Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
The evolution of codon usage in structural and non-structural viral genes: The case of Avian coronavirus and its natural host Gallus gallus

Paulo Eduardo Brandão*
Department of Preventive Veterinary Medicine and Animal Health, School of Veterinary Medicine, University of São Paulo, Brazil

1. Introduction

Codon usage (CU) refers to the frequency of the occurrence of each codon for at least two-fold degenerate codons (Hershberg and Petrov, 2008), i.e., it is an indication of the ‘preference’ of a genome for one or more codons if more than one codon is possible for the same amino acid.

Natural selection for efficient protein synthesis speed and folding and genetic drift based on mutation pressure that leads to a homogeneous genome and the 3rd codon’s GC% are the most evident forces under codon usage evolution that could lead to detectable codon usage bias (CUB) (Yang and Nielsen, 2008), which has been increasingly used in studies on virus and host molecular evolution.

Avian coronavirus (Nidovirales: Coronaviridae: Coronavirinae: Gamma coronaviruses), which originated approximately 4800 years ago (Woo et al., 2012) and has a large number of serotypes and genotypes, primarily infects the respiratory tract of laying hens, broilers and breeders but can also infect the kidneys, intestines and reproductive tracts of both females and males (Cook et al., 2012), depending on the pathotype. Though affinity to different classes of cell membrane glycans could be one of the explanations for the existence of the different viral pathotypes (Wickramasinghe et al., 2011), the exact mechanism for this level of diversity is still unknown.

The 27.6 kb single-stranded positive sense RNA of A. coronavirus encodes 23 proteins, and the first two-thirds of the genome contains ORF 1, which encodes 15 non-structural proteins involved in RNA transcription and replication (Masters, 2006; Ziebuhr and Snijder, 2007). Among these, the papain-like protease (PLpro) is the proteolytic processor of the N-proximal domain of polyproteins pp1a and pp1ab (Ziebuhr et al., 2000). Non-structural protein 2 (NSP2), the first in ORF 1 because the A. coronavirus lacks NSP1, has a still undefined role, though a role on global RNA synthesis has been suggested (Graham et al., 2005).

Of the structural proteins, the spike glycoprotein (S) has a strong interaction with the host immune system and is so highly polymorphic that mutations in only 10 amino acids on the amino terminal ectodomain (S1) could result in the loss of cross-reactivity (Cavanagh, 2007). While S1 allows the virus to attach to α2,3Sia, which is widespread in chicken cells (Winter et al., 2008), the carboxy terminal S2 has the capacity to fuse virus-to-cell and cell-to-cell membranes (Masters, 2006).

The nucleocapsid (N) protein binds to the genomic RNA due to its positively charged amino acid domains, and though under a more
strict mutation constraint than S, positive selection plays a role in N evolution (Kuo et al., 2013; Masters, 2006).

The codon usage of A. coronavirus has been reported to be highly to moderately biased but closer to that found in the respiratory tract of G. gallus when compared to other tissues (Brandão, 2012). However, that report was limited because codon usage was measured based on only the spike gene.

The aim of this study was to assess the evolution of codon usage in viral structural and non-structural genes and their molecular relationship with host codon usage using A. coronavirus and its natural host G. gallus as a model.

2. Materials and methods

2.1. Sequences

2.1.1. A. coronavirus

For A. coronavirus, sequences were chosen to promote diversity of geographic origin and serotype/genotypes, including the archetypical strains, with an effort to keep the same datasets if possible. Because the number of complete genomes and genes for A. coronavirus available in GenBank did not allow for the representation of such diversity, only partial genes were used in this study instead of complete ones to have the most diverse dataset possible. As the accuracy of codon usage measurements is lower for short sequences, sequences <100 codons in length (Roth et al., 2012) were not included. Sequence redundancy was avoided by keeping only one sequence if any 100% nucleotide identity was found.

Following these criteria, this study included 64 S protein sequences, codons 1–169 (14.6% of the 1162 S codons); 25 N protein sequences, codons 301–409 (26.7% of the 409 N codons); 18 NSP2 sequences, codons 1–245 (36.4% of the 673 NSP2 codons); and 15 papain-like protease sequences, codons 3–437 (99.5% of the 437 PIPro codons). The accession numbers are shown in Fig. 1. All indicated positions are relative to the complete genome of the Avian infectious bronchitis virus strain M41 (DQ834384.1).

2.1.2. G. gallus

Aiming to assess the codon usage of the different tissues in which A. coronavirus replicates in chicken, non-redundant complete codon sequences were retrieved from the GenBank database and from the G. gallus genome project for cholecystokinin, expressed in the duodenum (NM_001001741.1 and GFC_000002315.3); lung surfactant pulmonary-associated protein A1 (SFTPA1), expressed in the lungs (NM_204606.1 and GFC_000002315.3); surfactant protein A receptor, expressed in the kidneys (NM_205098.1 and GFC_000002315.3); and ovovomucin alpha subunit, expressed in the ovitod (AB046524.1 and GFC_000002315.3). As a reference, the complete G. gallus beta-actin gene (L08165 and GFC_000002315.3) was included in the analyses as a ubiquitously expressed gene.

All sequences used in this study can be found in Supplementary material 1.

2.2. Relative synonymous codon usage (RSCU)

RSCU, the relationship between the observed and the expected frequency of a codon if the synonymous codon usage is random (Roth et al., 2012), was calculated for 59 codons, excluding the single codons of methionine and tryptophan and the three stop codons, using the equation RSCU = Xij / (Σi Σj m) (Nei and Kumar, 2000), where Xij is the total count for a given codon, Σi Σj is the sum of the count for all synonymous codons regarding the amino acid under consideration and m is the number of possible isoacceptors for that amino acid, implemented in MEGA 5.0 (Tamura et al., 2011).

The continuous RSCU values from A. coronavirus and G. gallus genes were converted to binary data using the value 1 for RSCUs > 1, when a given codon was preferred for a specific amino acid, or 0 for RSCUs ≤ 1, when the codon was not preferred (RSCU < 1) or was neutral (RSCU = 1). Finally, the combined dataset of the four viral and five host genes was used to build a binary 59 characters × 132 sequences matrix (Supplementary material 2) for the presence or absence of a preferred codon, which was used to build a neighboring tree (1000 bootstrap replicates) using PAUP, version 4.1b (Swoford, 2000).

2.3. Codon adaptation index (CAI)

The CAI is a measure of codon usage derived from the geometric mean of the relative codon adaptiveness for each codon based on a set of translationally optimal codons used as a reference (Roth et al., 2012) and can be calculated according to the equation

\[
\text{CAI}_{g} = \prod_{k=1}^{61} W_{g,k}
\]

Here, \( W_{g,k} \) is the relative adaptiveness of the kth codon (61 codons; the three stop codons were excluded), and \( X_{g,k} \) is the fraction of the codon k relative to the total number of codons in the gene.

Values closer to 1 indicate a high fitness in terms of codon usage for a given codon sequence in relation to the reference system (Sharp and Li, 1987), i.e., a high adaptation of viral genes to the host.

The CAI was calculated for sequences from both A. coronavirus and G. gallus using a reference set of highly expressed G. gallus genes available in the ACUA 1.0 software (Vetrivel et al., 2007).

2.4. Effective number of codons (Nc)

\( Nc \) is a measure of the total number of different codons present in a sequence and shows the bias from equal use of all synonymous codons for a given amino acid, with each synonymous codon treated as an allele as in the calculation of the effective number of alleles in population genetics (Roth et al., 2012). \( Nc \) values range from 20 to 61, with values closer to 61 indicating a lower bias (Wright, 1990).

\( Nc \) was calculated according to the equation

\[
Nc = 2 + \left( \frac{9}{F2} + \frac{1}{F3} + \frac{5}{F4} + \frac{3}{F6} \right), \text{ where } F \text{ is the average homozgyosity for equal use of each synonymous codon for each class of degeneracy ranging from 2 to 6, using ACUA 1.0 (Vetrivel et al., 2007).}
\]

2.5. Codon selection test

The expected effective number of codons (ENC), a measure of codon usage affected only by the GC3k (the percentage of G or C at the third position of all codons in a sequence) as a result of mutation pressure and drift, was calculated using the equation

\[
\text{ENC}_{\text{exp}} = 2 + s + 29(s^2 + (1 – s)^2)^{-1}, \text{ where } s \text{ is the GC3k ranging from 0 to 100}.
\]

The ENC and simulated GC3k values were plotted as a curve together with the \( Nc \) and observed GC3k values; an \( Nc \times \text{observed plot lying on the } ENC \times \text{simulated curve indicates genetic drift/mutational bias, while plots outside the curve indicate natural selection (Wright, 1990).}

2.6. Analysis of conserved amino acids coded by preferred codons

To assess the significance of each preferred codon on the molecular evolution of A. coronavirus, 100% conserved amino acid positions coded by the preferred codon(s), i.e., those with RSCUs >1,
were counted for each gene, and the significance of the differences was assessed with Fisher’s exact test and the odds ratio (OR).

2.7. Protein selection test

To understand the relationship between codon and protein selection, the occurrence of purifying or positive selection on A. coronaviruses S, N, NSP2 and PLpro sequences was tested with Fisher’s exact test of neutrality for sequence pairs using the Nei–Gojobori method (Nei and Gojobori, 1986) for the difference between the synonymous and non-synonymous substitution distances (dS–dN) using Mega 5 (Tamura et al., 2011).

3. Results

3.1. RSCU phylogeny

Fig. 1 shows that G. gallus RSCUs segregate in a tissue-specific manner in a topology supported by bootstrap values of 100 for each gene analyzed.

Fig. 1. Neighbor-joining distance tree for the relative synonymous codon usage (RSCU) for the Avian coronavirus spike (S), nucleocapsid (N), non-structural protein 2 (NSP2) and papain-like protease (PLpro) genes and the Gallus gallus beta-actin, lung surfactant protein A (SFTPA1, gray background), intestinal cholecystokinin (CCK), oviduct ovomucin alpha subunit (OSA) and kidney vitamin D receptor genes. The tree was based on binary data using the value 1 for RSCUs > 1 (codon is preferred) or 0 for RSCUs ≤ 1 when the codon is not preferred (RSCU < 1) or is neutral (RSCU = 1). ENC (effective number of codons) values < 40 and > 45 are marked with an asterisk and a hash, respectively; sequences with ENC values between 40 and 45 have no marks. The arrow indicates the separation between G. gallus and Avian coronavirus clusters. Numbers at each node are bootstrap values (1000 replicates, only values > 50 are shown). The bar represents the codon usage preferences distance.
For the *A. coronavirus* RSCUs, all genes segregated in gene-specific clusters, except for the sequence EU26388.1 A2 PLpro, which segregated closer to the NSP2 cluster. All strains segregated in a cluster separated from *G. gallus*, with the internal nodes resulting in the genotype-specific sub-clusters for the S gene, including those for the archetypes Connecticut, Massachusetts and Arkansas, with two sub-clusters and the PLpro cluster between them. No pathotype-specific cluster was found.

Though the distinction between the *A. coronavirus* and *G. gallus* RSCUs clusters is also clear for the N, NSP2 and PLpro genes, a less resolved topology emerges because the distinction among the different genotypes is not sustained.

For all four genes, *A. coronavirus* clusters show an increasing distance from the *G. gallus* clusters, with them being closer to SFTPA1 (from the respiratory tract) and more distant from cholecystokinin (from the intestine) and with the ubiquitous beta-actin being the most distant from both *A. coronavirus* and the other *G. gallus* clusters.

### 3.2. Codon adaptation index (CAI)

Mean CAI values for the *A. coronavirus* S, N, NSP2 and PLpro genes were 0.66 (sd 0.01), 0.77 (sd 0.01), 0.69 (sd 0.01) and 0.7 (sd 0.01), respectively, while, for the *G. gallus* genes, the mean CAI was 0.81 (sd 0.06), ranging from 0.71 for the pulmonary gene SFTPA1 to 0.88 for the renal vitamin D receptor (mean values for two sequences).

A boxplot representation of *G. gallus* and *A. coronavirus* CAIs (Fig. 2) shows that, in relation to *G. gallus*, S has the lowest values (0.64–0.7) and N has the highest values (0.75–0.79), while NSP2 and PLpro have intermediate values (0.69–0.71), with non-overlapping medians.

### 3.3. Effective number of codons (Ne)

The mean Ne values for *A. coronavirus* S, N, NSP2 and PLpro were 43 (sd 2.31), 44.9 (sd 3.64), 51.33 (sd 1.56) and 43.79 (sd 0.86), respectively, and for *G. gallus*, the mean Ne values were 33.59 for vitamin D receptor, 40.03 for beta-actin, 46.48 for cholecystokinin, 50.21 for SFTPA1 and 53.01 for ovomucin.

### 3.4. Codon selection test

The Ne × GC3% graphs (Fig. 3) show that, regardless of the *A. coronavirus* gene under consideration (S, N, NSP2 or PLpro), all plots fall either just below or in the vicinity of the ENC × GC3% expected curve. This same pattern was also found for the *G. gallus* genes, though with plots dislocated to the right side of the graph due to a higher GC3% content.

### 3.5. Analysis of conserved amino acids coded by preferred codons

The number of 100% conserved amino acid positions coded by the preferred codons for genes S, N, NSP2 and PLpro was one, 20, 28 and 71, respectively. Fisher’s exact test showed that only the S gene presented a statistically significant lower number of occurrences (Table 1) when compared to the other 3 genes (p < 0.0001), with ORs

---

**Fig. 2.** Four graphs showing the expected (seen in the curves of each graph) and observed (seen in the points of each graph) effective number of codons (ENC and Ne, respectively) (Y axis) and the expected and observed GC3% (X axis) for (a) Avian coronavirus spike (S); (b) nucleocapsid (N); (c) non-structural protein 2 (NSP2) and (d) papain-like protease (PLpro) (dots) and Gallus gallus beta-actin, lung surfactant protein A, intestinal cholecystokinin, oviduct ovomucin alpha subunit and kidney vitamin D receptor (asterisks).
Table 1
Conserved amino acid (aa) positions in the *Avian coronavirus* spike (S), nucleocapsid (N), non-structural protein 2 (NSP2) and papain-like protease (PLpro) genes coded by a preferred codon and the preferred codon for each aa in the *Gallus gallus* beta-actin (B-act), lung surfactant protein A (SFTPA1), intestinal cholecystokinin (CCK), oviduct ovomucin alpha subunit (Ovo) and kidney vitamin D receptor (Vitamin D rec) genes. Tryptophan and methionine, coded by a single codon, were excluded. Codon preference was indicated by relative synonymous codon usage (RSCU) > 1. Positions are provided only for *Avian coronavirus* genes as *G. gallus* genes were used as the reference for comparison.

| G gallus | Avian coronavirus |
|----------|-------------------|
| aa       | B-act | CCK | Ovo | SFTPA1 | VitD rec | S | Position | N | Position | Nsp2 | Position | PLpro | Position |
| F        | UUC   | UUC | UUU/UUC | UUU | UUC | UUU 155 | UUU 313,390 | UUU 52,64,101,138,200,217 | UUU 28,54,57,102,144,211,220,270,326,369,393 |
| L        | CUC/CUG | CUC/CUG | CUG | UUG/CU/CAU/CUG | CUC/CUG | NC | NC | NC | NC | NC | NC | NC | CUU 16,47,58,94,111,129,273,288,315,413 |
| I        | AUU/AUC | AUU/AUC | AUU/AUC | AUU | AUU | AUU 319,397 | AUU 23,199,210 | AUU 52,133,194,226,383,429 |
| V        | GUG   | GUG | GUG/GUG | GUG/GUG | GUG/GUG | NC | NC | NC | NC | NC | NC | NC | GUU 191,192,317,323,371,394,430,435 |
| S        | UCU/UCC/AGC | UCC/AGC | UCU/UCC/UCAG/AGC | UCU/UCAG/AGC | UCC/AGC | NC | NC | UCA | 340,344 | NC | NC | NC | NC |
| P        | CCC   | CCC | CCC/CC/CCA | CCC | CCC | NC | NC | CCA | 338 | NC | NC | NC | CCA 178,294,338 |
| T        | ACC/ACA | ACA | ACA/ACC/ACA | ACA | ACC/ACA | NC | NC | NC | NC | ACU | 123,167,241 | NC | NC |
| A        | GCC   | GCU/GCG | GCU/GCC/GCA | GCU/GCA | GCG | NC | NC | GCA | 376 | NC | NC | NC | NC |
| Y        | UAC   | UAC | UAC/UAC | UAC/UAC | UAC | NC | NC | UAU | 316 | NC | NC | NC | NC |
| H        | CAC   | CAC | CAC/U/CAC | CAC | CAC | NC | NC | NF | NF | NC | NC | CAU 143,201 |
| Q        | CAG   | CAG | CAG/CAG | CAG | CAG | NC | NC | CAG | 312,369,387 | NC | NC | NC | NC |
| N        | AAC   | AAC | AAC | AAC | AAC | NC | NC | AAU | 315,385,307 | NC | NC | AAU 27,82,97,140,186,296,343 |
| K        | AAG   | AAA | AAA | AAA | AAA | NC | NC | NC | NC | AAA | 6,21,86 | NC | NC |
| D        | GAU   | GAU | GAU/GAC | GAU | GAC | NC | NC | GAU | 314,374 | NC | NC | GAU 5,105,160,176,182,218,227,288,281 |
| E        | GAG   | GAG | GAG | GAG | GAG | NC | NC | NC | NC | GAA | 98,136,142,165 | GAA 130,164,185,342 |
| C        | CGU    | CGU | CGU | CGU | CGU | NC | NC | UGU | 320,323 | UGU 68,242 | UGU 132,154,183,202,439 |
| R        | CGU/AGC | CGC/CGG/AGG | AGC/AGC | AGC/AGC | AGC/AGC | NC | NC | AGA | 349 | CGU 54,111 | NC | NC |
| G        | CGU/GGC | GCC | GCC | GCC | GCC | NC | NC | NC | NC | GCC | 56,86,177,319,402 |

NC: no 100% conserved amino acids positions coded by the preferred codon; NF: amino acid not found in the sequence.
of 21.7, 32.8 and 37.8 when compared to N, NSP2 and PLpro genes, respectively, while differences among N, NSP2 and PLpro were not significantly different (p > 0.05). The mean number of amino acids for each sequence is shown in Table 2.

The number of amino acids in the G. gallus proteins that presented the same codons used by at least one of the A. coronavirus genes in 100% conserved aa positions ranged from 1 (for vitamin D receptor) to 15 (for ovomucin alpha) and the most conserved preferred codon, found for all A. coronavirus genes, was UUU for F (Table 1). The positions of each of the conserved amino acids coded by preferred codons for A. coronavirus are also shown in Table 1.

3.6. Protein selection test

The sequences of N, NSP2 and PLpro from all the strains in this study were found to be under purifying selection as the p values from Fisher’s exact test were all above 0.05, with mean values of 0.99 for each gene and sd values of 0.06, 0.05 and 0.08, respectively. For S sequences, the mean p value was 0.97 (sd 0.13), but p values <0.05 were found between the groups of sequences FJ899690.1 Conn39528/FJ899689.1 Conn32062/FJ904716.1 Conn461996/FJ904717.1 Conn46197 and AY561711.1 M41/DQ834384.1 M41, indicating positive selection for these strains.

4. Discussion

Regardless of the gene being considered, all A. coronavirus sequences segregated in an exclusive cluster in the RSCU tree, which, despite being consistently separate from the G. gallus cluster, was closer to the SFTPA1 (a gene expressed in the respiratory tract of chicken) cluster. Taking the codon usage for these genes as a reflection of the codon usage in the respiratory tract, both structural and non-structural genes show a codon usage closer to the chicken respiratory tissue translational environment than to the reproductive, renal and enteric ones.

This similar codon usage could allow for an improved viral replication in the respiratory tract as a first site of viral replication, a feature common to all A. coronavirus strains in chickens, before the virus reaches other replication sites for each pathotype, as a result of the natural selection for codons and a more efficient translation of virus proteins, as already suggested for the S gene alone (Brandão, 2012).

Evidence of natural selection for codon usage as an evolutionary force acting upon A. coronavirus was found in the Nc × GC3% graphs (Fig. 2) because for all four viral genes, observed GC3% points fell outside the curve, indicating that codon usage for all the strains under analysis was not the sole result of the random accumulation of mutations.

Nonetheless, the Nc × GC3% plots show that A. coronavirus codon usage could also be a consequence of mutation pressure, as the points were in the vicinity of the curve, meaning that the GC% at the synonymous 3rd codon position follows the viral genomic GC% to some degree.

It must be considered that both genetic drift derived from the mutation pressure and natural selection detected for A. coronavirus could also harbor some relationship with genomic RNA secondary structure constraints and not only codon usage, as synonymous 3rd base mutations, though synonymous in terms of amino acid codification, could result in altered RNA secondary structure
The distribution of 100% conserved amino acid positions coded by the preferred codon is noteworthy when one compares the S gene with N, NSP2 or PLP, as a single position was found in a region outside antigenic and hypervariable regions (Cavanagh et al., 1988; Kant et al., 1992) in the S gene, while for the other three genes, these positions (n = 20, 28 and 71, respectively) were scattered throughout the regions considered, with statistically significant differences when compared to S (p < 0.0001, OR = 2.17–3.78).

This low number of conserved amino acid positions coded by the preferred codon in S could be an additional molecular evolutionary-ary mechanism for S antigenic diversity, as fine-tuning translation kinetics could result in high deoptimization of codon usage and a consequent increased fitness (Aragonés et al., 2010).

On the other hand, possibly due to strong structural and functional constraints, N, NSP2 and PLP have a higher number of amino acid positions coded by the preferred codon, which is the same codon preferred by the host (Table 1), which would allow higher fitness to the host transcription environment (Zhou et al., 2012) in a concerted virus–host molecular evolution.

Thus, taking conserved amino acid positions coded by the preferred codons as a selection unit, it follows from the above mentioned differences that natural selection could either be positive for these positions, leading a protein under purifying selection (e.g., N, NSP2 and PLP) to show the same codons as the host for that amino acid, or negative if a protein is under positive selection (as shown for S).

The most probable reason for the fact that non-100% conserved amino acid positions coded by a preferred codon for that amino acid (noted as NC in Table 1) were only found in A coronavirus genes and not in the G. gallus genes is that host genes are less susceptible to both the occurrence of putative amino acids and codon usage polymorphisms, contrary to what is observed and expected for virus genes.

NC might be considered to be an accurate indicator of codon usage bias because the frequency of amino acids is normalized during the analysis and does not add bias; however, similarly to the CAI, the outcome of the NC analysis is a single number, leading to a loss of deep evolutionary information similar to the loss of evolutionary information in nucleotide or amino acid distance-based phylogenetic analyses.

Taking into account informative sites during codon evolution studies, for instance, 100% conserved amino acid positions coded by the preferred codon for that amino acid, could unveil data that would otherwise be lost in the analysis and that could be used to gain a more comprehensive understanding of molecular evolution in association with the codon usage bias indicators and selection analysis.

It would be interesting to use the analyses presented herein not only for a better understanding of virus evolution but also as supporting predictors of spill-over events, such as influenza (Wahlgren, 2011) and the new human coronavirus (Kindler et al., 2013) now named MERS-CoV, for which the role of codon usage evolution in virus adaptation to new hosts has been widely ignored.

In conclusion, A. coronavirus codon usage evolves independently for each gene in a manner predictable by the protein function. Proteins with high functional and structural constraints are more adapted to G. gallus, its natural host, with a balance between natural selection and mutation pressure, giving further molecular basis for the virus’ ability to exploit the host’s environment.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.virusres.2013.09.033.
