ALGORITHM FOR SEARCHING FOR ALTERNATIVE SECONDARY RNA STRUCTURES

[*] ***Lyubetsky E.V., Lyubetsky V.A.***

Institute for Information Transmission Problems, RAS, Moscow, Russia, e-mail: Lin@iitp.ru

***Key words**: secondary structure hairpin, alternative structure, terminator, coordinates of hairpin loops, algorithm, statistical certainty*

Resume

Motivation: A new algorithm and a model of searching for alternative secondary RNA structures and their specific peculiarities in including hairpin loops were implemented.

Results: The algorithm has shown a good efficiency on more than 120 experimental natural RNA fragments.

Availability: The software is available on request directed to authors.

Introduction

The problem of prediction of alternative secondary RNA structures is considered. It was found out last years that such structures being considered at mRNA level play an unexpectedly important role taking part in regulation of biosynthesis processes in a cell (attenuation). In models of regulation, the essential role is allocated to just unpaired bases, in particular, to the nucleotides of hairpin loops. The problem arises to find a small number of candidates for the alternative secondary structure in a given fragment of an RNA sequence (or to declare that there is no such structure).

An algorithm is proposed that is apparently one of the first algorithms aimed to solve this problem. Therefore, we had no opportunity to compare its results with processing of other algorithms. The algorithm was tested on random sequences and also in situations when the alternative secondary structures were retrieved experimentally.

The algorithm was applied in rather general situation, but here we present only the results of search for “three hairpin” attenuation and T-box terminator–antiterminator structures as well as of search for coordinates of hairpin loops and other hairpin patterns in such structures.

The algorithm was tested on the following structures:

1) Transcription attenuators for pheA and trp genes (which are used in biosynthesis of aromatic amino acid in gamma-proteobacteria) and for pheS gene (coding for phenylalanine-tRNA synthetase). We tested 17 such structures: the results were near 100% (see the list the table below).

2) Transcription attenuators for pyrimidine biosynthesis genes in *B. subtilis.* We tested three such structures: loop coordinates of the secondary hairpin structure regulating them were retrieved (see rows 1–3 in table).

3) T-box terminator–antiterminator structure (taking part in regulation of genes pertaining to biosynthesis of amino acids and genes coding for aminoacyl-tRNA synthetase in gram-positive bacteria). Loop coordinates and terminator hairpin of the secondary hairpin structure regulating them were retrieved (see rows 4–43 in the table).

Let us remark that the algorithm does not use any specific parameter of the secondary structure; it was not supported by experimental answers anyway.

Methods and Algorithms

The algorithm is based on recursive operating with some set of hairpin and structure parameters and proceeds from parameters to hairpins and structures only at the last stage of the processing. The following hairpin parameters are used by the algorithm: power, the origin A and endpoint B of the left half-stem, the origin C and endpoint D of the right half-stem of a hairpin, free energy, and so on. A structure parameter linking power of two hairpins is also used. The algorithm generates a set of locally optimal parameters and used it to produce the set of locally optimal hairpins. The last of them is statistically analyzed to build a consensus hairpin and an alternative secondary structure from the obtained hairpins. A more detailed description of this rather logically intricate, but fast and efficient algorithm can be found in (Vereshchagin, Lyubetsky, 2000).

The algorithm was implemented as a program in Object Pascal language in Delphi 5 environment and also in ANSI C in a serial and parallel computing architectures. It showed a high efficiency on some natural mRNA sequences.

[*] Corresponding author.

Implementation and Results

Some results of the algorithm testing is shown in table and in the list following it. Here, Sh is a specifier hairpin, A is an antiterminator, and T is a terminator. Expression "exact" in the 4th column means that loops of hairpins from one alternative secondary structure are found in exact correspondence with experimental answer. "NF" (not found) means that the coordinates of hairpin loops differ from the experimental answer on more than 10 positions in the hairpin left or right half-stems. Expressions like "8,10" mean that difference between (B,C) values in algorithm results and experimental answers is 8 nucleotides in the left half-stem and 10 nucleotides in the right one for a hairpin under consideration. Expression like "W2" (without 2 pairs) in the 5th column means that at the end of the terminator hairpin, located by the algorithm, 2 nucleotide pairs from experimental answer are missing.

	Gene name	Number of biol. hairpes	Precision of locating of hairpin loop coordinates	Precision of term-or locating
1	Bs_pyrB	2	T – exact; A – exact	---
2	Bs_pyrP	2	T – 8,10; A – 4,1	---
3	Bs_pyrR	2	T – exact; A – exact	---
4	Be_serS	5	T, Sh, 2, 3, A – exact	Exact
5	Be_tyrS	5	T – exact; Sh - 2,5; 2 – NF; 3, A – exact	Exact
6	Bq_serS	5	T – exact; Sh – NF; 2 – 2,3;3 – exact; A – 0,2	Exact
7	Bq_tyrS1	5	T – exact; Sh – exact; 2,3 – NF; A – exact	NF
8	Bq_tyrS2	5	T – exact; Sh – exact; 2 – 0, 3; 3, A – exact	W3
9	Bs_serS	5	T – 1,1; Sh – 3,5; 2,3,A – exact	W1
10	Bs_tyrS	6	T – exact; Sh – NF; 2 – 2,2; 3 – 3,3; 4 – exact; A – 0,1	Exact
11	Bs_tyrZ	6	T – NF; Sh – exact; 2 – NF; 3 – 1,0; 4 – exact; A – NF	NF
12	Ca_tyrZ	3	T – NF; Sh – 1,4; A – 0,1	NF
13	Ca_yurG	5	T – 0,1; Sh – 4,1; 2,3 – NF; A – 2,0	NF
14	DF_serS	5	T – NF; Sh, 2, 3 – exact; A – 4,1	NF
15	DF_tyrZ	3	T – 1,0; Sh – 4,1; A – exact	W1
16	DHA_tyrZ	3	T – 0,1; Sh – 0, 2; A – 0,1	NF
17	EF_serS	3	T – 3,2; Sh – NF; A – 1,4	NF
18	EF_tyrS	5	T – exact; Sh – NF; 2 – exact; 3 – 6,2; A – exact	NF
19	HD_serS	5	T – exact; Sh – exact; 2 – 0,6; 3 – 5,3; A – 0,1	NF
20	HD_tyrZ	6	T – exact; Sh – 5,3; 2 – NF; 3 – 4,1; 4 – 1,1; A – exact	NF
21	LLX_serS	3	T – 1,2; Sh – 1,0; A – 0,1	NF
22	LO_serS	3	T – 0,1; Sh – 3,1; A – NF	Part. found
23	LO_tyrS	5	T – exact; Sh – 4,1; 2,3 – exact; A – 0,1	Exact
24	PN_serS	3	T – 1,2; Sh – 2,6; A – exact	NF
25	Sa_serS	5	T – exact; Sh – 5,3; 2 – 0,1; 3,A – NF	Exact
26	SEQ_serS	3	T – 1,1; Sh – NF; A – exact	NF
27	Bs_thrS	5	T – 1,0; Sh – NF; 2,3,A – exact	NF
28	LL_his	5	T – 5,4; Sh,2,3 – exact; A - NF	NF
29	LL_trp	5	T – exact; Sh – exact; 2 – NF; 3– 0,2; A – exact	Exact
30	Bs_purE	3	T – NF; Sh – exact; A – NF	NF
31	Bs_purM	1	T – 2,1	W2
32	Bs_tyrS	5	T – exact; Sh, 2 – NF; 3, A – exact	Exact
33	Ec_pyrB	2	T – 0,1; A – NF	NF
34	Ec_ilvG	3	T – exact; Sh, A – exact	NF
35	Ec_rpsJ	6	1 – NF; 2 – 1,1; 3,4 – NF; 5 - exact; 6 – 3,2	No term-r
36	Hi_rpsJ	5	1 – 3,1; 2 – 1,1; 3,4 – exact; 5 - 6,2	No term-r
37	Bs_ilv_leu	5	T – exact; Sh – exact; 2 – 2,6; 3 – NF; A – exact	NF
38	Bs_yczA	5	T – exact; Sh – NF; 2 – 1,1; 3 – exact; A – NF	Exact
39	Bs_trpE	2	1 – exact; 2 – 0,1	No term-r
40	Sa_ileS	4	T – exact; Sh, 2, A – NF	NF
41	Ec_rplK	2	1, 2 – exact	No term-r
42	Hi_rplK	2	1, 2 – exact	No term-r
43	Bs_valS	3	T – 0,1; Sh – 4,3; A – NF	NF

In the following list are names of genes followed by evaluations of likeness of alternative secondary structures found experimentally and by the algorithm. The first one is evaluation of the terminator hairpin; the second and third are evaluations of antiterminator and specifier hairpins of alternative secondary structures. Evaluation "2" means here that the distances between two answers on B and on C are less then 5, and on A and on D are less then 7. Similarly evaluation "1" means that these distances are less then 5 and more then 7. Finally, evaluation "0" means that these distances are more then 5 and more then 7. So, we have the following comparisons of the algorithm results with the experimental answers: Aa_aroma_pheA 210, 000; Ec_aroma_pheA 000, 000, 121; Ec_aroma_pheS 221, 000; Ec_aroma_trpE 210, 000; Hi_aroma_pheA 210; Hi_aroma_pheST 210; Hi_aroma_trpBA 000; Hi_aroma_trpE 000, 000; St_aroma_pheA 210, 000, 000; St_aroma_pheS 212, 000; St_aroma_trpE 210; Vc_aroma_pheA 211; Vc_aroma_trpE 210, 000, 000, 000; Yp_aroma_pheA1 221; Yp_aroma_pheA2 210; Yp_aroma_PheS 000, 211, 000, 000; and Yp_aroma_trpE 210. Let us remark that in three cases including two of them having incorrect answers, terminator hairpins are really more complicated (they include bulges). Now, we have developed an algorithm for locating such terminator hairpins.

Discussion

1) Loop coordinates of many hairpins (e.g. terminator hairpins) are located with a very high precision. Terminator hairpins of "three hairpin" attenuation structures are located in 15 cases of 17 (88%). Terminators are located in 29 cases (56%) of 52.

2) When testing 129 sequences, we found that the number of hairpin loop coordinates repeated more then 9 times, was rather small relative to the total number of hairpin loop coordinates that were located by the algorithm. The total number of locally optimal hairpins for each studied sequence was 300–600, the algorithm has selected only 10–25 loci in it for further processing.

3) Hairpin loop coordinates and other hairpin and configuration patterns found by the algorithm are almost independent of their locations in the original sequence and its length.

4) In most cases when the algorithm found an answer with a low precision, the biological answer was also "incorrect" in the sense that hairpins contained side subhairpins or sections of power 2 or pairs <G,T> at the endpoints of sections. An algorithm is now developed that locates such hairpins with the same efficiency.

5) In the case of exact retrieving of hairpin loop coordinates, hairpin's first section was located almost always correctly. In half of the cases, two sections were located correctly. In some cases, whole hairpin was retrieved (all its 1–5 sections).

The detailed results of research are placed in the online journal *Information Processes* at http://www.jip.ru.

Acknowledgements

The authors express a deep gratitude to Prof. M.S.Gelfand and A.A.Mironov for their help and explanations of the biological content of the problem. Computer program was implemented and run by L.A.Leontiev.

References

1. Vereshchagin N.K., Lyubetsky V.A. (2000) Algorithm for determination of alternative secondary RNA structures. Transact. Research Seminar of the Logical Center of the Institute of Philosophy RAS, 14, M.: Nauka, 99-109.