Molecular Biology, Vol. 39, No. 6, 2005, pp. 946–948. Translated from Molekulyarnaya Biologiya, Vol. 39, No. 6, 2005, pp. 1076–1078.

SHORT COMMUNICATIONS

UDC 577.053

Algorithm for a Large-Scale Search for T-Box Transcription Regulation in Bacteria

L. A. Leontiev, A. V. Seliverstov, and V. A. Lyubetsky

Institute of Information Transmission Problems, Russian Academy of Sciences, Moscow, 127994 Russia
e-mail: lyubetsk@ittp.ru

Received June 27, 2005

Key words: T box, transcription regulation, bacteria, multiple sequence alignment

The formation of alternative mRNA structures with the help of tRNA in response to changes in the content of an amino acid (or aminoacyl-tRNA synthase) is an important mechanism which regulates gene expression in the bacteria of many groups. T-box-mediated transcriptional regulation of gene expression has been studied only in a few bacteria, e.g., in the case of some aminoacyl-tRNA synthase genes of *Bacillus subtilis* [1].

We propose an algorithm for a large-scale search of 5'-leader regions of bacterial operons for elements involved in transcriptional regulation and associated with generation of a secondary structure involving the T box. The algorithm was used to examine all complete bacterial genomes available from GenBank (NCBI). In brief, the algorithm involves constructing a rooted tree for each leader region (nucleotide sequence), with the elements of a T-box-containing secondary structure of the given sequence ascribed to tops of the tree. At each top, such elements are inherited from the "father" of the given top and added from the original sequence. A conserved word, specific for the T box and containing a highly conserved binding site for the acceptor site of tRNA, is ascribed to the root of the tree. Terminal tops of the tree are ordered lexicographically, by increasing lengths of the internal stem of the specifying hairpin and the internal and external stems of the antiterminator hairpin, as well as by some additional numerical parameters. Then, the top highest in the order is selected. The secondary structure closest to the typal T-box structure is ascribed to this terminal top.

Additional parameters describe the conserved word of the antiterminator side loop, the quality of the uracil tract, the free energy of the structure, etc. In addition, interval limitations are imposed on the arrangements of these elements. Typical limitations include the following: the distance between the start of the specifying hairpin and the translation start is no more than 1000 nt, the terminator has a stem of no less than 6 nt and a loop of 3–20 nt, the antiterminator has stems of no less than 4 nt and a loop of 4–20 nt, and the specifying hairpin has a stem of no less than 4 nt and a loop of 4–20 nt. The computer program was developed by L.A. Leontiev.

Compared to the published algorithm [2], our algorithm takes into account more secondary structure elements and, consequently, has a greater prediction potential.

Using our algorithm, we identified the conserved RNA structures that involve the T box and predicted T-box transcriptional regulation for the corresponding genes. The results obtained using the algorithm were compared with the data available from the Rfam database for Firmicutes, since the greatest number of T boxes was found in this group (Table 1). The use of our algorithm allowed us to propose two new T box-containing RNA structures for proteobacteria, two for Chloroflexi, and ten for *Symbiobacterium thermophilum.* Only one T box, located upstream of *tna,* is characterized for *S. thermophilum* in the Rfam database.

In Firmicutes, our algorithm often predicted T boxes upstream of the genes involved in biosynthesis of aromatic amino acids (*aro*A and *aro*F), asparagine (*asn*A), cysteine (*cys*E), methionine (*met*A, *met*B, *met*E, *met*G, and *met*K), proline (*pro*B and *pro*I), branched amino acids (*ilv*B), threonine–methionine (*hom*), and tryptophan (*trp*E). This finding contradicts

Table 1. Number of T boxes found in Firmicutes

Phylogenetic group	Total T boxes	
	our algorithm	Rfam database
Bacillaceae	214	275
Staphylococcus	21	69
Listeria	27	43
Clostridia	47	31
Lactobacillales	102	52
Mollicutes	4	3

0026-8933/05/3906-0946

Table 2. Transporter protein genes with putative T boxes in the upstream regions

Group	Transporter function	*B. anthracis*	*B. cereus*	*B. thuringiensis*	Anticodon	Amino acid
1	Na^+ symporter	*GBAA1448* (282, 142)	*BCZK1312* (280, 75)	*BT9727_1311* (281, 75)	GGC	Glycine
2	Na^+ symporter	*GBAA1835* (465, 280)	*BCZK1657* (500, 277)	*BT9727_1682* (456, 276)	?	Isoleucine, valine
3	Na^+ symporter	*GBAA2216* (331, 86)	*BCZK1999* (332, 87)	*BT9727_2000* (332, 87)	CCU	Proline
4	Permease with DUF894	*GBAA2649* (258, 37)	*BCZK2393* (258, 34)	*BT9727_2432* (258, 34)	GAA	Glutamic acid
5	Na^+/H^+ antiporter	*nhaC3* (116, 37)	*nhaC* (486, 250)	*nhaC* (486, 253)	CGG	Arginine

Note: The distances from the start of the specifying hairpin and the end of the polyuridine tract to the initiator codon are indicated in parentheses.

the data that *trp*E expression is regulated via classical attenuation in many γ-proteobacteria and actinobacteira and agrees with the absence of such attenuation in Firmicutes [3–5]. In *Bacillus* (*B. anthracis* strain Ames Ancestor, *B. cereus* ZK, and *B. thuringiensis* serovar konkukian strain 97-27), the algorithm predicted 15 T boxes upstream of the genes for transporter proteins. These proteins form five orthologous groups.

In addition, T boxes were found upstream of the two orthologous *nha*C genes coding for Na^+/H^+ antiporters of *Lactobacillus plantarum* WCFS1 (323, 132) and *Enterococcus faecalis* V583 (319, 153). The arginine anticodons were detected in the T boxes of these antiporter genes.

T boxes were revealed upstream of the *Clostridium tetani* and *L. plantarum* genes coding for transporters of branched amino acids: the *CTC00787* gene coding for protein AAO35385.1 of *C. tetani* E88 (343, 121) and the *lp_1744* gene coding for protein CAD64164.1 of *L. plantarum* WCFS1 (236, 78). In *L. plantarum* WCFS1, a T box was found upstream of the gene for ABC transporter CAD65358.1 (257, 35). In *L. acidophilus* NCFM, a T box was detected upstream of the *opp*A (284, 20) gene for oligopeptide ABC transporter AAV42092.1.

In addition, a tryptophan T box was found upstream of *rtp*A coding for TRAP inhibitor AAU21914.1 of *B. licheniformis* ATCC 14580 (276, 86).

The algorithm predicted T box-containing structures for α and β-proteobacteria and for bacteria of the group Chloroflexi. These structures were found upstream of the *SMc00946* gene for hypothetical protein CAC46217.1 of the α-proteobacterium *Sinorhizobium meliloti* (375, 169), *leu*A for 2-isopropylmalate synthase of the δ-proteobacterium *Geobacter sulfurreducens* (266, 90), *ile*S for isoleucyl-tRNA synthase, and *trp*E for anthranylate synthase of *Dehalococcuides ethenogenes* (247, 18).

In *S. thermophilum*, the algorithm predicted T boxes upstream of *lys*S (239, 27), *tyr*S (216, 25), *ala*S (238, 20), *ile*S (217, 35), *leu*S (189, 21), *thr*S (140, 21), *trp*S (275, 151), *sth8* (335, 29), *sth2424* (238, 19), and *sth3130* (227, 33).

It should be noted that 53% of the predicted T boxes are upstream of the genes coding for aminoacyl-tRNA synthases.

Since our results are in good agreement with the data available from the Rfam database, we think that the algorithm of a large-scale search for T boxes is adequate.

Five families of orthologous transporters with completely coinciding amino acid sequences of orthologs were found in bacilli (Table 2). A T box is present upstream of each of the transporter genes. For these genes, it was possible to construct a high-quality multiple sequence alignment of all structural elements important for T-box regulation.

For the first three groups of the Pfam database, we found a homologous domain characteristic of the neurotransmitter symporter family (SNF) of eukaryotic transporter proteins, which includes glycine and proline transporters. Putative glycine and proline anticodons were detected in groups 1 and 3, respectively. The protein of group 4 belongs to the major facilitator superfamily (MFS, $E = 10^{-34}$). This protein is annotated as permease and contains domain DUF894 ($E = 0.009$) with an unknown function. The domain accounts for almost all protein length and has low homology to domains characteristic of the metabotropic glutamate receptor family (PF00003, $E = 6.7$). The glutamine anticodon is in the antiterminator loop of the specifying hairpin.

Protein CAC46217.1 of the α-proteobacterium *S. meliloti* contains a domain homologous to the YjeF family ($E = 10^{-49}$). A T box upstream of *leu*A, involved in leucine synthesis, was predicted for the δ-proteobacterium *G. sulfurreducens*. In γ-proteobacteria, expression of this gene is regulated via classical

attenuation [3]. In the case of *G. sulfurreducens leu*A, attenuation of transcription was not observed.

ACKNOWLEDGMENTS

We are grateful to M.S. Gelfand for his fruitful discussion of our research. This study was supported by an ISTC grant (no. 2766).

REFERENCES

1. Grundy F.J., Henkin T.M. 2003. The T box and S box transcription termination control systems. *Front. Biosci*. **8**, d20–d31.
2. Vitreshchak A.G. 2002. Computer-aided prediction of regulatory RNA sites. Regulation of the expression of genes involved in amino acid biosynthesis and genes coding for tRNA synthetases in Gram-positive bacteria. *Inform. Protsessy.* **2**, 91–95.
3. 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.
4. Loubatsko V.A., Salivarstiv A.V. 2004. Note on cliques and alignments. *Inform. Protsessy.* **4**, 241–246.
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. *BIN Microbiol*. (in press).