Review Article
Volume 9:1, 2021
DOI: 10.37421/jpgeb.2021.9.211

Journal of
Phylogenetics & Evolutionary Biology

ISSN: 2329-9002 Open Access

How Codon Usage Biases Affect Our Ability to Recover the Tree of Life

Justin B. Miller[1], Michael F. Whiting[1,2], John S.K. Kauwe[1] and Perry G. Ridge[1]*

[1]Department of Biology, Brigham Young University, Provo, UT 84602, USA
[2]M.L. Bean Museum, Brigham Young University, Provo, UT 84602, USA

Abstract

Many common phylogenomic algorithms that were well-adapted to classify limited numbers of species have become increasingly intractable as large whole-genome sequencing datasets have emerged. Various novel approaches use characteristics of DNA sequences, including variations in codon usage biases, to establish the phylogenetic relatedness of species. Codon choice affects transcription and translational efficiencies, which can lead to differential protein expression and phenotypic variation that may be a target of selection. Several functional biases exist within genes, including the number of codons that are used, the position of the codons, and the overall nucleotide composition of the genome. Although recent algorithms capitalize on specific codon usage biases to improve phylogenetic tree inference, the phylogenies produced by these algorithms vary significantly and indicate different evolutionary histories. Therefore, we propose that gene-specific analyses of the phylogenetic signal of specific codon usage biases are required to best incorporate these biases in phylogenomic models.

Keywords: Codon usage bias • Phylogenetics • Ortholog • Codon aversion • Codon pairing • Ramp sequence • Phylogenomics

The Continued Importance of Phylogenetic Systematics

Phylogenetic systematics explores the historical and hierarchical relationships among genes, individuals, populations, and taxa. Phylogenies allow biologists to infer similar characteristics in closely related species and provide an evolutionary framework for analyzing biological patterns [1]. Furthermore, phylogenies are statements of homology and are used to organize shared structures or patterns between species [2]. Originally, phylogenies were recovered using only morphological data. However, with the increased availability of molecular data, a combined approach using morphology and genetic markers is typically used in phylogenetic analyses [3]. Although genetic data provide researchers with access to more species, the datasets typically require significant data cleaning (e.g., alignment and annotation) before they become useful. Some of the greatest difficulties in recovering phylogenetic trees from molecular data (e.g., multiple substitutions at the same position between ancient terminal branches or no substitutions in a gene between short internal tree branches) are explored by Philippe, Brinkmann [4]. These issues have recently become more pertinent as sequencing costs have decreased and genomic data now largely span the Tree of Life.

Codon Usage Biases Span the Tree of Life

Codon usage biases are present throughout molecular datasets. There are 61 canonical codons plus three stop codons that indicate the incorporation of 20 amino acids and the stop signal [5]. Since there are more codons than amino acids, the term synonymous codon is used to describe how multiple codons encode the same amino acid and were presumably identical in function. However, an unequal distribution of synonymous codons occurs within genomes, and highly expressed genes have especially prominent biases that suggest synonymous codons might play different roles in species fitness [6]. Furthermore, an unequal distribution of tRNA anticodons directly coupling codons also varies between species, leading to the wobble hypothesis: tRNA anticodons do not need to latch onto all three codon nucleotides during translation [7]. Codon usage is highly associated with the most abundant tRNA present in the cell [8], and codon usage patterns affect gene expression [9]. Some phylogenetic differences in synonymous codon usage biases may be explained by non-random mutations or selection for phenotypic differences caused by differential gene expression. Although codon usages directly affect phenotypes by altering gene expression, common phylogenomic approaches typically ignore the subtle influences of codon usage biases when recovering a phylogeny. Common phylogenomic approaches are described below.

Overview of Common Phylogenomic Techniques that do not Utilize Codon Usage Biases

Homologous characters are often identified by aligning orthologous gene sequences and identifying character state changes of amino acid residues or nucleotides that are then used to recover a tree topology. This multi-step process is time-consuming and requires significant data preprocessing (e.g., orthologous gene annotations). Non-homologous sequence comparisons have also been explored in alignment-free methods and will subsequently be discussed.

Ortholog identification

Orthologs are genes within two or more species that usually share the same function because they are derived from the same ancestral gene in the most recent common ancestor [10]. In contrast, paralogs and xenologs may share the same function, but can arise from gene duplication or horizontal gene transfer. Paralogs may not be under the same evolutionary pressures and should not be compared in a direct positional alignment because these comparisons are often a poor indicator of phylogenetic relationships [10]. An in-depth evaluation of ortholog identification techniques is presented by

***Address for Correspondence:** Perry G. Ridge, Department of Biology, Brigham Young University, Provo, UT 84602, USA, E-mail: pgridge@bu.edu*

Received 17 November 2020; **Accepted** 29 December 2020; **Published** 05 January 2021

Tekaia [11]. Once an ortholog is identified, phylogenetic studies typically require a multiple sequence alignment to align homologous characters. Reviews of some common multiple sequence aligners such as T-coffee [12], MUSCLE [13], Clustal [14], Clustal Omega [15], and MAFFT [16] can be examined elsewhere [17,18].

Recovering the phylogenetic tree

Maximum parsimony: Maximum parsimony assumes that each character is equally important and minimizes the number of character state changes to recover the relatedness of species. Proponents of parsimony point to its explanatory power and ability to minimize *ad hoc* hypotheses [19]. However, parsimony can be misleading if unequal evolutionary rates between lineages exist because longer evolutionary branches have a tendency to form monophyletic groups even if the species have different phylogenetic histories [20]. PAUP [21] and TNT [22] are two popular software packages to identify phylogenies based on parsimony.

Maximum likelihood: Maximum likelihood requires specific models of evolution that show the probability of character state changes and can be used in the likelihood function. Maximum likelihood calculates the probability of obtaining the data given the model and tree topology. One of the main reasons that maximum likelihood estimates have gained traction is the mathematical property of consistency, which states that as more data (i.e., phylogenetically informative characters) are added, the likelihood function will converge to the correct tree, assuming the underlying model is correct [23,24]. Furthermore, maximum likelihood takes into account more complex modeling of datasets, and the modeling has become more computationally tractable through faster algorithmic design and faster computer processors [25]. However, in contrast to maximum parsimony, maximum likelihood is more likely to separate highly divergent species, leading to long branch repulsion [26]. MEGA X [27], RaxML [28], IQ-TREE [29] and PHYLIP [30] are commonly used to recover phylogenies using maximum likelihood.

Bayesian inference: Bayesian phylogenetic estimates use posterior probabilities of a distribution of trees calculated with Markov Chain Monte Carlo (MCMC) techniques to evaluate tree probabilities. Bayesian inference adds statistical support to phylogenies and produces more accurate trees in simulations. However, Bayesian inference is highly sensitive to prior probabilities [31]. How Bayesian techniques compare to other phylogenetic methods is addressed by Yang and Rannala [32], and popular Bayesian techniques are implemented in MrBayes [33,34] and BEAST2 [35].

Distance-based and alignment-free: Distance-based phylogenies use techniques such as neighbor-joining to quickly produce relatively good trees that are often used as a starting point for phylogenetic analyses using other methods. Neighbor-joining decomposes a star tree by taking the two closest taxa based on the number of character changes between them, pairing the taxa together to form a new node, recalculating weights based on the shortest distance between the new node and all other species (or nodes), and repeating this process until all taxa are paired. Although this technique is computationally fast, compressing the sequences into distances loses information and phylogenetic reliability is difficult to ascertain from highly divergent sequences [36]. However, distance-based methods are frequently used when sequence alignments are not available or in whole genome comparisons. Since genome assembly and multiple sequence alignments affect phylogenies more than the algorithm used to recover the phylogeny, alignment-free methods attempt to recover shared phylogenetic history without an alignment by comparing basic characteristics of genomes (i.e., GC content, k-mer counts, codon usages, etc.) [37]. Broadly, alignment-free approaches can be classified into three main groups. The first group analyzes the frequency of words with a certain length (e.g., FFP [38, 39] and CV Tree [40]). The second group matches lengths of overlapping sequences (e.g., ACS [41], KMACS [42], and Kr [43]). The last group calculates informational content between sequences (e.g., Co-phylog [44], FSWM [45], andi [46], CAM [47], and codon pairing [48]). These techniques are still being developed, and new software packages are updated to recover more robust trees.

Assessing the phylogenetic tree: Bootstrapping is a common technique to assess the robustness of a phylogeny by randomly sampling characters with replacement and determining the extent to which the recovered phylogenetic tree changes. Proponents of bootstrapping point to its ability to uncover the phylogenetic signal under the noise of phylogenetically uninformative characters. Bootstrapping also has statistical properties that allow a confidence value to be placed on clades [49]. On the other hand, critics of bootstrapping (and phylogenomic algorithms in general) point to the statistical assumptions that are violated in DNA characters because DNA characters cannot be considered independently and identically distributed [49]. Furthermore, a bootstrap proportion is generally unbiased but highly imprecise, meaning the bootstrap number can give high confidence that the data support a clade even if the clade is not real [50].

Biological Construct of Codon Usage Bias

Phylogenomic studies have recently used codon usage biases to recover species relationships with or without ortholog annotations. Various codon usage biases appear to track speciation events and can cause gene expression to either increase or decrease [51]. Furthermore, codon usage biases affect protein and RNA folding, which impacts transcription and translational efficiency, as well as gene expression. Although genetic drift drives global codon usages, the majority of codon usage biases within individual genes is influenced by translational selection [52]. Figure 1 outlines how codon biases affect protein levels.

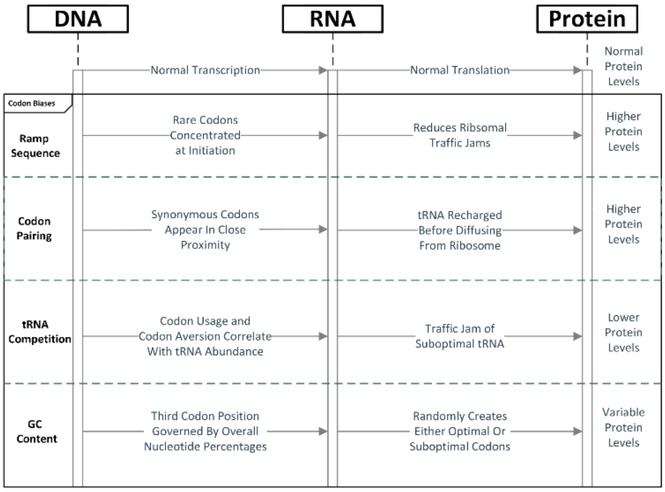
DNA
RNA
Protein
Normal Transcription
Normal Translation
Normal Protein Levels
Codon Biases
Ramp Sequence
Rare Codons Concentrated at Initiation
Reduces Ribsomal Traffic Jams
Higher Protein Levels
Codon Pairing
Synonymous Codons Appear In Close Proximity
tRNA Recharged Before Diffusing From Ribosome
Higher Protein Levels
tRNA Competition
Codon Usage and Codon Aversion Correlate With tRNA Abundance
Traffic Jam of Suboptimal tRNA
Lower Protein Levels
GC Content
Third Codon Position Governed By Overall Nucleotide Percentages
Randomly Creates Either Optimal Or Suboptimal Codons
Variable Protein Levels

Figure 1. How Codon Usage Biases Affect Protein Levels. Many types of codon usage biases directly affect DNA, RNA, and protein secondary structure. They also affect transcription and translational efficiency. The mechanisms by which ramp sequences, codon pairing, tRNA competition, and the GC nucleotide composition affect protein levels are depicted.

Codon usage metrics

Originally, the Codon Adaptation Index was used to compare the relative codon usage of the most commonly used codons within highly expressed genes [6]. This metric was soon replaced by the effective number of codons, which quantified the difference in codon usage versus the expected usage if all synonymous codons were used equally [53]. Because of their simplicity, the effective number of codons and codon adaptation index are still widely used techniques. However, those methods oversimplify the dynamics of codon usage. The tRNA adaptation index (tAI) takes into account the complex relationship between tRNA and codons by using tRNA copy number, gene length, number of codons, and the preponderance of tRNA wobble to determine codon optimality [54,55]. Building on tAI, the normalized translational efficiency (nTE) measurement balances tRNA supply and demand on codon usage and considers cellular tRNA dynamics. A codon is considered "optimal" if the relative supply of its cognate tRNAs exceeds the codon's usage [56]. Unfortunately, tAI and nTE require data that are not always available in a species and can vary between individuals and cell types, limiting their use across the Tree of Life.

Biological implications of codon usage bias

Selection toward decreased translational efficiency: Occasionally, suboptimal codons are beneficial to cells because they slow the ribosome (or polymerase) and allow for more precise, deliberate gene translation (or transcription). Codon usage biases affect mRNA secondary structure so strongly that local mRNA secondary structure can be used to predict codon usage in highly expressed genes [57]. Highly expressed genes also have a ramp of 30-50 slowly-translated, rare codons at the 5' end of most protein coding sequences [58] that serves to evenly space ribosomes [59] and reduce mRNA secondary structure [60] at translation initiation. These ramp sequences are population-specific and can also have disease implications [61]. A comprehensive analysis of ramp sequences from all domains of life, as well as a method to extract ramp sequences from individual genes is presented in Miller, Brase [62].

Additionally, the cell cycle impacts codon choice for suboptimal codons. Since tRNA expression levels are highest during the G2 phase, suboptimal codon usage for genes expressed during this phase is also highest. The G1 phase has the lowest tRNA expression, and genes expressed during G1 have a tendency toward optimal codon usage [63].

Codon usage biases in various bacteria are associated with species lifestyle [64,65]. For cyanobacteria (photosynthetic bacteria), selection toward sub-optimal codon usage produces the circadian clock conditionality, where the circadian clock is expressed only under certain environmental conditions where cyanobacteria are not intrinsically robust [66]. Similarly, the pathogenicity and habitat of *Actinobacteria* (High GC gram positive bacteria important for soil systems) also influence codon usage, where aerobic species vary significantly from anaerobic species, and pathogenic species vary significantly from non-pathogenic species [67]. In each case, codon usage alone explains bacterial adaptation to their environment.

Selection toward increased translational efficiency: Highly expressed genes tend to use more optimal codons after the ramp sequence to increase overall gene expression because once ribosomes (or polymerases) are evenly spaced; they can translate optimal codons more efficiently [51]. Faster translation is due to decreased wobble interactions, increased optimal tRNA composition, and decreased competition from synonymous codons within a gene [68]. Selective pressures for protein expression also act on mRNA sequences to optimize co-translational folding within polypeptides in over 90% of high expression genes and about 80% of low expression genes [56]. Furthermore, gene body methylation is strongly correlated with codon usage bias and appears to systematically replace CpG bearing codons, potentially influencing optimal codon establishment [69].

Recharging a tRNA while the ribosome is still attached to the mRNA strand is another strategy used to increase translational efficiency and decrease overall resource utilization. Co-tRNA codon pairing occurs when two non-identical codons that encode the same amino acid are located in close proximity to each other in a gene. Identical codon pairing occurs when identical codons are located in close proximity in a gene sequence. Co-tRNA and identical codon pairing are mechanisms to reuse a tRNA by recharging the tRNA with an amino acid before it diffuses from the ribosome, increasing translational speed by approximately 30% [70]. Although co-tRNA codon pairing occurs more prominently in eukaryotes and identical codon pairing occurs prominently in bacteria [71] and archaea [72], both co-tRNA and identical codon pairing are phylogenetically conserved in all domains of life [48].

Other systematic biases also influence codon choice. Background dinucleotide substitution biases from GC to AT and AT to GC often coincide with shifts in optimal codons [73]. Even under sustained selective pressure, GC content at the third codon position is highly correlated with overall GC content in a gene, suggesting that optimal codons are affected by genomic GC content [73]. In an analysis of 65 eukaryotes and prokaryotes, GC content accounted for 76.7% of amino acid variation [74]. A summary of mechanisms that affect codon usage bias are shown in Table 1.

Codon Usage Bias in Phylogenetic Systematics

Codon usage biases are less likely to be affected by random mutations than expected based on genomic mutation rates because codons often reside in conserved genomic regions [76]. Therefore, random mutations appear to play less of a role in phenotypic variation caused by codon usage, and the extent to which codon usage can be used in phylogenomics is currently being explored.

Codon usage in maximum likelihood

Limited codon substitution models have been used for decades in maximum likelihood estimates. However, until recently, a full 61 x 61 codon matrix was too computational intensive to apply to more than a few species and

Table 1. Mechanisms affecting codon usage biases.

Name	Location/ Domain	Description
Ramp Sequence	30-50 nucleotides downstream of start codon	The ramp sequence consists of rare, slowly translated codons that increase ribosomal spacing, reduce mRNA secondary structure, and slow initial translation.
Co-tRNA Codon Pairing	More prominent in eukaryotes. Phylogenetically conserved in all domains of life	tRNA are recharged with amino acids for synonymous codon translation when synonymous codons are in close proximity to each other. Recharging allows the tRNA to stay attached to the ribosome and significantly increases translation efficiency.
Identical Codon Pairing	All domains of life	tRNA are recharged with amino acids for identical codon translation when identical codons are in close proximity to each other. Recharging allows the tRNA to stay attached to the ribosome and significantly increases translation efficiency.
tRNA competition	Eukarya, bacteria, and archaea	Cognate, near-cognate, and non-cognate tRNA may attempt to bind to an mRNA codon. If relatively few cognate tRNA are available, translation will slow because other tRNA attempt to bind to the same codon. This process is essential for translation elongation, efficiency, and accuracy [75].
GC Content	All domains of life	Overall GC content in a gene is highly correlated with GC content at the third codon position. GC content influences over two-thirds of codon variation.

genes [77]. Somewhat surprisingly, after a 61 x 61 codon matrix became computationally viable, it was determined that the full matrix is not always optimal because models that use a fixed codon mutation rate for phylogenetic tree reconstruction fit the data better than a variable codon substitution rate. The apparent variation in codon substitution is actually caused by variable selection against amino acid substitutions in the regions used to develop the model, specifically mitochondria, chloroplast, and hemagglutinin proteins [78]. Maximum likelihood estimates that use codon models outperform a parsimony analysis only when codon usage is highly skewed and is not affected by asymmetry in substitution rates (approach validated using *Drosophila*) [79].

Because full codon models are computationally intensive and do not always elucidate more information than simpler models, common likelihood approaches use non synonymous to synonymous mutation rates per site (d_N/d_S) instead of the complete codon model. If the codon usage bias is strongly conserved, then d_S will decrease and d_N/d_S will increase within a population. The d_N/d_S ratio was used in *Drosophila* lineages, and helped determine that the *Notch* locus had evolved to include suboptimal codons [80]. Using 158 orthologous genes, maximum likelihood also detected a strong shift from suboptimal to optimal codons in two lineages of *Populus* [81]. Detecting the cause of such shifts in codon usage is important for determining the biological significance of mutations. SCUMBLE (Synonymous Codon Usage Bias Maximum Likelihood Estimation) uses a model inspired by statistical physics to identify different sources of codon bias including selection and mutation [82]. SCUMBLE is also used as a filter to identify regions with insufficient information for analysis. This technique helped determine that natural selection shaped codon biases in *Strongylocentrotus purpuratus* (purple sea urchin) by limiting the analysis to only regions with sufficient support [83]. Shifts in mutation and selection rates allow the evolutionary history of species to be recovered using this method.

Violations of maximum likelihood statistical properties in a codon model

Many assumptions of the statistical properties in maximum likelihood are violated by a codon model. For instance, species are constrained to taxon-specific pools of tRNA, and triplets in coding sequences are not independent. Algorithms with statistical properties that require character independence, such as maximum likelihood, violate that rule for genetic data [84]. Furthermore, the codon model assumption of homogeneity of codon composition leads to seriously biased phylogenetic estimations when that assumption is violated [85].

Horizontal gene transfer is another important mechanism in evolution and complicates phylogenetic analyses in bacteria because 81±15% of genes have been laterally transferred among bacteria at some point in their evolutionary history [86]. Common transposable elements in eukaryotes also arose from horizontal gene transfer, with over 50% of some mammalian genomes originally arising from horizontal gene transfer [87]. Detecting horizontal gene transfer has been challenging, and codon bias is a poor indicator of horizontal transmission, normally underestimating the effects of lateral transfer [88-90]. However, codon composition is an excellent indicator of whether a gene will become fixed in a species after a lateral transfer event [90]. The concept of horizontal gene transfer not only complicates a general phylogenetic analysis, but suggests that a standard bifurcating tree might not be the best choice in analyses of bacteria or archaea [91]. Although it is known that codons (and DNA in general) do not strictly follow many of the assumptions of phylogenetic analyses, the bifurcating tree is still the most widely used phylogenetic representation, and generally depicts statements of homology even when some assumptions are violated.

Codon usage in viruses

Phylogenies have also been used to predict the pathogenicity of viruses and viral interactions with their hosts. Bee-infecting viruses have strong correlations in their codon usages with their hosts, and the infected insects' codon usage similarity follows the insect phylogeny [92]. Furthermore, human-host viruses tend to share the same codon usages as proteins expressed in tissues that the viruses infect [93]. More specifically, the key determinant in codon patterns within herpes viruses were the overall GC content, GC content at the third codon position, and gene length [94]. In contrast, mutation played a larger role in Zika viruses, with higher frequencies of A-ending codons [95]. However, evidence of natural selection in Zika viruses also suggest that they evolved host- and vector-specific codon usage patterns to successfully replicate in various hosts and vectors [96]. In hepatitis C, preferred codon usages did not always match the phylogenetic histories of the viruses as determined by sequence similarity, indicating that codon usage might provide additional information not identified by common phylogenomic approaches [97].

Successful implementations of codon usage bias in phylogenetics

Beyond analyzing pathogenicity, phylogenetic inferences using codon usage biases from all domains of life have successfully uncovered several interesting biological principles. One study found compositional differences in codon usage between monocots (i.e., flowering plants whose seeds contain one embryonic leaf) and dicots (i.e., flowering plants whose seeds contains two embryonic leaves), where monocots had lower DNA background compositional bias, but higher codon usage bias than dicots [98]. Another technique used a distance-based clustering method of codon usage weighted by nucleotide base bias per position (i.e., the frequency of a codon over the product of the frequency of the nucleotide at the first, second, and third positions) to recover the phylogeny of closely related *Ectocarpales* (brown algae) [99]. The phylogenetic signal of codon usage was not limited to nuclear DNA, and mitochondrial synonymous codon usage in plants was associated with intron number that mirrored species evolution [100].

Creative attempts at analyzing codon usage have also proven fruitful. A binary representation of codon aversion (i.e., creating a character matrix based on codons which are not used in an ortholog) successfully recover the phylogeny of various tetrapods, showing that complete codon aversion is also conserved [101]. That study also found that stop codon usage had the highest phylogenetic signal [101], meaning a codon matrix of 64 x 64 (the probability of all codons including the stop codons transitioning to all other codons) might be better than the traditional 61 x 61 codon matrix in a likelihood framework. Codon aversion has also been used in an alignment-free context by comparing sets of codon tuples found in a genome, where each tuple is a list of codons not used in a gene [47]. A similar technique found that codon pairing (i.e., the same codon being used within a ribosomal window) is phylogenetically informative under both alignment-free and parsimony frameworks [48].

Other studies map codon usage in a particular gene across a reference phylogeny. This technique can produce meaningful representations of codon transitions across genes. Mapping the codon usage bias of a gene tree to a species tree revealed purifying selection among the actin-depolymerizing factor/cofilin (ADF/CFL) gene family [102]. This technique also showed that codon usage is significantly correlated with gene age within metazoan genomes [103]. Codon aversion in all domains of life was also mapped to the Open Tree of Life (OTL) [104] and showed that codon aversion follows established species relationships more closely than expected by random chance [105].

Contradictory Signals

At times, codon usage dynamics have contradictory signals that indicate different evolutionary histories. For instance, Miller, McKinnon [47], Miller, McKinnon [48], and Miller, McKinnon [105] used the same dataset to conclude that codon aversion can be used in an alignment-free algorithm, codon pairing can recover phylogenies using either parsimony or alignment-free techniques, and codon aversion is largely conserved within orthologs across the Tree of Life. However, the reported trees from those three studies vary significantly from each other (Figure 2), indicating codon aversion and

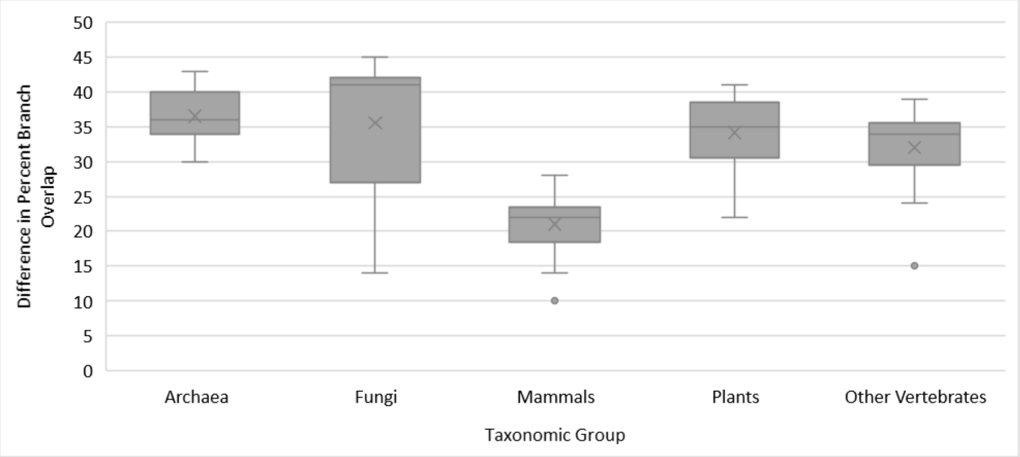

Figure 2. Difference in Percent Branch Overlap of Seven Phylogenies Recovered Using Codon Pairing or Codon Aversion. Mean differences in percent branch overlap are marked with an 'X', median differences for each taxonomic group are marked with a horizontal line, and outlier are individually displayed on the box plot.

codon pairing do not have the same evolutionary constraints. Even using the same codon usage bias, the model used to recover the phylogeny produced contradictory results, with recovered phylogenies differing by 10-45%. Therefore, gene selection appears to play a pivotal role in recovering the species tree, and more work needs to be done to identify which genes have the highest phylogenetic signal under each codon model. Perhaps a combination of different codon usage biases, or using certain biases in only highly expressed genes, may more adequately track speciation.

Conclusion

Codon usage biases continue to be widely studied in a phylogenetic construct. However, their application in phylogenomics remains limited by their incorporation in current phylogenomic techniques. While some applications attempt to include codon usage biases either as a singular character state in parsimony or in combination with the overall maximum likelihood model, many key attributes of codon biases remain unexplored. For instance, the cause of differing phylogenetic signals between codon aversion and codon pairing has yet to be identified. Additionally, although it is known that tRNA supply and demand is correlated to codon usage, a model does not currently exist to assess tRNA supply and demand in a maximum likelihood framework. Future codon analyses will necessitate more complete datasets with accurate tRNA expression values in different tissues and species. A more robust dataset of tRNA expression values would also facilitate more precise codon modeling. Furthermore, since codons are used to regulate gene translational efficiency, codon models might require gene expression data in addition to the full (or reduced) codon matrix, and some codon usage biases may track speciation only within certain genes.

Codon usage bias is an exciting biological principle that has not been fully utilized in phylogenetic systematics. Few likelihood methods incorporate specific codon usage biases in their models beyond nucleotide substitution rates, and many aspects of the ramp sequence, co-tRNA codon pairing, gene expression, and tRNA expression remain unknown. Although codon usage biases have been shown to be phylogenetically conserved, many of the biological principles surrounding codon usage bias have yet to be fully utilized in phylogenomics. Therefore, including specific codon usage biases in phylogenomic algorithms and identifying the gene-specific biological implications of each codon usage bias will enable future phylogenomic studies to identify more robust phylogenetic trees and aid in understanding nuanced phylogenetically conserved mechanisms affecting gene expression and overall species fitness.

References

1. Soltis DE and Soltis PS. “The Role of Phylogenetics in Comparative Genetics.” *Plant Physiol* 132 (2003): 1790-800.
2. Haszprunar G. “The types of homology and their significance for evolutionary biology and phylogenetics.” *J Evol Biol* 5 (1992): 13-24.
3. Bertolani R, Guidetti R, Marchioro T, Altiero T, Rebecchi L and Cesari M. “Phylogeny of Eutardigrada: New molecular data and their morphological support lead to the identification of new evolutionary lineages.” *Molecular Phylogen Evol* 76 (2014): 110-26.
4. Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M and Wörheide G, et al. “Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough.” *PLOS Biol* 9 (2011): e1000602.
5. Crick FH, Barnett L, Brenner S and Watts-Tobin RJ. “General nature of the genetic code for proteins.” *Nature* 192 (1961): 1227-1232.
6. Sharp PM and Li WH. “An evolutionary perspective on synonymous codon usage in unicellular organisms.” *J Mol Evol* 24 (1986): 28-38.
7. Crick FH. “Codon--anticodon pairing: the wobble hypothesis.” *J Mol Biol* 19 (1966): 548-55.
8. Post LE, Strycharz GD, Nomura M, Lewis H and Dennis PP. “Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit beta in Escherichia coli.” *Proc Natl Acad Sci U S A* 76 (1979): 1697-701.
9. Gutman GA and Hatfield GW. “Nonrandom utilization of codon pairs in Escherichia coli.” *Proc Natl Acad Sci U S A* 86 (1989): 3699-703.
10. Koonin EV. “Orthologs, paralogs, and evolutionary genomics.” *Annu Rev Genet* 39 (2005): 309-338.
11. Tekaia F. “Inferring Orthologs: Open Questions and Perspectives.” *Genom Insigh* 9 (2016): 17-28.
12. Magis C, Taly JF, Bussotti G, Chang JM, Di Tommaso P and Erb I, et al. “T-Coffee: Tree-based consistency objective function for alignment evaluation.” *Methods Mol Biol* 1079 (2014): 117-29.
13. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32 (2004): 1792-1797.
14. Sievers F and Higgins DG. “Clustal omega.” *Curr Protoc Bioinformatics* 48 (2014):3 13 1-6.
15. Sievers F, Higgins DG. “Clustal Omega for making accurate alignments of many protein sequences.” *Protein Sci* 27 (2018): 135-45.
16. Katoh K and Standley DM. “MAFFT: iterative refinement and additional methods.” *Methods Mol Biol* 1079 (2014): 131-46.
17. Pais FS, Ruy Pde C, Oliveira G and Coimbra RS. “Assessing the efficiency of multiple sequence alignment programs.” *Algorithms Mol Biol* 9 (2014): 4.

18. Daugelaite J, Driscoll OA, Sleator RD. "An Overview of Multiple Sequence Alignments and Cloud Computing in Bioinformatics." *ISRN Biomathematics* 2013 (2013): 14.

19. Farris JS. "Parsimony and explanatory power." *Cladistics* 24 (2008): 825-47.

20. Felsenstein J. "Cases in which Parsimony or Compatibility Methods will be Positively Misleading." *Syst Biol* 27 (1978): 401-410.

21. Wilgenbusch JC and Swofford D. "Inferring evolutionary trees with PAUP." *Curr Protoc Bioinformatics* 2003; 6.

22. Goloboff PA, Farris JS and Nixon KC. "TNT: Tree Analysis Using New Technology 54 (2005): 176-178.

23. Rogers JS. "On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences." *Syst Biol* 46(1997)354-357.

24. Wald A. Note on the Consistency of the Maximum Likelihood Estimate. (1949):595-601.

25. Paninski L, Pillow JW, and Simoncelli EP. "Maximum likelihood estimation of a stochastic integrate-and-fire neural encoding model." *Neural Comput* 16 (2004)2533-2561.

26. Siddall ME. Success of Parsimony in the Four-Taxon Case: Long-Branch Repulsion by Likelihood in the Farris Zone. *Cladistics* 14 (1998):209-220.

27. Kumar S, Stecher G, Li M and Knyaz C, et al. "MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms." *Mol Biol Evol* 35 (2018):1547-1549.

28. Stamatakis A. "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies." *Bioinformatics* 309 (2014):1312-1313.

29. Nguyen LT, Schmidt HA, von Haeseler A and Minh BQ. "IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies." *Mol Biol Evol* 32 (2015): 268-74.

30. Retief JD. "Phylogenetic analysis using PHYLIP." *Methods Mol Biol* 132 (2000): 243-58.

31. Huelsenbeck JP, Larget B, Miller RE and Ronquist F. "Potential applications and pitfalls of Bayesian inference of phylogeny." *Syst Biol* 51 (2002): 673-88.

32. Yang Z and Rannala B. "Molecular phylogenetics: principles and practice." *Nat Rev Genet* 13 (2012): 303-314.

33. Ling C, Hamada T, Gao J and Zhao G, et al. "MrBayes tgMC3++: A High Performance and Resource-Efficient GPU-Oriented Phylogenetic Analysis Method." *IEEE/ACM Trans Comput Biol Bioinform* 13 (2016): 845-854.

34. Ronquist F, Teslenko M, van der Mark P and Ayres DL, et al. "MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space." *Syst Biol* 61 (2012): 539-542.

35. Bouckaert R, Heled J, Kuhnert D and Vaughan T, et al. "BEAST 2: a software platform for Bayesian evolutionary analysis." *PLoS Comput Biol* 10 (2014): e1003537.

36. Holder M and Lewis PO. "Phylogeny estimation: traditional and Bayesian approaches." *Nat Rev Genet* 4 (2003): 275-84.

37. Chan CX, Bernard G, Poirion O and Hogan JM, et al.. "Inferring phylogenies of evolving sequences without multiple sequence alignment." *Sci Rep* 4 (2014): 6504.

38. Jun S-R, Sims GE, Wu GA and Kim S-H. "Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution." Proceedings of the National Academy of Sciences 107 (2010): 133-138.

39. Sims GE, Jun SR, Wu GA and Kim SH. "Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions." *Proc Natl Acad Sci U S A* 106 (2009): 2677-82.

40. Zuo G and Hao B. "CVTree3 Web Server for Whole-genome-based and Alignment-free Prokaryotic Phylogeny and Taxonomy." *Genom Prot Bioinformatics* 13 (2015): 321-31.

41. Ulitsky I, Burstein D, Tuller T and Chor B. "The average common substring approach to phylogenomic reconstruction." *J Comput Biol* 13 (2006): 336-50.

42. Leimeister CA and Morgenstern B. "Kmacs: the k-mismatch average common substring approach to alignment-free sequence comparison." *Bioinformatics* 30 (2014): 2000-8.

43. Haubold B, Pfaffelhuber P, Domazet-Loso M and Wiehe T. "Estimating mutation distances from unaligned genomes." *J Comput Biol* 16 (2009): 1487-500.

44. Yi H and Jin L. "Co-phylog: an assembly-free phylogenomic approach for closely related organisms." *Nucleic Acids Res* 41 (2013): e75.

45. Leimeister CA, Sohrabi-Jahromi S and Morgenstern B. "Fast and accurate phylogeny reconstruction using filtered spaced-word matches." *Bioinformatics* 33 (2017): 971-9.

46. Haubold B, Klotzl F and Pfaffelhuber P. "andi: fast and accurate estimation of evolutionary distances between closely related genomes." *Bioinformatics* 31 (2015): 1169-75.

47. Miller JB, McKinnon LM, Whiting MF and Ridge PG. "CAM: An alignment-free method to recover phylogenies using codon aversion motifs." *Peer J Preprints* 7 (2019): e27756v1.

48. 48 Miller JB, McKinnon LM, Whiting MF and Kauwe JSK, et al. "Codon Pairs are Phylogenetically Conserved: A comprehensive analysis of codon pairing conservation across the Tree of Life." *Plos one* 15 (2020): e0232260.

49. Sanderson MJ. "Objections to Bootstrapping Phylogenies: A Critique." *Systematic Biol* 44 (1995): 299-320.

50. Hillis DM and Bull JJ. "An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis." *Sys Biol* 42 (1993): 182-92.

51. Quax TE, Claassens NJ, Soll D and van der Oost J. "Codon Bias as a Means to Fine-Tune Gene Expression." *Mol Cell* 59 (2015): 149-61.

52. Labella AL, Opulente DA, Steenwyk JL and Hittinger CT, et al. "Variation and selection on codon usage bias across an entire subphylum." *PLOS Genet* 15 (2019): e1008304.

53. Wright F. "The 'effective number of codons' used in a gene." *Gene* 87 (1990): 23-29.

54. dos Reis M, Savva R and Wernisch L. "Solving the riddle of codon usage preferences: a test for translational selection." *Nucleic Acids Res* 32 (2004): 5036-5044.

55. dos Reis M, Wernisch L and Savva R. "Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome." *Nucleic Acids Res* 31 (2003): 6976-6985.

56. Pechmann S and Frydman J. "Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding." *Nat Struct Mol Biol* 20 (2013): 237-243.

57. Trotta E. "Selection on codon bias in yeast: a transcriptional hypothesis." *Nucleic Acids Res* 41 (2013): 9382-9395.

58. Tuller T, Carmi A, Vestsigian K and Navon S, et al. "An evolutionarily conserved mechanism for controlling the efficiency of protein translation." *Cell* 141 (2010): 344-354.

59. Shah P, Ding Y, Niemczyk M and zsaxKudla G, et al. “Rate-limiting steps in yeast protein translation.” *Cell* 153 (2013): 1589-1601.

60. Goodman DB, Church GM and Kosuri S. “Causes and effects of N-terminal codon bias in bacterial genes.” *Sci* 342 (2013): 475-479.

61. Hodgman MW, Miller JB, Meurs TE and Kauwe JSK. “CUBAP: an interactive web portal for analyzing codon usage biases across populations." *Nucleic Acids Res* 48 (2020): 11030-11039.

62. Miller JB, Brase LR and Ridge PG. “ExtRamp: a novel algorithm for extracting the ramp sequence based on the tRNA adaptation index or relative codon adaptiveness.” *Nucleic Acids Res* 47 (2019): 1123-1131.

63. Frenkel-Morgenstern M, Danon T, Christian T and Igarashi T, et al. “Genes adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels.” *Molecular Syst Biol* 8 (2012): 572

64. Carbone A, Kepes F and Zinovyev A. “Codon bias signatures, organization of microorganisms in codon space, and lifestyle.” *Mol Biol Evol* 22 (2005): 547-561.

65. Botzman M and Margalit H. “Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles.” *Genome Biol* 12 (2011): R109.

66. Xu Y, Ma P, Shah P and Rokas A, et al. “Non-optimal codon usage is a mechanism to achieve circadian clock conditionality.” *Nat* 495 (2013): 116-120.

67. Lal D, Verma M, Behura SK and Lal R. “Codon usage bias in phylum Actinobacteria : relevance to environmental adaptation and host pathogenicity.” *Res Microbiol* 167 (2016): 669-677.

68. Brule CE and Grayhack EJ. “Synonymous Codons: Choose Wisely for Expression.” *Trends Genet* 33 (2017): 283-297.

69. Dixon GB, Bay LK and Matz MV. “Evolutionary Consequences of DNA Methylation in a Basal Metazoan.” *Mol Biol Evol* 33 (2016): 2285-2293.

70. Cannarozzi G, Schraudolph NN, Faty M and von Rohr P, et al. “A role for codon order in translation dynamics.” *Cell* 141 (2010): 355-67.

71. Shao ZQ, Zhang YM, Feng XY and Wang B, et al. “Synonymous codon ordering: a subtle but prevalent strategy of bacteria to improve translational efficiency." *PLoS One* 7 (2012): e33547.

72. Zhang YM, Shao ZQ, Yang LT, Sun XQ, Mao YF and Chen JQ et al. “Non-random arrangement of synonymous codons in archaea coding sequences.” *Genom* 101 (2013): 362-367.

73. Sun Y, Tamarit D and Andersson SGE. “Switches in Genomic GC Content Drive Shifts of Optimal Codons under Sustained Selection on Synonymous Sites.” *Genome Biol Evol* 9 (2017): 2560-2579.

74. Li J, Zhou J, Wu Y and Yang S, et al. “GC-Content of Synonymous Codons Profoundly Influences Amino Acid Usage.” *G3* 5 (2015): 2027-2036.

75. Zur H and Tuller T. “Predictive biophysical modeling and understanding of the dynamics of mRNA translation and its evolution.” *Nucleic Acids Res* 44 (2016): 9031-9049.

76. Castle JC. “SNPs occur in regions with less genomic sequence conservation.” *PLoS One* 6 (2011): e20660.

77. Anisimova M and Kosiol C. “Investigating protein-coding sequence evolution with probabilistic codon substitution models.” *Mol Biol Evol* 26 (2009): 255-271.

78. Miyazawa S. “Superiority of a mechanistic codon substitution model even for protein sequences in Phylogenetic analysis." *BMC Evol Biol* 13 (2013): 1-10.

79. Akashi H, Goel P and John A. “Ancestral inference and the study of codon bias evolution: implications for molecular evolutionary analyses of the Drosophila melanogaster subgroup.” *PLoS One* 2 (2007): e1065.

80. Nielsen R, Bauer DuMont VL, Hubisz MJ and Aquadro CF. “Maximum likelihood estimation of ancestral codon usage bias parameters in Drosophila.” *Mol Biol Evol* 24 (2007): 228-235.

81. Ingvarsson PK. “Molecular evolution of synonymous codon usage in Populus.” *BMC Evol Biol* 8 (2008): 307.

82. Kloster M and Tang C. “SCUMBLE: a method for systematic and accurate detection of codon usage bias by maximum likelihood estimation.” *Nucleic Acids Res* 36 (2008): 3819-3827.

83. Kober KM and Pogson GH. Genome-Wide Patterns of Codon Bias Are Shaped by Natural Selection in the Purple Sea Urchin, Strongylocentrotus purpuratus.” *G3* 3 (2013):1069.

84. Christianson ML. "Usage patterns distort phylogenies from or of DNA sequences." *Am J Bot* 92 (2005): 1221-1233.

85. Inagaki Y and Roger AJ. “Phylogenetic estimation under codon models can be biased by codon usage heterogeneity.” *Mol Phylogenet Evol* 40 (2006): 428-434.

86. Dagan T, Artzy-Randrup Y and Martin W. “Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution.” *Proceedings of the National Academy of Sciences* 105 (2008): 10039-44.

87. Ivancevic AM, Kortschak RD, Bertozzi T and Adelson DL. “Horizontal transfer of Bov B and L1 retrotransposons in eukaryotes.” *Genome Biol* 19 (2018): 85.

88. Koski LB, Morton RA and Golding GB. “Codon bias and base composition are poor indicators of horizontally transferred genes.” *Mol Biol Evol* 18 (2001): 404-412.

89. Friedman R and Ely B. “Codon usage methods for horizontal gene transfer detection generate an abundance of false positive and false negative results.” *Curr Microbiol* 65 (2012): 639-642.

90. Tuller T. “Codon bias, tRNA pools and horizontal gene transfer.” *Mob Genet Elements* 1 (2011): 75-77.

91. Koonin EV and Wolf YI. “Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world.” *Nuclei Acids Res* 36 (2008): 6688-6719.

92. Chantawannakul P and Cutler RW. “Convergent host-parasite codon usage between honeybee and bee associated viral genomes.” *J Invertebr Pathol* 98 (2008): 206-210.

93. Miller JB, Hippen AA, Wright SM and Morris C, et al. "Human viruses have codon usage biases that match highly expressed proteins in the tissues they infect." *Biomed Genet Genomics* 2 (2017):1-5.

94. Roychoudhury S and Mukherjee D. “A detailed comparative analysis on the overall codon usage pattern in herpesviruses.” *Virus Res* 148 (2010): 31-43.

95. Cristina J, Fajardo A, Sonora M, and Moratorio G, et al. “A detailed comparative analysis of codon usage bias in Zika virus.” *Virus Res* 223 (2016): 147-152.

96. Butt AM, Nasrullah I, Qamar R and Tong Y. “Evolution of codon usage in Zika virus genomes is host and vector specific.” *Emerg Microbes Infect* 5 (2016): e107.

97. Mortazavi M, Zarenezhad M, Alavian SM and Gholamzadeh S, et al. “Bioinformatic Analysis of Codon Usage and Phylogenetic Relationships in Different Genotypes of the Hepatitis C Virus.” *Hepat Mon* 16 (2016): e39196.

98. Camiolo S, Melito S and Porceddu A. "New insights into the interplay between codon bias determinants in plants." *DNA Res* 22 (2015): 461-470.

99. Das S, Chakrabarti J, Ghosh Z, Sahoo S and Mallick B. “A new measure to study phylogenetic relations in the brown algal order Ectocarpales: The codon impact parameter". *J Biosci* 30 (2005): 699-709.

100. Xu W, Xing T, Zhao M and Yin X, et al. "Synonymous codon usage bias in plant mitochondrial genes is associated with intron number and mirrors species evolution." *PLoS One* 10 (2015): e0131508.

101. Miller JB, Hippen AA, Belyeu JR and Whiting MF, et al. "Missing something? Codon aversion as a new character system in phylogenetics." *Cladistics* 33 (2017): 545-556.

102. Roy-Zokan EM, Dyer KA and Meagher RB. “Phylogenetic Patterns of Codon Evolution in the ACTIN-DEPOLYMERIZING FACTOR/COFILIN (ADF/CFL) Gene Family.” *PLoS One* 10 (2015): e0145917.

103. Prat Y, Fromer M, Linial N and Linial M. “Codon usage is associated with the evolutionary age of genes in metazoan genomes.” *BMC Evol Biol* 9 (2009): 285.

104. Hinchliff CE, Smith SA, Allman JF and Burleigh JG, et al. “Synthesis of phylogeny and taxonomy into a comprehensive tree of life.” *Proc Natl Acad Sci USA* 112 (2015): 12764-12769.

105. Miller JB, McKinnon LM, Whiting MF and Ridge PG. “Codon use and aversion is largely phylogenetically conserved across the tree of life.” *Mol Phylogen Evol* 144 (2020): 106697.

How to cite this article: Justin B. Miller, Michael F. Whiting, John S.K. Kauwe and Perry G. Ridge. How Codon Usage Biases Affect Our Ability to Recover the Tree of Life. *J Phylogenetics Evol Biol* 8 (2021) doi: 10.37421/jpgeb.2021.8.211