Available online at www.sciencedirect.com
ScienceDirect
Biochimie 89 (2007) 1276–1288
BIOCHIMIE
www.elsevier.com/locate/biochi

New computational methods reveal tRNA identity element divergence between Proteobacteria and Cyanobacteria

Eva Freyhult[1], Yuanyuan Cui[1], Olle Nilsson, David H. Ardell*

The Linnaeus Centre for Bioinformatics, Uppsala University, Box 598, 751 24 Uppsala, Sweden

Received 11 May 2007; accepted 6 July 2007
Available online 31 July 2007

Abstract

There are at least 21 subfunctional classes of tRNAs in most cells that, despite a very highly conserved and compact common structure, must interact specifically with different cliques of proteins or cause grave organismal consequences. Protein recognition of specific tRNA substrates is achieved in part through class-restricted tRNA features called tRNA identity determinants. In earlier work we used TFAM, a statistical classifier of tRNA function, to show evidence of unexpectedly large diversity among bacteria in tRNA identity determinants. We also created a data reduction technique called function logos to visualize identity determinants for a given taxon. Here we show evidence that determinants for lysylated isoleucine tRNAs are not the same in Proteobacteria as in other bacterial groups including the Cyanobacteria. Consistent with this, the lysylating biosynthetic enzyme TilS lacks a C-terminal domain in Cyanobacteria that is present in Proteobacteria. We present here, using function logos, a map estimating all potential identity determinants generally operational in Cyanobacteria and Proteobacteria. To further isolate the differences in potential tRNA identity determinants between Proteobacteria and Cyanobacteria, we created two new data reduction visualizations to contrast sequence and function logos between two taxa. One, called Information Difference logos (ID logos), shows the evolutionary gain or retention of functional information associated to features in one lineage. The other, Kullback–Leibler divergence Difference logos (KLD logos), shows recruitments or shifts in the functional associations of features, especially those informative in both lineages. We used these new logos to specifically isolate and visualize the differences in potential tRNA identity determinants between Proteobacteria and Cyanobacteria. Our graphical results point to numerous differences in potential tRNA identity determinants between these groups. Although more differences in general are explained by shifts in functional association rather than gains or losses, the apparent identity differences in lysylated isoleucine tRNAs appear to have evolved through both mechanisms.

Keywords: tRNA identity; Function logos; tRNA identity determinants; Lysylated isoleucine tRNA; TilS; Aminoacyl-tRNA synthetase; Proteobacteria; Cyanobacteria; Kullback–Leibler Divergence

1. Introduction

1.1. tRNA identity: recent computational advances

Both the basic function of protein synthesis and its fidelity depend on the specificities with which tRNAs are selected in two independent and separate processes: by ribosomes at the mRNA-programmed A-site, and by various enzymes involved in specialized biosynthesis and maturation reactions in the cytoplasm. Some classes of tRNAs play special roles in translation (like initiator tRNAs), are selectively modified

Abbreviations: kIle model, lysylated isoleucine tRNA TFAM model; ID, Information Difference; KLD, Kullback–Leibler divergence difference; TFAM, tRNA family; tRNA, transfer RNA; tDNA, tRNA gene; CTD2, second carboxy-terminal domain in TilS; MetRS, methionyl-tRNA synthetase; IleRS, isoleucyl-tRNA synthetase.

* Corresponding author. Tel.: +46184716694; fax: +46184716698.
E-mail address: david.ardell@lcb.uu.se (D.H. Ardell).

[1] These authors contributed equally to this manuscript.

0300-9084/$ - see front matter © 2007 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.biochi.2007.07.013

post-translationally (like lysylated isoleucine tRNAs), or are used for non-canonical translation (such as selenocysteine tRNAs). Therefore there are more than 20 functional classes of tRNAs that must be distinguished by cytoplasmic enzymes despite their uniformly small size and remarkable secondary and tertiary structural similarity [1].

The structural features that promote recognition of tRNA substrates by specific enzymes are called tRNA identity determinants. Structural elements also exist in tRNAs to prevent indiscriminate interactions with enzymes, which are called tRNA identity antideterminants. Together identity determinants and antideterminants are called tRNA identity elements [2].

Although many experimental advances continue to be made defining tRNA identity elements, these have mostly been restricted to a few model organisms, and are usually pursued for one tRNA functional class, or perhaps a few classes, in a single investigation. This reduces perspective on two important aspects of the tRNA identity problem in general: first, that the identity elements for different tRNA functional classes must co-exist and operate compatibly as a system within cells, and second, that (for reasons discussed below) tRNA identity elements should diverge among lineages for specific functional classes and coevolve within lineages for different functional classes. In the investigation of these two aspects − one systems biological and the other evolutionary − computational analysis of genomic tDNA data has certain advantages.

To begin to address the systems biology aspect of tRNA identity using computational methods, we introduced new methods to predict and visualize potential identity elements for all tRNA functional classes simultaneously, called function logos [3]. Using function logos we recapitulated known identity elements and were able to predict novel identity elements that have since been confirmed by others [4].

To address the bioinformatic and evolutionary aspects of tRNA identity, we introduced a statistical classifier of tDNA function called TFAM [5]. TFAM provides family-specific models of tRNA functional classes (originally in bacteria only) built from a largely pre-genomic database of experimentally characterized tDNA and tRNA sequence data deriving mostly from γ-Proteobacteria and Firmicutes [6]. TFAM models exploit the information of identity elements templated in tDNAs to classify tDNAs independently of the anticodons they template. They also provide positive or negative scores of any tDNA sequence against every family-specific model, that are useful for detecting functional sequence variation and other abnormalities in tDNAs. This feature allowed us to discover that a clade of α-Proteobacteria has lost or altered a $tRNA^{His}$ identity element nearly universally conserved in prokaryotes and may have lost another universally conserved $tRNA^{His}$ identity element [5]. The loss of the universally conserved identity element has since been experimentally confirmed [7].

In a statistical analysis of TFAM score variation across bacterial phyla, we showed evidence of substantial, significant and widespread functionally significant variation in tDNA sequences among Bacteria [5]. The overall functional sequence variation that we found among bacterial tDNAs could be largely explained by taxonomic variation in average tDNA base-content [5]. However, many other factors and evolutionary phenomena likely contribute to diversity in tRNA identity elements across the Tree of Life. Verification and rationalization of these results require deeper and more specific investigations.

1.2. Lysylated isoleucine tRNAs

In bacteria, the cytidine in position 34 (C34), at the first anticodon position of a special class of isoleucine tRNAs, are post-transcriptionally modified to lysidine (symbolized as L [8]) by a recently characterized enzyme called TilS [9]. This modification simultaneously changes both the codon reading specificity of this usually minor isoleucine isoacceptor from AUG to AUA and its amino acid charging specificity from methionine to isoleucine in keeping with the genetic code [10].

Among its other applications, TFAM is intended to improve the annotation of tDNAs that are routinely misclassified in genome sequencing projects. Common tDNA misannotations stem from the fact that the two tRNA gene-finders in widespread use, tRNAscan-SE [11] and ARAGORN [12], classify tDNAs on the basis of their anticodons. Genes templating lysylated isoleucine tRNAs (symbolized $tDNA^{Ile}_{CAT}$) and initiator methionine tRNAs (symbolized $tDNA^{fMet}_{CAT}$) share the same CAT anticodon template with methionine elongator tDNAs ($tDNA^{Met}_{CAT}$), so that the two classes of tDNAs are routinely misannotated in bacterial genome projects. The initial release of TFAM contained a model to annotate initiator tDNAs in bacteria [5]. A more recent release (starting with version 1.0) now provides models to annotate initiator tDNAs in eukaryotes and archaea, as well as a preliminary model for $tDNA^{Ile}_{CAT}$ based on proteobacterial data [13].

Because an unmodified $tRNA^{Ile}_{CAU}$ is charged with methionine [10], this class of tRNA shares methionine identity elements with methionine elongator tRNAs. The high similarity of the respective tDNAs templating these classes combined with erroneous classifications in training data made it difficult to define distinct models for them in earlier versions of TFAM. However, the elements that identify $tRNA^{Ile}_{CAU}$ for post-transcriptional modification by TilS are themselves genetically templated in the tRNA gene [9,14], and it is possible for TFAM to distinguish this class of tDNA with good confidence when suitably trained [15].

Nonetheless, both experimental [14] and bioinformatic [15] evidence shows that the determinants for this class have diverged in bacteria. In *Aquifex aeolicus*, TilS is missing a second C-terminal domain (CTD2) present in *Escherichia coli* TilS [14]. This domain binds and recognizes identity elements in the acceptor stems of $tRNA^{Ile}_{CAU}$. Consistent with the differences in their domain organization, *A. aeolicus* and *E. coli* TilS recognize different features in their respective $tRNA^{Ile}_{CAU}$ substrates: *E. coli* TilS recognizes major identity elements in the acceptor and anticodon stems while *A. aeolicus* TilS recognizes only the anticodon stem element or elements [14].

There is bioinformatic evidence for divergence in $tRNA^{Ile}_{CAU}$ identity determinants beyond those between *E. coli* and *A. aeolicus*. It was recently shown that certain bacterial phyla such as