Research in Microbiology 167 (2016) 669–677

www.elsevier.com/locate/resmic

Original article

Codon usage bias in phylum *Actinobacteria*: relevance to environmental adaptation and host pathogenicity

Devi Lal [a,b], Mansi Verma [a,c], Susanta K. Behura [d], Rup Lal [a,*]

[a] *Molecular Biology Laboratory, Department of Zoology, University of Delhi, Delhi 110007, India*
[b] *Department of Zoology, Ramjas College, University of Delhi, Delhi 110 007, India*
[c] *Department of Zoology, Sri Venkateswara College, University of Delhi South Campus, New Delhi 110 021, India*
[d] *ARSC, Division of Animal Sciences, Room # S128, University of Missouri, Columbia, MO 65211, USA*

Received 2 September 2015; accepted 10 June 2016
Available online 24 June 2016

Abstract

Actinobacteria are Gram-positive bacteria commonly found in soil, freshwater and marine ecosystems. In this investigation, bias in codon usages of ninety actinobacterial genomes was analyzed by estimating different indices of codon bias such as Nc (effective number of codons), SCUO (synonymous codon usage order), RSCU (relative synonymous codon usage), as well as sequence patterns of codon contexts. The results revealed several characteristic features of codon usage in *Actinobacteria*, as follows: 1) C- or G-ending codons are used frequently in comparison with A- and U ending codons; 2) there is a direct relationship of GC content with use of specific amino acids such as alanine, proline and glycine; 3) there is an inverse relationship between GC content and Nc estimates, 4) there is low SCUO value (<0.5) for most genes; and 5) GCC−GCC, GCC−GGC, GCC−GAG and CUC−GAC are the frequent context sequences among codons. This study highlights the fact that: 1) in *Actinobacteria*, extreme GC content and codon bias are driven by mutation rather than natural selection; (2) traits like aerobicity are associated with effective natural selection and therefore low GC content and low codon bias, demonstrating the role of both mutational bias and translational selection in shaping the habitat and phenotype of actinobacterial species.

Keywords: Actinobacteria; Codon usage; Pathogenicity; Codon adaptation index; Effective number of codons; Highly expressed genes

1. Introduction

It is well known that synonymous codons are not used at equal frequencies in coding amino acids during protein synthesis [1,2]. Some are preferred codons which are used more frequently than alternate synonymous codons. These are under the constant influence of mutational bias and translational selection [3,4]. The effective number of codons (Nc) are usually contoured by the GC content at the 3rd base position, leading to domination of mutational bias over translational selection [5]. Nevertheless, the genomic GC content has a considerable impact on the GC content of first and second codon positions as well, contributing to translational selection [6]. Translational selection also operates on highly expressed genes, leading to strong codon bias [7]. Apart from gene expression, codon bias is known to be affected by various factors like GC content, lifestyle, generation time and genome size [8–11]. Bias in codon usages is also known to confound molecular phylogenetic reconstruction among species [12].

Various statistical methods have been developed to evaluate codon bias. These include the CAI (codon adaptation index) [13], the Nc (effective number of codons) plot [14], the SCUO (synonymous codon usage order) plot [15], etc. Highly expressed genes (such as ribosomal protein genes) tend to have a

* Corresponding author.
E-mail addresses: dev83.sharma@gmail.com (D. Lal), mansiverma20@gmail.com (M. Verma), behuras@missouri.edu (S.K. Behura), ruplal@gmail.com (R. Lal).

http://dx.doi.org/10.1016/j.resmic.2016.06.003
0923-2508/

high CAI [13]. Although the CAI remains the gold standard for codon bias indices, the drawback with CAI is that it uses a set of reference genes [16]. On the other hand, the Nc value measures the extent of codon preference, which is independent of gene length and amino acid composition [14]. However, Nc is a non-directional measure of codon usage: it will vary if there is selection on G/C (bias in codon usages) as well as if there is mutational bias towards A/T. Thus, it is essential to compare Nc with a directional measure of codon bias (such as the CAI) to determine if genes are truly biased in codon usages [17].

The SCUO is an information-based measure and has been used for estimating non-random usages of synonymous codons [15]. It has also been observed that the efficiency of translation of some codons depends on the nucleotides adjacent to that codon, a condition called the codon context [18]. This may be due to the presence of such codon pairs in the ribosome A and P decoding sites [19].

With the availability of whole genome sequences, coding sequences of all genes predicted from genome sequences are now used to understand evolution and codon bias among related species [20–22]. *Actinobacteria* is one of the largest bacterial phyla, and is known for its variations in GC content and genome size [23]. GC content among sequenced actinobacterial genomes varies from 42 to 74.4% [23]. Comparative genome analyses among *Actinobacteria* have provided new insight into the structure and evolution of species composition within the phylum [23,24]. *Actinobacteria* are commonly found in soil, freshwater and marine environments. These bacteria play an important role in the decomposition of organic materials, including cellulose and chitin. *Actinobacteria* are producers of extracellular enzymes [25] and many secondary metabolites [26]. However, some genera of *Actinobacteria* inhabit plants and animals, and often act as disease-causing pathogens. Among these, the genus *Mycobacterium* includes the species *Mycobacterium tuberculosis* and *Mycobacterium leprae*, well known pathogens toward humans [27]. Some Corynebacterial species are known to cause diseases in plants and animals [28]. Such host diversity of *Actinobacteria* is thought to be due to the diverse lifestyle of the constituent species, many of which are aerobic, whereas other are either anaerobic or facultative. Although studies have shown possible associations of codon bias with the respective lifestyle of microorganisms and their generation time [9–11], it is not known whether the genome-wide pattern of codon bias can have a confounding effect on their habitat and pathogenicity. To test this hypothesis, we investigated coding sequences of 90 *Actinobacteria* to understand the genome-wide pattern of codon usage bias and the usage of codon context sequences, and we tested the hypothesis that codon bias can be a predictor of aerobic and pathogenic phenotypes of *Actinobacteria*.

2. Materials and methods

2.1. Genome selection

Ninety completely sequenced and annotated actinobacterial genomes (the number of genomes that were available at the time of analysis; see Supplementary Sheet 1) were analyzed in the present study. For calculation of different measures of codon bias, CDS within each genome were used. Both estimated genome size and the GC content of each genome were obtained from the NCBI database (http://www.ncbi.nlm.nih.gov/). The positional GC content in each of the genomes was obtained using the EMBOSS program 'cusp' [29].

2.2. RSCU (relative synonymous codon usage) analysis and amino acid usage

The RSCU value for a codon is the observed frequency of codon usage divided by the expected frequency. Therefore, RSCU > 1 represents codons that are used more frequently than expected, whereas RSCU < 1 represents codons which are used less frequently than expected [4]. RSCU values were calculated using the Anaconda program [30]. Cluster analysis of RSCU was performed using Cluster 3.0 software [31]. The similarity matrix of RSCU values based on rank order relation was used to obtain clusters which were viewed using the TreeView program (http://www.eisenlab.org/eisen/). Amino acid usage was calculated for all available coding sequences in each genome using the GCUA program [32].

2.3. Codon usage bias

For calculation of codon usage bias, Nc values (the effective number of codons) and SCUO were measured. The Nc values determine the departure of the codon usage from equal usage of synonymous codons, and range between 20 and 61, where lower values suggest extreme codon bias, and higher values indicate less or no codon bias. Nc values are independent of gene length and amino acid composition. Nc values were obtained from the DAMBE program [33]. Average Nc values were taken for each genome. SCUO is a newer approach to analyzing codon usage bias which is based on the 'Shannon Information Theory' [34]. The SCUO value ranges from '0' to '1', where 0 means no bias and 1 means highly biased. The SCUO was calculated using the CodonO software [35]. The Nc difference was calculated as: Nc (genome) − Nc (HEG)/Nc (genome), where HEG denotes highly expressed genes.

2.4. Codon context analysis

Codon context sequences correspond to the A (binds to an aminoacylated tRNA) and P (peptidyl-tRNA) sites of the ribosome during translation and may influence translational selection of genes. The Anaconda program was used to analyze codon context patterns and amino acid pair usage of all actinobacterial genomes. The Anaconda program generates 64 × 64 codon context contingency tables and can also be used to compare large genome sequences [30]. The clustering of *Actinobacteria* based on codon context patterns was also performed using the same software.