Available online at www.sciencedirect.com
SCIENCE DIRECT®
BioSystems 80 (2005) 145–154
BioSystems
www.elsevier.com/locate/biosystems

Identification of conserved regulatory RNA structures in prokaryotic metabolic pathway genes

Elena A. Lesnik[a], Gary B. Fogel[b], Dana Weekes[b], Timothy J. Henderson[a], Harold B. Levene[a], Rangarajan Sampath[a,*], David J. Ecker[a]

[a] *Ibis Therapeutics, 1891 Rutherford Road, Carlsbad, CA 92008, USA*
[b] *Natural Selection Inc., 3333 N. Torrey Pines Ct., Suite 200, La Jolla, CA 92037, USA*

Received 8 September 2004; received in revised form 4 November 2004; accepted 5 November 2004

Abstract

A combination of algorithms to search RNA sequence for the potential for secondary structure formation, and search large numbers of sequences for structural similarity, were used to search the 5′UTRs of annotated genes in the *Escherichia coli* genome for regulatory RNA structures. Using this approach, similar RNA structures that regulate genes in the thiamin metabolic pathway were identified. In addition, several putative regulatory structures were discovered upstream of genes involved in other metabolic pathways including glycerol metabolism and ethanol fermentation. The results demonstrate that this computational approach is a powerful tool for discovery of important RNA structures within prokaryotic organisms.

Keywords: Evolutionary computation; RNAMotif algorithm; RNA structure discovery; Thiamin regulation; Thi-box

1. Introduction

RNA structures found in the untranslated regions (UTRs) of genes are often involved in gene regulation both at the transcriptional and translational levels through interaction with regulatory proteins, other RNAs, or small molecule ligands. These RNA structures are found in the UTRs of genes encoding similar products from different organisms (Bonner et al., 2001; Diwa and Belasco, 2002). Moreover, similar RNA structures function as *cis*-regulators for sets of genes involved in a particular metabolic pathway within a given species (Miranda-Ríos et al., 2001; Stormo and Ji, 2001; Winkler et al., 2002; Rodionov et al., 2002; Mandal et al., 2003; Winkler et al., 2003). Biochemical and genetic evidence has proven the importance of RNA sequence and structure for regulatory function in many cases, but computational approaches have also been used to analyze the link between structure and function (Van Helden et al., 1998; Gorodkin et al., 2001a,b; Macke et al., 2001; Lesnik et al., 2001; Unniraman et al., 2001, 2002). We have previously

* Corresponding author. Tel.: +1 760 603 2652; fax: +1 760 603 4653.
E-mail address: rsampath@isisph.com (R. Sampath).

0303-2647/$ – see front matter © 2004 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2004.11.002

demonstrated that the RNAMotif algorithm (Macke et al., 2001) can be used successfully to mine nucleic acid sequence databases for known RNA structures such as the signal recognition particle (SRP) domain IV stem loop and the rho-independent transcription terminator (Lesnik et al., 2001, 2002; Fogel et al., 2002). We have further shown that evolutionary computation can be used for the discovery of conserved RNA structures across whole bacterial genomes (Fogel et al., 2002).

There are examples of translational regulation by small molecule–mRNA interactions in prokaryotes (Stormo and Ji, 2001). The "thi-box" RNA structures of *Escherichia coli* are conserved RNA structures that co-regulate expression of genes involved in thiamin metabolism. These structures are found within the 5′UTRs of three *E. coli* thiamin operons, *thiC*, *thiM*, and *tbpA*, and are also conserved in other prokaryotes (Miranda-Ríos et al., 1997, 2001; Rodionov et al., 2002). Cobalamin and riboflavin metabolic pathways also appear to be regulated by interaction between mRNA and the metabolite (Stormo and Ji, 2001; Gelfand et al., 1999; Nou and Kadner, 2000). These examples lead to the speculation that RNA structures will bind to and regulate other vitamins and cofactors. In this paper, we describe a strategy to search for regulatory structures within *E. coli* that share structural geometry similar to the thi-box hairpin. This strategy makes use of RNAMotif and evolutionary computation, in addition to a variety of post-processing filters, to arrive at a set of putative transcriptional regulatory structures, some of which match previously known riboswitches with experimental validation. This new approach helps reduce the time required to scan genomes for putative regulatory features, and very likely decreases the number of false positives in the resulting output, leading to faster downstream experimental verification of a putative list of features with high probability of importance.

2. Experiments

The complete genome for *E. coli* K-12 MG1655 (NC_000913) (Blattner et al., 1997) was downloaded from GenBank. The algorithm RNAMotif makes use of RNA structure descriptors as hypotheses to be tested for their abundance and positions relative to a nucleotide sequence. RNAMotif descriptors take the form of single strand, stem, bulge, internal loop, terminal loop characteristics with length and/or sequence specificity for each region in the structure hypothesis. An RNAMotif descriptor similar to the *E. coli* thi-box structure (Fig. 1a), but containing relaxed structural constraints (Fig. 1b), was used in initial experiments. Nucleotide sequences can be threaded through these descriptors with any locations that conform to the structure hypothesis saved to file. When the complete *E. coli* genome and the descriptor in Fig. 1b are used in the context of RNAMotif, more than 4×10^5 locations ("hits") result, with only a small fraction of these hits located in 5′UTRs (Table 1, database 1). Elements outside of the 5′UTRs were considered to have less likelihood of involvement in mRNA regulation. Thus, additional filters as described below and in Table 1 were required to in-

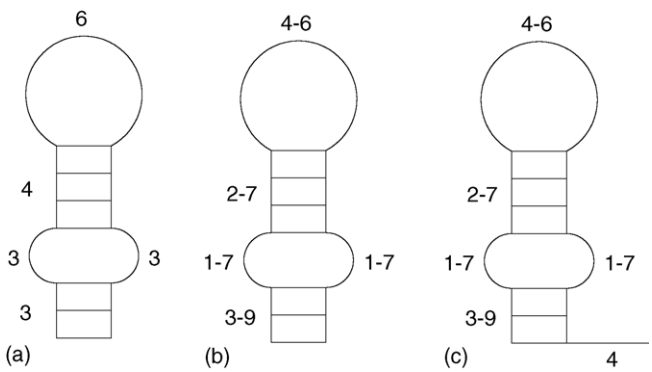

Fig. 1. Schematic of descriptors used in RNAMotif algorithm search: (a) the *E. coli* thi-box structure (Miranda-Ríos et al., 2001) descriptor; (b) descriptor with relaxed structural constraints; (c) structure descriptor with four additional single stranded nt at the 3′ end used to eliminate rho-independent transcription terminators. Ranges in stem, loop, and internal loop length are noted on the figure.