BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Prediction models

Volume 8(3)

Schematic for efficient computation of GC, GC3, and AT3 bias spectra of genome

Ahsan Z Rizvi, T Venu Gopal, C Bhattacharya*

Department of Electronics Engineering, Defence Institute of Advanced Technology, Girinagar, Pune, 411025, India; C Bhattacharya- Email: cbhat0@diat.ac.in; * Corresponding author

Received December 16, 2011; Accepted January 07, 2012; Published February 03, 2012

Abstract:
Selection of synonymous codons for an amino acid is biased in protein translation process. This biased selection causes repetition of synonymous codons in structural parts of genome that stands for high N/3 peaks in DNA spectrum. Period-3 spectral property is utilized here to produce a 3-phase network model based on polyphase filterbank concepts for derivation of codon bias spectra (CBS). Modification of parameters in this model can produce GC, GC3, and AT3 bias spectra. Complete schematic in LabVIEW platform is presented here for efficient and parallel computation of GC, GC3, and AT3 bias spectra of genomes alongwith results of CBS patterns. We have performed the correlation coefficient analysis of GC, GC3, and AT3 bias spectra with codon bias patterns of CBS for biological and statistical significance of this model.

Keywords: Period -3, Codon bias spectra, GC and GC3 bias spectra, LabVIEW Schematic.

Background:
Bias in the selection of synonymous codons for an amino acid is termed as codon bias. Codon bias is dominant in structural parts of genome like exon, t-RNA locations **[1, 2].** Codon bias enhances the speed of protein translation with high accuracy in fast growing organisms like E.coli **[3].** Codon bias locations have high degree of repetition of codons with GC and GC at third codon position (GC3) in these organisms. Predominance of GC, GC3, and AT3 bias are the parameters those influence the codon bias patterns in genome **[4].** The phenomenon of codon bias is corroborated by presence of a strong spectral peak magnitude in Fourier domain at a frequency sample N/3, where N is the length of windowed discrete Fourier transform (DFT) of structural parts of genome **[5,6].** Several recent studies **[6-8]** utilized this period-3 spectral property of genomes to determine exon locations but without identification of its accessories like stop-start codons, splice donor-acceptors, etc. Although spectral methods to determine exon locations in genes are there, no attempt is yet made to utilize the DNA spectra in evaluating GC, GC3, and AT3 bias spectra. The 3-phase network model in the paper demonstrates a parallel and efficient way of computing codon bias spectra (CBS). We have utilized the 3-phase network model to estimate GC, GC3, and AT3 bias in spectral domain. Amplitude patterns of CBS in **(Figure 3, 4)** are showing the strength of codon bias at nucleotide positions in genome while peaks in GC and GC3 spectra are showing the GC, GC3 bias regions in genome. The schematics for GC, GC3, and AT3 bias spectra along with CBS are implemented in LabVIEW and MATLAB platforms. Structural parts of genome like genes, t-RNA locations, etc., have high codon bias those are demonstrated by high amplitude peaks of CBS, GC and GC3 bias spectra.

In this paper, analysis of the correlation coefficients of GC, GC3, and AT3 bias spectra with CBS brings out the predominance of these factors in the total codon bias patterns of genome. Selection of synonymous codons in structural parts of genome is demonstrated here by correlating codon spectra (CS) with CBS. It is shown that the values of correlation coefficients for the synonymous codons with GC and GC3 contents are higher

in comparison to those of synonymous codons ending with AT3. This observation shows natural selection of GC and GC3 containing synonymous codons in structural parts of genome.

Methodology:

We analyze the 3-phase network model with polyphase filter bank **[9, 10]** over a set of bacterial genomes listed in **Table 1 (see supplementary material**). These genomic sequences are downloaded in FASTA format from online genome database of National Centre for Biotechnology Information (NCBI), USA. Downloaded genomic sequences are represented as an array of characters l∈ {A, T, C, G}. Genome character sequences are mapped into four binary indicator sequences such that a nucleotide base is represented as 1 and all others are 0 **[6].** These indicator sequences serve here as input to the 3-phase network model of genome to produce CBS, GC, GC3, and AT3 bias spectra. Complete flow diagram for this 3-phase network model is shown in **(Figure 1).** Three arms shown in this flow diagram calculate preponderance of nucleotides at three codon positions in parallel manner.

GC, GC3, AT3 bias spectra are correlated with CBS and results are tabulated in **Table 1 (see supplementary material)** or twenty one genome sequences. Values of correlation coefficients close to one show the strength of association of GC, and GC3 bias with codon bias of genome. Analysis of correlation coefficients of CS with codon bias patterns in CBS in **Table 2 (see supplementary material)** shows the preference of synonymous codons toward protein translation.

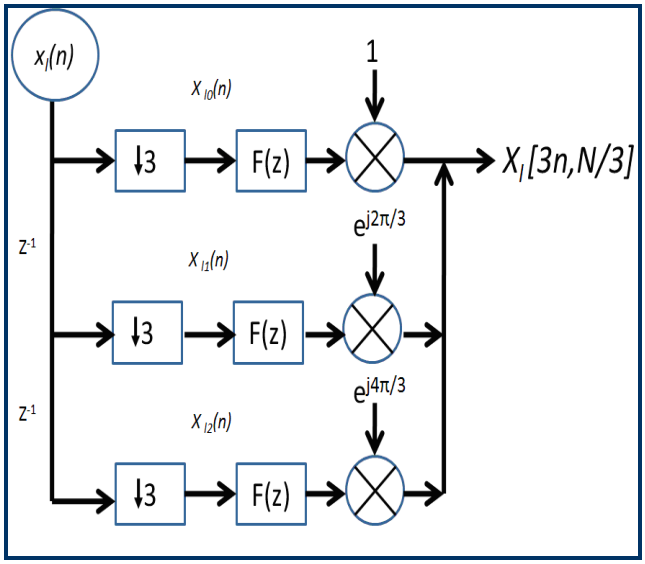

Figure 1: Schematic for the 3-phase network model of genome, where $X_l(n)$ is binary nucleotide sequence and ↓3 is down sampling after delay element Z^{-1} for codon positions. F (z) is rectangular window function before complex multiplier {1, $e^{j2\pi/3}$, $e^{j4\pi/3}$}. Summation of X_l [3n, N/3] for l ∈ {A, T, G, C} construct decimated CBS S [3n, N/3].

LabVIEW schematic model for GC, GC3, and CBS spectra:

The complete LabVIEW schematic for 3-phase network model of genome is shown in **(Figure 2).** In this schematic, connected blocks are executed through data flow programming. Here the schematic is shown for nucleotide G. Similarly schematics are developed for the rest three nucleotides A, T, and C. The length of the rectangular window function f (n) is L = 117, and the size of moving window for generation of DFT is N = 351. The counter i is an index for array. Delay elements or codon positions are indicated in the schematic by blocks marked as {0,-1,-2}, and the blocks labeled as {Xg0, Xg1, Xg2} are the convolution outputs. DFT of nucleotide G sequence Xg [3n] is generated by summation of {Xg 0, Xg 1, and Xg 2} after complex multiplier. Power spectrum of genome or CBS S [3n, N/3] with period-3 properties is generated by summation of absolute squared magnitude of X_l [3n, N/3]. CBS of genome thus obtained are in decimated form, and we have performed the cubic spline interpolation to match with exact length of genome. GC bias spectra is obtained by fixing l ∈ {G, C} while GC3 and AT3 bias spectra are obtained by fixing index j= 2 with l ∈ {G, C} or l ∈ {A, T} in the schematic shown in **(Figure 2).**

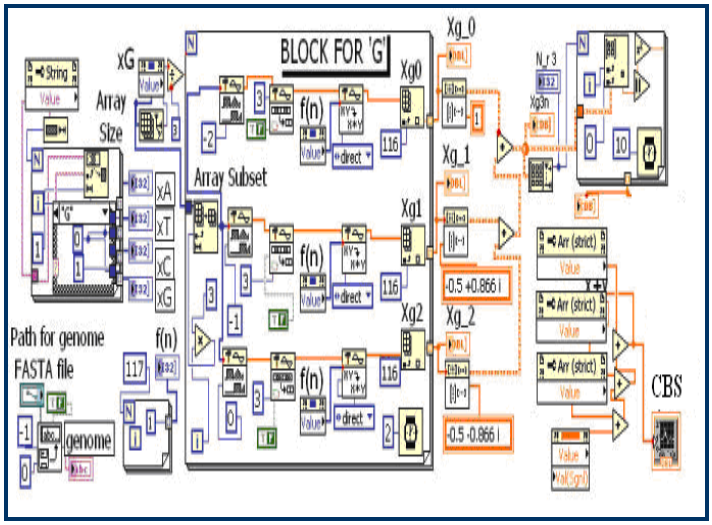

Figure 2: LabVIEW schematic for computation of decimated CBS S [3n, N/3] of genome.

Discussion:

Generated patterns of CBS, GC, GC3, and AT3 bias spectra through the 3-phase network model are shown in **(Figure 3, 4)** for the first 15000 nucleotide bases of *E. coli*, and *Y. pestis* genomes. High peaks in the amplitude spectrum of CBS indicate the codon bias strength at nucleotide positions in genome. It is visually shown in these figures that the spectral patterns of CBS, GC, and GC3 bias are associated. GC and GC3 spectral peaks are increasing with increase of peaks in CBS while AT3 spectral peaks are not correlated with CBS. These visual observations are statistically validated by correlation coefficient analysis of CBS over GC, GC3, and AT3 spectra and results are listed in **Table 1 (see supplementary material)** for twentyone bacterial genomes. High values of correlation coefficients listed in this table for GC, GC3 bias in comparison to AT3 bias indicate the selectivity of GC, GC3 bias for the total codon bias pattern generation. **Table 2 (see supplementary material)** is showing the correlation coefficient values of synonymous codons for amino acids Pro, Ala, and Gly those are obtained by correlating their CS with CBS. High values of correlation coefficients shown in this table demonstrate preference towards synonymous codons ending with GC3. These observations explain natural selection of GC3 ending in synonymous codons.

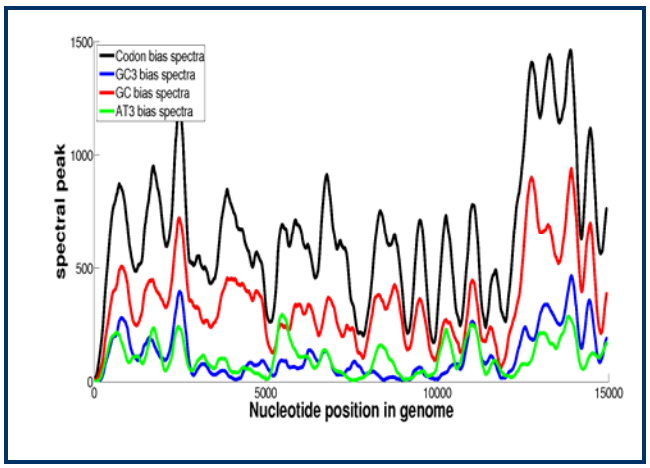

Figure 3: Plots of CBS, GC, GC3, and bias AT3 spectra with first 15000 nucleotide bases of *E. coli* 536 genome.

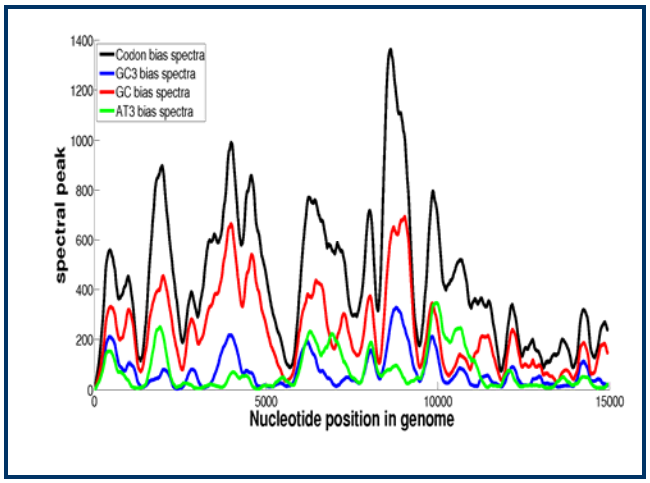

Figure 4: Plots of CBS, GC, GC3, and bias AT3 spectra with first 15000 nucleotide bases of *Y. pestis* C092 chr. genome.

Conclusion:

There is requirement of fast computational schemes for producing GC, GC3, and AT3 bias spectra along with codon bias spectra. 3-phase network model of genome has shown the fitness in extracting GC, GC3, and AT3 spectra along with parallel computation. LabVIEW schematic of this model is a gateway for hardware implementation. High correlation coefficient values are observed for the synonymous codons which are richer in GC, and GC3 contents and correlation coefficients for GC and GC3 bias are higher than AT3 bias. These statistical observations state that the synonymous codons with GC and GC3 contents are less prone to mutations.

Acknowledgment:

The authors acknowledge the financial support for the research work from DIAT (DU), Pune, India.

References:

[1] Hershberg R & Petrov DA, *PLoS Genet*. 2010 **6**: e1001115 [PMID: 20838599]
[2] Plotkin JB & Kudla G, *Nat Rev Genet*. 2011 **12**: 32 [PMID: 21102527]
[3] Novey R & Drott D, *InNovation*. 2001 **12**:1
[4] Palidwor GA *et al. PLoS One.* 2010 **5**: e13431 [PMID: 21048949]
[5] Sánchez J, *Bioinformation*. 2011 **6**: 327 [PMID: 21814388]
[6] Vaidyanathan PP. *IEEE Cir and Syst Magz*. 2004 **4**: 6
[7] Wang L & Stein LD, *BMC Bioinformatics*. 2010 **11**: 550 [PMID: 21059240]
[8] Yin C & Yau SS, *J Theor Biol*. 2007 **247**: 687 [PMID: 17509616]
[9] Tuqan J & Rushdi A, *IEEE J of Sel Top Signal Process*. 2008 **2**: 343
[10] Vadyanathan PP, *Proc of IEEE*. 1990 **78**: 56

Edited by P Kangueane
Citation: Rizvi *et al*. Bioinformation 8(3): 163-166 (2012)

BIOINFORMATION

open access

Supplementary material:

Table 1: Spectral correlation coefficient values of genomes are tabulated with their first 15000 nucleotide bases. Correlation coefficients of GC, GC3, and AT3 bias are obtained in this table by correlating GC, GC3, and AT3 bias spectra with CBS of genomes.

Genome	Accession	Reading frame no.	Correlation Coefficient		
			GC Bias	GC3 Bias	AT Bias
A. tumefaciens	NC 003062	03	0.9415	0.8053	0.7463
A. haemolyticum	NC 014218	01	0.9196	0.7476	0.3939
A. arilaitensis	NC 014550	02	0.9809	0.8706	0.7710
B. faecium	NC 013172	02	0.9680	0.7707	0.5767
B. fragilis	NC 003228	03	0.8310	0.7699	0.6101
C. glutamicum	NC 003450	03	0.9646	0.8040	0.1732
C. diptheriae	NC 002935	01	0.9155	0.7546	0.6346
E. coli	NC 008253	03	0.9096	0.7933	0.5782
H. butylicus	NC 008818	03	0.8334	0.5228	0.2012
M. fervens	NC 013156	03	0.9722	0.5882	0.6127
M. tuberculosis	NC 000962	01	0.9846	0.9728	0.7588
M. testaceum	NC 015125	01	0.9844	0.9480	0.8441
P. acnes	NC 006085	01	0.8310	0.7699	0.6101
S. suis	NC 012926	01	0.9222	0.6569	0.3511
S. avermitilis	NC 003155	01	0.9654	0.6773	0.5589
S. violaceusniger	NC 015957	03	0.9842	0.4723	0.3391
S. scabiei	NC 013929	01	0.9653	0.8963	0.1690
T. sibiricus	NC 012883	02	0.8148	0.7594	0.4046
T. whippleis	NC 004551	01	0.9007	0.8470	0.7319
Y. pestis	NC 003142	01	0.9077	0.6955	0.4215
Z. galactanivorans	NC 015844	01	0.9796	0.8749	0.7759

Table 2: Correlation coefficient values of synonymous codons present in *T. sibiricus*. These correlation coefficient values have obtained by correlating CS with CBS. High correlation coefficient values of synonymous codons with GC3 endings are tabulated in bold fonts.

Amino Acid	Synomymous codon	Correlation Coefficient
Pro	CCU	0.3053
	CCC	**0.7690**
	CCA	0.4425
	CCG	**0.5457**
Ala	GCU	0.3515
	GCC	**0.6127**
	GCA	-0.4512
	GCG	**0.5122**
Gly	GGU	-0.4257
	GGC	**0.4002**
	GGA	-0.0402
	GGG	**0.6510**