2006

An Investigation of Codon Usage Bias Including Visualization and Quantification in Organisms Exhibiting Multiple Biases

Douglas W. Raiford
Wright State University - Main Campus

Travis E. Doom
Wright State University - Main Campus, travis.doom@wright.edu

Dan E. Krane
Wright State University - Main Campus, dan.krane@wright.edu

Michael L. Raymer
Wright State University - Main Campus, michael.raymer@wright.edu

An Investigation of Codon Usage Bias Including Visualization and Quantification in Organisms Exhibiting Multiple Biases

Douglas W. Raiford, Travis E. Doom, Dan E. Krane, and Michael L. Raymer

Abstract

Prokaryotic genomic sequence data provides a rich resource for bioinformatic analytic algorithms. Information can be extracted in many ways from the sequence data. One often overlooked process involves investigating an organism's codon usage. Degeneracy in the genetic code leads to multiple codons coding for the same amino acids. Organism's often preferentially utilize specific codons when coding for an amino acid. This biased codon usage can be a useful trait when predicting a gene's expressivity or whether the gene originated from horizontal transfer. There can be multiple biases at play in a genome causing errors in the predictive process. For this reason it is important to understand the interplay of multiple biases in an organism's genome. We present here new techniques in the measurement and analysis of multiple biases in prokaryotic genomic data. Included is a visualization technique aimed at demonstrating genomic adherence to a set of discrete biases.

1 Introduction

Recent advances in genomic sequencing techniques have caused a rapid increase in the amount of available whole genome sequence data. At the time of this writing the the National Center for Biotechnology Information (NCBI) [14] has sequence information for 318 complete microbial genomes. This represents over one billion base-pairs of sequence information. Sequence data can provide a great deal of valuable information, including gene location prediction, gene ancestral origins, and taxonomic relationships between species. Another source of information is the degeneracy in the genetic code. There are 64 amino acid coding triplets, or codons, that code for only twenty common amino acids. This means that multiple codons sometimes code for the same amino acid (degeneracy). It has long been known that organisms preferentially utilize one or more of these synonymous codons in their coding sequences [6, 7, 9, 15]. This bias in codon usage can be exploited to predict such things as how often a gene is expressed [5] or whether a gene is a recent addition to the genome [4].

Selective pressure to enhance translational efficiency is thought to be the underlying cause of the bias used in predicting gene expressivity [10, 16]. There is an amino acid carrying molecule (tRNA) associated with each codon that is used in translating the mRNA transcripts into the protein macromolecules. When the codon associated with the highest tRNA abundance is utilized, efficiencies in translation can be realized due to the higher relative availability.

Biases associated with translational efficiency are not the only biases found in prokaryotic and small eukaryotic genomes. They can also be affected by such biases as those introduced by high or low GC-content [2]. In some cases these biases can coexist with translation bias [2, 8]. When this occurs translation bias can be obscured, making gene expression levels difficult to predict.

Several approaches have been employed to identify and measure codon usage biases [1, 3, 5, 9, 11, 13, 17–20]. Some methods, such as codon adaptation index [17], require prior knowledge of a set of genes known to be highly expressed. Others, such as the updated codon adaptation index (CAI) algorithm [3] attempt to identify the bias using coding sequence information only. Algorithms that take the latter approach (using sequence information only) can be confounded by other biases that exist within the target genome (e.g. GC or strand bias) [2, 3, 12].

Identified biases may not be those associated with translational efficiency. Figure 1(a) reveals the location of the reference set (small set of genes that are the most highly biased) for *Nostoc sp. PCC 7120*. These genes would be assumed to be the most highly expressed in the absence of a confounding bias. Figure 1(b) depicts the location of ribosomal protein coding genes. One would normally expect these genes to be highly expressed. It is of concern that they are not in the same region of the genome as the predicted reference set. The region of the codon usage space where the CAI identified reference set resides is also the region where the high AT-content genes are located. This is an indication that the predicted reference set (Fig. 1(a)) is more likely to be the set of genes identified due to a high AT-content bias.

We present new techniques for measuring and visualizing a genome's adherence to codon usage biases. To this end, the *Methods* section begins by describing how to isolate the dominant bias in an organism's codon usage space (CAI algorithm). Following the description of CAI, we will present a measure of genomic adherence to an identified bias, followed by a visualization technique useful in gaining insights into this genomic adherence.

2 Methods

2.1 Codon Adaption Index

Codon adaptation index (CAI) [3] is an algorithm that isolates the dominant bias in an organism's genome. Once the bias is identified, the algorithm computes a score representative of each gene's adherence to that bias. CAI is calculated through the use of an iterative algorithm that first locates a reference set of genes (small set of the most highly biased genes) that is then used to calculate weights that determine a CAI score for each gene. It starts with a reference set of *all* genes and assigns a weight to each codon based upon the codon usage in that reference set. It then iteratively reduces the reference size by one half until it achieves the correct reference set size (1% of all genes).

The algorithm assigns a weight to each codon based upon the codon usage in the current reference set. The weight for a given codon is equal to the count of that codon (within the subset of genes currently considered the reference set) divided by the count of its sibling with the highest count (the maximal sibling will have a weight of one). Equation (1) describes the weight w of the ith codon for the jth amino acid. The x in the numerator is the count for that codon and the denominator (y) is the count of the maximal sibling for the amino acid in question.

$$w_{ij} = \frac{x_{ij}}{y_{imax}} \tag{1}$$

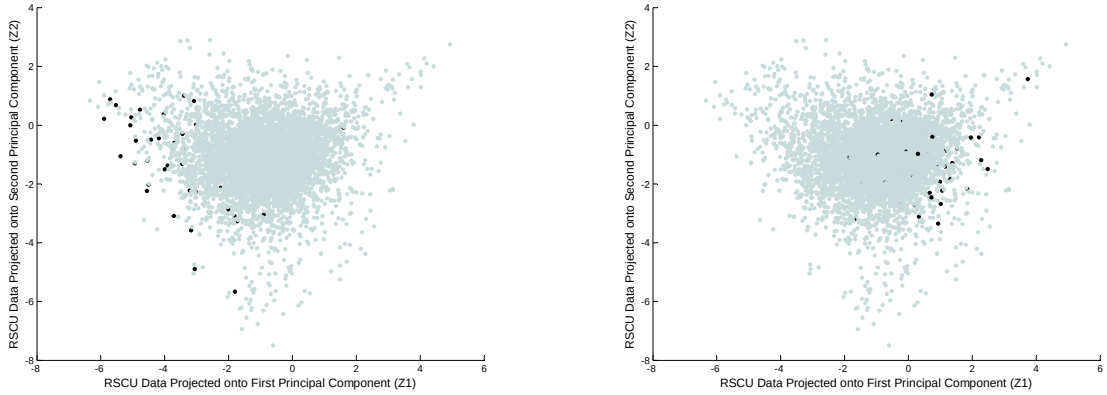
Given these weights a CAI score is calculated for each gene in the genome (2).

$$CAI(g) = \sqrt[L]{\prod_{i=1}^{L} w_i} \tag{2}$$

L is the length of the gene (number of codons). The CAI value for a gene is a geometric average of codon usage within that gene. The list of genes is sorted by CAI score, and the genes in the top half of the list are kept as the new reference set. New w values are calculated, followed by new CAI values for the genes. This process is repeated until the reference set of genes equals one percent of the original number of genes.

2.2 Locating Second Bias

In organisms where the CAI algorithm is confounded – i.e. ribosomal protein coding genes are in disparate locations from identified reference sets (as in Fig. 1) – a second search must be performed. This second search is localized to the region where the ribosomal protein coding genes reside and is similar to the random search employed by [3]. Once a suitable reference set is located in the appropriate region of the codon usage space, CAI scores are generated for all



(a) Reference Set

(b) Ribosomal Protein Coding Genes

Figure 1: *Nostoc sp. PCC 7120*. 1(a) Reference Set. Small set (1% of genome) of genes identified by CAI algorithm as being highly expressed. Each point represents a gene. The dark genes comprise the reference set. 1(b) Ribosomal protein coding genes. Genes known to be, generally, highly expressed. Each point represents a gene. The dark genes are ribosomal protein coding genes. Ribosomal protein coding genes are distant from the region identified by the reference set. This indicates that the bias identified by the CAI algorithm is confounded and is not representative of translation bias. RSCU is relative synonymous codon usage [16], a normalized codon frequency. A gene is represented by a 64 dimensional vector of frequencies.

genes representing their adherence to this new, secondary, bias.

2.3 Genomic Adherence

Once a bias has been isolated it is useful to determine how strongly the genome adheres to that bias. This can be determined by aggregating the bias adherence of the individual genes. CAI is a measure of a particular gene's adherence to the bias identified by the CAI algorithm. A gene that displays perfect adherence achieves a CAI value of 1. CAI values depicted graphically (Fig. 2) form a characteristic curve. The area under this curve (a summation of the discrete CAI values for all genes) is representative of the genome's adherence to the bias. If all genes adhere perfectly to the bias their sum will equal the number of genes in the organism's genome (N). This allows for the use of N as a normalizing value.

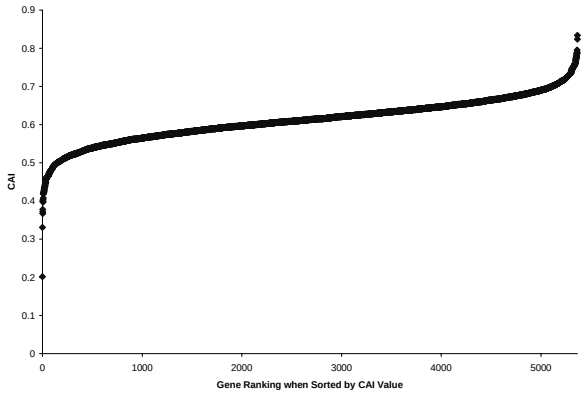


Figure 2: *Nostoc sp. PCC 7120* CAI values. The X axis is a listing of genes arranged by ascending CAI score. Y axis is CAI score of each corresponding gene. The area under this characteristic curve represents the genomic adherence to a specified bias.

$$GASB = \sum_{i=1}^{N} CAI_i \qquad (3)$$

$$GASB_{max} = \sum_{i=1}^{N} CAI_i = \sum_{i=1}^{N} 1 = N \quad (4)$$

$$GASB_{norm} = \frac{GASB}{GASB_{max}} \quad (5)$$

Because genomic adherence to a specified bias (GASB) is normalized by N, the adherence metric also describes the organism's average CAI score. This makes other related quantities, such as variance and standard deviation, available and useful. This is especially true since organismal CAI scores generally adhere to a normal distribution (Fig. 3). This also implies that a t-test can be employed to verify whether one genomic adherence score is significantly greater than another (say, between a translation bias and a GC bias).

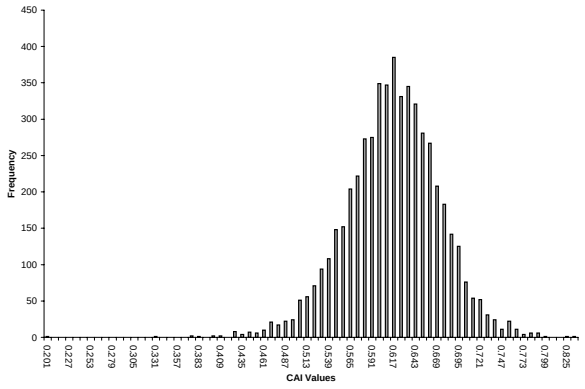


Figure 3: *Nostoc sp. PCC 7120* Distribution of CAI values. Gene CAI scores generally adhere to a normal distribution making such measures as standard deviation and t-tests meaningful.

2.4 Polar Bias View

Visualizing the genome's adherence to multiple biases can be useful in gaining insights into how the biases interrelate. Our adherence visualization method treats two competing biases as point-sources (poles) of attractive force exerted on the individual genes within the genome. A gene that is strongly attracted to one bias but not the other is shown as being very close to the one pole and distant from the other. Genes pulled equally by both poles are shown as equally distant from both. Figure 4 is an example of this data view. Each point is a gene, and the distance between the gene and a pole is described by $1 - CAI(g)$. The maximum CAI score is 1 so $1 - CAI(g)$ will be high for genes that are far from the pole and low for genes that are close. Note that the genes appear more drawn to the translational efficiency bias even though the CAI algorithm is confounded by the GC bias.

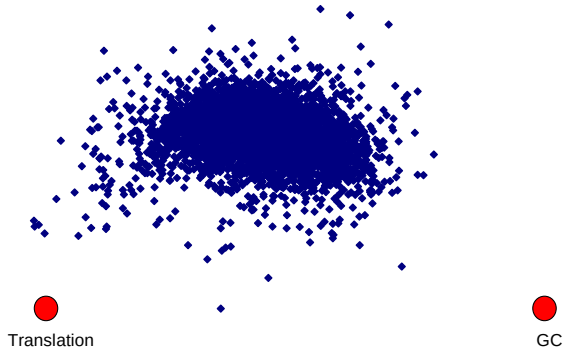


Figure 4: Example of Polar Bias View of Translational and GC bias for *Nostoc sp. PCC 7120*. Each point represents a gene and the distance from that point to a bias pole is $1 - CAI(g)$ for that gene as defined by the reference set associated with that bias. Even though AT-content confounds the CAI algorithm, once isolated, translation bias exhibits stronger genomic adherence (i.e. on average the the genes are closer to the translational bias pole than to the content pole).

The procedure for generating the polar bias view is to first determine the location of the poles (the two end points of b in Fig. 5). The magnitude of b is determined by finding the smallest $(1 - CAI(g)_1) + (1 - CAI(g)_2)$. This can be accomplished by storing both CAI values in a listing of genes, calculating the result of the equation for each gene, and then sorting the gene list by that value. Trigonometric techniques are employed to evaluate x_1 and y values (6 and 7). The law of cosines is used to determine θ_1 and θ_2.

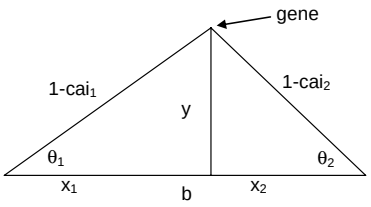


Figure 5: Geometric Representation of Codon Usage Polar View. Bias poles located at opposite ends of base b. Gene located at apex of triangle (top). Gene location relative to two biases defined by $[x_1, y]$ coordinates. To build the polar bias view for an organism, x and y are calculated for each gene.

$$x_1 = \Big(1 - CAI(g)_1\Big) cos(\theta_1) \quad (6)$$

$$y = \Big(1 - CAI(g)_1\Big) sin(\theta_1) \quad (7)$$

3 Results

An example of a polar bias view was presented in Fig. 4. That view was of an organism whose two biases were in disparate regions of the codon usage space. An example of a polar bias view for an organism whose biases are very close can be seen in Fig. 6. In these depictions the degree to which a different ordering of genes occurs, when sorted by CAI, is indicated by the horizontal spread of the gene cloud. The wider the spread the more dissimilar the ordering.

4 Discussion

Algorithms developed to determine gene bias levels tend to find the gene's adherence to the *dominating* bias. This can be problematic if the intent is to find translational bias levels indicative of expressivity. Previous work has indicated that multiple biases can coexist in a genome [2, 8]. With the use of our genomic adherence measure and polar bias view we have extended our understanding of genomic codon usage in the presence of multiple biases.

The polar bias view is useful in visualizing the stronger adherence of genomes to the translation bias. Figures 4 and 6 clearly show a general tendency of the genes to be more strongly *attracted* to the translational bias than the GC(AT)-content bias. *Nostoc sp. PCC 7120* (Fig. 4) and *Streptomyces coelicolor A3(2)* (Fig. 6) are characterized by AT and GC-content, respectively. It is hoped that analyses such as these will lead to a better understanding of how and why bias identification algorithms become confounded in the first place, and how we can avoid this problem in the future.

Genomic adherence measures and visualization techniques such as our polar bias view are excellent tools for investigating and understanding the forces at work in the universe of codon usage. They provide insights into the nature of the biases at play and lead to advances in the discovery and isolation of secondary biases within an organism's codon usage space.

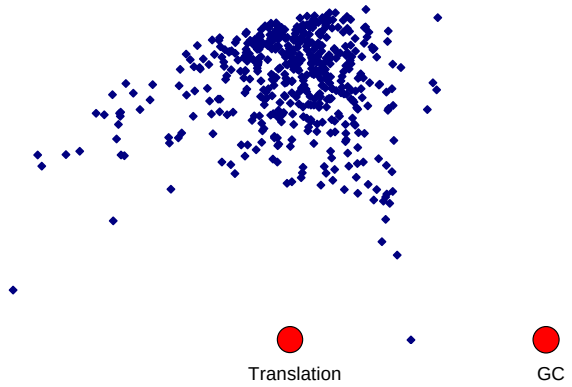


Figure 6: *Streptomyces coelicolor A3(2)* Polar Bias View. Indicative of biases that are in close proximity. Each point is a gene and the distance from the gene to either pole (bias) is $1 - CAI(g)$. Previous polar view (Fig. 4) was of biases in disparate regions of the codon usage space.

References

[1] J. Bennetzen and B. Hall. Codon selection in yeast. *J. Biol. Chem.*, 257(6):3026–3031, 1982.

[2] A. Carbone, F. Képès, and A. Zinovyev. Codon bias signatures, organization of microorganisms in codon space, and lifestyle. *Mol Biol Evol*, 22(3):547–61, Mar 2005.

[3] A. Carbone, A. Zinovyev, and F. Kepes. Codon adaptation index as a measure of dominating codon bias. *Bioinformatics*, 19(16):2005–2015, 2003.

[4] S. Garcia-Vallvé, A. Romeu, and J. Palau. Horizontal Gene Transfer in Bacterial and Archaeal Complete Genomes. *Genome Res.*, 10(11):1719–1725, 2000. http://www.fut.es/˜debb/HGT/.

[5] M. Gouy and C. Gautier. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.*, 10 (22):7055–7074, 1982.

[6] R. Grantham, C. Gautier, M. Gouy, M. Jacobzone, and R. Mercier. Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucl. Acids Res.*, 9(1):r43–74, 1981.

[7] R. Grantham, C. Gautier, M. Gouy, R. Mercier, and A. Pav. Codon catalog usage and the genome hypothesis. *Nucl. Acids Res.*, 8(1):r49r62, 1981.

[8] R. J. Grocock and P. M. Sharp. Synonymous codon usage in *Pseudomonas aeruginosa* PA01. *Gene*, 289(1-2):131–139, May 2002.

[9] T. Ikemura. Correlation between the abundance of *Escherichia coli* transfer rnas and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.*, 146:1–21, 1981.

[10] T. Ikemura. Correlation between the abundance of *Escherichia coli* transfer rnas and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J. Mol. Biol.*, 151:389–409, 1981.

[11] S. Kanaya, Y. Kudo, Y. Nakamura, and T. Ikemura. Detection of genes in *Escherichia coli* sequences determined by genome projects and prediction of protein production levels, based on multivariate diversity in codon usage. *Comput. Appl. Biosci*, 12:213–225, 1996.

[12] A. C. McHardy, A. Phler, J. Kalinowski, and F. Meyer. Comparing expression level-dependent features in codon usage with protein abundance: An analysis of 'predictive proteomics'. *Proteomics*, 4(1):46–58, 2004.

[13] A. McLachlan, R. Staden, and D. Boswell. A method for measuring the non-random bias of a codon usage table. *Nucl. Acids Res.*, 12(24):9567–9575, 1984.

[14] NCBI. National center for biotechnology information. (May 20) http://www.ncbi.nih.gov/, May 2005.

[15] P. M. Sharp, E. Cowe, D. G. Higgins, D. C. Shields, K. H. Wolfe, and F. Wright. Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. *Nucleic Acids Res*, 16(17):8207–11, Sep 1988.

[16] P. M. Sharp and W. H. Li. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol*, 24(1-2):28–38, 1986.

[17] P. M. Sharp and W. H. LI. The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, 15:1281–1295, 1987.

[18] D. Shields, P. M. Sharp, D. G. Higgins, and F. Wright. "Silent" sites in Drosophilia genes are not neutral: evidence of selection among synonymous codons. *Mol. Biol. Evol.*, 5:704–716, 1988.

[19] D. C. Shields and P. M. Sharp. Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. *Nucl. Acids Res.*, 15(19):8023–8040, 1987.

[20] F. Wright. The effective number of codons used in a gene. *Gene*, 87:23–29, 1990.