ISSN 0026-8933, Molecular Biology, 2007, Vol. 41, No. 4, pp. 670–673. © Pleiades Publishing, Inc., 2007.
Original Russian Text © E.V. Lyubetskaya, A.V. Seliverstov, V.A. Lyubetsky, 2007, published in Molekulyarnaya Biologiya, 2007, Vol. 41, No. 4, pp. 739–742.

SHORT COMMUNICATIONS

UDC 577.053

The Number of Long Hairpins in Intergenic Trailer Regions of Actinobacteria Is Far Greater Than in Other Genomic Regions

E. V. Lyubetskaya, A. V. Seliverstov, and V. A. Lyubetsky

Kharkevich Institute of Information Transmission Problems, Russian Academy of Sciences, Moscow, 127994 Russia; e-mail: slvstv@iitp.ru

Received June 30, 2006

Accepted for publication August 31, 2006

DOI: 10.1134/S002689330704022X

Key words: DNA secondary structure, gene expression

Bacterial genomes frequently contain genes arranged alternately in the complementary DNA strands and directed towards one another (so-called convergons). If such genes are intensely transcribed, it is natural to assume that the region between them contains a transcription terminator, which is necessary to stop transcription and relieve the conformational stress caused at this site by DNA supercoiling. Hereinafter, this region is called the convergent trailer region. In the genomes of *Bacillus subtilis* and *Escherichia coli*, the role of such a terminator is played by the so-called classical transcription terminator, i.e., a relatively short (5–10 pairs of complementary nucleotides in the paired stem) GC-rich hairpin having a small loop and followed by a T-rich region. The operons of *B. subtilis* and *E. coli* usually end with this particular terminator. However, classical terminators are rather rare in many Actinobacteria, including mycobacteria [1, 2].

We propose a new high-capacity algorithm allowing a massive and very rapid search for hairpins with specified parameters of the stem and loop in an arbitrary nucleotide sequence. The algorithm was applied to search for hairpins in convergent trailer sequences and other regions of actinobacterial genomes. This allowed us to count the long (from 17 bp in the stem) hairpins and the hairpins capable of forming a cruciform structure (a certain pair of hairpins on complementary DNA strands) [3] in the genome regions of the types listed below. We assume that such cruciform structures form a noncanonical secondary DNA structure, which serves for relief of the conformational stress and transcription termination. In addition, the algorithm detected essentially different numbers of long hairpins in different-type regions of the Actinobacteria genome. These hairpins were especially abundant in convergent trailer regions. The hairpins do not belong to the known end transcription terminators or transcription initiation regulators described earlier [4], differing from them in both the length and position in the genomic sequence. The detected hairpin pairs forming cruciform structures in the actinobacterial genome are considerably more abundant compared with similar pairs in the regions of the same type in the *B. subtilis* and *E. coli* genomes. The majority of hairpins in Actinobacteria are not accompanied by a T-rich region and are considerably longer than the classical transcription terminator.

It is known that the expression of genes involved in amino acid biosynthesis in Actinobacteria is regulated at the translational level [5], unlike γ- and α-proteobacteria, where expression of these genes is presumably regulated at the level of transcription [6]. Therefore, in this work we analyzed the regions of the actinobacterial genome that can potentially contain the secondary structures connected with transcription termination. In this respect, the regions located downstream of tRNA genes and the genes expressed to a high level are of special interest. Study of these genes is helpful for predicting the boundaries of operons and distinguishing between the genes with a high expression level from their paralogs with a low transcription level [7].

The nucleotide sequences of bacterial genomes were extracted from GenBank (ftp.ncbi.nlm.hih.gov/genbank). Sequences from the completely sequenced genome of one strain were taken for each actinobacterial species represented in GenBank, and only the species that had such a strain were considered. Strains *E. coli* K12 and *B. subtilis* 168 were used

Frequency of hairpins with a specified stem length (l) in various genomic regions of Actinobacteria, *B. subtilis*, and *E. coli*

l	1	2a	2b	3	4
			Corynebacterium efficiens		
25	0.23	0.38	1 (0)	0.00	0.00
23	0.23	0.76	2 (0)	0.00	0.00
20	0.68	6.06	16 (11)	0.00	0.00
17	2.60	13.64	37 (29)	0.17	0.31
15	4.98	21.59	57 (48)	0.43	1.85
10	20.61	45.83	121 (95)	31.79	8.31
			Mycobacterium tuberculosis		
25	0.31	3.29	11 (11)	0.07	0.18
23	0.38	4.19	14 (14)	0.07	0.18
20	0.54	5.09	17 (17)	0.07	0.36
17	1.30	5.99	20 (20)	0.14	0.53
15	1.83	6.29	21 (21)	0.57	0.89
10	15.44	10.78	36 (35)	35.83	5.35
			Propionibacterium acnes		
25	0.00	0.00	0	0.00	0.00
23	0.00	0.00	0	0.00	0.00
20	0.00	0.96	2 (2)	0.04	0.00
17	0.45	8.17	17 (16)	0.09	0.00
15	1.20	12.98	27 (23)	0.13	0.00
10	15.19	30.29	63 (57)	24.29	6.34
			Streptomyces coelicolor		
25	0.08	0.95	8 (8)	0.00	0.00
23	0.15	2.25	19 (18)	0.01	0.00
20	0.54	3.08	26 (25)	0.02	0.17
17	1.73	9.60	81 (78)	0.09	0.41
15	3.61	14.45	122 (117)	0.35	1.00
10	31.07	37.32	315 (306)	23.09	14.36
			Bifidobacterium longum		
25	0.45	1.51	3 (3)	0.05	0.00
23	0.45	2.02	4 (4)	0.05	0.00
20	1.05	5.56	11 (10)	0.05	0.00
17	1.79	13.64	27 (23)	0.43	0.45
15	4.19	21.21	42 (36)	0.65	1.80
10	28.25	47.47	94 (75)	28.42	11.26
			Leifsonia xyli		
25	0.11	0.00	0	0.00	0.00
23	0.11	0.00	0	0.00	0.00
20	0.65	2.78	4 (4)	0.05	0.00
17	1.63	4.86	7 (7)	0.34	0.00
15	2.28	7.64	11 (11)	1.01	0.47
10	21.93	22.92	33 (33)	39.28	9.48

Table (Contd.)

l	1	2a	2b	3	4
			Escherichia coli		
25	0.00	0.00	0	0	0.00
23	0.00	0.00	0	0	0.00
20	0.00	0.00	0	0	0.00
17	1.38	0.22	1 (1)	0.02	0.19
15	4.15	2.91	13 (8)	0.13	0.56
10	77.67	47.65	213 (159)	18.56	11.52
			Bacillus subtilis		
25	0.00	0.00	0	0	0.00
23	0.00	0.00	0	0	0.00
20	0.13	0.51	2 (1)	0	0.00
17	0.89	3.55	14 (6)	0.02	0.20
15	2.74	13.96	55 (12)	0.11	0.20
10	25.86	67.77	267 (97)	18.71	13.64

Note: The table shows the ratio (%) of the number of hairpins with a specified minimal stem length *l* as found by our algorithm to the number of all regions of this type in the genome. Column 1, all leader regions; 2a and 2b, all convergent trailer regions; 3, all coding regions; and 4, all divergent regions. Column 2b lists the number of hairpins found in all convergent trailer regions; the number of hairpins lacking a T-rich tract is shown in parentheses.

for comparison. The long hairpins in question were detected downstream of intensely transcribed protein genes or the tRNA genes. The table summarizes part of the data obtained.

Genomic regions of four types are considered: all leader regions, all convergent trailer regions (convergons), all coding regions, and all divergent leader regions (divergons). Only one hairpin with the longest stem was selected in each region. The hairpins selected had to meet the following requirements: the stem length is equal to or exceeds the threshold value *l*, measured in base pairs; the stem may contain unilateral bulges no longer than 2 nt; and the length of the loop should not exceed 15 nt.

The longest hairpins are typical of convergons of protein-coding genes. The maximal stem length in convergent trailer regions is 31 bp in *Mycobacterium tuberculosis, M. bovis*, and *Streptomyces coelicolor*; 29 bp in *S. avermitilis*; and 27 bp in *Corynebacterium diptheriae*. Unlike in Actinobacteria (table), the hairpins with a stem length over 17 bp are considerably rarer in *B. subtilis* and especially *E. coli*. The convergon column of the table lists the frequency (fraction) of the hairpins that lack a downstream T-rich sequence with a length of at least 7 bp and no more than two exceptions. The T-rich sequence was searched for in region –5 to +12 from the last nucleotide of the hairpin stem.

The results of our analysis suggest the following conclusions. If at least one of the two actinobacterial genes directed towards one another is actively transcribed, it is often followed by a potential long cruciform DNA structure. This structure can perform two putative functions. First, it can relieve the stresses arising during intensive transcription (with or without the involvement of topoisomerase). Second, a change in DNA conformation may provide transcription termination.

The majority of detected hairpins are considerably longer than conventional classical transcription terminators; moreover, they have a number of specific features, for example, they lack downstream T-rich regions. The resulting secondary structure is symmetric and can play a role of a terminator for RNA polymerase sliding along each DNA strand. A higher frequency of these structures observed particularly in the convergent trailer regions compared with regions of the other types is an indirect confirmation favoring this hypothesis.

ACKNOWLEDGMENTS

The computer realization of our algorithm was performed by L.I. Rubanov. We are grateful to M.S. Gelfand for discussions and valuable criticism, and M.A. Shirshin for his contribution to the development of the software.

The work was supported by the International Science and Technology Center (project no. 2766).

REFERENCES

1. Washio T., Sasayama J., Tomita M. 1998. Analysis of complete genomes suggests that many prokaryotes do not rely on hairpin formation in transcription termination. *Nucleic Acids Res*. **26**, 5456–5463.
2. Unniraman S., Prakash R., Nagaraja V. 2002. Conserved economics of transcription termination in eubacteria. *Nucleic Acids Res*. **30**, 675–684.
3. Knorre D.G., Myzina S.D. 2003. *Biologicheskaya khimiya* (Biological Chemistry). Moscow: Vysshaya Shkola. p. 107.
4. Wadkins R.M. 2000. Targeting DNA secondary structures. *Curr. Med. Chem*. **7**, 1–15.
5. Seliverstov A.V., Putzer H., Gelfand M.S., Lyubetsky V.A. 2005. Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria. *BMC Microbiol*. **5**, 54.
6. Vitreschak A.G., Lyubetskaya E.V., Shirshin M.A., Gelfand M.S., Lyubetsky V.A. 2004. Attenuation regulation of amino acid biosynthetic operons in proteobacteria: Comparative genomics analysis. *FEMS Microbiol Lett*. **234**, 357–370.
7. Ishchukov I.I., Likhoshvai V.A., Matushkin Yu.G. 2004. A new algorithm for recognizing the operon structure of procaryotes. *Proc. 4th Int. Conf. on Bioinformatics of Genome Regulation and Structure,* Novosibirsk: Inst. Tsitol. Genet., SO RAN, pp. 73–76.