Phylogenetic Analysis of Coxsackie B Viruses Reveals Genomic Plasticity and Adaptation as Studied by Codon Usage Patterns

Ricardo Recarey1, Gonzalo Moratorio1,2, Rodney Colina2, Mónica Cappetta4,5, Rosario Uriarte4, Héctor Musto6 and Juan Cristina1*

1Laboratorio de Virología Molecular, Centro de Investigaciones Nucleares, Facultad de Ciencias, Universidad de la República, Uruguay
2Unidad de Biofísica de Proteínas, Instituto Pasteur de Montevideo, Uruguay
3Laboratorio de Virología Molecular, Regional Norte, Sede Salto, Universidad de la República, Uruguay
4Laboratorio de Biología Molecular, Asociación Española Primera de Socorros Mutuos, Uruguay
5Departamento de Genética, Facultad de Medicina, Universidad de la República, Uruguay
6Laboratorio de Organización y Evolución del Genoma, Instituto de Biología, Facultad de Ciencias, Universidad de la República, Uruguay

Abstract

Coxsackie B viruses (CVB) are associated with serious illnesses in humans. In this study, the patterns of synonymous codon usage in CVB have been studied through multivariate statistical methods. Effective number of codons (ENC) indicates that the overall extent of codon usage bias in CVB is not significant. The relative dinucleotide abundances suggest that codon usage bias in CVB genomes is influenced by underlying biases of dinucleotide frequencies. The distribution of CVB ORFs along the plane defined by the first two axes of correspondence analysis (COA) showed that different genotypes, as well as strains known to infect different cell types, are located at different places in the plane suggesting that CVB codon usage is reflecting an evolutionary process. The results of these studies suggest that CVB genomic biases are the result of co-evolution of translation adaptation to different cell environments and probably the need to escape anti-viral cell defenses.

Keywords: Coxsackie B virus; Evolution; Codon usage; Adaptation

Introduction

The genus Enterovirus of the family Picornaviridae are among the most common human viral pathogens, with more than 50 serotypes of enteroviruses that cause illness in humans [1]. Enteroviruses are the most common cause of aseptic meningitis, which is the most frequent central nervous system infection worldwide [2]. There is an increasing concern of the impact of this disease in South America [3,4]. Coxsackie viruses belong to this genus of the family and are divided into serogroups A and B, including six types in the B serogroup [5]. While Coxsackie A viruses are most commonly associated with skin exanthema, the Coxsackie B viruses (CVB) are associated with serious illnesses such as myocarditis, pericarditis, meningitis, diabetes, among other diseases [6,7]. All six types of CVB bind and enter cells through a common receptor protein, the coxsackie and adenovirus receptor (CAR) [8,9].

Due to the degeneracy of the genetic code, most amino acids are coded by more than one codon (synonymous codon usage) , and those that code for the same amino acid have been observed to be used unequally in most species. Two major models have been proposed to explain codon usage, the translation related (or selective) model and the mutational (or neutral) model. In the translation related model, there is a co-adaptation of synonymous codon usage and tRNA abundance to optimize translational efficiency, and a correlation between codon usage and gene expression is found at levels of speed and accuracy [10]. In the mutational model, genomic compositional constraints influence the probability of mutational fixation. This has been found in many species [11,12]. However, these two models are not mutually exclusive [10].

Understanding the extent and causes of biases in codon usage is essential to the comprehension of viral evolution, particularly the interplay between viruses and the host cell, as well as between viruses and the immune response [13].

In order to gain insight into the processes governing the evolution and host adaptation of CVB, we have performed a detailed analysis of codon usage biases of CVB strains, including strains isolated from Uruguayan children with diagnosis of aseptic meningitis, as well as strains isolated elsewhere from different patients and disease conditions. The results of these studies provide clues into the variation of codon usage pattern among the CVB genomes and the genomic plasticity of CVB in relation to adaptation to different cell hosts.

Materials and Methods

Patients

Cerebrospial fluid (CSF) samples from Uruguayan children with diagnosis of aseptic meningitis were obtained from Hospital Asociación Española Primera de Socorros Mutuos, Montevideo, Uruguay. All procedures were performed and approved under the ethics protocols and standards required and approved by hospital authorities.

RNA extraction, cDNA synthesis and amplification

RNA was extracted from 140 μl CSF samples with the QIAamp viral RNA Kit (QIAGen) according to the manufacturer’s instruction. The extracted RNA was eluted from the columns with 50 μl RNase free water. cDNA synthesis and PCR amplification of the CVB polymerase (3D pol) were carried out as previously described [14]. Amplicons were purified using the DNA extraction Kit (Fermentas).

*Corresponding author: Juan Cristina, Laboratorio de Virología Molecular, Centro de Investigaciones Nucleares, Facultad de Ciencias, Uruguay, Tel: +596 2525 09 01; Fax: +598 2525 08 95; E-mail: cristina@cin.edu.uy
Received February 06, 2012; Accepted March 07, 2012; Published March 09, 2012


Copyright: © 2012 Recarey R, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
Sequences

DNA sequencing was performed with an ABI Prism Big Dye Terminator Cycle Sequencing Ready Reaction Kit® and an ABI Prism 3730 Genetic Analyzer (both from Applied Biosystems, Foster City, CA, USA) at Institute Pasteur Montevideo, Uruguay. Sequences from CVB polymerase (3D pol), from positions 5907 through 6287 (according to reference strain X05690) were obtained. Complete CVB genome sequences were obtained for all available CVB strains by means of the use of ARSA at DDBJ database (available at: http://arsa.ddbj.nig.ac.jp/html and EMBL database (available at: http://www.ebi.ac.uk/embl/Access/index.html). For strain names, accession numbers, and genotypes, see Supplementary Information Table 1.

Sequence alignment

Sequences were aligned using the MUSCLE program [15].

Phylogenetic analysis

In order to assign CVB strains isolated from Uruguayan patients, a phylogenetic analysis was performed. For this purpose, the FindModel program [16] was used to identify the optimal evolutionary model that best fitted our sequence datasets. Akaite Information Criteria and the hierarchical likelihood ratio test indicated that the GTR+I model was the best fit to the dataset. Maximum likelihood phylogenetic trees were constructed under the GTR+I model using software from the PhyML program [17]. As a measure of the robustness of each node, we used an approximate Likelihood Ratio Test (aLRT), which demonstrates that the branch studied provides a significant likelihood against the null hypothesis that involves collapsing that branch of the phylogenetic tree but leaving the rest of the tree topology identical [18]. aLRT was calculated using a Shimodaira-Hasegawa-like procedure (SH-like) [19,20]. All Uruguayan strains were assigned to genotype B4 (see Supplementary Information Figure 1).

Codon usage analysis

In order to investigate the extent of codon usage bias in CVB, we constructed two different datasets, one composed of partial 3D pol sequences and another composed by all available CVB complete genome sequences. The effective number of codons (ENC) and the frequency of use of ARSA at DDBJ database (available at: http://arsa.ddbj.nig.ac.jp/html and EMBL database (available at: http://www.ebi.ac.uk/embl/Access/index.html). For strain names, accession numbers, and genotypes, see Supplementary Information Table 1.

Correspondence analysis

COA is an ordination technique that identifies the major trends in the variation of the data and distributes genes along continuous axis in accordance with these trends. COA creates a series of orthogonal axis to identify trends that explain the data variation, with each subsequent axis explaining a decreasing amount of the total variation across the sample [23]. Each ORF is represented in a 59-dimensional space and each dimension is related to the relative synonymous codon usage (RSCU) value of each triplet (excluding AUG, UGG and stop codons). This was also done using with CodonW.

Statistical analysis

Correlation analysis was carried out using Spearman’s rank correlation analysis method [24] (available at: www. wessa. net).

Nucleotide sequence accession numbers

The nucleotide sequences obtained for the 3D pol region of CVB strains isolated in Uruguay were deposited in the EMBL Database under accession numbers FR865902 to FR865910.

Results

Base compositional constraints on CVB codon usage bias

An ENC-GC₃S plot (the effective number of codons, ENC, plotted against G+C content at the synonymous third positions, GC₃S) can be used as a strategy to investigate patterns of synonymous codon usage. Genes, whose codon choices are constrained only by a G+C mutation bias, will lie on or just below the curve of predicted values if codon usage is only due to mutational bias [21]. To investigate whether the G+C compositional variation may determine differences in codon usage among CVB strains, ENC and GC₃S values were calculated and plotted for 3D pol sequences from all CVB strains enrolled in these studies, which accounts for 5,080 codons (for strain included in these studies, see Supplementary Information Table 1). The results of these studies are shown in figure 1. Mean values for ENC of 56.36 ± 4.63 were obtained. Since all ENC values for CVB genes are > 40, the results found suggests that the extent of codon usage bias in CVB may be low. When the GC₃S values were calculated and the ENC-GC₃S plots constructed all spots lie just on or bellow the expected curve, indicating that the codon usage bias may be influenced by the G+C compositional constraints (see Figure 1).

When the same studies were repeated using other dataset, exclusively composed by complete genomic sequences of all available CVB strains, which accounts for 61,171 codons, roughly the same results are found (for strains included in these studies, see Supplementary Information Table 1). A mean value for ENC of 54.54 ± 0.55 was obtained and all spots lie bellow the expected curve (see Figure 1). Moreover, since all
ENC values are > 50, we can again conclude that the codon usage bias in CVB is probably very slight.

Since codon usage by its very nature is multivariate, it is necessary to analyze the data using multivariate statistical techniques (i.e. Correspondence Analysis, COA) in order to confirm these findings. Moreover, COA has the advantage that it does not assume that the data falls into discrete clusters and therefore can represent continuous variation accurately [23]. The correlation between the position on the first axis generated by COA for each gene and the respective GC₃S values of each strain was analyzed using the complete CVB codes dataset. We have found that the position of the sequences on the first axis from COA are not correlated with the GC₃S values (r = -0.030, P = 0.865, respectively).

Taking all together, these results suggest that factors other than gene composition contribute to codon usage among CVB strains.

**Codon usage variation among different CVB genomes**

In order to detect the possibility of codon usage variation in different CVB genomes, the complete code ORFs of all available CVB strains and enrolled in these studies were divided according to their genotype (CVB 1 to 6, see Supplementary Information Table 1). A COA was performed on the RSCU values of each CVB ORF and the distribution of the six genotypes along the first two principal axes generated by the analysis was determined (Figure 2).

Surprisingly, the distribution of the six genetic groups in the plane defined by the first two major axes showed that different genotypes were located at different places, suggesting that different CVB genotypes exhibit differences in their codon usage patterns. Moreover, CVB strains isolated from different cell types and strains previously known to be associated with different diseases manifestations are also located in different sectors of the plane (Figure 2).

**Dinucleotides frequencies and codon usage in CVB genomes**

It has been suggested that dinucleotide frequencies can affect codon biases [25]. To study the possible effect of dinucleotide frequencies in CVB codon usage, the relative abundances of the 16 dinucleotides in the ORFs of the CVB strains were determined for both partial 3D pol and complete genomes (Table 1).

Roughly similar frequencies are obtained using partial 3D pol or complete CVB genomes. Interestingly, the relative abundance of CpG showed a strong deviation from the “normal range” (mean ± S.D. = 0.526 ± 0.121 and 0.484 ± 0.020 for partial 3D pol and complete CVB codes, respectively) and was markedly underrepresented. On the other hand, the frequency of CpA was above the expected value (mean ± S.D. = 1.472 ± 0.092 and 1.429 ± 0.031 partial 3D pol and complete CVB codes, respectively) (Table 1). Moreover, when complete CVB codes are used, 10 out of the 16 dinucleotides frequencies are highly correlated with the first axis in COA (Table 2). These observations indicate that the composition of dinucleotides plays a role in the variation found in synonymous codon usage among CVB ORFs.

Besides, the position of each codon in each of the four major axes of COA was determined for complete CVB ORFs. Table 3 shows the codons for which the maximum and minimum values were obtained for each of the axes studied (i.e. the most divergent codons values), indicating bias in their use by CVB. As it can be seen in the table, most of the divergent codons were triplets coding for Arg and Leu.

**Discussion**

The results of these studies revealed that the 3 D pol region of the CVB genome contains robust phylogenetic information, in agreement with previous results [26]. Partial 3D sequences are useful for rapid genotype assignment (Supplementary Information Figure 1). Comparable results were found using partial 3D pol region or full, complete genome sequences of CVB (Figure 1 and Table 1).

The results of this work are in agreement with previous results found for other viruses such as H5N1 Influenza A Virus (mean ENC = 50.91) [27,28]; SARS (mean ENC = 48.99) [29]; foot-and-mouth disease virus (mean ENC = 51.42) [30]; classical swine fever virus (mean ENC = 51.7) [26] and Duck Enteritis virus (mean ENC = 52.17) [31]. The ENC values found for CVB are comparatively similar to the ones found for these viruses (mean ENC value of 56.36), indicating that the overall extent of codon usage bias in CVB is not significant.

No significant correlations between the first axis of COA and GC₃S values were obtained using complete CVB genomes. This result suggest that although mutational pressure is present (Figure 1) does not significantly contribute to the codon usage bias in CVB strains, suggesting that other additional factors may be also contributing to the codon usage bias observed in CVB strains [13].

Unexpectedly, the distribution of the six genetic groups in the plane generated by the two major axes of COA showed that different genotypes are distantly located (Figure 2). Since species with a close genetic relationship always present a similar codon usage pattern [32], these findings suggest that codon usage in CVB is undergoing an evolutionary process, which is probably the result of natural selection to re-adapt its codon usage to different environments (Figure 2). This is particularly important since all six types of CVB bind and enter cells through a common receptor protein (CAR) [8,9], present in different cell types that CVB can infect. Indeed, CVB is known to be associated with different cell types and disease manifestations are distantly located in the plane, suggesting a codon usage adaptation to the infected cell (Figure 2). This is also in agreement with COA analysis, where most of
the divergent codons were triplets coding for Arg and Leu (Table 3). This reveals that the use of Arg and Leu codons play also a role in the evolution and the variability observed among CVB strains.

CpG containing codons were markedly underrepresented in CVB ORF’s (Table 1). This result is in agreement with previous results found for Coronavirus [33], vertebrate-infecting members of the family Flaviviridae [34], Poliovirus [35] and Hepatitis A virus [36]. CpG deficiency has been related to the immunostimulatory properties of unmethylated CpG, which are recognized by the host’s innate immune system as a pathogen signature [13,32]. Since the vertebrate immune system relies on unmethylated CpG recognition in DNA molecules as a sign of infection, the under-representation in RNA viruses of this dinucleotide, has been so far exclusively observed in vertebrate viruses [33]. Thus, escaping from the host antiviral response may act as another selective pressure contributing to the multifactorial codon usage shaping of CVB strains [37].

The results of these studies suggest that CVB genomic biases are the result of co-evolution of translation adaptation and probably the need to escape anti-viral cell defenses. This is in agreement with the evolution rhetoric theory proposed by Vetsigian & Goldenfeld [37] in which genome biases emerge by the need to increase communication with the ever changing cell environment without changing the message.

Acknowledgements

The authors acknowledge support from Agencia Nacional de Investigación e Innovación (ANII) through project PE_ALI_2009_1_1603 and PEDECIBA, Uruguay.

References