ORIGINAL RESEARCH

Factors Affecting Synonymous Codon Usage Bias in Chloroplast Genome of *Oncidium Gower Ramsey*

Chen Xu, Xiaoning Cai, Quanzhan Chen, Hongxia Zhou, Yao Cai and Ailing Ben

School of Biochemical and Environmental Engineering, Nanjing Xiaozhuang University, Nanjing 211171, Jiangsu, China.
Corresponding author email: xuchenidea@hotmail.com; benimuru@hotmail.com

Abstract: *Oncidium Gower Ramsey* is a fascinating and important ornamental flower in floral industry. In this research, the complete nucleotide sequence of the chloroplast genome in *Oncidium Gower Ramsey* was studied, then analyzed using Codonw software. Correspondence analysis and method of effective number of codon as Nc-plot were conducted to analyze synonymous codon usage. According to the corresponding analysis, codon bias in the chloroplast genome of *Oncidium Gower Ramsey* is related to their gene length, mutation bias, gene hydropathy level of each protein, gene function and selection or gene expression only subtly affect codon usage. This study will provide insights into the molecular evolution study and high-level transgene expression.

Keywords: codon usage, chloroplast, molecular evolution, mutation, synonymous codon

Introduction

Generally, the alternative synonymous codons for any amino acid are not randomly used.[1,2] Studies of the synonymous codon usage reveal information on molecular evolution of individual genes, which provides data to improve gene recognition algorithms, are utilized to design DNA primers, and detects horizontal transfer events.[3,4] It has been reported that a multiplicity of factors account for the biased usage of synonymous codons, such as CpG islands,[5] gene length,[6] gene expression level,[7] proteins secondary structure[8] and gene density[9,10] and so on. Codon usage variation is represented by two major paradigms. Codon usage is determined by either mutational bias or natural selection. The unified theory for codon usage has not been provided so far. We consider it desirable for more codon usage patterns to be revealed, and more information about the molecular evolution of individual genes be founded after more sequences are determined.

The chloroplast genome has long been a focus of research in plant molecular evolution due to its small size, high copy number, conservation and extensive characterization at the molecular level. More recently, with technical advances in DNA sequencing, the number of completely sequenced chloroplast genomes has grown rapidly (http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=2759&opt=plastid#pageTop). Since 2006, more than twenty chloroplast genomes have been sequenced every year. The translation machinery in chloroplasts is known to be structurally similar to those in prokaryotes, leading to the suggestion that the translation mechanism and patterns of codon usage in chloroplasts might be similar to those in *Escherichia coli*.[11] However, most studies in plants were mainly focused on codon bias in nuclear genomes. Nevertheless, Morton found that the asymmetry of two strands of the chloroplast genome from *Euglena gracilis* was the major factor contributing to codon bias.[12] In addition, it was considered that context dependent mutation played some roles in shaping codon usage of the chloroplast genomes of grass species,[13] although there was also evidence that the codon usage of certain chloroplast genes was influenced by selection.[14]

Oncidium Gower Ramsey is a popular indoor and florist orchids for a very good reason: their large sprays of flowers often sag with dozens of blooms, which is often referred to as the "dancing lady". It is an important cut-flower orchid. *Oncidium Gower Ramsey* has been widely used in commerce and industry throughout the world. The biochemistry and physiology of *Oncidium Gower Ramsey* generates a lot of curiosity among the biologists. Codon usage in *Oncidium Gower Ramsey* has not been investigated in any detail, and it is not clear how (or even if) different genes should vary.[15] The complete chloroplast genome sequence of *Oncidium Gower Ramsey* been determined.[16] Therefore it is of interest to understand the factors that shape codon usage in this species. In this study, the analysis of codon usage bias in chloroplast genome of *Oncidium Gower Ramsey* is reported using methods of multivariate statistical analysis and correlation analysis, and the optimal codons are also determined.

Materials and Methods
Sequence data
The complete chloroplast genome sequence from *Oncidium Gower Ramsey*(NC_014056) was obtained from GenBank. To minimize sampling errors, only CDS sequences that are more than or equal to 100 codons and that have correct initial and termination codons were included in the dataset. A total of 47 genes were combined for codon usage analysis. Write C program was used to complete sequence processing, statistics and computation.

Indices of codon usage
The effective Number of Codons (NC) was often used to measure the magnitude of codon bias for an individual gene, yielding values ranging from 20 for a gene with extreme bias using only one codon per amino acid, to 61 for a gene with no bias using synonymous codons equally.[17] We further investigated the relationship between nucleotide content and codon usage by NC-plot. Wright[18] suggested the NC-plot (NC plotted against GC3s) as part of a general strategy to investigate the patterns of synonymous codon usage. Genes whose codon choice is constrained only by a G+C mutation bias will lie on or just below the curve of the predicted values.[18] The Codon Adaptation Index (CAI) was used to estimate the extent of bias toward codons that were known to be preferred in highly expressed genes. A CAI value is between

0 and 1.0, and a higher value means a stronger codon usage bias and a higher expression level. The set of sequences used to calculate CAI values in this study were the genes coding for ribosomal proteins.[19] G+C value is the frequency of nucleotides that are guanine or cytosine.

Hydropathicity value is the General AVerage HYdropathicity or GRAVY score, for the hypothetical translated gene product. It is calculated as the arithmetic mean of the sum of the hydropathic indices of each amino acid. Length value is equivalent to the length of one gene.

COA analysis

Correspondence analysis (COA) has become the method of choice for multivariate statistical analysis of codon usage patterns.[20,21] Since there are a total of 59 synonymous codons (61 sense codons, less the unique methionine and tryptophan codons), this analysis partitions the variation along 59 orthogonal axes, with 41 degrees of freedom. The first axis is the one that captures most of the variation in codon usage, with each subsequent axis explaining a diminishing amount of the variance.

Software implementation

The program CodonW 1.4 was used for calculating the indices of codon usage. The statistical analysis was implemented using SPSS 13.0 for Windows.

Results

The size of the *Oncidium Gower Ramsey* chloroplast genome is 146,484 bp. The overall GC content of the chloroplast genome is 37.0%. Coding regions make up 49.94% of the chloroplast genome (41.86% protein-coding genes, 8.08% RNA genes) and non-coding regions, which contain intergenic spacer (IGS) regions and introns, comprising 50.06%. Among the full 128 coding genes of the *Oncidium Gower Ramsey* chloroplast genome, we identified 74, 37 and 8 protein-coding, transfer RNA, and ribosomal RNA genes, respectively.

Initially we observe the Nc-plot distribution, which ENC and GC3s values were calculated (Fig. 1). ENC values vary from 37.39 to 57.34 with a mean of 47.05 and standard deviation of 4.49. The heterogeneity of codon usage was further confirmed from the GC3s values ranging from 14% to 37% with a mean of 24.86% and standard deviation of 4.19%. Wright suggested that a plot of Nc versus GC3s could be used effectively to explore the codon usage variation among the genes.[28] He argued that the comparison of actual distribution of genes, with the expected distribution under no selection could be indicative if codon usage bias of genes have some other influences other than compositional constraints. If the codon usage bias is completely dictated by GC3s the values of Nc should fall on the expected curve between GC3s and Nc-plot of the *Oncidium Gower Ramsey* chloroplast genome shown in Figure 1 shows that a considerable number of points are lying on the expected curve towards the GC poor region, which certainly originates from the extreme compositional constraints. But it is also interesting to note that a majority of the points with low Nc values are lying well below the expected curve. This result suggests that some genes in this organism have
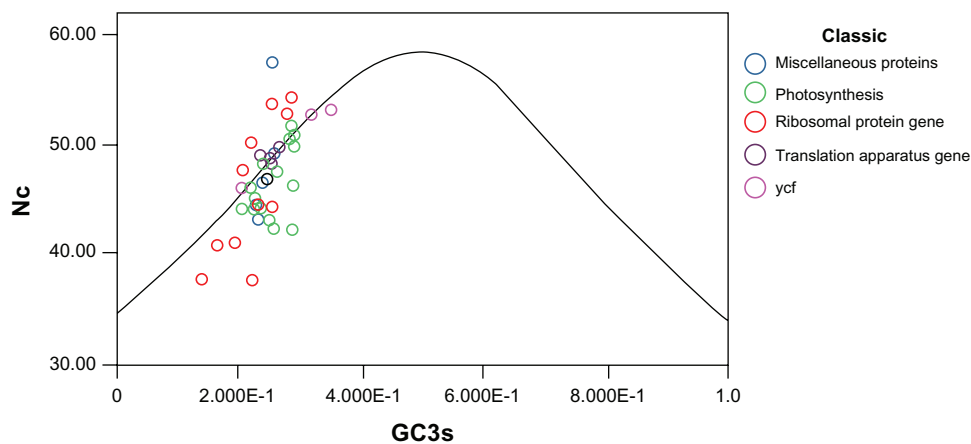

Figure 1. Effective number of codons(Nc) used in each gene plotted against GC content at synonymously variable third positions of codons(GC3s). The continuous curve plots the relationship between Nc and GC3s in the absence of selection.

additional codon usage bias, which are independent of compositional constraints.

We did further correspondence analysis (Table 1). The first axis can explain 24.12% of the total variation, and the next three axes explained 12.49%, 7.37%, 5.65%. The ordination of genes on the first four COA axes was examined for correlations with indices of codon usage and amino acid composition (eg, ENc, GC3s, GC and GRAVY). On one hand, the first major axis is negatively correlated with GC (r = −0.659, $P < 0.01$) and GRAVY (r = −0.440, $P < 0.01$), while correlated positively with gene length (r = 0.464, $P < 0.01$) and L_sym (r = 0.467, $P < 0.01$). Despite the lack of a correlation between GC3s and axis 1, there is a significant correlation between axis 1 ordination and the frequency of G ending synonymous codons (r = 0.686, $P < 0.01$), G+C content (r = −0.659, $P < 0.01$). On the other hand, it is not correlated with CAI (r = −0.035), Fop (r = −0.178), ENC (r = 0.149). There is no significant correlation between codon usage bias and gene expression level in the chloroplast genome of *Oncidium Gower Ramsey*, indicating that synonymous codon usage in this genome is not constrained by translational selection. These results suggest that mutational bias is the main factor in shaping the codon usage variation among the genes whereas gene expression plays a minor role.

From Table 1, we find the second important fact about the codon usage variation of chloroplast genes of *Oncidium Gower Ramsey*: all four axes are significantly correlated with GRAVY (r = 0.440, $P < 0.01$, r = 0.728, $P < 0.01$, et al), an index of amino acid hydrophobicity. This trend is apparently due to selection pressure on codons encoding hydrophobic amino acids. In Figure 2, low GRAVY scores are signed by up and down green triangle, which is putative genes encoding integral membrane proteins, psa and psb. The major photosynthetic complexes PSI, PSII, the cytochrome b/f complex and the ATP synthase complex are located at the thylakoid membrane. In their roles as a matrix for hydrophobic and amphiphilic proteins, membranes can modulate protein conformational changes and determine protein motility in the plane of the membrane. This is essential for the dynamics of forming protein domains and complexes that perform such functions as trans-membrane in transport.

Table 1. Correlation between the codon usage and amino acid usage indices, ENc GC3s, GC, GRAVY, CBI Aromaticity and Fop and the first four correspondence analysis axes from an analysis of *Oncidium Gower Ramsey* chloroplast.

	ENc	CAI	Fop	T3s	C3s	A3s	G3s	GC3s	GC	num	L_sym	Gravy
axis1	0.149	−0.035	−0.178	−0.023	−0.012	0.219	0.686(**)	0.213	−0.659(**)	0.464(**)	0.467(**)	−0.440(**)
axis2	−0.179	−0.473(**)	−0.231	−0.421(**)	−0.533(**)	0.786(**)	−0.243	−0.519(**)	−0.132	−0.362(*)	−0.354(*)	−0.728(**)
axis3	0.426(**)	−0.068	0.057	−0.528(**)	0.306(*)	−0.095	0.003	0.330(*)	0.795(**)	0.060	0.061	−0.380(*)
axis4	0.135	−0.132	0.198	−0.026	0.198	−0.090	0.093	0.170	0.145	0.197	0.199	0.446(**)

Notes: *Correlation is significant at the 0.05 level; **Correlation is significant at the 0.01 level.
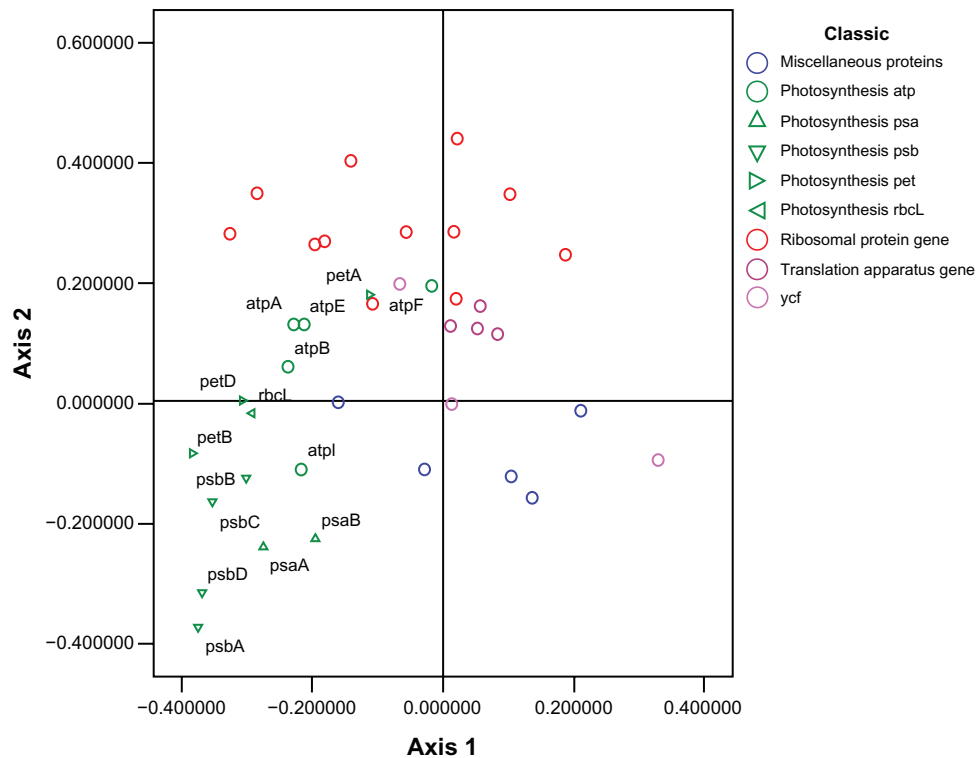
Figure 2. Correspondence analysis of the relative synonymous codon usage in 46 genes from chloroplast genome of *Oncidium Gower Ramsey*.

From the above discussion we can conclude that codon usage in complete chloroplast genome data of *Oncidium Gower Ramsey* is inferred to be the result of the hydropathy level of each protein.

According to orchid chloroplast gene function, 46 sequences can be classified six categories: The number of the first classified gene is 13, which encoding ribosomal protein. The rpl and rps genes encode large and small subunit ribosomal protein; the number of the second classified genes with photosynthesis protein coding genes is 15, including of psa gene, psb gene, atp gene, pet gene and nadh gene, also rbcL and rbcS gene; The third category is a conservative gene: ycf. The fourth category is translation apparatus genes, including the rpo gene of the RNA polymerase gene family. The fifth type is the miscellaneous proteins gene, for example accD, clpP, matK, cemA, ccsA; the sixth category is unknown function and hypothetical protein gene. Correspondence analysis of codon usage varied among the above classified genes (Fig. 2). Each gene is plotted using its gene classification name at its coordinate on the first two axes produced by the analysis. Genes with encoding photosynthesis are plotted in green, those genes encoding ribosomal protein production are plotted in

red, and those ycf genes are plotted in pink. Genes of ribosomal protein gene are plotted on the first axis left, and those of unknown function and hypothetical protein genes are mainly concentrated on the far right of the first axis. We can infer that gene function and codon usage are closely related.

Codon bias is also correlated with gene length (r = 0.464, $P < 0.01$). These results are similar to those observed in a number of different organisms *(Populustremula[22]; A.thaliana[23,24]; Physcomitrella[25,26]; C. elegans[27]; D. melanogaster[28]).*

Discussion
Previous studies were mainly focused on codon bias in nuclear genomes. It has been shown that codon bias reflect a balance between mutational biases and natural selection for translational optimization.[29] Optimal codons in fast-growing microorganisms, like *Escherichia coli*[30] or *Saccharomyces cerevisiae*,[31] reflect the composition of their respective genomic tRNA pool. It is thought that optimal codons help to achieve faster translation rates and higher accuracy. As a result of these factors, translational selection is expected to be stronger in highly expressed genes, as is indeed the case for the above-mentioned organisms.

In other organisms that do not show high growing rates or that present small genomes, codon usage optimization is normally absent, and codon preferences are determined by the characteristic mutational biases seen in that particular genome. Examples of this are *Homo sapiens* and *Helicobacter pylori*.[32] Organisms that show an intermediate level of codon usage optimization include *Drosophila melanogaster*, *Caenorhabditis elegansor*, *Arabidopsis thaliana*. It is not clear whether codon usage drives tRNA evolution or vice versa.[34]

Several earlier discussions of plant codon usage focussed on the differences between codon choice in plant nuclear genes and in chloroplasts.[32] Chloroplasts differ from the nuclear genome of higher plants in that they encode only 30 tRNA species. Since chloroplasts have restricted their tRNA genes, the use of preferred codons by chloroplast encoded proteins appears more extreme. However, a positive correlation has been reported between the level of isoaccepting tRNA for a given amino acid and the frequency with which this codon is used in the chloroplast genome.[33] In this study, AT contents in chloroplast genes are higher than that in nuclear genes in *Oncidium Gower Ramsey*. Codon use by chloroplast genes is biased toward a high representation of NNT and NNA codons and appears to reflect a mutational bias. The codon usage patterns of chloroplast genes more conserved in GC content and influenced by translation level. Maybe chloroplast and nuclear genes shared particularly different features of codon usage and evolutionary constraints. In the future, Further research on a comparative analysis of codon bias and factors in shaping the codon usage patterns among mitochondrion, chloroplast and nuclear genes may help clarify the relationship between nuclear genomes and mitochondria, chloroplast genome in *Oncidium Gower Ramsey*. This perhaps will provide detailed research about endosymbiont hypothesis and a framework from which to build more robust models to improve our understanding of not only molecular evolution in general, but also how we interpret molecular data for reconstructing phylogenies.

In this study, we present evidence suggesting that codon bias in the chloroplast genome of *Oncidium Gower Ramsey* is closely related to their gene length, mutation bias, gene hydropathy level of each protein,

gene function and selection or gene expression. Why is it related with gene length? We argue that it is consequence of selection to maximize translational speed, minimize the costs of proofreading, or maximize the accuracy of translation; by using codons which match common tRNAs or which bind the tRNA efficiently, it is thought that the time to find and bind the correct tRNA is minimized, along with the probability of misincorporating an incorrect tRNA. Both missense and processivity errors can potentially lead to a positive correlation between synonymous codon bias and gene length. Since the cost of producing a protein in terms of the energy used and the resources consumed is proportional to gene length, selection to avoid missense errors should be stronger in longer genes. We thought the strength of selection maximizes the rate of translation or minimizes the costs of proofreading. Selection is thought to be acting upon the elongation rate to reduce the time the ribosome is bound to the mRNA by increasing the rate at which each ribosome proceeds along the message; this leads to an increase in the concentration of free ribosomes and hence to an overall increase in the rate of protein production. Since the effect of delaying a ribosome during elongation on the concentration of free ribosomes is the same in long and short genes. the strength of selection should be independent of gene length. Similarly, the cost of proofreading is expected to be independent of gene length. When a charged tRNA binds to the mRNA in the A site of the ribosome, GTP is cleaved to GDP and elongation factor G is released. The time taken to perform this reaction allows the ribosome extra time to reject incorrect tRNAs; this is a process known as kinetic proofreading. The process consumes energy in terms of the GTP consumed, and time, and therefore exerts a cost on the cell.

In this study, codon usage is positively correlated with the hydrophobicity of each protein in *Oncidium Gower Ramsey*. Hydrophobicity scales are ranking lists for the relative hydrophobicity of amino acid residues. In previous studies, hydrophobicity scales were used to predict the preservation of the genetic code,[35] increases in the hydropathy of the encoded proteins may be explained an increase in protein stability, namely by a structurally and functionally meaningful change. The significance of the change is

stressed by the need for compensating for the lower hydrophobicity of eukaryotic proteins by an increase in another stabilizing factor, namely disulfide bridges.[36] At this point, it becomes obvious that the driving forces of the nucleotide changes are the changes of the encoded proteins, which reverses the neutralist view that the mutational bias is responsible for them.[37]

This study indicates that codon usage of the chloroplast genome of *Oncidium Gower Ramsey* may be mainly related with mutational bias, gene hydropathy level of each protein, gene length and gene function and gene expression. The orchid genome project (http://www.genomics.cn/en/research.php?type=show&id=329) focuses on sequencing and assembling the draft sequence of the orchid genome, which is also expected to complete the transcriptomes sequencing on *Pahipopedilum armeninacum, Olcoglossum amesianum, Ophrys apitera*, entirely representative orchids.[11] As more complete orchid genomes are analyzed different factors shaping the pattern of codon usage might be found. This study has provided a basic understanding of the mechanisms for codon usage bias, which could be useful in further studies of their molecular evolution, cloning and heterologous expression of its chloroplast genetic engineering.

References

1. Ghosh T. Studies on codon usage in Entamoeba histolytica. *Int J Parasitol*. 2000;30:715–22.
2. Grantham R, Gautier C, Gouy M. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res*. 1980;8:49–62.
3. Fickett JW. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res*. 1982;10:5303–18.
4. Peng J, Xiao S, Zuhong Lu. Analysis of Synonymous Codon Usage in Aeropyrum pernix K1 and Other Crenarchaeota Microorganisms. *JGG*. 2007;34:275–84.
5. Ponger L, Duret L, Mouchiroud D. Determinants of CpG islands: expression in early embryo and isochore structure. *Genome Res*. 2001;11:1854–60.
6. Duret L, Mouchiroud D. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proc Natl Acad Sci U S A*. 1999;96:4482–7.
7. Carlini DB, Chen Y, Stephan W. The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes Adh and Adhr. *Genetics*. 2001;159:623–33.
8. Gu W, Zhou T, Ma J, Sun X. The relationship between synonymous codon usage and protein structure in Escherichia coli and Homo sapiens. *Biosystems*. 2004;73:89–97.
9. Kahali B, Basak S, Ghosh TC. Reinvestigating the codon and amino acid usage of S. cerevisiae genome: a new insight from protein secondary structure analysis. *Biochem Biophys Res Commun*. 2007;354:693–9.
10. Versteeg R, van Schaik BD, Van Batenburg. The human transcriptome map reveals extremes in gene density, intron length, GC content and repeat pattern for domains of highly and weakly expressed genes. *Genome Res*. 2003;13:1998–2004.
11. Sugiura M. The chloroplast genome. *Plant Mol Biol*. 1992;19:149–68.
12. Morton BR. Strand asymmetry and codon usage bias in the chloroplast genome of Euglena gracilis. *Proc Natl Acad Sci U S A*. 1999;96:5123–8.
13. Morton BR. The role of context-dependent mutations in generating compositional and codon usage bias in grass chloroplast. *DNA J Mol Evol*. 2003;56:616–29.
14. Morton BR. Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. *Mol Evol*. 1998;46:449–59.
15. http://www.genomics.cn/data/attachment/Chinese%20Scientists%20Announced%20the%20Completion%20of%20%E2%80%9COrchid%20Genome%20Draft%20Map%E2%80%9D%20 in%20Shenzhen.doc.
16. FuHui W, MingTsair C, DeChih L. Complete chloroplast genome of *Oncidium Gower Ramsey* and evaluation of molecular markers for identification and breeding in Oncidiinae. *BMC Plant Biology*. 2010;doi: 10.1186/1471-2229-10-68.
17. Comeron JM, Aguade M. An evaluation of measures of synonymous codon usage bias. *J Mol Evol*. 1998;47:268–74.
18. Wright F. The "effective number of codons" used in a gene. *Gene*. 1990;87:23–9.
19. Hou ZC, Yang N. Factors affecting codon usage in Yersinia pestis. *Acta Biochim Biophys Sin*. 2007;35:580–6.
20. Greenacre MJ. Theory and applications of correspondence analysis. Academic Press: London; 1984.
21. Maria D, Ermolaeva. Synonymous codon usage in bacteria. *Curr Issues Mol Biol*. 2001;3:91–7.
22. Pär K, Ingvarsson. Gene expression and protein length influence codon usage and rates of sequence evolution in populus tremula. *Mol Bio and Evol*. 2007;24:836–44.
23. Duret L, Mouchiroud D. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol*. 2000;17:68–74.
24. Marais G, Mouchiroud D, Duret L. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. *Proc Natl Acad Sci U S A*. 2001;98:5688–92.
25. Marais G, Domazet-Loso T, Tautz D. Correlated evolution of synonymous and non synonymous sites in Drosophila. *J Mol Evol*. 2004;59:771–9.
26. Bierne N, Eyre-Walker A. The problem of counting sites in the estimation of the synonymous and non synonymous substitution rates: implications for the correlation between the synonymous substitution rate and codon usage bias. *Genetics*. 2003;165:1587–97.
27. Tenoien HK. Adaptive basis of codon usage in the haploid moss Physcomitrella patens. *Heredity*. 2005;94:87–93.
28. Duret L. tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet*. 2000;16:287–9.
29. Ruth H, Dmitri A. Selection on Codon Bias. *Annu Rev Genet*. 2008;42:287–99.
30. Eduardo PC, Rocha. Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Research*. 2004;14:2279–86.
31. Coghlan A, Wolfe KH. Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae. *Yeast*. 2000;16:1131–45.
32. Mario dos Reis, Renos S, Lorenz W. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Research*. 2004;17:5036–44.
33. Bulmer M. The selection-mutation-drift theory of synonymous codon usage. *Genetics*. 1991;129:897–907.
34. Elizabeth EM, Jeff L, Mary E. Codon usage in plant genes. *Nucleic Acids Research*. 1989;17(2):477–98.
35. Trinquier G, Sanejouand YH. Which effective property of amino acids is best preserved by the genetic code? *Protein Eng*. 1998;11(3):153–69.

36. Giuseppe DO, Kamel J, Hector M, Giorgio B. The correlation of protein hydropathy with the base composition of coding sequences. *Gene*. 1999; 238:3–14.

37. Kinshuk CN. Mutational bias and Gene expression level shape codon usage in Thermobifida fusca YX. *In Silico Biology*. 2009;9:0027.