Computational Molecular Biology

2014, Vol.4, No.2, 18-25 http://cmb.sophiapublisher.com

BioPublisher©

Research Report Open Access

GC2 Biology Dictates Gene Expressivity in *Camellia sinensis*

Supriyo Chakraborty, Prosenjit Paul

Department of Biotechnology, Assam University, Silchar-788011, Assam, India

Corresponding Author email: supriyoch_2008@rediffmail.com; Author

Computational Molecular Biology, 2014, Vol.4, No.2 doi: 10.5376/cmb.2014.04.0002

Abstract The effectiveness of the gene expression is influenced by the nature of codons used throughout the coding sequence (cds) of the gene. This is due to the fact that most genes and organisms do not use synonymous codons uniformly. Certain synonymous codons are used preferentially and this phenomenon is called codon usage bias (CUB). We analyzed normalized AT and GC frequency at each codon site. We observed that the correlations between gene expression (measured by CAI) and GC content at any codon site were very weak except GC2s showed moderate positive correlation with gene expression. We also measured the correlations between CAI and AT content at three codon sites. AT2s showed moderate negative correlation with gene expression. We further observed a strong correlation between RCBS (a measure of gene expression) and cds length indicating that natural selection is probably operating in favor of shorter genes to be expressed at higher level. For this analysis, we initially downloaded 350 coding sequences of *Camellia sinensis*, out of which only ten cds were found to begin with the initiator codon ATG, and length as exact multiple of three bases and devoid of N (any unknown base). Our analysis on these ten cds revealed that the second position of synonymous codons in *Camellia sinensis* possibly plays a more prominent role than the third position in determining the gene expressivity as evident from the CUB and the correlation analyses.

Keywords Gene expression; Relative codon usage bias (RCBS); Codon adaptation index (CAI); Codon usage bias (CUB)

Introduction

The effectiveness of the gene expression is influenced by the nature of codons used throughout the gene. Since the course of evolution, there are few genes in a coding sequence which remains unchanged, i.e. conserved throughout. This is due to the fact that most genes and organisms do not use synonymous codons uniformly; certain synonymous codons are used preferentially, a phenomenon called codon usage bias (CUB).

Codon bias, the unequal usage of synonymous codons, varies widely between species and in some cases, it has also been reported that there is significant variation of codon usage bias among different genes within the same organism (Bernardi, 1993). Previous codon usage analyses showed that codon usage bias is very complicated and associated with various biological factors, such as gene expression level (Gouy and Gautier, 1982; Sharp and Li, 1986; Sharp et al., 1986; Sharp and Li, 1987), gene length (Bains, 1987; Eyre-Walker, 1996), gene translation initiation signal (Ma, 2002), protein amino acid composition (Lobry and Gautier, 1994), protein structure (D'Onofrio et al., 2002), tRNA abundance (Ikemura, 1981, 1982), mutation frequency and patterns (Sueoka, 1999), and GC composition (Sueoka and Kawanishi, 2000). The influence of GC bias has a major impact on codon bias, resulting in a close association between GC% at the third codon position, also called GC3 and GC bias (Sueoka, 1988). As all amino acids (with the exception methionine and tryptophan) allow GC-changing synonymous substitutions in the third position, this has led to a common belief that the use of synonymous G/C-ending codons should increase in frequency with increasing GC bias, while usage of A/T-ending codons should decrease (Wan et al., 2004).

Preferred citation for this article:

Chakraborty and Paul, 2014, GC2 Biology Dictates Gene Expressivity in *Camellia sinensis*, Computational Molecular Biology, Vol.4, No.2 10-17 (doi: 10.5376/cmb.2014.04.0002)

Received: 27 Dec., 2013 | Accepted: 15 Jan., 2014 | Published: 27 Feb., 2014

Tea (*Camellia sinensis*) is one of the most popular beverages in the world owing to the availability of diverse cultivars and qualities, its taste, the stimulative effect, but also for its health benefits. It has received much attention for its attractive aroma, pleasant taste, and numerous medicinal benefits, and has been socially and habitually consumed by people since 3000 B.C. (Kliman and Bernal, 2005). Many secondary metabolites, such as polyphenols, alkaloids (e.g. caffeine), vitamins (A, B1, B2, E, C), polysaccharides, volatile oils, and minerals are found in tea leaves (Lin et al., 2003).

Despite its fundamental importance in several areas of genetics, there has been a long period of struggle to measure CUB. Advances in sequencing technology have provided an abundance of genomic data from different organisms. The study of CUB is gaining rehabilitated attention with the advent of whole genome sequencing of numerous organisms. In this paper, we propose to study the CUB for *Camellia sinensis* by analyzing the codon adaptation index (CAI), relative codon usage bias (RCBS), frequency of optimal codon (Fop), relative synonymous codon usage value (RSCU), effective number of codons (ENc), GC content, GC skew and AT skew.

1 Results

1.1 Overall codon usage analysis

Since the whole genome sequence is not available for *Camellia sinensis* only ten (10) genes were used in this study. Table 1 shows the selected genes with their accession number along with the overall RCBS, CAI, GC%, GC1s, GC2s and GC3s values. It was found that the coding sequences of *Camellia sinensis* are rich in A and/or T. But in the case of *P. aeruginosa* it is evident that codons ending in G and/or C are predominant in the entire coding region (Gupta and Ghosh, 2001). However, the overall codon usage values may obscure some heterogeneity of codon usage bias among the genes that might be superimposed on the extreme genomic composition of this organism.

Table 1 RCBS, CAI, CDS length, GC content analysis and accession number for *Camellia sinensis* genes

Sl. no	Gene name	Accession number	CDS length (bp)	CAI	RCBS	GC content (%)			
						GC	GC1	GC2	GC3
1	Acetyl CoA carboxylase	DQ366599	1800	0.6351	0.0413	47.6	55.0	42.3	45.3
2	Polyphenol oxidase (PPO)	FJ656220	1800	0.6118	0.0408	49.1	52.3	40.5	54.3
3	pRB mRNA for retinoblastoma related protein	AB247284	3078	0.5455	0.0252	42.9	47.9	43.9	37.1
4	cycD3-2 mRNA for cyclin D3-2	AB247283	1119	0.3605	0.0637	43.3	50.4	34.3	45.0
5	cycD3 mRNA for cyclin D3-1	AB247282	1116	0.5152	0.0629	46.7	53.0	38.2	48.9
6	cycb mRNA for cyclin B	AB247280	1323	0.4688	0.0541	45.4	53.7	39.5	43.1
7	*Stearoyl acyl carrier protein desaturase*	KC242133	1191	0.3339	0.0599	45.3	54.2	37.8	44.1
8	Cultivar Longjing43 glycerol-3-phosphate acyltransferase	KC920896	1353	0.4935	0.0529	46.4	52.5	42.1	44.6
9	Omega-3 fatty acid desaturase (FAD8)	KC847167	1359	0.5694	0.0536	46.7	53.0	42.2	44.8
10	AMP deaminese	KC700025	2571	0.5943	0.0298	44.2	52.5	37.5	42.6

1.2 Codon usage variation

The effective number of codons used by a gene (Nc) and (G+C) percentage at the third synonymous codon positions (GC3s) are used to study the codon usage variation among the genes of *Camellia sinensis*. Wright (1990) suggested that a plot of Nc against GC3s could be effectively used to explore the codon usage variation among the genes. It was demonstrated by Wright (1990) that the comparison of the actual distribution of genes with the expected distribution under no selection could be indicative, if the codon usage bias of the genes had some influence other than the genomic GC composition.

Figure 1 shows the Nc distribution of different genes in *Camellia sinensis*. The mean and standard deviation value of Nc are 15.2 and 0.42637 respectively, indicating that there is a wide variation of codon usage bias among the genes. The variation of codon usage biases among the genes is further confirmed from the distributions of (G+C) at the third synonymous codon positions, shown in Figure 2. These results indicate that apart from compositional constraints, other trends might influence the overall codon usage variation among the genes in *Camellia sinensis*.

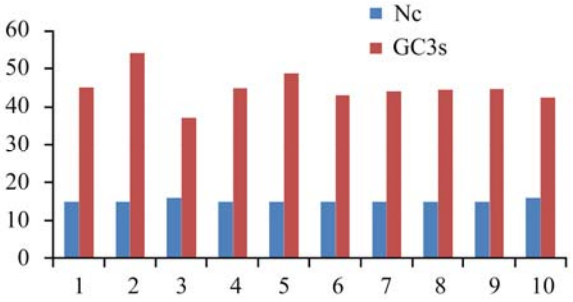

Figure 1 Nc distribution of *Camellia sinensis* genes

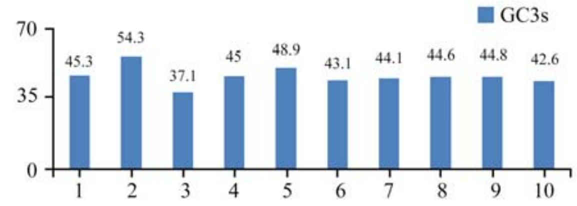

Figure 2 GC3s distribution of *Camellia sinensis* genes

1.3 Relationship between RCBS and CAI values

Each gene has evolved a codon usage pattern accommodating gene expression level, and RCBS value >0.5 and CAI value >0.5 exhibits favorable codon usage. So, we chose these two indices as effective expression measures based on literary evidence. The expression level of genes has shown both CAI and RCBS. From our analysis we have found that six genes out of ten have RCBS and CAI values each greater than 0.5, suggesting that these six genes of *Camellia sinensis* could qualify as highly expressed genes.

The RCBS and CAI showed similar pattern when we plotted them on the graph (Figure 3). We analyzed further the relationship between the length of the coding region and the expression level of genes. In agreement with previous other studies (Ikemura, 1981; Ikemura, 1982; Moriyama and Powell, 1998), our data support the smaller size of highly expressed genes. We observed that RCBS decreases with the length of the encoded proteins. A significant negative correlation was observed between RCBS and protein length. In Figure 4 we plotted RCBS as a function of the gene length.

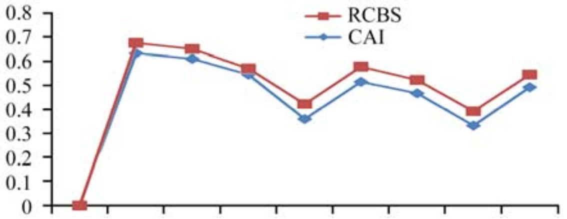

Figure 3 The relationship between RCBS and CAI values

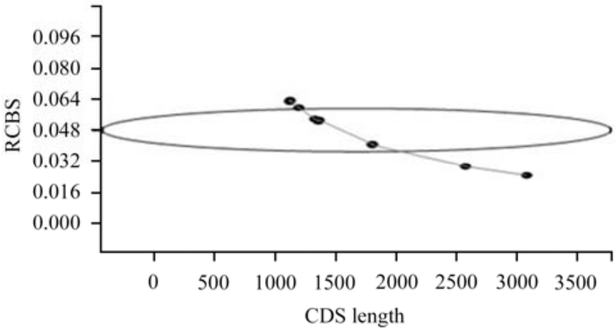

Figure 4 Relationship between RCBS and protein length

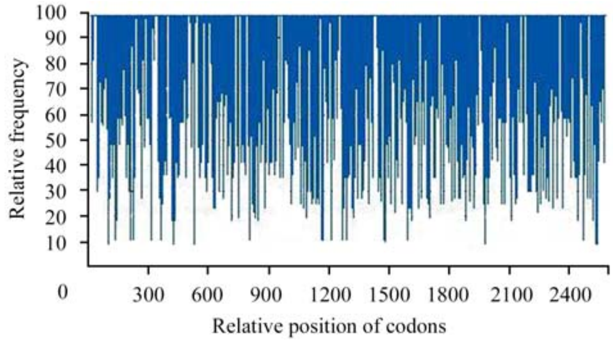

Figure 5 The distribution of codon usage frequency along the length of the CDS for the gene AMP deaminese

The ideal percentage range of GC content is between 30% to 70% (Figure 5 and Figure 6). Any peak outside this range adversely affects transcriptional and translational efficiency. Out of ten genes it was found that the gene Polyphenol oxidase (PPO) has GC content outside of the ideal range of 32% to 75%. These results suggest that the GC content present in the CDS sequence of the gene affects the expressiveness of that particular gene. The differences in GC content among the genes were again confirmed by plotting CAI value against the GC content and it was found that two genes showed negative correlation with gene expressiveness due to their GC content (Figure 7).

1.4 Relationship between GC and AT content and the expression patterns of genes

We analyzed normalized AT and GC frequency at each codon site. We observed that correlations between gene expression as measured by CAI and GC content at any codon site are very weak (rGC1=0.069, rGC2=0.604 and rGC3=0.186) (Figure 8). Thus, in contradiction with others GC content at the third codon position comes out to be a very poor predictor of gene expression in *Camellia sinensis* (Sharp and Lloyad, 1993; Gerton et al., 2000; Marin et al., 2003). But at the second codon position the GC content shows moderate positive correlation with gene expression. Since the coding sequences of *Camellia sinensis* are rich in AT, we also analyzed the AT frequency at each codon site and it was found that the second codon position showed a moderate negative correlation with gene expression.

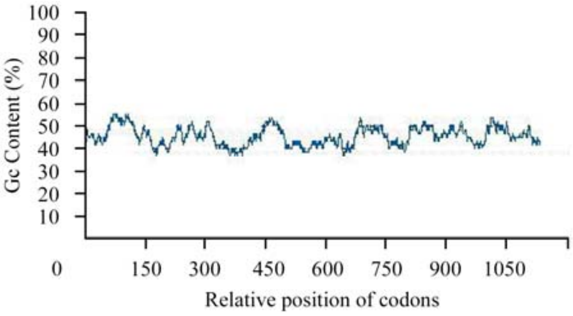

Figure 6 The percentage range of GC content for the Stearoyl acyl carrier protein desaturase gene

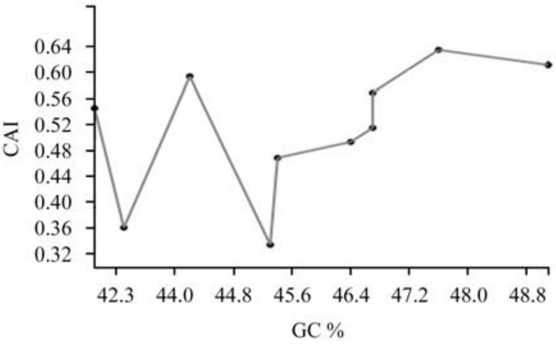

Figure 7 CAI plotted against the GC content for *Camellia sinensis* genes

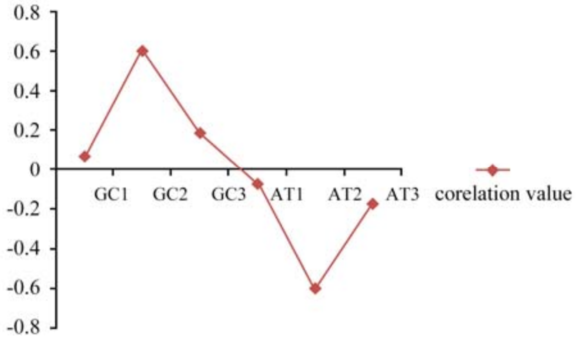

Figure 8 Correlation between CAI and GC/AT content at different codon positions

2 Discussion

In brief, we have presented an expression measure of a gene, devised to predict the level of gene expression from relative codon bias and codon adaptation index. 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. We have outlined a simple approach to assess the strength of codon bias index in genes as a guide to their likely expression level and illustrate this with an analysis of *Camellia sinensis* genes.

The present study was carried out with the objectives: (a) To analyze the CAI, RCBS, GC skew, GC content, Relative position of codon for the genes of *Camellia sinensis*. (b) To correlate above mentioned parameters with the gene expression pattern. As per our mentioned objectives in this present study, we selected ten genes from *Camellia sinensis* for CUB analysis. The accurate coding sequences having correct initial and termination codons were retrieved using a program in perl, developed by us. To minimize sampling errors we have taken only those coding sequences which are greater than or equal to 1000 bp. All the above-mentioned parameters for CUB analysis were calculated by using a PERL-based program developed by us.

After analyzing the coding sequences for *Camellia sinensis* it was found that genes are rich in AT. But in the case of *P. aeruginosa* it is evident that codons ending in G and/or C are predominant in the entire coding region. We also predicted the heterogeneity of codon usage by analyzing the effective number of codons (Nc). The mean and standard deviation value of Nc are 15.2 and 0.42637 respectively, indicating that there is a wide variation of codon usage bias among the genes of *Camellia sinensis*. The variation of codon usage bias among the genes is further confirmed from the distributions of (G+C) 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 genes in *Camellia sinensis*.

Each gene has evolved a codon usage pattern accommodating gene expression level, and RCBS value >0.5 and CAI value >0.5 exhibits favorable codon usage. We calculated the CAI and RCBS values for the genes and it was found that six out of ten genes of *Camellia sinensis* qualify as highly expressed genes. We also analyzed the GC content distribution on relative position of codons; our results revealed that except for the gene PPO all other genes have ideal GC percentage.

We analyzed normalized AT and GC frequency at each codon site. We observed that correlations between gene expression as measured by CAI and GC content at any codon site are very weak. GC2s showed moderate positive correlation (0.604) with gene expression. We also measured the correlations between CAI and AT content at any codon site. AT2s showed moderate negative correlation (-0.604) with gene expression.

Moreover, our analysis further revealed that the second position of synonymous codons in *Camellia sinensis* played a more prominent role than the third position, as indicated by the positive correlation coefficient (0.064) between CAI and GC2s as compared to correlation coefficients (0.069 and 0.187) of CAI with GC1s and GC3s, in determining the level of gene expression. This contradicts the fact that the third position of codon in *E. coli* plays a major role in determining gene expression although both *Camellia sinensis* and *E. coli* are AT rich. This was further confirmed by the highest negative correlation between CAI and AT2s (-0.064) in comparison to the correlation coefficients (-0.069 and -0.172) of CAI with AT1s and AT3s. This might be due to small number of coding sequences and only the genes with high CAI and RCBS taken for the present investigation.

The compositional bias of cds plays a crucial role in shaping the codon usage. GC content has a major influence on codon usage bias, resulting in a close association between GC% at the third codon position, also called GC3 biology. As all amino acids (except methionine and tryptophan) allow GC-changing synonymous substitutions in the codon third position, this has led to a common belief that the use of synonymous G/C-ending codons could increase the expressivity of genes, while the usage of A/T-ending codons could decrease the level of gene expression. For this analysis, we initially downloaded 350 coding sequences of *Camellia sinensis*, out of which only ten cds were found to begin with the initiator codon ATG, and length as exact multiple of three bases and devoid of N (any unknown base). As evident from the CUB analysis of the cds and correlation analysis between GC/AT content at three codon sites with the CAI value, our results suggest that the 2nd position of synonymous codons in *Camellia sinensis* possibly plays a more prominent role than the 3rd position of codons in determining the gene expressivity.

3 Materials and Methods

3.1 Datasets

The coding sequences (cds) of *Camellia sinensis* were downloaded from NCBI (www.ncbi.nlm.nih.gov). To minimize sampling errors we have taken only those cds which are greater than or equal to 1000 bp and have the correct initial and termination codons, devoid of N (any unknown base). The accurate coding sequences were retrieved using a program in perl developed by us. Finally, ten (10) sequences were selected for CUB analysis.

3.2 Models

Relative codon usage bias and codon adaptation index were used to study the overall codon usage variation among the genes. RCBS is the difference of observed frequency of a codon from the expected frequency under the hypothesis of random codon usage where the base compositions are biased at three sites in the sequence, divided by the expected frequency. RCBS is the overall score of a gene indicating the influence of RCB of each codon in a gene. RCB reflects the level of gene expression. The expression measure of a gene is denoted by RCBS (Hertog et al., 1993). RCBS value close to 0 indicates a lack of bias for the codons and is thus useful for comparing different sets of genes.

Gene expression level is related to codon usage difference of a gene with respect to biased nucleotide composition at the three codon sites. Let f(x,y,z) be the normalized codon frequency for the codon triplet (x,y,z) of a gene. Then the relative codon bias (RCB) of a codon triplet (x,y,z) in a gene is defined as:

$$dxyz = \frac{f(x,y,z) - f1(x)f2(y)f3(z)}{f1(x)f2(y)f3(z)}$$

Where, f1(x) is the normalized frequency of x at the first codon position, f2(y) is the normalized frequency of y at the second codon position, and f3(z) is the normalized frequency of z at the third codon position of the gene. The frequencies f1, f2, f3 have been derived from the set of codon samples of a gene and the normalization of frequency is done over the gene length in codons, in an attempt to compensate for the expected increase of RCB with the total number of codons quantified the degree of codon bias of a gene in such a way that comparisons can be made both within and between genomes. As defined earlier, d_{xyz} contains somewhat more quantitative information than others, since it considers codon usage as well as the base compositional bias. Then the expression measure of a gene is:

$$RCBS = (\prod_{i=1}^{L} (1 + d_{xyz}^{i})^{1/L} - 1$$

Where, d^{i}_{xyz} is the codon usage difference of ith codon of a gene. L is the number of codons in the gene.

Gene expressivity was again measured by calculating the parameter codon adaptation index (Sharp and Li, 1986). It essentially measures the distance from a given gene to a reference gene with respect to their amino-acid codon usages. CAI defines translational optimal codons as those that appear frequently in highly expressed genes i.e.

$$CAI(L(g)) = \exp(\frac{1\sum_{l=1}^{L} \log w_{c(l)}}{L}) = (\prod_{l=1}^{L} w_{c(l)})^{1/L}$$

Where, *L* is the length of gene g and w_c (*l*) is the relative adaptiveness of the codon *c* in the reference genes (not g). Relative adaptiveness is defined:

$$w_c = \frac{f_c}{\max(f_s)}, 8 \in \{c_a\},$$

Where, fc is the frequency of codon c which is the *l*th codon in gene g. a is the amino acid encoded by *c* and $\{C_a\}$ is the set of synonymous codons encoding amino acid *a*. Certain codons will appear multiple times in the gene. Hence we can rewrite the equation to sum over codons rather than length, and use counts rather than frequencies. This makes the dependence on the actual gene clearer. The more usual form is:

$$CAI(o(g)) = \exp(\frac{1}{o_{tot}} \sum_{c \in C} o_c \log w_c) = (\prod_{c \in C} o_c \log w_c)^{\frac{1}{o_{tot}}}$$

The effective number of codons (Nc) is the total number of different codons used in a sequence (Wright, 1990). The values of Nc range from 20, where only one codon is used per amino acid, to 61 (for standard genetic code), where all possible synonyms codons are used with equal frequency. Nc measures bias toward the use of a smaller subset of codons, away from equal use of synonymous codons. For example, as mentioned above, highly expressed genes use fewer codons due to selection. The underlying idea of Nc is similar to the concept of zygosity from population genetics, which refers to the similarity for a gene from two organisms.

In the context of codon usage, multiple synonymous codons are treated analogously to multiple alleles. Homozygosity for an amino acid Za measures the degree of similarity and is computed based on the relative codon frequencies fac:

$$Z_a = \frac{O_a \sum_{c \in Ca} \int_{ac}^{2} - 1}{O_a - 1}$$

The number of effective codons for an amino acid is the inverse of homozygosity:

Na =Za-1

The value of Na ranges from 1 to the number of synonymous codons ka (the codon degeneracy). With equal codon usage, homozygosity is minimal and the value of Na is the number of synonymous codons. The overall number of effective codons for a gene (Nc) is a sum of average homozygosities Za for different redundancy classes k (in set K of all redundancy classes):

$$Nc = \sum_{k=K} n_k \overline{N}_{a=k}$$

Where for each redundancy class:

$$\overline{N}_a = \frac{1}{n_k} \sum_{a \in K_k} N_a$$

When the codon usage pattern is more uniform than expected, it is possible to obtain Nc > 61, in which case it is readjusted to 61. If an amino acid is not observed, or is very rare, then the value is replaced by the average homozygosity of the amino acids in the same redundancy class. If Ile amino acid is missing (the only member in the redundancy class with three synonymous codons), then the corresponding Z is estimated from the average homozygosity of the other redundancy classes.

For example, in the case of isoleucine:

$$\overline{Z}_{k=3} = \frac{1}{3}(\left(\frac{2}{\overline{Z}_{k=3}} - 1\right)^{-1} + \left(\frac{2}{3\overline{Z}_{k=4}} - \frac{1}{3}\right)^{-1} + \left(\frac{2}{5\overline{Z}_{k=6}} - \frac{3}{5}\right)^{-1})$$

When there is a large discrepancy among the amino acids for a gene, the sum of Nc for all individual amino acids can be used instead of taking the sum of the averages of each redundancy class:

$$Nc = \sum_{a \in A} N_a$$

GC3s is the frequency of (G+C) and A3s, T3s, G3s, and C3s are the distributions of A, T, G and C at the synonymous third positions of codons (Gupta and Ghosh, 2001). GC skew and AT Skew are defined as the ratio of (G - C) to (G+C) and (A - T) to (A + T) respectively along the DNA sequences (Wright, 1990).

3.3 Analysis

All the above-mentioned parameters were calculated by using a PERL program developed by us. After which we have measured the correlations between all the above mentioned parameters with the gene expressivity for *Camellia sinensis*.

Authors' contributions

S.C conceived the idea, prepared the software for analysis. P.P. analyzed the data set and prepared the manuscript, figures and tables. All authors read and approved the final manuscript.

Acknowledgment

We are thankful to Assam University, Silchar, Assam, India for providing the necessary facilities in carrying out this research work. We sincerely acknowledge the help rendered by Dr. A. Sen, Director and other staff members, Computer Centre, Assam University, Silchar for their kind support in providing internet access for this research work.

References

Bains W., 1987, Codon distribution in vertebrate genes may be used to predict gene length, J Mol. Biol., 197(3): 379-388 http://dx.doi.org/10.1016/0022-2836(87)90551-1

Bernardi G., 1993, The vertebrate genome: isochores and evolution, Mol. Biol. Evol., 10: 186-204

D'Onofrio G., Ghosh T.C., and Bernardi G., 2002, The base composition of the genes is correlated with the secondary structures of the encoded proteins, Gene, 300(1-2): 179-187 http://dx.doi.org/10.1016/S0378-1119(02)01045-4

Eyre-Walker A., 1996, Synonymous codon bias is related to gene length in *Escherichia coli:* selection for translational accuracy? Mol Biol Evol., 13(6): 864-872 http://dx.doi.org/10.1093/oxfordjournals.molbev.a025646

Gerton J.L., DeRisi J., Shroff, R., Lichten M., Brown P.O., and Petes T.D., 2000, Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*, Proc. Natl. acad. Sci. USA, 97(21), 11383-11390 http://dx.doi.org/10.1073/pnas.97.21.11383

Gouy M., and Gautier C., 1982, Codon usage in bacteria: correlation with gene expressivity, Nucleic Acids Res., 10: 7055-7074 http://dx.doi.org/10.1093/nar/10.22.7055

Gupta S.K., and Ghosh T.C., 2001, Gene expressivity is the main factor in dictating the codon usage variation among the genes in *Pseudomonas aeruginosa*, Gene, 273: 63-70 http://dx.doi.org/10.1016/S0378-1119(01)00576-5

Hertog H.G., Hollman P.C., Katan M.B., and Kromhout D., 1993, Intake of potentially anticarcinogenic flavonoids and their determinations in adults in the Netherlands, Nutr. Cancer, 20(1): 21-29
http://dx.doi.org/10.1080/01635589309514267

Ikemura T., 1981, 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(3): 389-409
http://dx.doi.org/10.1016/0022-2836(81)90003-6

Ikemura T., 1982, Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs, J Mol. Biol., 158(4): 573-597
http://dx.doi.org/10.1016/0022-2836(82)90250-9

Kliman R.M., and Bernal C.A., 2005, Unusual usage of AGG and TTG codons in humans and their viruses, Gene, 352: 92-99
http://dx.doi.org/10.1016/j.gene.2005.04.001

Lobry J.R., and Gautier C., 1994, Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes, Nucleic Acids Res., 22(15): 3174-3180
http://dx.doi.org/10.1093/nar/22.15.3174

Ma J., Campbell A., and Karlin S., 2002, Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures, J Bacteriol., 184(20): 5733-5745
http://dx.doi.org/10.1128/JB.184.20.5733-5745.2002

Marin A., Gallardo M., Kato Y., Shirahige K., Gutiérrez G., Ohta K., and Aguilera A., 2003, Relationship between G+C content, ORF length and mRNA concentration in *Saccharomyces cerevisiae*, Yeast, 20(8): 703-711
http://dx.doi.org/10.1002/yea.992

Moriyama E.N., and Powell J.R., 1998, Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*, Nucleic Acids Res., 26(13): 3188-3193
http://dx.doi.org/10.1093/nar/26.13.3188

Roymondal U., Das S., and Sahoo S., 2009, Predicting gene expression level from relative codon usage bias: an application to *Escherichia coli* genome, DNA Res., 16(1): 13-30
http://dx.doi.org/10.1093/dnares/dsn029

Sharp P.M., and Li W.H., 1986, An evolutionary perspective on synonymous codon usage in unicellular organisms, J Mol. Evol., 24(1-2): 28-38
http://dx.doi.org/10.1007/BF02099948

Sharp P.M., and Li W.H., 1987, The codon Adaptation Index – a measure of directional synonymous codon usage bias, and its potential applications, Nucleic Acids Res., 15(3): 1281-1295
http://dx.doi.org/10.1093/nar/15.3.1281

Sharp P.M., and Lloyad A.T., 1993, Regional base composition variation along yeast chromosome III: evolution of chromosome primary structure, Nucleic Acids Res, 21(2): 179-183
http://dx.doi.org/10.1093/nar/21.2.179

Sharp P.M., Tuohy T.M., and Mosurski K.R., 1986, Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes, Nucleic Acids Res., 14: 5125-5143
http://dx.doi.org/10.1093/nar/14.13.5125

Sueoka N., 1988, Directional mutation pressure and neutral molecular evolution, Proc. Natl. Acad. Sci., 85(8): 2653-2657
http://dx.doi.org/10.1073/pnas.85.8.2653

Sueoka N., 1999, Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra-strand rule of A = T and G = C, J Mol. Evol., 49(1): 49-62
http://dx.doi.org/10.1007/PL00006534

Sueoka N., and Kawanishi Y., 2000, DNA G+C content of the third codon position and codon usage biases of human genes, Gene, 261(1): 53-62
http://dx.doi.org/10.1016/S0378-1119(00)00480-7

Wan X.F., Xu D., Kleinhofs A., and Zhou J., 2004, Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes, BMC Evolutionary Biology, 4: 19
http://dx.doi.org/10.1186/1471-2148-4-19

Wright F., 1990, The ‘effective number of codons’ used in a gene, Gene, 87(1): 23-29
http://dx.doi.org/10.1016/0378-1119(90)90491-9

Y.S. Lin, Y.J. Tasi, J.S. Tsay, and J.K. Lin, 2003, Factors affecting the levels of tea polyphenols and caffeine in tea leaves, J. Agric. Food Chem., 51(7): 1864-1873
http://dx.doi.org/10.1021/jf021066b