BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 10(5)

A cross talk between codon usage bias in human oncogenes

Tarikul Huda Mazumder, Supriyo Chakraborty* & Prosenjit Paul

Department of Biotechnology, Assam University, Silchar 788011, Assam, India; Supriyo Chakraborty – E mail: supriyoch_2008@rediffmail.com; *Corresponding author

Received April 03, 2014; Accepted April 15, 2014; Published May 20, 2014

Abstract:
Background: Oncogenes are the genes that have the potential to induce cancer. The extent and origin of codon usage bias is an important indicator of the forces shaping genome evolution in living organisms. **Results**: We observed moderate correlations between gene expression as measured by CAI and GC content at any codon site. The findings of our results showed that there is a significant positive correlation (Spearman's r= 0.45, P<0.01) between GC content at first and second codon position with that of third codon position. Further, striking negative correlation (r = -0.771, P < 0.01) between ENC with the GC3s values of each gene and positive correlation (r=0.644, P<0.01) in between CAI and ENC was also observed. **Conclusions**: The mutation pressure is the major determining factor in shaping the codon usage pattern of oncogenes rather than natural selection since its effects are present at all codon positions. The results revealed that codon usage bias determines the level of oncogene expression in human. Highly expressed oncogenes had rich GC contents with high degree of codon usage bias.

Keywords: Synonymous codon, Oncogene, Codon usage pattern.

Background:
Nature has gifted the genetic code that provides the basic instructions and information to direct efficient protein synthesis and folding. There are sixty-one codons that specify for only twenty amino acids found commonly in protein sequences; most of these amino acids (building blocks of protein) can be encoded by more than one codon (i.e., a triplet of nucleotides); such codons are described as being synonymous, and mostly differ by one nucleotide in the third position **[1].** The term codon bias or more preferably codon usage bias represents the unequal usage of synonymous codons for encoding amino acids which may vary significantly between genomes, between genes in the same genome, and within a single gene **[2-3].** Since the 1970s, the unequal use of synonymous codons has been confirmed in many organisms. To date, the codon usage patterns in many organisms have been interpreted for diverse reasons. Recently, it has been reported that two major factors are involved in the continuation of codon usage bias: weak natural selection and mutational pressure **[4]**. The selection associated with translational efficiency/accuracy is often termed as 'translation selection'. Moreover, scientific investigation also reported that synonymous codon usage pattern varied at distinct sites along a coding sequence, balances of strong versus weak base pair bonding, maintenance of DNA and RNA secondary structure, and translational efficiency and fidelity **[5].** That is why codon usage bias among different organisms or within the genes of the same organism has invited much attention and various works on the subject have been published in recent years.

Lavner and Kotlar (2005) suggested that there are three possible ways in which selection may act on codon bias in the human genome: (1) Increasing translation efficiency in highly expressed genes; (2) Regulating translation efficiency of some proteins that can be a disadvantage at high levels; and (3) Improving translation efficiency and reducing the rate of amino acid misincorporation in the production of biosynthetically expensive proteins **[3].** Many genomic analyses have been done

on oncogenes but till date very little is known about the codon usage patterns and the factors that influence them. Codon usage patterns are important for bringing out molecular mechanism and evolutions of oncogenes. In this paper we have analyzed the key genetic factors playing crucial role in determining the codon usage pattern in fifty (50) oncogenes. To the best of our knowledge, it is the first systematic study to verify and insist that the synonymous codon usage pattern is one of the factors affecting the codon usage in oncogenes..

Methodology:

Retrieval of Sequence data

A list of human oncogenes was compiled from the web site (http://cbio.mskcc.org/CancerGenes/Select.action). Complete nucleotide coding sequence of each of the concerned gene, was obtained from NCBI nucleotide database website (http://www.ncbi.nlm.nih.gov). Codon usage bias was measured in the 50 oncogenes listed in **Table 1 (see supplementary material)**. The complete coding sequence (cds) of each oncogene was analyzed using PERL program developed by us.

Analysis of synonymous codon usage bias

We measured the non-uniform usage of synonymous codons for the oncogenes by analyzing several genetic indices given below:

Nucleotide composition

The frequency of the nucleotide G+C at the synonymous third codon position (GC3s) is a good indicator of the extent of base composition bias **[6].** The frequencies of the nucleotides A, C, U (T), G in the complete coding sequences of each oncogene and the occurrence of overall (G+C)% content at the different codon positions GC1%, GC2%, GC3% was calculated to study the relationship between codon usage variation and compositional constraints.

Effective number of codons

The effective number of codons (ENC) is generally used to measure the codon usage bias of a gene that is independent of the gene length and number of amino acids **[7]**. The ENC value ranges from 20-61. For a gene in which only one codon is used for each amino acid, this value would be 20 while all codons are used equally the value would be 61 **[7].** The ENC value closer to 20 indicates, strong codon usage bias in the gene and these biased genes are expressed highly **[8].** The ENC values for all cds sequence were computed as per Wright (1990) **[7].** In addition, to examine the influence of GC content on codon usage, the relationship of ENC and GC3s content of each gene was plotted according to the equation described by Wright (1990) **[7].**

Codon adaptation index

Codon adaptation index (CAI) is used to estimate the degree of bias toward codons in highly expressed genes and thus assesses the effective selection which helps in shaping the codon usage pattern **[9-10].** The CAI ranges from 0 to 1, for a gene in which all synonymous codons are used equally, the value would be 0 for no bias while only optimal codons are used, value will be 1 for strongest bias **[11]**. The CAI value was measured as per Sharp PM *et al*. **[12].**

Frequency of optimal codons

Frequency of optimal codons (Fop) is used to measure codon usage bias in a gene **[11].** Fop is calculated as the ratio of the number of optimal codons used to the total number of synonymous codons **[13].** The Fop value ranges from 0.36, for a gene in which codon usage pattern is uniform, to 1 for a gene in which codon usage is highly biased **[11].** We used the formula given by Lanver & Kotlar to calculate the Fop values for each of the cds selected for the present study **[3].**

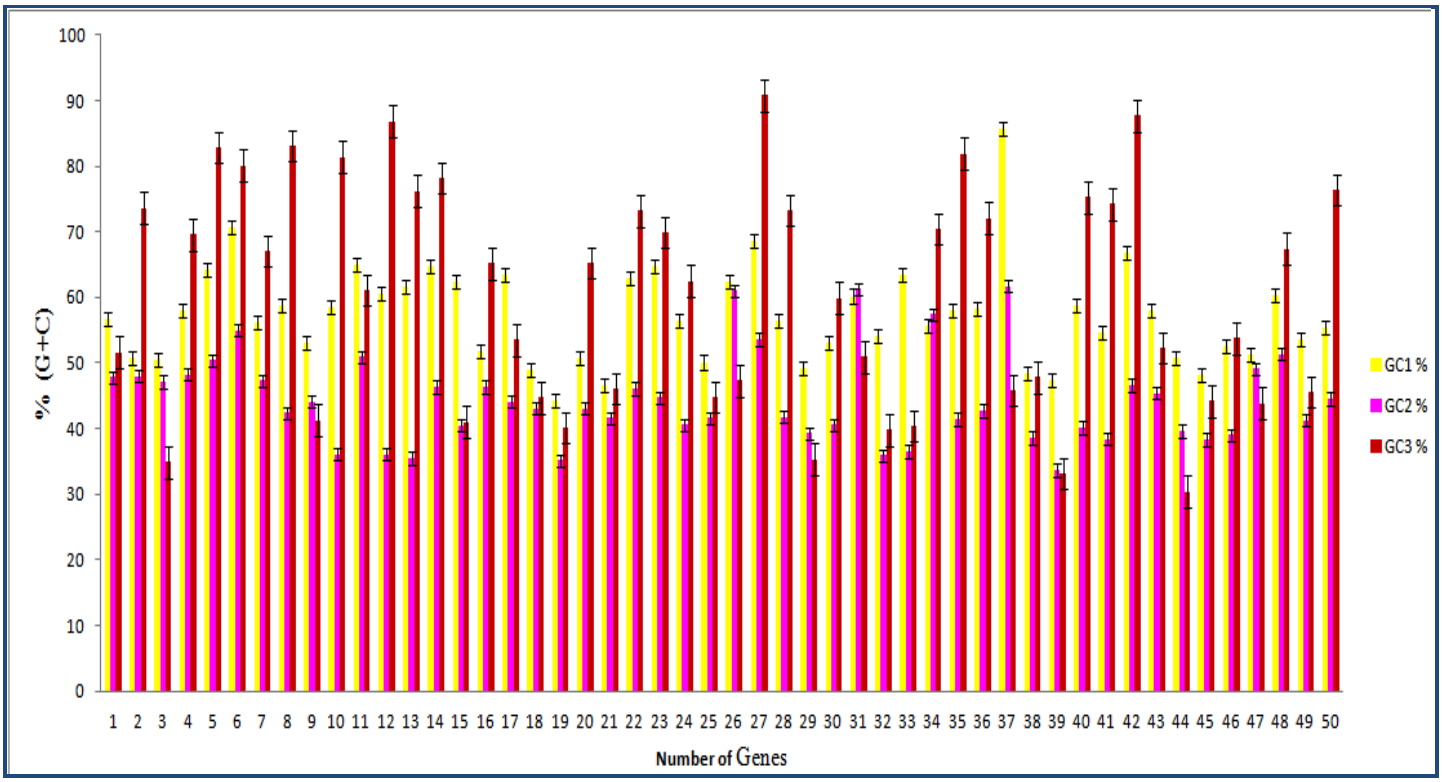

Figure 1: Percentage of GC content at three codon positions.

Relative synonymous codon usage

Relative synonymous codon usage (RSCU) is calculated by dividing the observed frequency of a codon by the expected if all synonymous codons for that amino acid were used equally **[14].** Thus, an RSCU value close to 1 indicates a lack of bias, RSCU >1 indicates a codon used more frequently than expected randomly, and RSCU <1 indicates a codon used less frequently than expected randomly **[14].**

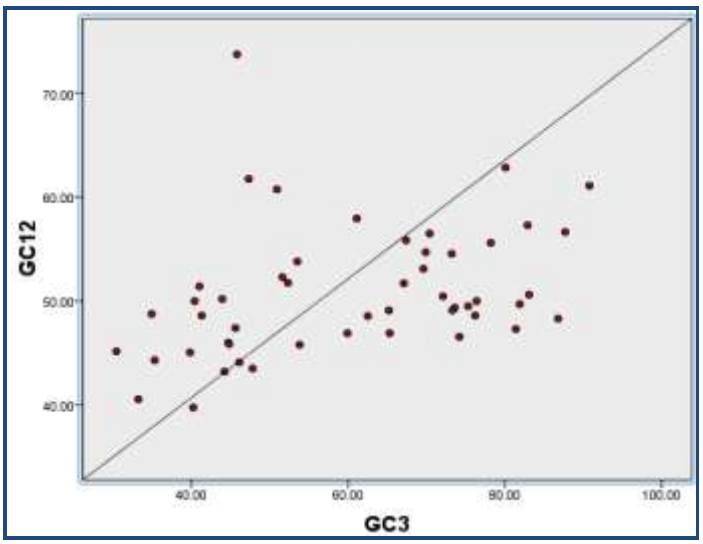

Figure 2: Correlation between GC content at first and second codon positions (GC1 & GC2) with that at synonymous third codon positions (GC3s). GC12: average GC content at first and second codon positions.

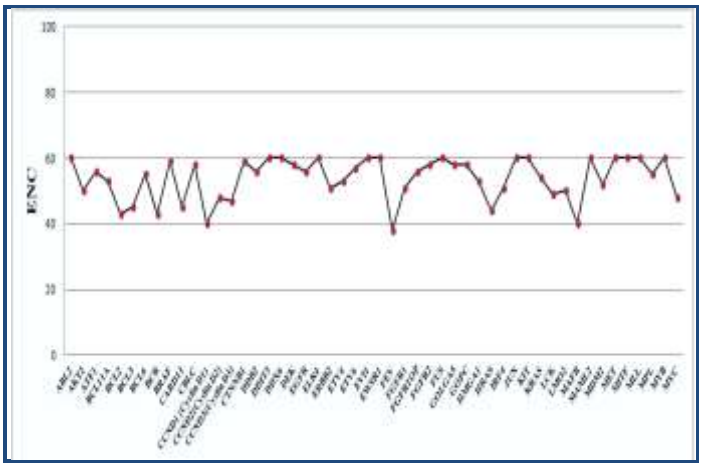

Figure 3: ENC distribution of 50 selected oncogenes.

Correlation analysis

Correlation analysis was used to identify the relationship between the pattern of synonymous codon usage and the genetic indices used for the present study. This analysis was implemented based on the Spearman's rank correlation analysis. All statistical analyses were carried out by using software SPSS.

Results:

In this present study, the selected oncogenic cds sequences were downloaded from NCBI nucleotide database using a perl program. The program was written in such a way that it selects only those cds sequences which have perfect start and stop signal and devoid of any unknown bases (N). We found fifty cds sequences in correct format for codon bias study. The extent of codon usage bias was determined in these fifty oncogenes **Table 1.** Two amino acids methionine and tryptophan coded by single codon ATG and TGG, respectively and three stop codons (TAA, TAG, and TGA) would not reveal any usage bias and therefore discarded from the calculation.

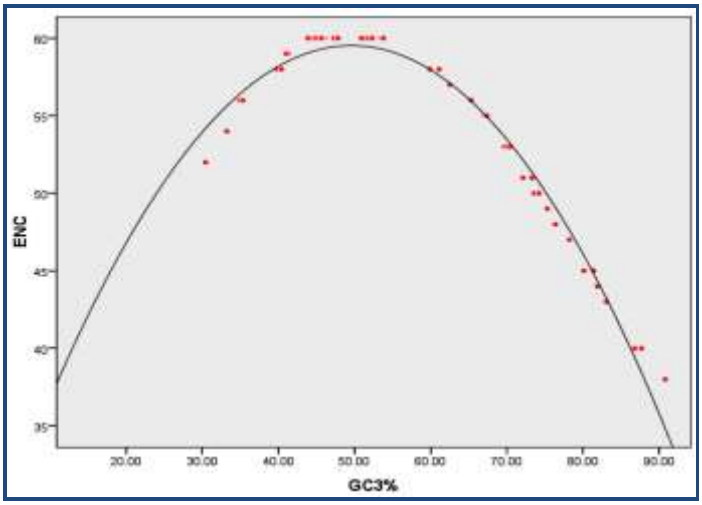

Figure 4: Distribution of ENC and GC content of the third codon position of 50 different oncogenes. The continuous curve represents the expected curve between ENC and GC contents under random codon usage.

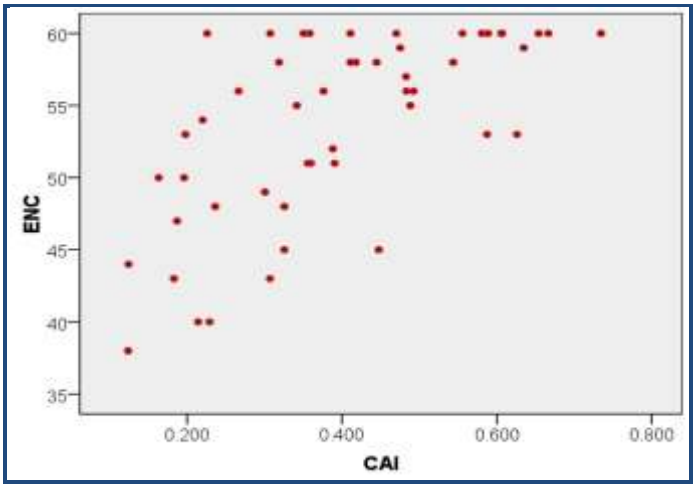

Figure 5: Correlation between Effective number of codons (ENC) and Codon adaptation index (CAI).

Codon usage bias and correlation with GC3s

The overall percentage of guanine and cytosine contents GC% and adenine and thymine contents AT% on the first, second, and third codon positions of the 50 target oncogenes of human were investigated **Table 2 (see supplementary material).** It can be assumed that the evolution of codon usage might be either controlled by natural selection or by mutation pressure. To determine the extent of the role of these two evolutionary forces on the codon usage pattern of human oncogene, we performed correlation analysis between different nucleotide constraints. First we calculated the GC content at different codon positions **(Figure 1)** and it was found that the GC content at each codon position varies among the genes. Finally, compared GC content

at first codon position (GC1) and second codon positions (GC2) with that of third codon positions (GC3s) and observed a significant positive correlation (r=0.45, P<0.01) **(Figure 2)**, that reveals base compositions are prone to the result of mutation pressure rather than natural selection, since at all codon positions its effects are present.

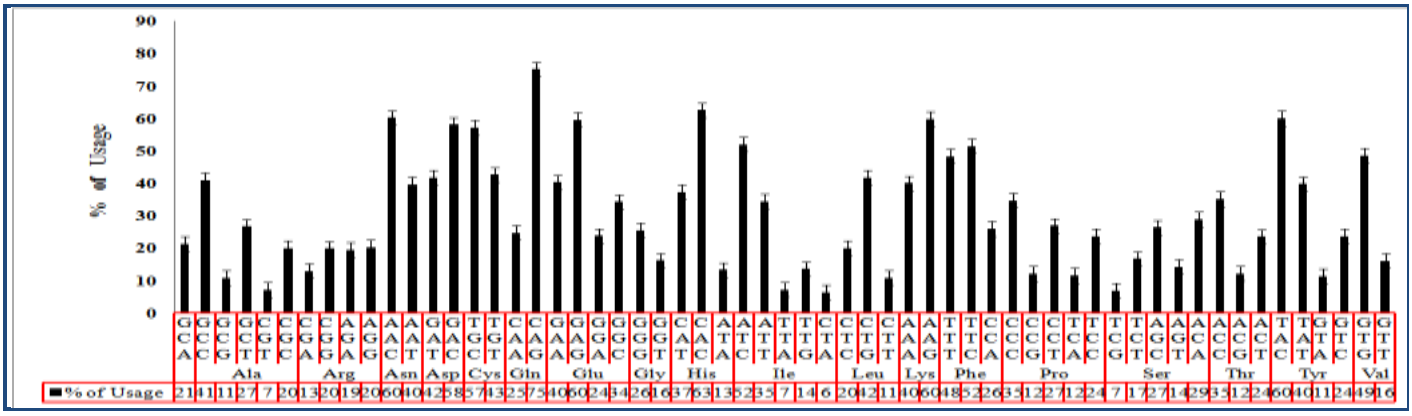

Figure 6: Frequency of highest and least used codons among the 50 cds selected for the present study.

Effective number of codons and its relationship with GC3s values

The average ENC value used by the oncogenes was found to be 53.74 with a range of 38 to 60. Thirty eight oncogenes had ENC values in the range 50-60, 11 in the range 40-50 and 1 between 38 and 40 **(Figure 3).** Therefore, codon usage bias is in most cases little, although some variation is evident. Moreover the GC3 values were found to range from 0.3 to 0.1. We calculated the correlation coefficient between ENC and GC3s values. The results showed that the ENC value was strongly negatively correlated with the GC3s values of each gene (r = -0.771, P < 0.01). These calculations suggested that genes with higher GC3s values and lower ENC values had strong bias. Finally, we plotted the ENC against the GC3% values to investigate the general codon usage variation with different GC content of each gene **(Figure 4) [7].** The continuous curve represents the expected positions of genes where GC3 values are the only determinant factor shaping the codon usage pattern. Most genes were found to be located on or above the reference line, representing that the codon usage pattern was only determined by GC3 values. Moreover, some genes located above the reference line, indicates that GC3 is not the only factor for shaping the codon usage pattern other factors like nucleotide composition, may be involved for these genes.

Level of oncogene expression and codon bias

The level of expression of oncogene was measured through codon adaptation index (CAI) values **[10, 15],** which varied from 0.124 to 0.735 with the mean of 0.395 and standard deviation of 0.159. The CAI value indicates that most of the genes selected for the present study are highly expressive in nature. Moreover, a significant negative correlation was observed between CAI & GC3s (r=-0.489, P<0.01) and CAI & GC content (r=-0.463, P<0.01). Furthermore, significant positive correlation was also observed in between ENC and CAI (r= 0.644, P<0.01) **(Figure 5).** The results revealed that codon usage pattern determines the level of all expression in human and highly expressed genes have high GC contents and a greater extent of codon usage bias. We also calculated the frequency of the occurrence of synonymous codons for the amino acids. The frequency was allied with statistical analysis to find out the highest and lowest frequently used codon (**Figure 6).** Relative synonymous codon usage (RSCU) values for each synonymous codon were calculated to find out the highest and least abundant codons. The results of our analysis indicate that the highest abundant codon is CTG for Leucine and GTT for Valine. Least abundant codons are GTC, ACT, TCG, CTA, and ATA for amino acid valine, threonine, serine, leucine and isoleucine, respectively **(Figure 7).**

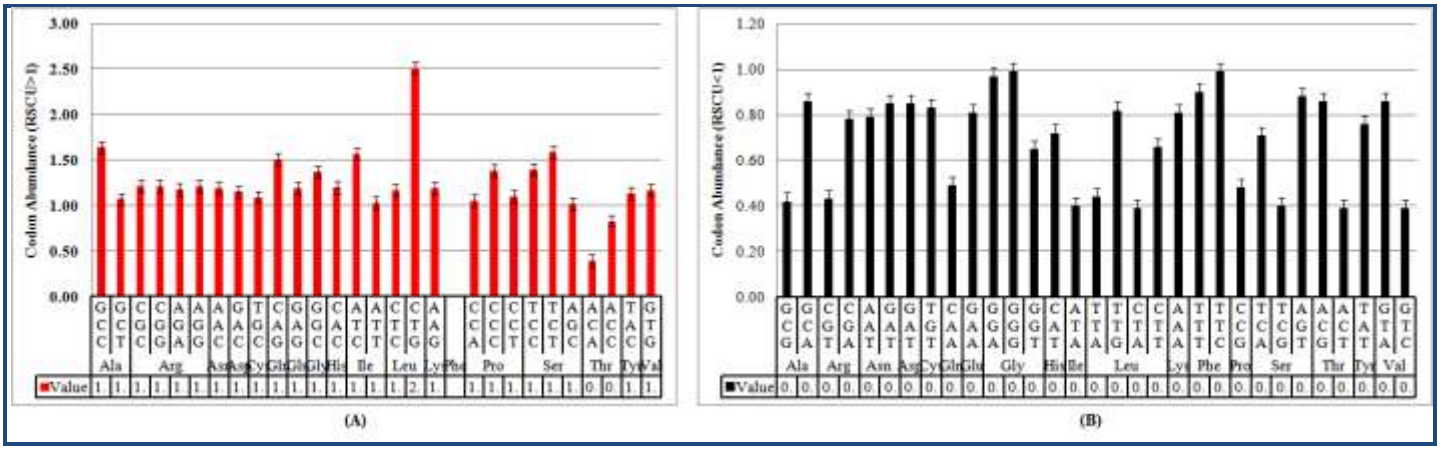

Figure 7: Relative synonymous codon usage and codon usage bias among the selected 50 cds. (A): Most abundant codons, RSCU > 1. (B): Least abundant codons, RSCU < 1.

Discussion:

In brief, we analyzed the codon usage pattern and the key genetic factors playing decisive role in determining the pattern of codon usage for the fifty oncogenes. Based on the hypothesis that gene expressivity and codon composition are strongly correlated, the codon adaptation index has been defined to provide an intuitively meaningful measure of the extent of the codon preference in a gene. The present study was carried out to analyze the CAI, Fop, ENC, RSCU, base composition for the oncogenes, and also to find out the level at which the above mentioned genetic factors are involved in the formation of codon usage pattern. As per our mentioned objectives in this present study, we selected fifty oncogenes from *Homo sapiens* for CUB analysis. The accurate coding sequences having correct initial and termination codons were retrieved using a program in perl, developed by us. After analyzing the cds sequences it was found that 70% of the cds selected for the study are rich in GC. We also predicted the heterogeneity of codon usage by analyzing the effective number of codons (ENC). We also measured the variation of codon usage bias among the oncogenes, further confirmed by the distributions of GC content at the third synonymous codon positions. These results indicate that apart from compositional constraints, other trends might influence the overall codon usage variation among the oncogenes. We calculated the CAI values for the oncogenes and it was found that seventy five percent of the cds selected from *Homo sapiens* qualify as highly expressed genes. We analyzed normalized AT and GC frequency at each codon site. Significant correlation was observed between gene expression as measured by CAI and GC content at any codon site. Among all GC3s showed highest correlation (-0.489) with gene expression. The frequency of the occurrence of each synonymous codon for the amino acids was calculated. The frequency was allied with statistical analysis to find out the highest and lowest frequently used codon. At the end of our frequency analysis we found that AAC, GAC, TGC, CAG, GAG, CAC, AAG and TAC are the codons used most frequently among cds sequence of oncogenes.

Conclusion:

The mutation pressure is the major determining factor in shaping the codon usage pattern of oncogenes rather than natural selection since its effects are present at all codon positions. The results revealed that codon usage bias determines the level of oncogene expression in human. Highly expressed oncogenes had rich GC contents with high degree of codon usage bias.

Acknowledgments:

We are thankful to Assam University, Silchar, Assam, India for providing the necessary facilities in carrying out this research work.

References:

[1] Angov E, *Biotechnol J.* 2011 **6**: 650 [PMID: 21567958]
[2] Hooper SD & Berg OG, *Nucleic Acids Res.* 2000 **28**: 3517 [PMID: 10982871]
[3] Lavner Y & Kotlar D, *Gene* 2005 **345**: 127 [PMID: 15716084]
[4] Hershberg R & Petrov DA, *Annu Rev Genet.* 2008 **42**: 287 [PMID: 18983258]
[5] Cai MS *et al. Intervirology* 2009 **52**: 266 [PMID: 19672100]
[6] Zhou T *et al. Biosystems* 2005 **81**: 77 [PMID: 15917130]
[7] Wright F, *Gene* 1990 **87**: 23 [PMID: 2110097]
[8] Li ZP *et al. Virol Sin*. 2010 **25**: 329 [PMID: 20960179]
[9] Naya H *et al. FEBS Lett.* 2001 **501**: 127 [PMID: 11470270]
[10] Gupta SK *et al. J Biomol Struct Dyn*. 2004 **21**: 527 [PMID: 14692797]
[11] Stenico M *et al. Nucleic Acids Res.* 1994 **22:** 2437 [PMID: 8041603]
[12] Sharp PM & Li WH, *Nucleic Acids Res*. 1987 **15**: 1295 [PMID: 3547335]
[13] Ikemura T, *J Mol Biol*. 1981 **151**: 389 [PMID: 6175758]
[14] Sharp PM & Li WH, *Nucleic Acids Res*. 1986 **14**: 7749 [PMID: 3534792]
[15] Behura SK & Severson DW, *PLoS One* 2012 **7**: e43111 [PMID: 22912801]

Edited by P Kangueane

Citation: **Mazumder *et al*.** Bioinformation 10(5): 256-262 (2014)

BIOINFORMATION

open access

Supplementary material:

Table 1: The information of 50 Oncogenes used in this study with accession number and gene length

SL.NO.	GENES	ACCESSION NO.	GENE LENGTH Input CDS bp)
1	ABL2	DQ009672.1	3549
2	AKT2	BC063421.1	444
3	ATF1	BC029619.1	816
4	BCL11A	GU324937.1	2508
5	BCL2	AY220759.1	720
6	BCL3	M31732.1	1341
7	BCL6	EU883531.1	1953
8	BCR	U07000.1	3816
9	BRAF	EU600171.1	2301
10	CARD11	BC111719.1	3444
11	CBLC	BC006122.1	678
12	CCND1 (Cyclin D1)	M64349.1	888
13	CCND2(Cyclin D2)	M90813.1	870
14	CCND3(Cyclin D3)	M90814.1	879
15	CTNNB1	AB451264.1	2350
16	DDB2	AY220533.1	1284
17	DDIT3	AY880949.1	510
18	DDX6	BC039826.1	564
19	DEK	BC035259.1	1128
20	EGFR	U48722.1	1218
21	ELK4	BC063676.1	1218
22	ERBB2	AY208911.1	3768
23	ETV4	BC016623.1	1455
24	ETV6	BC043399.1	1359
25	EVI1	GQ352634.1	3156
26	EWSR1	BC011048.1	1968
27	FEV	BC023511.2	717
28	FGFR1	AY585209.1	2469
29	FGFR1OP	BC037785.1	450
30	FGFR2	M97193.1	2469
31	FUS	CR456747.1	1581
32	GOLGA5	BC023021.1	2196
33	GOPC	KF420123.1	1224
34	HMGA1	BC067083.1	324
35	HRAS	EF015887.1	513
36	IRF4	BC015752.1	1356
37	JUN	J04111.1	997
38	KIT	U63834.1	2931
39	KRAS	JX512447.1	570
40	LCK	M36881.1	1530
41	LMO2	BC034041.1	477
42	MAFB	BC036689.1	972
43	MAML2	AY040322.1	3462
44	MDM2	GQ848196.1	1401
45	MET	J02958.1	4227
46	MITF	BC065243.1	1260
47	MLL	AY373585.1	11910
48	MPL	M90103.1	1740
49	MYB	AF104863.1	1923
50	MYC	AY214166.1	1320

Table 2: GC content and the AT contents at different codon positions in the complete coding regions of 50 oncogenes

SL.NO.	GENES	GC %	GC1 %	GC2 %	GC3 %	AT1 %	AT2	AT3 %
1	ABL2	52.1	56.8	47.8	51.6	43.2	52.2	48.4
2	AKT2	57.4	50.7	48	73.6	49.3	52	26.4
3	ATF1	44.1	50.4	47.1	34.9	49.6	52.9	65.1
4	BCL11A	58.6	57.9	48.3	69.6	42.1	51.7	30.4
5	BCL2	65.8	64.2	50.4	82.9	35.8	49.6	17.1
6	BCL3	68.6	70.7	55	80.1	29.3	45	19.9
7	BCL6	56.8	56.1	47.3	67.1	43.9	52.7	32.9
8	BCR	61.4	58.8	42.4	83.1	41.2	57.6	16.9
9	BRAF	46.2	53.1	44.1	41.3	46.9	55.9	58.7
10	CARD11	58.7	58.5	36.1	81.4	41.5	63.9	18.6

11	CBLC	59	65	50.9	61.1	35	49.1	38.9
12	CCND1 (Cyclin D1)	61.1	60.5	36.1	86.8	39.5	63.9	13.2
13	CCND2(Cyclin D2)	57.8	61.7	35.5	76.2	38.3	64.5	23.8
14	CCND3(Cyclin D3)	63.1	64.8	46.4	78.2	35.2	53.6	21.8
15	CTNNB1	48	62.3	40.5	41	37.7	59.5	59
16	DDB2	54.4	51.9	46.3	65.2	48.1	53.7	34.8
17	DDIT3	53.7	63.5	44.1	53.5	36.5	55.9	46.5
18	DDX6	45.6	48.9	43.1	44.7	51.1	56.9	55.3
19	DEK	39.9	44.4	35.1	40.2	55.6	64.9	59.8
20	EGFR	53	50.7	43.1	65.3	49.3	56.9	34.7
21	ELK4	44.7	46.6	41.6	46.1	53.4	58.4	53.9
22	ERBB2	60.8	63	46.1	73.2	37	53.9	26.8
23	ETV4	59.8	64.7	44.7	69.9	35.3	55.3	30.1
24	ETV6	53.2	56.5	40.6	62.5	43.5	59.4	37.5
25	EVI1	45.5	50.1	41.6	44.8	49.9	58.4	55.2
26	EWSR1	56.9	62.5	61	47.3	37.5	39	52.7
27	FEV	71	68.6	53.6	90.8	31.4	46.4	9.2
28	FGFR1	57.1	56.4	41.8	73.3	43.6	58.2	26.7
29	FGFR1OP	41.3	49.3	39.3	35.3	50.7	60.7	64.7
30	FGFR2	51.2	53.2	40.6	59.9	46.8	59.4	40.1
31	FUS	57.4	60.2	61.3	50.9	39.8	38.7	49.1
32	GOLGA5	43.3	54.2	35.9	39.8	45.8	64.1	60.2
33	GOPC	46.8	63.5	36.5	40.4	36.5	63.5	59.6
34	HMGA1	61.1	55.6	57.4	70.4	44.4	42.6	29.6
35	HRAS	60.4	57.9	41.5	81.9	42.1	58.5	18.1
36	IRF4	57.7	58.2	42.7	72.1	41.8	57.3	27.9
37	JUN	64.5	85.8	61.7	45.8	14.2	38.3	54.2
38	KIT	44.9	48.4	38.6	47.8	51.6	61.4	52.2
39	KRAS	38.1	47.4	33.7	33.2	52.6	66.3	66.8
40	LCK	58.1	58.8	40.2	75.3	41.2	59.8	24.7
41	LMO2	55.8	54.7	38.4	74.2	45.3	61.6	25.8
42	MAFB	67	66.7	46.6	87.7	33.3	53.4	12.3
43	MAML2	51.9	58.1	45.4	52.3	41.9	54.6	47.7
44	MDM2	40.3	50.7	39.6	30.4	49.3	60.4	69.6
45	MET	43.5	48.1	38.3	44.2	51.9	61.7	55.8
46	MITF	48.5	52.6	39	53.8	47.4	61	46.2
47	MLL	48.1	51.3	49.1	43.9	48.7	50.9	56.1
48	MPL	59.7	60.3	51.4	67.4	39.7	48.6	32.6
49	MYB	48.6	53.5	41.3	45.6	46.5	58.7	54.4
50	MYC	58.8	55.5	44.5	76.4	44.5	55.5	23.6

Supplementary material contains two tables (Table 3 and Table 4). Table 3 contains the frequency of optimal codons (Fop) in the complete coding region of 50 oncogenes. Table 4 contains relative synonymous codon usage values (RSCU) of 50 cds selected in this study.