ORIGINAL ARTICLE 10.1111/j.1469-0691.2012.03864.x

Elucidating metabolic pathways and digging for genes of unknown function in microbial communities: the riboswitch approach

A. Gutiérrez-Preciado *and* E. Merino

Department of Molecular Microbiology, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, México

Abstract

In the current post-genomic era, only 3% of all genes have been annotated based on experimental evidence. Even though functions can readily be predicted for many genes, 25% of these are likely to be wrong. The most widely used methods for function prediction rely on sequence similarity, which might be misleading in many cases. Other methods such as genomic context or phylogenetic profiles have been developed to increase gene annotation accuracy; nevertheless these are only efficient when complete genome sequences are available. Here we propose a new approach based on riboswitch identification. Riboswitches are highly conserved regulators of gene expression located in the 5′ untranslated region of certain genes. When transcribed they adopt three-dimensional structures that recognize their ligands with great affinity and specificity. This specificity is a key issue for our method, allowing functional assignment with great accuracy.

Keywords: Gene function, gene regulation, genome annotation, riboswitches, T box
Original Submission: 22 November 2011; **Revised Submission:** 20 March 2012; **Accepted:** 20 March 2012
Editors: A. Moya, Rafael Cantón, and D. Raoult

Clin Microbiol Infect 2012; **18** (Suppl. 4): 35–39

Corresponding author: Enrique Merino, Instituto de Biotecnología, UNAM, Av. Universidad #2001, Col. Chamilpa, C.P. 62210, Cuernavaca, Morelos, México
E-mail: merino@ibt.unam.mx

Introduction: function assignation of recently sequenced genes

A common task in life sciences is the assignment of a biological function to a recently sequenced gene. Massive genomic and metagenomic sequencing projects have unveiled a large collection of genes whose function remains to be determined. On average, one-third of the genes from a given genome have poorly understood or unknown functions. The most common method for assigning a biological function to a new gene is through homology inference. In the best scenario, a similarity search (e.g. using BLAST) will find a clear homologue with a known function. In these cases, it is likely that both genes have the same or a similar function. However, many proteins have high sequence similarity despite performing different functions (e.g. paralogues TrpE and PabA [1]). The converse is also true, where proteins with low similarity can present the same structure and function (e.g. AroE and YdiB [2]). Only 3% of today's sequences are annotated based on experimental evidence and it has been estimated that over 25% of existing sequences are annotated incorrectly [3].

The riboswitch approach

Knowing how a particular gene is regulated can provide insights on its overall nature, the metabolic pathway in which it participates and even, in some cases, the conditions in which it is expressed. Since gene regulation in bacteria mainly takes place at the transcriptional level, identification of regulatory elements in the upstream region of transcription units is of crucial relevance. The first models of bacterial regulation were dependent on regulatory proteins, the binding sites of which tend to be short and with a low degree of conservation. Thus, gene function assignment based on the prediction of regulatory elements has been poorly explored. Recently, a new family of regulatory elements has gained importance: riboswitches. In contrast to regulatory proteins, riboswitches are not free molecules; they are part of the

transcription unit that carries the genes that will be regulated, either by activating or repressing their expression. Since 2001, a total of 14 riboswitch families have been described and experimentally characterized. Most of them bind small metabolites with high affinity and specificity, such as amino acids, vitamins or nucleotides. By definition, riboswitches recognize their target molecules in complete absence of proteins, thanks to their complex RNA structure that imposes a high level of conservation at both the three-dimensional level and the underlying nucleotide sequence [4]. These features contribute to placing riboswitches as excellent candidates to be identified with great accuracy in comparative genomic studies. Furthermore, a significant number of genes in a genome are regulated by a riboswitch. For example in *Bacillus subtilis*, 110 out of 4105 genes are regulated by a riboswitch.

As a representative example of the scope of our riboswitch approach, we present an analysis of gene function assignation using the T box, which was the first class of riboswitch to be identified. In contrast to most riboswitches, which recognize small metabolites, the T box recognizes specific tRNAs and distinguishes between their charged/uncharged states. This is due to Watson–Crick interactions that occur between the T box and the anticodon on the tRNA (Fig. 1). For each operon that is regulated by a T box, a specific tRNA is recognized and the relative levels of its charged/uncharged state determine if transcription will proceed (Fig. 1) [5,6]. Many genes that are under the control of this regulatory element are involved in increasing the intracellular levels of a specific amino acid, e.g. amino acid biosynthetic genes, transporter genes and regulatory genes. Proteins that aminoacetylate a specific tRNA can also be regulated by the T box mechanism [11].

Aims and impact

Given the specific recognition by the T box riboswitch for each of the different tRNAs, the aim of this project was to develop a new method to infer the biochemical or metabolic function of genes in metagenomic sequences. The method can also be applied to any other set of genes for which functional evidence is lacking (e.g. complete genomes or whole-genome shotgun projects).

Methods

The main steps of our method for functional inference are as follows.

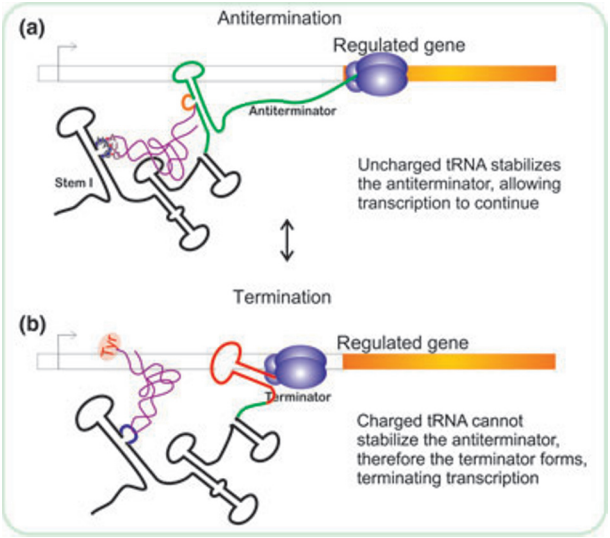

FIG. 1. Model of the T box regulatory mechanism. Structural model of the *Bacillus subtilis tyrS* T box leader RNA, as originally described by Henkin *et al.* [16]. The standard T box leader RNA arrangement consists of three major elements, stem I, stem II and stem III, plus the stem IIA/stem IIB pseudoknot and the competing terminator and antiterminator structures; the drawing is simplified. The Specifier loop, shown in blue, is an internal bulge in stem I and contains the Specifier sequence (zoomed in (a)): UAC residues complementary to the anticodon sequence of $tRNA^{Tyr}$. The antiterminator structure (green) has a bulge (orange) that interacts with the unpaired residues at the acceptor end of an uncharged tRNA. During the transcription of the leader region by RNA polymerase (purple ovals), the nascent RNA folds into a structure competent for binding of the cognate tRNA at two sites. The binding of uncharged tRNA (a) to both the Specifier sequence and the antiterminator bulge stabilizes the antiterminator (green RNA segment), preventing the formation of the terminator. This allows transcription to proceed into the downstream coding sequence (orange box). Charged tRNA (b) (represented by Tyr attached to the 3′ end of the tRNA) can interact with the Specifier sequence but cannot interact with the antiterminator; a failure to stabilize the antiterminator allows the formation of the terminator helix (red RNA segment), and transcription is terminated before the downstream coding region can be transcribed.

1 *Operon predictions.* Completely sequenced genomes as well as metagenomes were used as input. The transcription units (operons) in each of the genomic and metagenomic sequences were predicted based on the intergenic distances and on the functional relationships of the protein products of contiguous genes, obtained from the STRING database (as described previously [7]).
2 *T box identification.* The upstream region of each operon was used to identify the T box regulatory element by two complementary bioinformatic approaches: sequence