Genomics 96 (2010) 92–101

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Gene functionality's influence on the second codon: A large-scale survey of second codon composition in three domains

Sen-Lin Tang [a,*], Bill C.H. Chang [b], Saman K. Halgamuge [c]

[a] Biodiversity Research Center, Academia Sinica, Taipei, Taiwan
[b] Institute of Plant and Microbial Biology, Academia Sinica, Taipei, Taiwan
[c] Biomedical Engineering, Department of Mechanical Engineering, Melbourne School of Engineering, The University of Melbourne, Australia

ARTICLE INFO

Article history:
Received 9 June 2009
Accepted 7 April 2010
Available online 21 April 2010

Keywords:
Second codon
Nucleotide frequency
Directional mutation

ABSTRACT

The second codon of a transcript, besides encoding for an amino acid, is now known to also have multiple molecular functions and is involved in translation efficiency and protein turn-over and maturation processing. These multiple purposes therefore make the selection constraints on this codon's composition more complex. To examine the biological significance of various permutations of the second codon, we conducted a systematic survey of second codon composition from 442 selected genomes across three domains. The amino acid bias of the second codon is associated with specific protein functions. The most common amino acids (S, A, K and T) are significantly avoided in Cell Envelope-related genes but preferred in Translation or Energy Metabolism-related genes, suggesting that the function of a gene product is a significant factor influencing the composition of the second codon.

Introduction

The second codon (alternatively, the *+2 codon*) has long received less attention than its more glamorous cousins, the start codon or the stop codon, especially with respect to its biological function. The second codon does, in fact, have more molecular features than just being a part of a gene and encoding for an amino acid. One striking feature is that this codon may influence the efficiency of translation by participating in the mRNA molecule–ribosomal machine interaction. The second codon joins the downstream region (i.e. the short sequence after the start codon; also called the *downstream box*) and regulates gene expression, as suggested by several reports [1–7]. The presence of A-, AT-rich or CA-repeat areas in the downstream region results in increased levels of gene expression in *Escherichia coli* [4,5,7–11]. The effect or the molecular mechanism of gene expression variation attributed to the second codon remains unclear [5]. Nonetheless, Brock et al. [11] suggest that the level of adenine richness in the downstream sequence affects the mRNA–ribosome association rate and the amount of ternary complex formed. However, it is worth noting that the translation efficiency enhanced by the second codon does not seem to be the result of mRNA–16S rRNA base-pairing, as indicated by Sato et al. [6]. In eukaryotic genomes, the nucleotide distribution in the second codon also differs markedly to the distribution of nucleotides in the third to the sixtieth codons [12]. This suggests that the second codon has a unique role in eukaryotic genes even though its function is unknown.

Although the nucleotide sequence of an mRNA is likely to be a significant factor influencing gene expression [7], particular amino acids encoded by the mRNA, such as alanine, glycine or serine, are frequently found in the second codon position of highly-expressed genes in several microbial genomes [13]. This suggests that, for some genes at least, the specific amino acid encoded by the second codon position is crucial. Similarly the distribution of amino acids in the N-terminal of a protein was found in an earlier report [14] to be non-random and greatly biased towards methionine, alanine and serine. So far, much evidence shows that the amino acid encoded in the second codon plays at least several important functional roles in proteins. For example, the second amino acid participates in the mechanism of the N-end rule pathway, which is relevant to the metabolic stability of a protein [15–17]. This amino acid is also involved in the methionine aminopeptidase cleaving reaction [18,19]. Moreover, the amino acid in the second position of encoded proteins is somehow correlated with specific secondary structures, a phenomenon that has not been associated with amino acids in other positions [20]. Considering all the above, it is clear that the second codon may influence gene expression as well as protein maturation and degradation.

With all these molecular features, it would not be hard to imagine that the selection constraints on the composition of the second codon in a gene have been reasonably complex during the evolutionary process. Hence our question: what factors are critical for the composition of a second codon? Although there is no complete answer available, the following constraints at least should be

[*] Corresponding author. Fax: +886 2 27890844.
E-mail address: sltang@gate.sinica.edu.tw (S.-L. Tang).

0888-7543/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2010.04.001

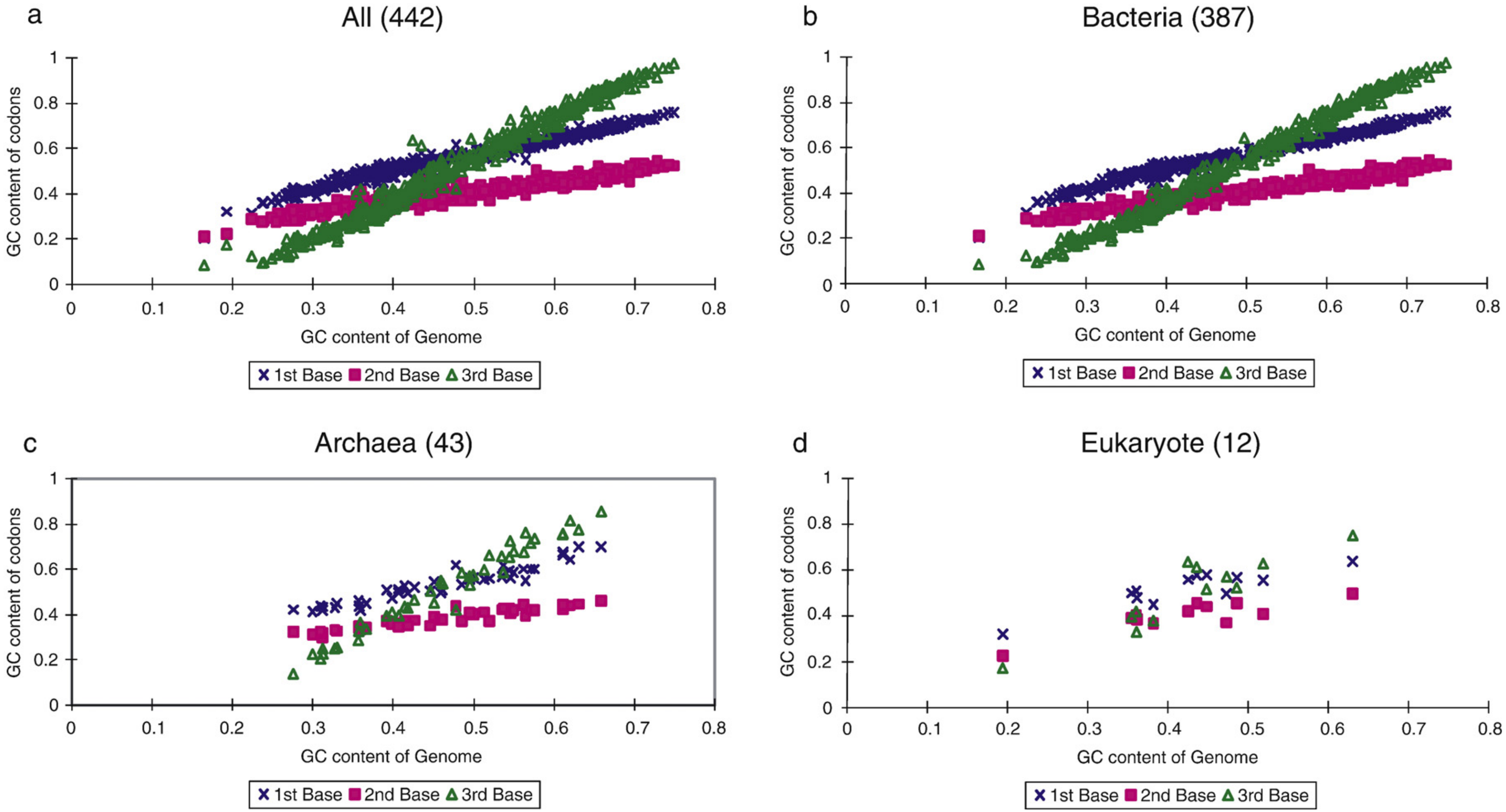


Fig. 1. Scatter plots of GC content between genome and base 1, 2 or 3 of codons for (a) all species (442 genomes), (b) Bacteria, (c) Archaea and (d) Eukaryota. The GC contents of the first, second and third bases are represented by different symbols, [x], [□] and [Δ], respectively.

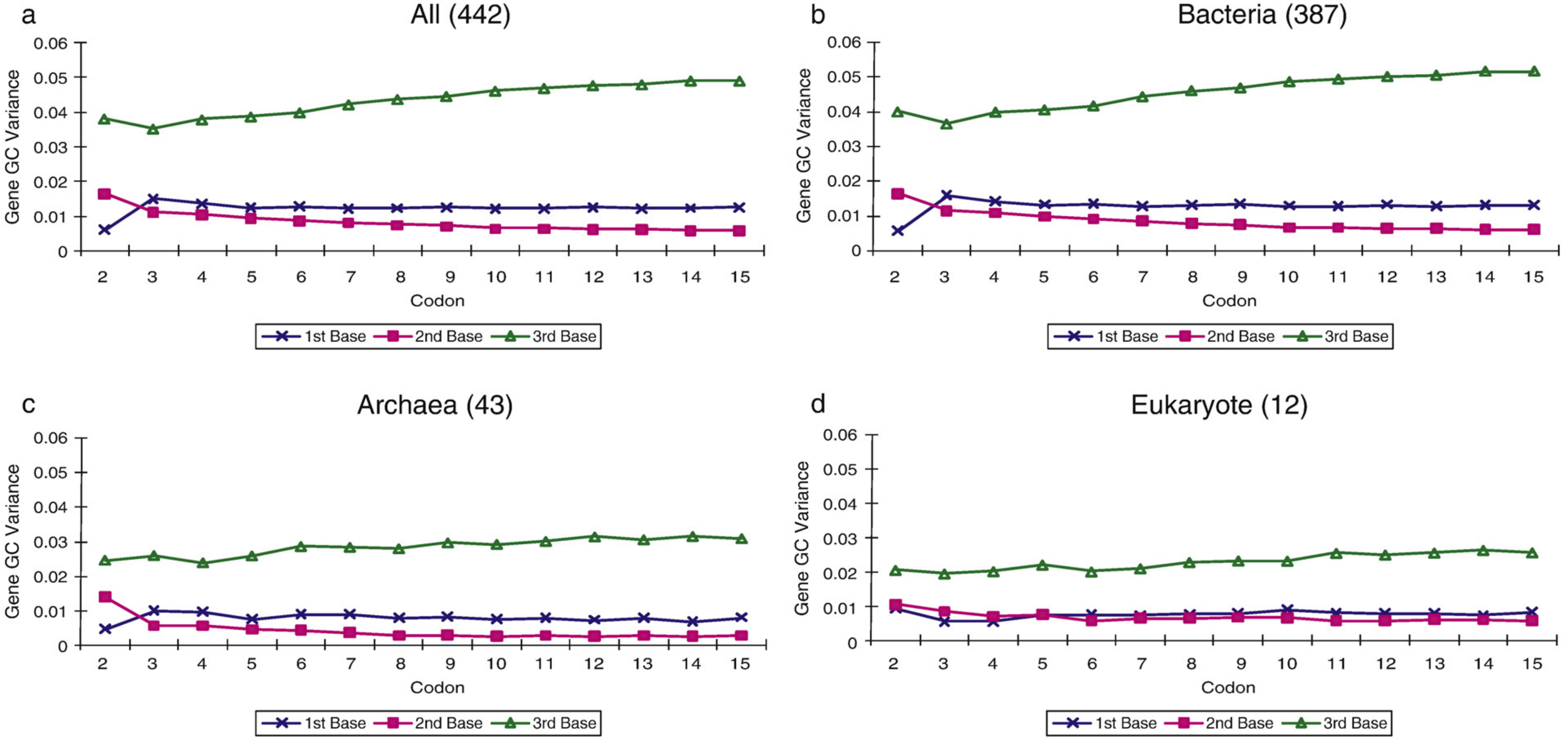


Fig. 2. GC variance for various codons for (a) all species, (b) Bacteria, (c) Archaea and (d) Eukaryota. The GC variance of the first, second and third bases are represented by different symbols, [-x-], [-□-] and [-D-], respectively. For (a), (b), and (c), the second base is the most conserved out of the 3 bases in all but the second codon. Eukaryotic genomes differ slightly from that rule for the third and fourth codons, such that the second base is less conserved.

considered, namely directional mutation pressure (AT/GC pressure), translational selection, codon usage bias (including any restriction from availability of charged tRNA), and specific function requirements for the N-terminal of protein. All these factors would influence the second codon composition to a varying degree, depending on the type of gene. For example, the effect of the mutational pressure on a second codon's composition is relatively low for highly-expressed genes [13,21]. However, to date, there have been only a few reports which have specifically addressed evidence of multiple selection constraints on the second codon.

In this study, we have looked for more evidence to comprehend the complex and compounded interactions of these constraints on the second codon, and to detect new and consistent molecular features across different species. We conducted a large-scale survey of the second codon of annotated genes from 442 species, including archaea, bacteria and eukaryotes. We examined second codon nucleotide distributions, amino acid distributions and their correlations to the genes of various biological processes. Two novel and intriguing molecular features of the second codon were detected. The first is that the first base of the second codon is universally more conserved than the second base in all archaeal, bacterial and eukaryotic genomes examined. Such a molecular feature is uniquely different from the well-established concept that the second base of a codon is always the most conserved [22,23], directly suggesting that the direction of mutational force plays a weaker role in this codon. The second molecular feature uncovered is that four of the most common amino acids, serine (S), threonine (T), alanine (A) and lysine (K), are frequently encoded in the second codon for Translation and Energy-related genes, whereas these four amino acids are avoided in Cell Envelope-related genes in most of the selected genomes. For the first time, our result clearly reveals that the second codon is differentially selected during the evolutionary process, a selection dependent on biological functions. Furthermore, we reveal strong C nucleotide bias in the second base position of this codon in the Cell Envelope-related genes in most of the genomes studied, which suggests that this molecular feature is specifically related to function. The discovery of these new features indicates that (i) adding to the previous understandings, a gene's functionality is also a significant factor influencing the composition of the second codon; (ii) the second codon plays a significant role not only in gene expression but also in protein function.

Results

The effect of directional mutation on the second codon

The effect of directional mutation in various micro-organisms was clearly observed by Muto and Osawa [23] although only approximately 59 genes from 11 micro-organisms were examined in their study. We evaluated this phenomenon with a much more comprehensive dataset of 442 genomes. The relationship between GC content at each nucleotide base and genome GC content we found is similar to Muto and Osawa's results [23]. The GC content of the first and the third base vary proportionally with genome GC content, whereas there is relatively less variation in the second base compared to the GC content of genomes (Fig. 1). Furthermore, we were interested in examining whether nucleotides of an individual codon, particularly

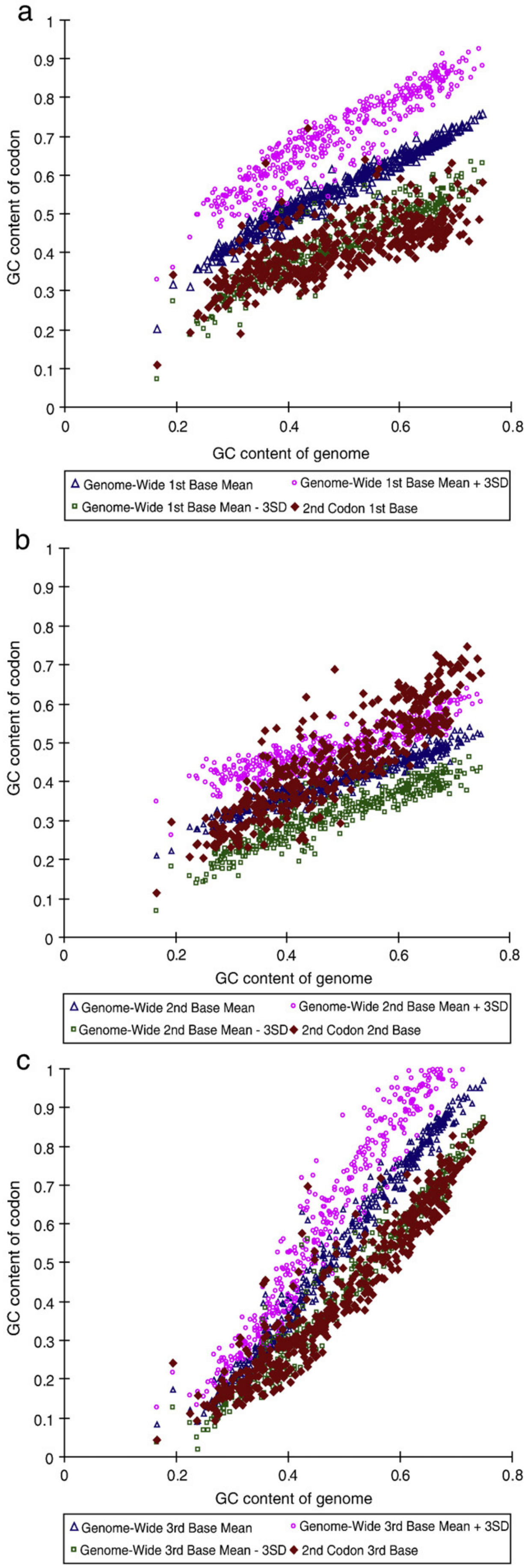


Fig. 3. Comparisons of nucleotide preference for (a) 1st base, (b) 2nd base and (c) 3rd base. The mean ± 3 standard deviation values are also plotted for visualisation. Clearly, the difference between base mean distributions (calculated from all codons) and second codon base distributions is statistically significant in most of genomes; this difference in GC-content distribution for the second codon 1st base in 395 genomes is beyond 2 standard deviations (i.e. $P<0.05$) and for 280 genomes is beyond 3 standard deviations ($P<0.01$); for the second codon 2nd base in 213 genomes ($P<0.05$) and 159 genomes ($P<0.01$); for the second codon 3rd base in 370 genomes ($P<0.05$) and 253 genomes ($P<0.01$).

the second codon, have a similar genome GC relationship to that described in the previous report. Second codons and other contiguous codons, from all annotated genes in 442 genomes, were retrieved and their GC contents at each position calculated. Surprisingly, the relationship between the GC content of the second codon and genome is different from other codons (Fig. 2). The second base in the second codon is less conserved than the first base, which is not the case for other codons, i.e. the third to the fifteenth codons (Fig. 2). The trends for these codons are generally similar to Fig. 1. In different domains, bacterial and archaeal genomes display a similar pattern from the second to the fifteenth positions (Fig. 2). Eukaryotic genomes are slightly different in the third and fourth codons, such that the second base is less conserved. After the fifth codon, the degree of conservation of each nucleotide becomes similar to that of the total codons.

In order to confirm this observed difference in the second codon, we tested the significance of differences between the GC content of each base of the second codon and that of the total codon average. Clearly, as shown in Fig. 3, the difference is statistically significant in most genomes; this difference in GC-content distribution in many of the genomes is beyond three standard deviations (i.e. $P<0.01$). The influence of selection pressure of the GC-directional mutation seems to be reduced for the first and third base, whereas for the second base, the effect is considerably larger in higher GC-content genomes. This suggests that the selection pressure on the second codon is uniquely different from other codons.

Adenine and cytosine preference in the second codon

Apart from the effect of directional mutation, there are other constraining forces influencing the nucleotide composition of the second codon, as suggested by our observations and other previous reports. In order to detect the effect of other constraints, we looked at the nucleotide bias for each base of the second codon and compared this to other contiguous codons (down to the thirtieth codon). We separated the genomes into three groups, with GC content less than 45%, between 45% and 55%, and larger than 55%, to minimise the effect of the GC-content directional mutation. Average of difference (AOD) is defined as the difference between all codons and the second codon's first, second and third base frequencies (refer to the Materials and methods section below for more details). In AOD analysis, nucleotide preference or avoidance trends are observed. The AOD results in Table 1 suggest that adenine (A) is preferred in the first and third bases, which can explain why the GC content of these two bases is lower than the average, as shown in Fig. 3. Cytosine is uniquely preferred in the second base of the codon, particularly in the two higher GC-content groups, which also supports the result in Fig. 3 that a higher than average GC content of the second base of the second codon is observed in higher GC-content genomes.

To confirm these molecular features, we conducted a more comprehensive comparison in which the first thirty codons, including the second codon, are visualised on a single heat map. From this analysis, a unique preference in the second codon can be clearly observed (Fig. 4; Table S1). A clear preference for A was seen in the first and third bases in the beginning of sequences. This echoes the observation in previous reports that the pattern for the second base appears to be slightly different. A very strong signal for C preference was detected in the second base of the second codon, and T avoidance is apparent in the first and second bases of the second codon. These nucleotide features are likely to be the result of the compounded effect from more constraining forces, rather than just the directional mutation.

Amino acid profiles of the second codon in different biological process-related genes

Little research has thus far focused on amino acid bias in the downstream sequence. There are two exceptions: the preference for specific amino acids in the second codon is related to the cleavage function [15–19], and there is an implied connection between the second codon and protein function in *Vibrio* genes [26]. Apart from these, there has been no further investigation of the relationship between the amino acid bias in the second position and protein functionality. Based on the functionality of different biological processes, we conducted an extensive analysis using genes from 333 genomes (including 45 eukaryotes, 284 bacteria and 4 archaea). We then examined whether these amino acid biases and various codon frequencies are observed equally in each of the biological

Table 1
Average of difference (AOD) between All Base frequencies and Second Codon frequencies for the first, second and third base of the second codon. In AOD analysis, the nucleotide preference or avoidance trends are observed. The table suggests that adenine (A) is preferred in the first and third bases. Also, the cytosine is uniquely preferred in the second base of the codon, particularly in the higher GC-content groups.

Letter	1st base			2nd base			3rd base		
GC contents	GC≤45%	45%<GC≤55%	GC>55%	GC≤45%	45%<GC≤55%	GC>55%	GC≤45%	45%<GC≤55%	GC>55%
All									
A	0.159	0.178	0.215	0.076	0.052	0.005	0.088	0.106	0.092
C	−0.031	−0.050	−0.047	0.016	0.047	0.106	−0.034	−0.071	−0.051
G	−0.079	−0.101	−0.150	−0.011	−0.005	0.009	−0.024	−0.065	−0.110
T	−0.049	−0.027	−0.018	−0.082	−0.094	−0.120	−0.030	0.030	0.069
Bacteria									
A	0.171	0.195	0.222	0.083	0.058	0.006	0.098	0.117	0.096
C	−0.030	−0.053	−0.049	0.016	0.046	0.109	−0.035	−0.074	−0.049
G	−0.092	−0.113	−0.155	−0.013	−0.007	0.006	−0.030	−0.073	−0.115
T	−0.049	−0.029	−0.019	−0.086	−0.097	−0.121	−0.033	0.029	0.068
Archaea									
A	0.121	0.133	0.123	0.068	0.046	−0.004	0.040	0.064	0.051
C	−0.029	−0.039	−0.028	−0.021	0.027	0.067	−0.034	−0.074	−0.082
G	−0.028	−0.066	−0.086	0.001	0.007	0.051	0.004	−0.029	−0.055
T	−0.065	−0.029	−0.010	−0.048	−0.080	−0.113	−0.011	0.039	0.086
Eukarya									
A	−0.013	−0.017	0.009	−0.039	−0.053	−0.034	−0.017	0.032	0.020
C	−0.046	−0.027	0.017	0.113	0.145	0.082	−0.020	0.004	−0.031
G	0.060	0.003	−0.092	0.010	−0.007	0.009	0.036	−0.052	−0.004
T	−0.001	0.041	0.066	−0.084	−0.084	−0.057	0.000	0.016	0.015

process categories; and, if not, which amino acid bias is preferred in different categories? The amino acid frequencies in the second position were calculated and the most frequent amino acids amongst all species were used for further analysis: serine (TCN, S), lysine (AAN, K), alanine (GCN, A), and threonine (ACN, T) (Fig. S2). Analysis of the frequency of members of this quartet (SKAT) was used to discover whether these four amino acids are equally popular across different biological process-related genes. To verify whether there is ubiquity for amino acids with C preference in the second base of the second codon, the NCN codons were also specifically selected for further analysis, including alanine, serine, threonine, and proline (CCN, P).

The heat map (Fig. 5a; Table S2) clearly shows that the frequencies of the four most common amino acids in the second position vary in different biological process-related genes. Translation and Energy Metabolism-related genes generally have a considerably higher SKAT proportion than other functional genes. This observation is also species-dependent for some functional categories. For example, the Fatty Acid and Phospholipid Metabolism-related genes in eukaryotic genomes prefer to encode these four amino acids in the second codon position; on the other hand, only a few of the prokaryotic genomes do.

Since three of the four most frequent codons as shown above are NCN, we therefore investigated the NCN codons serine (TCN, S), threonine (ACN, T), alanine (GCN, A) and proline (CCN, P). A more intriguing result was observed in analysis of the frequency of amino acids in this quartet, STAP (Fig. 5b; Fig S3; Table S3). The Cell Envelope-related genes in the surveyed organisms commonly avoid encoding for these amino acids in the second codon position. This suggests that there are specific factors which functionally or structurally affect these genes to prevent the use of NCN-encoded amino acids. On the other hand, the Translation-related genes mostly prefer STAP amino acids, which is generally rather species-dependent for other functional categories.

Discussion

A consensus feature of the second codon in three domains

In the present study, we have detected two novel molecular features of the second codon in genes. The most significant is the correlation between the GC contents of the second codon and the genome. The second codon sequences collected from a total of 442 genomes across three domains show that there is universally less conservation of the second base than the first base. To our knowledge, this has never been reported before in the literature. It has long been believed that the order of conservation of bases in any codon is that the second base is the most conserved, followed by the first and then the third; however, for the second codon, at least, this is not observed. This molecular feature also clearly indicates that the evolutionary selection pressures for the second codon are different to other codons in a given gene. In addition, the third, fourth and fifth codons of the eukaryotic genes also display a similar conservation pattern as the second codon. The reason for this difference in eukaryotic genes remains unknown. The Kozak sequence ((gcc) gccRccAUGG, AGNNAUGN, ANNAUGG, ACCAUGG, or GACACCAUGG; the underlined bases are the start codon) is one of only a few consensus sequences which include parts of the second codon's bases [27,28];

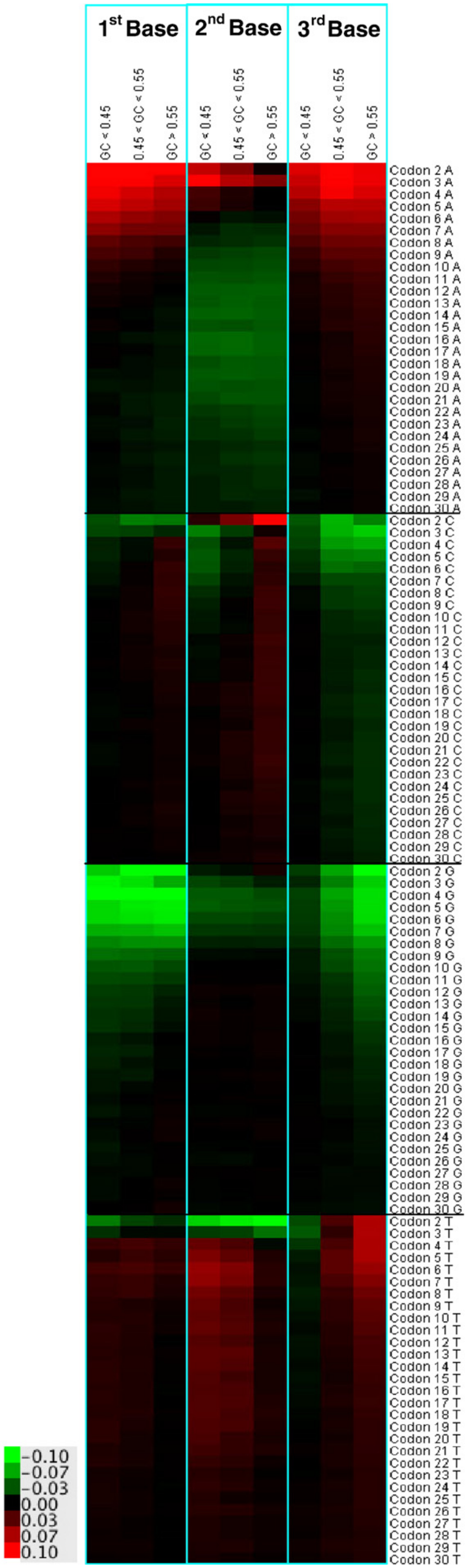


Fig. 4. Heat map of nucleotide bias at base 1, 2 and 3 for codons 2–30, separated into three genome GC-content groups. A red colour indicates preference whereas a green colour indicates avoidance. The colour scale indicates the value of normalised frequencies. Preference for A was clearly seen in the first and third bases in the beginning of sequences. A very strong signal indicating C preference is detected in the second base of the second codon, and T avoidance is apparent in the first and second bases of the second codon.

however, consensus sequences are not detectable in the third, fourth or fifth codons.

Other selective pressures interfere with directional mutation's influence on second codon usage

The direction mutation pressure (alternatively called A + T/G + C selective pressure or biased mutation pressure) has long been known to play a major role in the diversification of genomic DNA and codon usage in evolution. The correlation of G + C contents between genome and codon shows an unambiguous, linear, and positive relationship. Moreover, this pressure was suggested to predominate over other selective forces in determining codon usages [23]. However, in this study, we have presented an exception whereby the second codon is universally different from the genome average and from other codons, which suggests that the composition of the second codon is also significantly affected by other types of selective forces. What are these other forces? According to much of the relevant literature, one force should be translational selection; however, we have provided clear evidence here to show that not only translational selection but also protein functional selection both play a role in the composition of the second codon (Fig. 5).

Variable function-driven selection forces on the second codon of different genes

The previous studies on protein-function constraints in the second codon were mainly related to protein turn-over processing and maturation. Some preferred amino acids were detected, which were suggested to be involved in these two biochemical activities [15–19]. In contrast, our large-scale *in silico* analysis of different functional group genes shows that the amino acid bias of the second codon vary for different functional group genes. This suggests that protein-function constraints are not only applicable to protein turn-over and maturation but also to other functional genes as well. In other words, the influence on protein function of mutation in the second codon has been underestimated.

Although the same functional group genes from various organisms may have different preferences regarding their second codon composition, two functional groups, Cell Envelope-related and Translation-related genes, have amino acid biases that are clearly distinct from the patterns observed for most of the analysed genomes across three domains. These two groups demonstrate opposite biases: the Cell Envelope-related genes avoid using NCN codons and the popular amino acids (serine, lysine, alanine and threonine), whereas the Translation-related genes strongly prefer these same four amino acids in the second position. Why do these two groups differ in this way? What factors contribute to such a difference? We may be unable to find answers directly from this study or from previous reports. However, at the very least, these molecular features clearly suggest that the second codons of these two gene groups have been considerably affected by certain function-specific constraints over a reasonably long course of evolution. This has caused the bias patterns to be consistent in all three domains.

The Cell Envelope-related genes mostly encode for surface proteins, such as lipoproteins, cell wall-related proteins, membrane-bound proteins, transporters and defence proteins (antibiotic-sensitivity or resistance genes). We initially speculated that the avoidance of NCN codons may be related to Cell Envelope proteins' translocation or some chemical signal interaction between the cytoplasm and outside environments which may require some specific structural pattern, so we looked carefully at the Cell Envelope genes in *E. coli*. We manually checked for correlation between secondary structures and amino acids in 1212 genes including the Cell Envelope genes, using the RasMol program (http://rasmol.org/); however, no significant bias was detected (data not shown). Furthermore, we also inspected for any special amino acid motif in these *E. coli* genes using sequence logo analysis; however no significant bias was detected in this case either (data not shown).

Nonetheless, the avoidance, particularly the NCN avoidance, should reasonably be logically correlated to some aspect of protein function. To take an example: we examined 347 lipoprotein genes from 12 *Mycoplasma* genomes which were not included in the analysis due to the small size of their annotated Cell Envelope genes (see Materials and methods); however, most of the genes, 327 out of 347, do not have any NCN codons, and 243 of them encode for lysine in the second codon (see Supplementary material). Although it remains unclear why alanine, serine and threonine are avoided at this position, lysine is much preferred and encoded in the second codon in lipoprotein genes. The same phenomenon is also observed in other Firmicutes-related organisms and some eukaryotes (Figs. 5a and b). In addition, although lysine (AAA or AAG) preference is believed to facilitate translation, according to studies of highly-expressed genes of *E. coli*, three transcription profiles of different *Mycoplasma* species showed that merely two of 31 lipoprotein genes were up-regulated in the microarray analyses [29–31]. As a result, we argue that lysine preference may be more related to protein-function selection than translation selection.

In addition, we investigated the extent of the effect of the translation selection force, where the more highly-expressed genes are composed of the most popular codons [7,10,11], by analysing the codon usage bias of the genes in various functional groups. We found no difference in codon usage frequencies between the second codon of the Cell Envelope genes and other functional group genes in an analysis of *Salmonella enterica* subsp. Arizonae (see Table S5). This suggests that the observed codon usage bias in the Cell Envelope genes is more correlated to their particular function than the translational effect or other selection factors. However, more experimental evidence is required to explain this bias.

Materials and methods

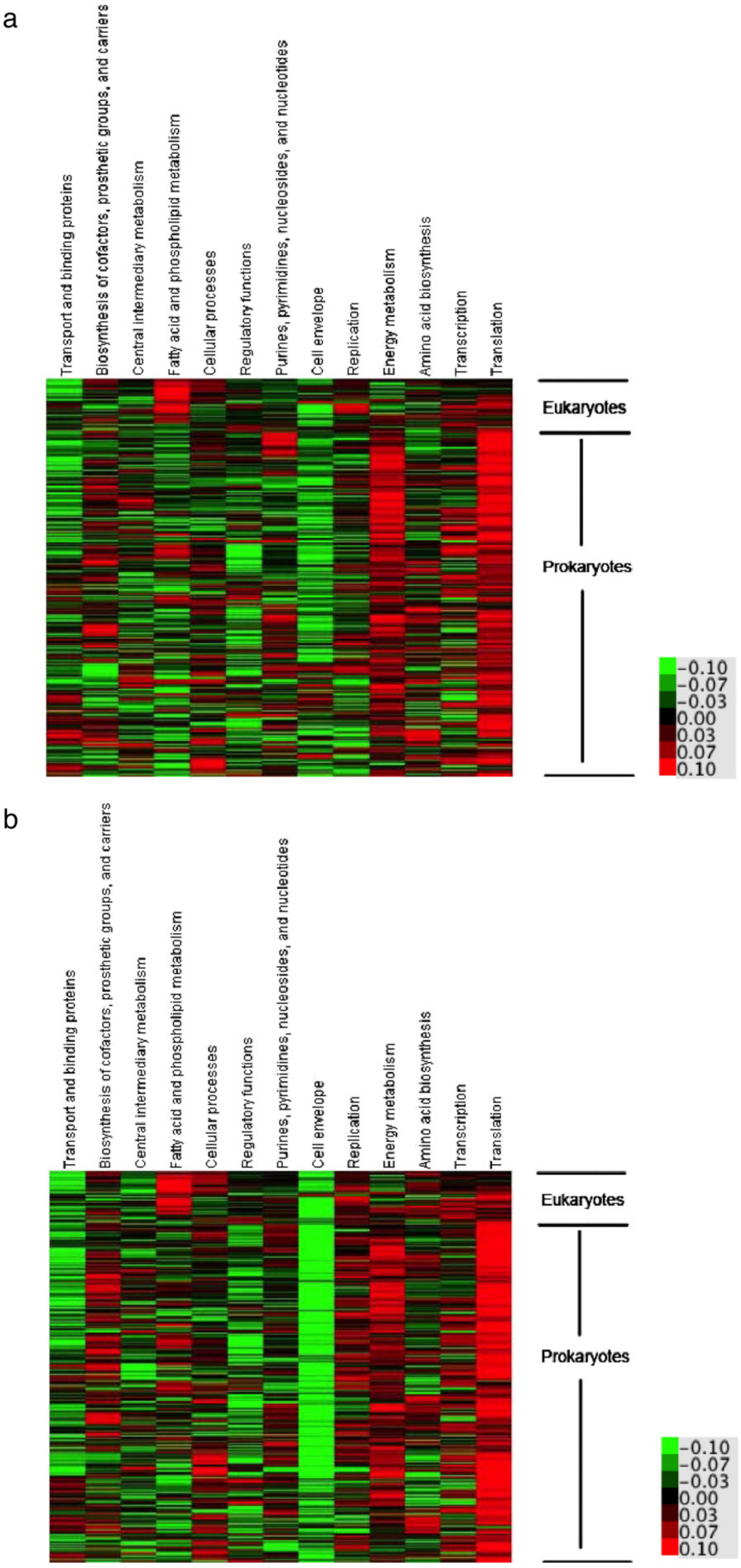
Database

Genome sequence data was obtained in 2007 from the KEGG database containing 659 genome sequences. This data was trimmed so that each species would only have one representative strand where their genome GC contents are also available (hence the completed genome sequence). This process yielded 442 individual genome sequences (442 species: 12 eukaryotes, 387 bacteria and 43 archaea), and this list is available in Supplementary material (genome list). Both Genome and Gene GC contents were calculated from the original sequence file, if available.

Codon versus gene GC-content variance

For each codon, we were interested in the variations of nucleotide bias (GC content), as this may give us an estimate of the conservation of nucleotide usage for all three base positions. Firstly, for each genome sequence, we aligned all genes together based on their

Fig. 5. Heat map of amino acid base across different functional groups. (a) SKAT bias and (b) STAP bias. A red colour indicates preference whereas a green colour indicates avoidance. (a) Translation and Energy Metabolism-related genes generally have a considerably higher SKAT proportion than other functional genes. (b) Cell Envelope-related genes in the surveyed organisms commonly avoid encoding for these amino acids in the second codon position. A two-tailed *t*-test showed that the normalised frequency distributions for Cell Envelope-related genes are significantly different to genes in other functional categories (Table S4). The colour scale indicates the value of normalised frequencies.


a
Transport and binding proteins
Biosynthesis of cofactors, prosthetic groups, and carriers
Central intermediary metabolism
Fatty acid and phospholipid metabolism
Cellular processes
Regulatory functions
Purines, pyrimidines, nucleosides, and nucleotides
Cell envelope
Replication
Energy metabolism
Amino acid biosynthesis
Transcription
Translation
Eukaryotes
Prokaryotes
-0.10
-0.07
-0.03
0.00
0.03
0.07
0.10
b
Transport and binding proteins
Biosynthesis of cofactors, prosthetic groups, and carriers
Central intermediary metabolism
Fatty acid and phospholipid metabolism
Cellular processes
Regulatory functions
Purines, pyrimidines, nucleosides, and nucleotides
Cell envelope
Replication
Energy metabolism
Amino acid biosynthesis
Transcription
Translation
Eukaryotes
Prokaryotes
-0.10
-0.07
-0.03
0.00
0.03
0.07
0.10

positions (i.e., start codon aligned with start codons, first codon aligned with first codons, and so on). From this alignment of genes, we calculated the nucleotide frequencies at each base position. After obtaining the nucleotide frequencies at each base position, we then calculated the variances of these nucleotide frequencies from all genome sequences.

Nucleotide preference

To measure the possible nucleotide preferences for certain base positions in the second codon, we first calculated the average nucleotide frequencies for Base 1, Base 2 and Base 3 for all the genes in each genome sequence. Then, we separated genome sequences into three categories according to their Genome GC content: Low GC ($GC<0.45$), Med GC ($0.45<GC<0.55$) and High GC ($GC>0.55$). This separation of genomes according to GC content was performed because we believed that genomes with different levels of GC content would have different GC-directional mutation effects. In this way we minimised the effect of directional mutation in our analysis. For each nucleotide j, we estimated its preference by calculating the average of difference (AOD) between frequencies in base i in codon x, and average frequency in base i for all codons, where $i=1$, 2 or 3:

$$AOD_j = \frac{\sum_m (f_i(x) - F_i)}{m},$$

where $f_i(x)=$ nucleotide frequency of base i in x^{th} codon, $F_i=$ average nucleotide frequency of base i for all codons, and $m=$ number of genome sequences.

Amino acid bias analysis

Amino acid frequency analysis in various peptide positions was calculated in a similar way to the nucleotide frequency analysis, and it is possible that any nucleotide frequency bias in the second codon position may cause the subsequent amino acid bias in the second peptide position.

To examine the relationship between the amino acid bias and biological functions, we downloaded functional group information from GeneQuiz [available before 2005; 24] which divided all genes into thirteen functional categories: Amino Acid Biosynthesis, Biosynthesis of Cofactors, Prosthetic Groups and Carriers, Cell Envelope, Cellular Processes, Central Intermediary Metabolism, Energy Metabolism, Fatty Acid and Phospholipid Metabolism, Purines, Pyrimidines, Nucleosides, and Nucleotides, Regulatory Functions, Replication, Transcription, Translation, and Transport and Binding Proteins. Since GeneQuiz only has a limited number of species available (around twenty genomes), we used BLAST to find similar genes (with an E-value of $1\times e^{-5}$) for all thirteen functional categories in the KEGG database. By accepting species that had at least 100 protein sequences in each of the functional categories, we obtained a total of 435 genome sequences, of which 333 genomes (45 eukaryotes, 284 bacteria and 4 archaea) were chosen so that each species would only have one genome sequence for the subsequent amino acid bias analysis.

SKAT and STAP bias

The most frequent amino acids in the second peptide position amongst all species are: serine (TCN, S), lysine (AAN, K), alanine (GCN, A), and threonine (ACN, T). SKAT analysis was performed to determine whether these four amino acids are all equally popular across different biological process-related genes.

For investigation of C nucleotide preference in the second base of the second codon, the frequencies of amino acids serine (TCN, S), threonine (ACN, T), alanine (GCN, A), and proline (CCN, P) were examined (STAP analysis). For each species, we separated genes into functional categories as described in the previous section. The protein sequences were aligned, based on positions, and the amino acid frequencies were calculated. We obtained the frequency values of 'SKAT' and 'STAP' at the second protein base position for each functional category, and each of these values was normalised in such a way that for each genome, the sum of 'SKAT' and 'STAP' frequencies across all functional groups was zero. This normalisation allowed us to compare SKAT and STAP patterns across the different functional categories in all species.

Heat map construction

In order to visualise the amino acid bias (the SKAT and STAP analyses), average of difference (AOD) and nucleotide bias analyses, we used the Gene Cluster program, version 3.0 (http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/), to perform k-means clustering analysis with 100 training runs. The Tree View program, version 1.1.0 (http://jtreeview.sourceforge.net/), was used to draw heat maps and the clustering result.

Codon usage frequency analysis

To analyse the codon usage bias at the second codon position, we first downloaded the codon usage tables from the Codon Usage Database (http://www.kazusa.or.jp/codon/). Based on these codon usage tables, we then replaced the DNA sequences with numerical frequency values for each of the functional groups. A t-test was then performed to check for significant differences between the various codons, and also between codons in various functional categories.

Mycoplasma lipoprotein and sequence logo analysis

One way of determining the degree of dominance of a particular base over the others is by using sequence logo graphs [25]. The graph is plotted using variables fa_k, fc_k, fg_k, ft_k and r_k where fa_k, fc_k, fg_k and ft_k are frequencies of A, C, G and T nucleotides respectively at position k. r_k is calculated as:

$$r_k = 2 + \sum_{i=a,c,g,t} fi_k \times \log_2(fi_k)$$

For each position k, a column of length r_k is plotted. The order of A, G, C, T is determined by their respective frequency values, with the highest frequency nucleotide drawn at the top and the lowest on the bottom. The size of each nucleotide's font is proportional to its frequency value.

Potential errors

As there were only twelve eukaryote genomes in this study, frequency analysis may not be representative for the eukaryote domain. The inclusion of 43 archaea genomes in the analysis could lead to a similar error, but the results here should be more representative than that of eukaryotes. Also, we could have also introduced a bias into our protein-functional analysis with our gene collection methodology. Genes in various functional groups were collected by performing BLAST analysis using GeneQuiz-annotated genes as queries, which could lead to a less diverse gene set. In addition, alternative translation start sites were detected in some eukaryotic genes. The number of such genes is still very small, and should not affect the significance of our observation; however, the second codons of such protein isoforms were excluded from our analyses [32,33]. Alternative transcripts were excluded from our analysis. We therefore have assessed how much the alternative transcripts might affect the result of the analysis. Human transcripts from the Alternative Splicing and Transcript Diversity (ASTD) database and KEGG and NCBI refseq databases were downloaded. For the second amino acid, the frequencies for STAK are 40.6% for KEGG, 36.6% for ASTD and 40.0% for NCBI refseq. This indicates a maximum 4% difference in our result due to the alternative transcription. The result indicates that the observation is consistent even though alternative transcripts would cause small variation.

Acknowledgments

This work was supported by the Biodiversity Research Center and the Institute of Plant and Microbial Biology, Academia Sinica, Taiwan. We would like to show our appreciation to Mr. Yu-Hsiung Lin in assisting with data collection and Ms. Hsiao-Chi Chen and Dr. Harry Wilson for proofreading and scientific suggestions and comments.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2010.04.001.

References

[1] N.V. Tzareva, V.I. Makhno, I.V. Boni, Ribosome-messenger recognition in the absence of the Shine–Dalgarno interactions, FEBS Lett. 337 (1994) 189–194.

[2] M. Faxen, J. Plumbridge, L.A. Isaksson, Codon choice and potential complementarity between mRNA downstream of the initiation codon and bases 1471–1480 in 16 S ribosomal RNA affects expression of glnS, Nucleic Acids Res. 19 (1991) 5247–5251.

[3] M. O'Connor, T. Asai, C.L. Squires, A.E. Dahlberg, Enhancement of translation by the downstream box does not involve base pairing of mRNA with the penultimate stem sequence of 16S rRNA, Proc. Natl. Acad. Sci. U S A 96 (1999) 8973–8978.

[4] C.M. Stenstrom, H. Jin, L.L. Major, W.P. Tate, L.A. Isaksson, Codon bias at the 3′-side of the initiation codon is correlated with translation initiation efficiency in *Escherichia coli*, Gene 263 (2001) 273–284.

[5] C.M. Stenstrom, E. Holmgren, L.A. Isaksson, Cooperative effects by the initiation codon and its flanking regions on translation initiation, Gene 273 (2001) 259–265.

[6] T. Sato, et al., Codon and base biases after the initiation codon of the open reading frames in the *Escherichia coli* genome and their influence on the translation efficiency, J. Biochem. 129 (2001) 851–860.

[7] C.M. Stenstrom, L.A. Isaksson, Influences on translation initiation and early elongation by the messenger RNA region flanking the initiation codon at the 3′ side, Gene 288 (2002) 1–8.

[8] H. Chen, L. Pomeroy-Cloney, M. Bjerknes, J. Tam, E. Jay, The influence of adenine-rich motifs in the 3′ portion of the ribosome binding site on human IFN-gamma gene expression in *Escherichia coli*, J. Mol. Biol. 240 (1994) 20–27.

[9] J. Martin-Farmer, G.R. Janssen, A downstream CA repeat sequence increases translation from leadered and unleadered mRNA in *Escherichia coli*, Mol. Microbiol. 31 (1999) 1025–1038.

[10] T. Nishikubo, N. Nakagawa, S. Kuramitsu, R. Masui, Improved heterologous gene expression in *Escherichia coli* by optimization of the AT-content of codons immediately downstream of the initiation codon, J. Biotechnol. 120 (2005) 341–346.

[11] J.E. Brock, R.L. Paz, P. Cottle, G.R. Janssen, Naturally occurring adenines within mRNA coding sequences affect ribosome binding and expression in *Escherichia coli*, J. Bacteriol. 189 (2007) 501–510.

[12] Y. Niimura, M. Terabe, T. Gojobori, K. Miura, Comparative analysis of the base biases at the gene terminal portions in seven eukaryote genomes, Nucleic Acids Res. 31 (2003) 5195–5201.

[13] A. Tats, M. Remm, T. Tenson, Highly expressed proteins have an increased frequency of alanine in the second amino acid position, BMC Genomics 7 (2006) 28.

[14] D. Pal, P. Chakrabarti, Terminal residues in protein chains: residue preference, conformation, and interaction, Biopolymers 53 (2000) 467–475.

[15] J.W. Tobias, T.E. Shrader, G. Rocap, A. Varshavsky, The N-end rule in bacteria, Science 254 (1991) 1374–1377.

[16] T.E. Shrader, J.W. Tobias, A. Varshavsky, The N-end rule in *Escherichia coli*: cloning and analysis of the leucyl, phenylalanyl-tRNA-protein transferase gene aat, J. Bacteriol. 175 (1993) 4364–4374.

[17] A. Varshavsky, The N-end rule: functions, mysteries, uses, Proc. Natl. Acad. Sci. U.S.A. 93 (1996) 12142–12149.

[18] R.A. Bradshaw, W.W. Brickey, K.W. Walker, N-terminal processing: the methionine aminopeptidase and N alpha-acetyl transferase families, Trends Biochem. Sci. 23 (1998) 263–267.

[19] A. Serero, C. Giglione, A. Sardini, J. Martinez-Sanz, T. Meinnel, An unusual peptide deformylase features in the human mitochondrial N-terminal methionine excision pathway, J. Biol. Chem. 278 (2003) 52953–52963.

[20] M.L. Chiusano, et al., Second codon positions of genes and the secondary structures of proteins: relationships and implications for the origin of the genetic code, Gene 261 (2000) 63–69.

[21] A. Pan, C. Dutta, J. Das, Codon usage in highly expressed genes of *Haemophillus influenzae* and *Mycobacterium tuberculosis*: translational selection versus mutational bias, Gene 215 (1998) 405–413.

[22] N. Sueoka, On the genetic basis of variation and heterogeneity of DNA base composition, Proc. Natl. Acad. Sci. U.S.A. 48 (1962) 582–592.

[23] A. Muto, S. Osawa, The guanine and cytosine content of genomic DNA and bacterial evolution, Proc. Natl. Acad. Sci. U.S.A. 84 (1987) 166–169.

[24] S. Hoersch, C. Leroy, N.P. Brown, M.A. Andrade, C. Sander, The GeneQuiz web server: protein functional analysis through the Web, Trends Biochem. Sci. 25 (2000) 33–35.

[25] T.D. Schneider, R.M. Stephens, Sequence logos: a new way to display consensus sequences, Nucleic Acids Res. 18 (1990) 6097–6100.

[26] J. Wang, F.B. Guo, Base frequencies at the second codon position of *Vibrio cholerae* genes connect with protein function, Biochem. Biophys. Res. Commun. 290 (2002) 81–84.

[27] M. Kozak, An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs, Nucleic Acids Res. 15 (1987) 8125–8148.

[28] M. Kozak, Pushing the limits of the scanning mechanism for initiation of translation, Gene 299 (2002) 1–34.

[29] J. Weiner III, C.U. Zimmerman, H.W. Gohlmann, R. Herrmann, Transcription profiles of the bacterium *Mycoplasma pneumoniae* grown at different temperatures, Nucleic Acids Res. 31 (2003) 6306–6320.

[30] K.R. Cecchini, T.S. Gorton, S.J. Geary, Transcriptional responses of *Mycoplasma gallisepticum* strain R in association with eukaryotic cells, J. Bacteriol. 189 (2007) 5803–5807.

[31] M.L. Madsen, S. Puttamreddy, E.L. Thacker, M.D. Carruthers, F.C. Minion, Transcriptome changes in *Mycoplasma hyopneumoniae* during infection, Infect. Immun. 76 (2008) 658–663.

[32] A.V. Pisarev, V.G. Kolupaeva, V.P. Pisareva, W.C. Merrick, C.U. Hellen, T.V. Pestova, Specific functional interactions of nucleotides at key -3 and + 4 positions flanking the initiation codon with components of the mammalian 48S translation initiation complex, Genes Dev. 20 (2006) 624–636.

[33] A.V. Kochetov, Alternative translation start sites and hidden coding potential of eukaryotic mRNAs, Bioessays 30 (2008) 683–691.