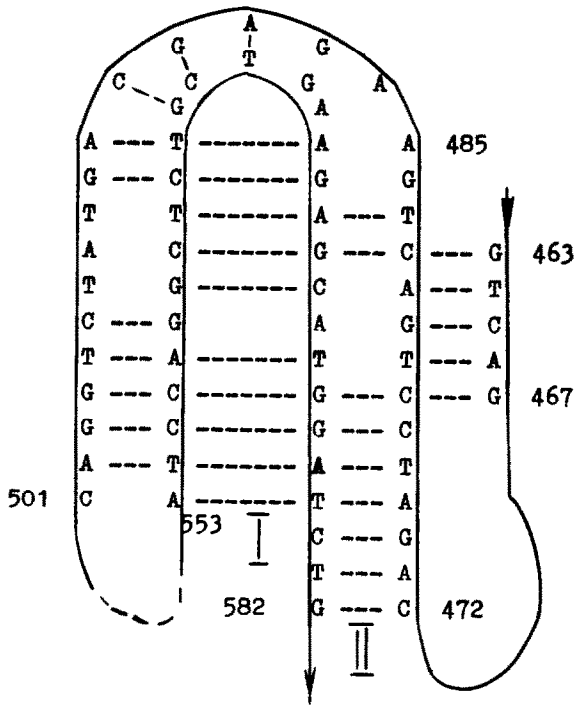

Fig.5. Detailed picture of the gene segment of *E. coli* RNA polymerase β'-subunit, capable of forming the competing secondary structures, indicated by Roman figures.

REFERENCES

[1] Pipas, J.M. and McMahon, J.E. (1975) Proc. Natl. Acad. Sci. USA 72, p.2017–2021.

[2] Studnicka, G.M., Rahn, G.M., Cummings, I.W. and Salser, W. (1978) Nucleic Acids Res. 5, 3365–3387.

[3] Bokhonov, V.B. and Kolchanov, N.A. (1979) in: Mathematical Models of Molecular Genetic Regulatory Systems (Ratner, V.A. ed.) pp.124–162, ICG-Press, Novosibirsk (in Russian).

[4] Omelyanchuk, L.V., Bessonov, Yu.E. and Kolchanov, N.A. (1981) in: Computer Systems (Zagoruyko, N.G. ed.) vol.88, pp.136–146, IM-Press, Novosibirsk (in Russian).

[5] Tinoko, I., Borer, R.N., Dengler, B., Levine, M.D., Uhlenbeck, O., Grothers, D.N. and Gralla, J. (1973) Nat. New Biol. 246, 40–41.

[6] Kolchanov, N.A. and Omelyanchuk, L.V. (1982) Stud. Biophys. 83, 115–116.

[7] Nussinov, R., Tinoko, I. and Jacobson, A.V. (1982) Nucleic Acids Res. 10, 351–363.

[8] Kolchanov, N.A., Solovyov, V.V. and Zharkikh, A.A. (1983) Dokl. Akad. Nauk SSSR, in press.

[9] Ovchinnikov, Yu.A., Momastyrskaya, G.S., Gubanov, V.V., Guryev, S.O., Salamatina, I.S., Schuvalova, T.M., Lipkin, N.M. and Sverdlov, E.D. (1981) Dokl. Akad. Nauk SSSR 261, 763–768.

[10] Ovchinnikov, Yu.A., Monastyrskaya, G.S., Gubanov, V.V., Guryev, S.O., Chertov, O.Y., Modianov, N.N., Grinkevich, V.A., Makarova, I.A., Marchenko, T.V., Polovnikova, I.N., Lipkin, V.M. and Sverdlov, E.D. (1981) Eur. J. Biochem. 116, 621–629.

[11] Barton, Z., Burgess, K.R., Lin, J., Moore, D., Holder, S. and Gross, C. (1981) Nucleic Acids Res. 9, 2889–2903.

[12] Grachev, M.A. and Pletnev, A.G. (1981) FEBS Lett. 127, 53–56.

[13] Yanofsky, C., Platt, T., Crawford, I.P., Nichols, B.P., Christie, G.E., Horowitz, H., Yan Cleamput, M. and Wu, A.M. (1981) Nucleic Acids Res. 9, 6647–6668.