

Research article

Open Access

## Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes

Xiu-Feng Wan<sup>1,2</sup>, Dong Xu<sup>2</sup>, Andris Kleinhofs<sup>3</sup> and Jizhong Zhou\*<sup>1</sup>

Address: <sup>1</sup>Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA, <sup>2</sup>Digital Biology Laboratory, Department of Computer Science, University of Missouri, Columbia, MO 65211, USA and <sup>3</sup>Department of Genetics and Cell Biology, Washington State University, Pullman, WA 99164, USA

Email: Xiu-Feng Wan - wanx@missouri.edu; Dong Xu - xudong@missouri.edu; Andris Kleinhofs - andyk@wsu.edu; Jizhong Zhou\* - zhouj@ornl.gov

\* Corresponding author

Published: 28 June 2004

Received: 28 January 2004

BMC Evolutionary Biology 2004, 4:19 doi:10.1186/1471-2148-4-19

Accepted: 28 June 2004

This article is available from: <http://www.biomedcentral.com/1471-2148/4/19>

© 2004 Wan et al; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article's original URL.

### Abstract

**Background:** Codon usage bias has been widely reported to correlate with GC composition. However, the quantitative relationship between codon usage bias and GC composition across species has not been reported.

**Results:** Based on an informatics method (SCUO) we developed previously using Shannon informational theory and maximum entropy theory, we investigated the quantitative relationship between codon usage bias and GC composition. The regression based on 70 bacterial and 16 archaeal genomes showed that in bacteria,  $SCUO = -2.06 * GC3 + 2.05*(GC3)^2 + 0.65$ ,  $r = 0.91$ , and that in archaea,  $SCUO = -1.79 * GC3 + 1.85*(GC3)^2 + 0.56$ ,  $r = 0.89$ . We developed an analytical model to quantify synonymous codon usage bias by GC compositions based on SCUO. The parameters within this model were inferred by inspecting the relationship between codon usage bias and GC composition across 70 bacterial and 16 archaeal genomes. We further simplified this relationship using only GC3. This simple model was supported by computational simulation.

**Conclusions:** The synonymous codon usage bias could be simply expressed as  $1 + (p/2)\log_2(p/2) + ((1-p)/2)\log_2((1-p)/2)$ , where  $p = GC3$ . The software we developed for measuring SCUO (codonO) is available at <http://digbio.missouri.edu/~wanx/cu/codonO>.

### Background

All amino acids except Met and Trp are encoded by more than one codon. DNA sequence data from diverse organisms have shown that synonymous codons for any amino acid are not used with equal frequency, even though choices among codons should be equivalent in terms of protein sequences [1-6]. Previous codon usage analyses showed that codon usage bias is very complicated and is associated with various biological factors, such as gene expression level [7-10], gene length [11-13], gene translation initiation signal [14], protein amino acid composi-

tion [6,15], protein structure [16,17], tRNA abundance [18-21], mutation frequency and patterns [22,23], and GC composition [24-27]. In this paper, we further explore the relationship between codon usage and GC composition.

GC composition may be described at three levels: 1) Overall GC content. The overall genome GC content in living organisms varies from 25–75% [28]. However, within a single gene, the overall GC content is 7–95%. 2) Local GC composition. Local GC composition is defined based on

the positions on the genetic codons. GC1 is the GC composition at the first site of codons, GC2 the GC composition at the second site of codons, and GC3 the GC composition at the third site of codons. 3) The ratio of G/C or A/T within a single strand of DNA. Based on Watson-Crick base pairing rules, the overall GC content is the same between the plus strand and minus strand of the DNA sequence [29]. However, within a single strand, the ratio of G to C and A to T may not be 1. Bacterial genomes were found to be relatively enriched in G over C and T over A, and slightly depleted in G+C, in their weakly evolution-selected positions (intergenic regions and third codon positions) in the leading strand compared with the lagging strand [28]. Although the overall GC content for a genome is reduced by AT-rich intergenic regions of the genome, the gene GC composition is tightly correlated with the genome's overall GC composition. The higher the overall GC content bias, the higher the local GC composition (GC1, GC2, and GC3).

It has been generally accepted that genome GC content is correlated with amino acid usage and codon usage [30]. A very low or very high GC composition is associated with a large codon usage bias. Recently, through a simple mutational model, Knight et al. [31] showed that it was the GC composition that drives codon and amino-acid usage although both mutation and selection play important roles. By using the corresponding analysis of codon usage over 32 bacterial and 8 archaeal genomes, Lynn et al. [32] further showed that codon usage bias was affected by GC composition and environment (e.g., temperature).

However, there has been no systematic, quantitative evaluation of the relationship between codon usage bias and GC composition. In addition, the theoretical basis underlying the relationship between codon usage bias and GC composition has not been illustrated. In this paper, we applied an informatics method [13], which is based on Shannon informatics theory and the entropy theory, to explore the relationship between GC composition and synonymous codon usage bias among 70 bacterial and 16 archaeal genomes. We presented an analytical model to quantify the non-linear relationship between GC3 and a measurement of codon usage bias (synonymous codon usage order, or SCUO), which reveals that GC3 is the key factor driving synonymous codon usage and that this mechanism is independent of species. These results were supported by our simulations. Our results also showed that the asymmetric distribution of G over C and A over T at the third codon position may increase codon usage bias. The underlying mechanisms behind GC composition and codon usage bias are discussed.

### Informatics method

We recently developed an informatics method [13] to provide an estimate for the orderliness of synonymous codon usage (SCUO) and the amount of synonymous codon usage bias. This method was based on the Shannon informatics theory and the entropy theory and allows the comparison of codon usage bias within and across genomes.

To calculate SCUO, we created a codon table for the amino acids that have more than one codon, indexed in an arbitrary way, so that we could unambiguously refer to the *j*-th (degenerate) codon of amino acid *i*,  $1 \leq i \leq 18$ . In mycoplasmas, Trp was also included into the codon table since a standard stop codon TGA encodes Trp in this specific species so that  $1 \leq i \leq 19$ . For simplicity, the following description of the method is only based on the standard genetic codon table although the actual SCUO computation considered special cases for different organisms.

Let  $n_i$  represent the number of degenerate codons for amino acid *i*, so  $1 \leq j \leq n_i$ ; for example,  $1 \leq j \leq 6$  for leucine,  $1 \leq j \leq 2$  for tyrosine, etc. For each sequence, let  $x_{ij}$  represent the number of times that synonymous codon *j* of amino acid *i* is present,  $1 \leq i \leq 18$ ,  $1 \leq j \leq n_i$ . Normalizing the  $x_{ij}$  by their sum over *j* gives the frequency of the *j*-th degenerate codon for amino acid *i* in each sequence.

$$p_{ij} = \frac{x_{ij}}{\sum_{j=1}^{n_i} x_{ij}} \quad 1$$

According to the information theory, we define the entropy  $H_{ij}$  of the *i*-th amino acid of the *j*-th codon in each sequence by

$$H_{ij} = -p_{ij} \log p_{ij} \quad 2$$

Summing over the codons representing amino acid *i* gives the entropy of the *i*-th amino acid in the each sequence

$$H_i = -\sum_{j=1}^{n_i} p_{ij} \log p_{ij} \quad 3$$

If the synonymous codons for the *i*-th amino acid were used at random, one would expect a uniform distribution of them as representatives for the *i*-th amino acid. Thus, the maximum entropy for the *i*-th amino acid in each sequence is

$$H_i^{\max} = -\log \frac{1}{n_i} \quad 4$$

If only one of the synonymous codons is used for the *i*-th amino acid, i.e., the usage of the synonymous codons is biased to the extreme, then the *i*-th amino acid in each sequence has the minimum entropy:

$$H_i^{\min} = 0 \quad 5$$

Unlike Shannon's definition of information, Gatlin [33] and Layzer [34] define the information as the difference between the maximum entropy and the actual entropy as an index of orderliness. The greater the information, the more ordered the sequence will be [35]. In our case, this information measures the nonrandomness in synonymous codon usage and therefore describes the degree of orderliness for synonymous codon usage for the *i*-th amino acid in each sequence.

$$I_i = H_i^{\max} - H_i \quad 6$$

Let  $O_i$  be the normalized difference between the maximum entropy and the observed entropy for the *i*-th amino acid in each sequence, i.e.,

$$O_i = \frac{H_i^{\max} - H_i}{H_i^{\max}} \quad 7$$

Obviously,  $0 \leq O_i \leq 1$ . When the synonymous codon usage for the *i*-th amino acid is random,  $O_i = 0$ . When this usage is biased to the extreme,  $O_i = 1$ . Thus,  $O_i$  can be thought of as a measure of the bias in synonymous codon usage for the *i*-th amino acid in each sequence. We designate the statistics  $O_i$  as the synonymous codon usage order (SCUO) for the *i*-th amino acid in each sequence.

Let  $F_i$  be the composition ratio of the *i*-th amino acid in each sequence:

$$F_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{\sum_{i=1}^{18} \sum_{j=1}^{n_i} x_{ij}} \quad 8$$

Then the average SCUO for each sequence can be represented as

$$O = \sum_{i=1}^{n_i} F_i O_i \quad 9$$

A software package called codonO was written using the C programming language to calculate SCUO for each open reading frame (ORF). This program is available at <http://digbio.missouri.edu/~wanx/cu/codonO/>. The 86 unicellular genomes explored in this paper can be found at our

Web site <http://digbio.missouri.edu/~wanx/cu/genomelist.htm>.

### Theoretical model between GC3 and codon usage

From the standard genetic code for synonymous codons, each amino acid was encoded by  $xyU/C$  and/or  $xyA/G$ . The third site of each codon should be purine (R, either U or C) or pyrimidine (Y, either A or G). For those amino acids with double genetic codons, they can only be coded by  $xyU/C$  or  $xyA/G$ . We define  $\bar{G} = A$ ,  $\bar{C} = U$ , and *vice versa*. We assume the probability of G+C at the GC3 is  $p$ . Then the probability of A+T at the GC3 is  $1-p$ . Let  $\alpha$  be  $P(G)$ , i.e., the probability of G, and  $\beta$  be  $P(A)$ , i.e., the probability of A. Hence,  $0 < \alpha \leq p$ ,  $0 < \beta \leq 1 - p$ . If we apply Equation 2, we will obtain entropy for each amino acid  $-\alpha \log_2 \alpha$  if the third site is G or  $-(p - \alpha) \log_2 (p - \alpha)$  if the third position is C (we use base 2 through our theoretical model). The entropy for each amino acid will be  $-\beta \log_2 \beta$  if the third site is A or  $-(1-p-\beta) \log_2 (1-p-\beta)$  if the third position is T. Let us use Gln as an example, which is coded by CGA/G. If the sequence was coded only by this single amino acid,  $\alpha = p$  and  $\beta = 1-p$ . Therefore, by applying Equation 3, we can calculate the overall entropy for synonymous codons,  $-p \log_2 p - (1-p) \log_2 (1-p)$ . In this case, the maximum entropy will be  $\log_2 2 = 1$ . Thus, the overall codon bias for Gln will be  $1 + p \log_2 p + (1-p) \log_2 (1-p)$ . For all of the amino acids with 2 genetic codons, we can use the same equation:

$$SCUO_{n_2} = 1 + p \log_2 p + (1-p) \log_2 (1-p) \quad 10$$

We can apply the similar deducing process for amino acids with 3, 4 or 6 genetic codes. Based on our definition of SCUO (Equation 9), the overall codon usage bias will be the sum of each amino acid. Thus, we obtain this first-order approximation for the overall codon bias:

$$SCUO = 1 + \mu * SCUO_{n_2} + \nu * SCUO_{n_3} + \omega * SCUO_{n_4} + \gamma * SCUO_{n_6} \quad 11$$

where  $\mu$ ,  $\nu$ ,  $\omega$ , and  $\gamma$  are associated adjusted weights, which are the amino acid usage frequency of 2, 3, 4 or 6 genetic codes, respectively. The adjusted weights in archaea and bacteria are shown in Table 1.

To simplify the analysis process, we simplify the analytical model as a binary selection model. Among the 18 synonymous codons, only Ile has three genetic codons. All other synonymous codons are encoded in pairs with  $xyZ$  and  $xy\bar{Z}$ . Assuming no additional codon bias results from a biased distribution between A and T or between C and G, i.e.,  $P(A) = P(T)$ ,  $P(G) = P(C)$ , the synonymous codons bias will be simplified to be a binary selection model:

**Table 1: Adjusted factors in archaea and bacteria**

	$\mu$	$\nu$	$\omega$	$\gamma$
Archaea	0.43	0.08	0.33	0.16
Bacteria	0.42	0.06	0.35	0.17

$$\text{SCUO} = 1 + (p/2)\log_2(p/2) + ((1-p)/2)\log_2((1-p)/2) \quad 12$$

In this case, the codon bias is solely determined by GC3,  $p$ . This simplified estimation between codon usage bias and GC3, which will reflect the lowest boundary of the codon usage bias for a specific GC3.

## Results

### GC composition affects SCUO within a single genome

We examined the relationships of SCUO over GC content, GC1, GC2 and GC3 in *E. coli* (Fig. 1). We only included genes with >200 codons, based on a comparison between gene length and SCUO in *E. coli* [13] that showed the SCUOs of genes with <200 codons have large fluctuations in codon usage bias. These results demonstrated that the overall content of GC and GC3 in *E. coli* have the largest impact on SCUO. This was also shown to be universal within other genomes (data not shown). The *E. coli* genome exhibited three horns (Figs. 1a and 1d). A lower or higher GC or GC3 over the center GC (50.8%) was associated with a relatively higher SCUO. GC1 showed two horns, whereas GC2 did not show this trend.

### Relationship between GC composition and SCUO across archaeal and bacterial genomes

The comparison between GC composition and SCUO across different genomes showed that the non-linear relationships between SCUO and GC3 among the unicellular organisms were independent of species (Figs. 2 and 3). The dependency of SCUO on GC3 was also strong and non-linear. Based on the quadratic regression, in bacteria,  $\text{SCUO} = -2.06 * \text{GC3} + 2.05 * (\text{GC3})^2 + 0.65$ ,  $r = 0.91$ ,  $P < 0.0001$ . In archaea,  $\text{SCUO} = -1.79 * \text{GC3} + 1.85 * (\text{GC3})^2 + 0.56$ ,  $r = 0.89$ ,  $P < 0.0001$ . To investigate the impact of GC3 on GC, we combined both archaeal and bacterial data and plotted the relationships between GC3 and GC (Fig. 4a). GC was linearly correlated ( $\text{GC3} = 1.78 * \text{GC} - 0.35$ ,  $r = 0.87$ ) with GC3 for 86 unicellular genomes. We also plotted the relationship between GC1 vs. GC3 and GC2 vs. GC3 (Figs. 4b and 4c). We found that GC3 was linearly correlated with both GC1 and GC2. Interestingly, the slope for GC3 vs. GC1 (0.42) was slightly higher than that for GC3 vs. GC2 (0.27).

We plotted the simplified binary selection model in the GC3 vs. SCUO of both archaea and bacteria. The binary

selection model closely fit the lowest boundary of the correlation between GC3 and SCUO. The codon usage bias was generally above the plotted curve because in most cases,  $P(A)$  is not equal  $P(T)$  and  $P(G)$  is not equal to  $P(C)$  in the natural world. The distribution of Ile may also increase the codon usage bias.

### Simulation of codon usage bias and GC composition

We carried out numerical simulations to further study the relationship between codon usage bias and GC composition. The simulation process was implemented to mimic the unicellular GC composition. For example, the linear relationship of GC3 vs. GC1 (Fig. 5b) and GC3 vs. GC2 (Fig. 5c) in the pseudo ORFs followed those in archaea and bacteria (Figs. 5b and 5c). The simulation showed that the correlation between GC and GC3 (Fig. 5a) was similar to that in archaea and bacteria (Fig. 4a). Moreover, the relationships of GC1 vs. SCUO (Fig. 5d), GC2 vs. SCUO (Fig. 5e), and GC vs. SCUO (Fig. 5g) in the simulation were also similar to those in archaea and bacteria (Figs. 2 and 3). As with the unicellular data, our simulation results demonstrated that there was good agreement between the non-linear relationship of codon usage bias and GC3 (Fig. 5f), and the binary selection model.

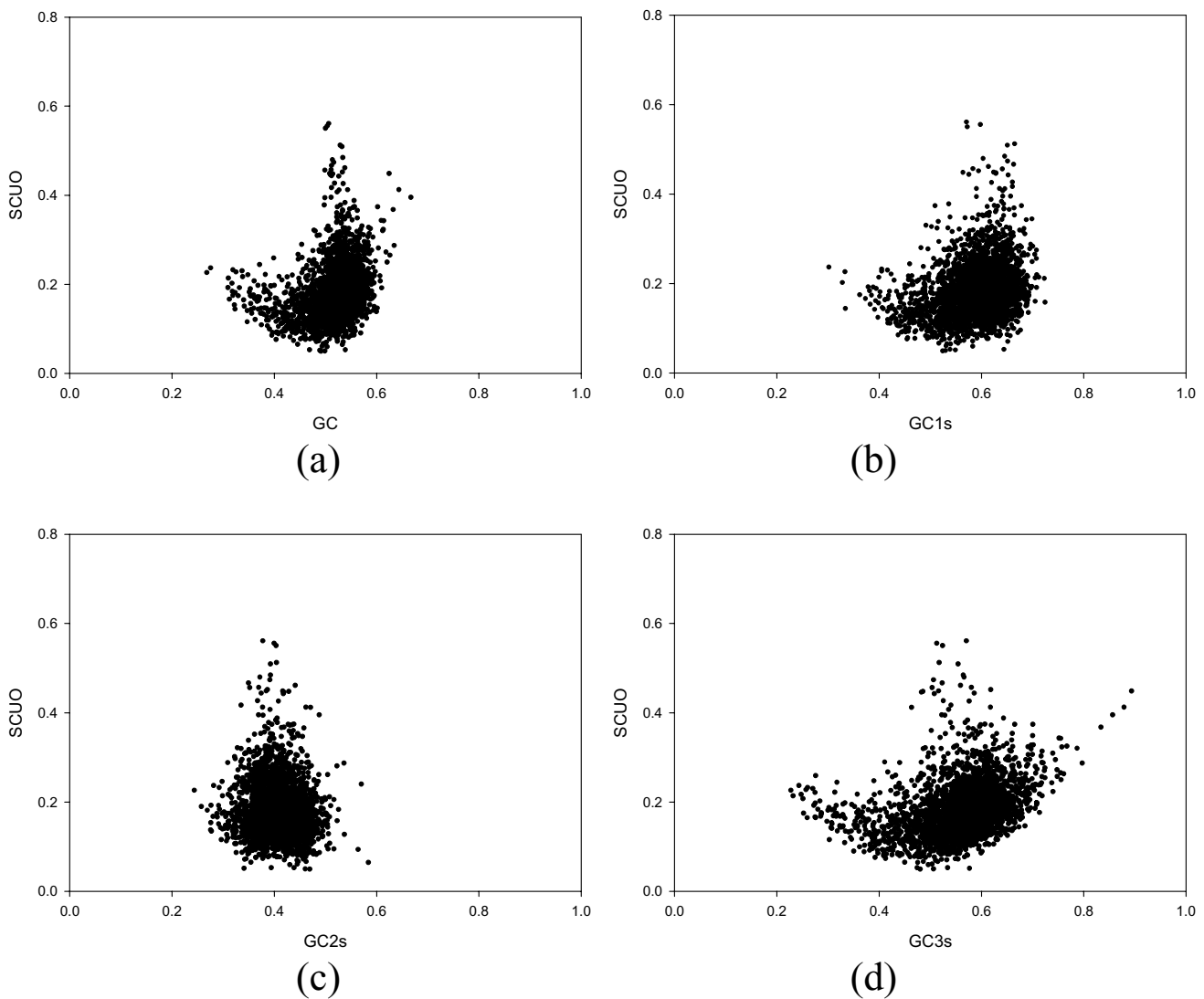
### Impact of asymmetric distribution of G/C and A/T on codon usage bias

We defined  $P(A) = P(T)$  and  $P(G) = P(C)$  for the third codon positions at step 4 in the simulation process. The correlation between codon usage bias and GC3 is shown in Figure 6. Compared with the random distribution of  $P(A)$  and  $P(G)$ , both the SCUO range and SCUO values of the pseudo ORFs are smaller during symmetry distribution of G over C and A over T at the third codon positions. Compared with simulation results shown in Figure 5, this simulation showed that asymmetric distribution of mutational pressure at the third codon positions may increase codon usage bias.

## Discussion

In this paper, we applied the informatics method [13] to measure SCUO and GC composition within and across unicellular genomes. We first explored the relationship between SCUO and GC composition in *E. coli*. Our results demonstrated that the relationship between codon usage bias and GC3 formed a "U" shape, which may reflect a

## Figure 1



### Figure 1

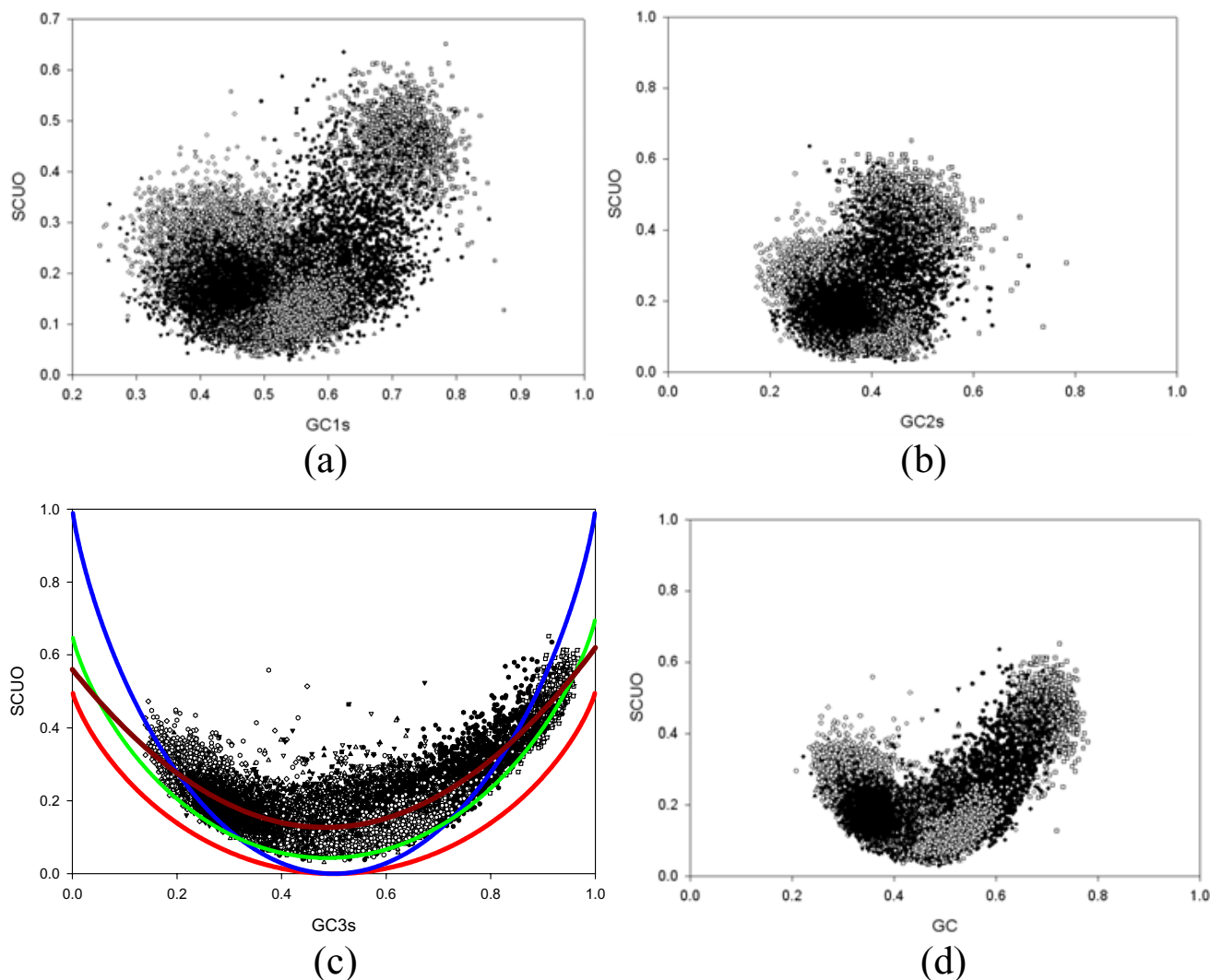
**Relationship between SCUO and GC composition in *E. coli* K12.** (a) Relationship between SCUO and the overall GC composition. (b) Relationship between SCUO and GC1. (c) Relationship between SCUO and GC2. (d) Relationship between SCUO and GC3.

direct mathematic consequence of GC3 over codon usage bias. It is interesting that there was a weak horn in the center of "U" shape (Fig. 1), which is similar to previous reports [36]. The genes in the central horns were discussed in Karlin *et al.* [36]. We can detect some weak horns in some other genomes, such as *Lactococcus lactis*. However, the central horns were not shown within some other genomes we explored (data not shown). The comparison

between GC composition and SCUO confirmed previous reports that GC3 was the most important factor in codon bias among GC, GC1, GC2, and GC3.

The comparison of codon usage and GC composition across 16 archaeal and 70 bacterial genomes showed that a similar, non-linear relationship existed between GC3

## Figure 2



**Figure 2**  
**Correlation between SCUO and GC composition in 16 archaeal genomes.** (a) Relationship between SCUO and GC1. (b) Relationship between SCUO and GC2. (c) Relationship between SCUO and GC3, where the red line denotes the binary selection model (Equation 12), the blue line denotes the codon bias resulting from a single amino acid with binary codons (Equation 10) and the green line denotes the mathematical model for 2, 3, 4 and 6 codons (Equation 11), and the maroon line reflects the quadratic regression line  $SCUO = -1.79 * GC3 + 1.85*(GC3)^2 + 0.56$ ,  $r = 0.89$ ,  $P < 0.0001$ . (d) Relationship between SCUO and overall GC.

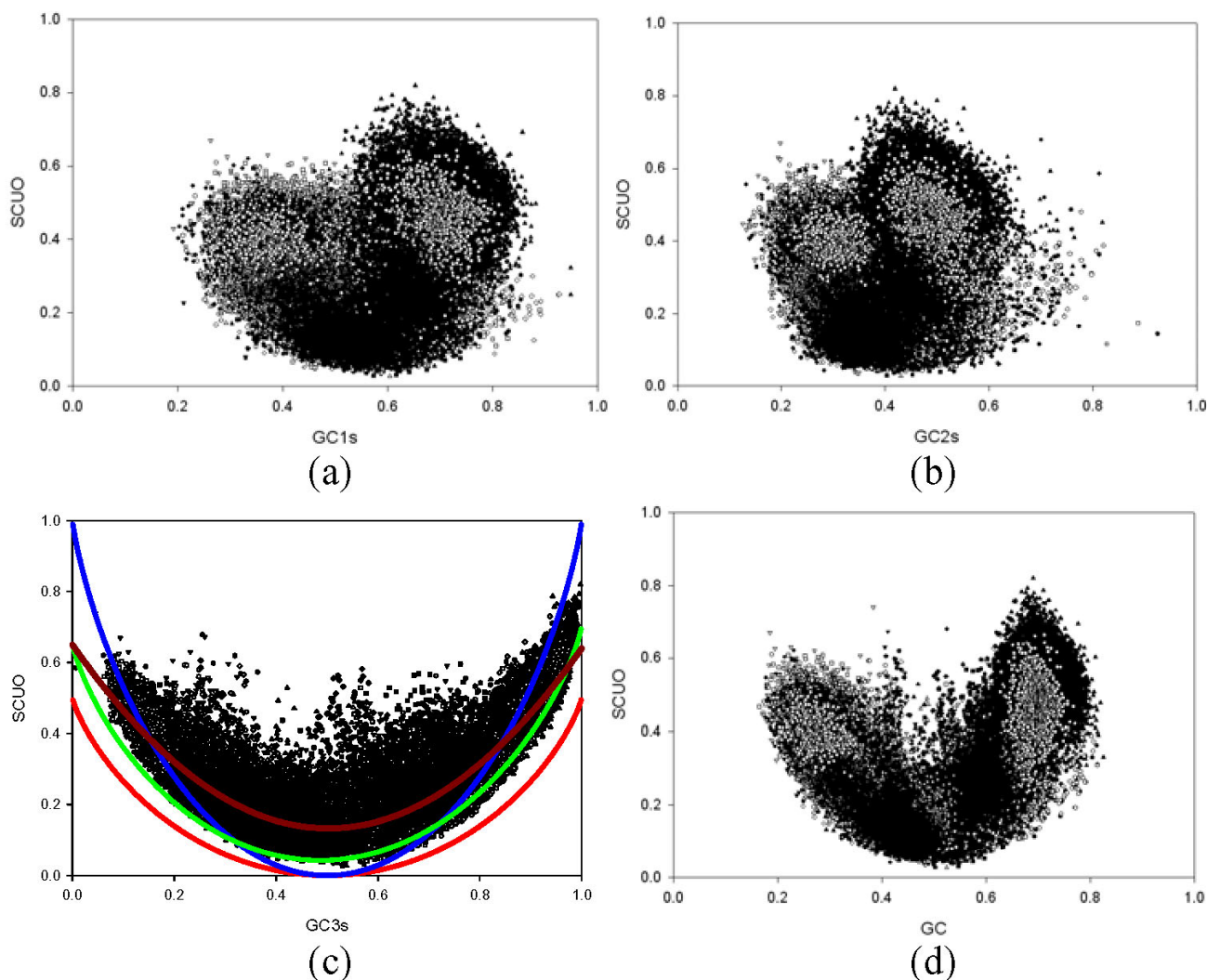
and SCUO across the unicellular cells. This also confirmed that GC3 is the main driving force for the codon usage bias. These results were simulated using pseudo ORFs.

To understand the theoretical basis for the relationship between GC composition and codon usage bias, we

derived a binary selection model of GC3 based on the Shannon informatics theory and the entropy concept. The non-linear relationship agreed very well with the theoretical equations (Figs. 3 and 4).

Similar to our results with unicellular organisms, the non-linear relationship between GC3 and codon usage bias

**Figure 3**

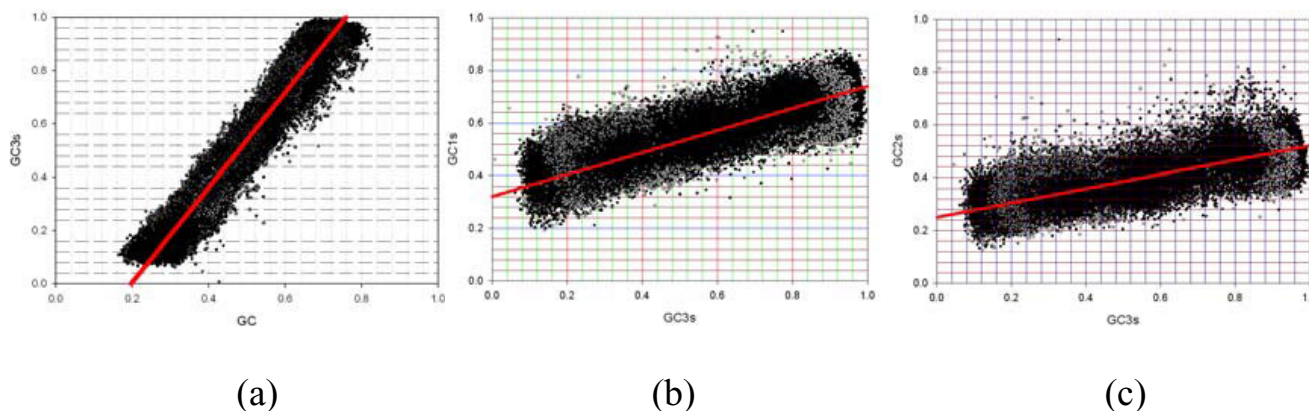


**Figure 3**  
**Correlation between SCUO and GC composition in 70 bacterial genomes.** (a) Relationship between SCUO and GC1. (b) Relationship between SCUO and GC2. (c) Relationship between SCUO and GC3, where the red line denotes the binary selection model (Equation 12), the blue line denotes the codon bias resulting from a single amino acid with binary codons (Equation 10) and the green line denotes the mathematical model for 2, 3, 4 and 6 codons (Equation 11), and the maroon line reflects the quadratic regression line  $SCUO = -2.06 * GC3 + 2.05*(GC3)^2 + 0.65$ ,  $r = 0.91$ . (d) Relationship between SCUO and overall GC.

has been reported in human and mouse [37]. This indicates that the relationship is independent of species. However, compared to genes in the same species, it is not known whether the features associated with codon usage bias, such as gene expression level, protein structure, etc., are associated with codon usage bias across species. It may

be interesting to investigate the relationships between those reported gene features [7-27] and codon usage bias of those horizontally transferred genes in their new adapted hosts. The methodology described here provides a simple way (SCUO) to compare the codon usage bias across species.

## Figure 4



**Figure 4**

**The relationship between GC vs. GC3, GC3 vs. GC1, and GC3 vs. GC2.** (a) The relationship between GC and GC3 ( $GC3 = 1.78*GC - 0.35$ ,  $r = 0.87$ ). (b) The relationship between GC3 and GC1 ( $GC1 = 0.42*GC3 + 0.32$ ,  $r = 0.82$ ). (c) The relationship between GC3 and GC2 ( $GC2 = 0.27*GC3 + 0.25$ ,  $r = 0.73$ ) in unicellular genomes.

GC3 were demonstrated to be tightly associated with cell functional significance [38,39]. Sequence analysis of human receptor tyrosine kinase genes demonstrated that functionally important transmembrane hydrophobic amino acids are specified by codons containing a higher GC frequency at the third bases than are transmembrane neutral amino acids.

We have demonstrated that the asymmetric distribution of G over C and T over A at the third codon positions increased both the codon usage bias values and ranges. The asymmetric distribution of mutational pressure at the third codon positions provides more flexible selection ability during the environmental adaptation process. The asymmetric distribution has been observed in many bacterial genomes, which were found to be relatively enriched in G over C and T over A at GC3 in the leading strand compared with the lagging strand [28]. The effect of asymmetric mutation pressure on the amino-acid composition of proteins has been reported elsewhere [40-42]. Thus, the asymmetric distribution of mutational pressure contributes to the codon usage bias besides the GC3, although we did not include such an effect in our simplified binary selection model.

It is interesting that the GC3 is not asymmetrically distributed (Fig. 2,3). One possible reason might be that the GC content of protein coding region is overall higher than the non-coding region. The other possible reason might be the length of the distribution. Within a single genome, we

found that the shorter sequences have a wider distribution of GC3s (data not shown). Within our simulation, we assumed the even distribution of the gene length between 200 and 1000. This might result in the symmetric distribution GC3 in the simulation.

### Conclusions

In summary, we developed a simple binary selection model that mimicked the quantitative relationship between the codon usage bias and GC composition in the unicellular organisms, which was supported by systematically characterizing the relationship between codon usage bias and GC composition among 86 unicellular organisms. Our simulation results support this finding by demonstrating that the asymmetric distribution of mutational pressure at the third codon positions has an impact on codon usage bias.

### Methods

#### Genome database

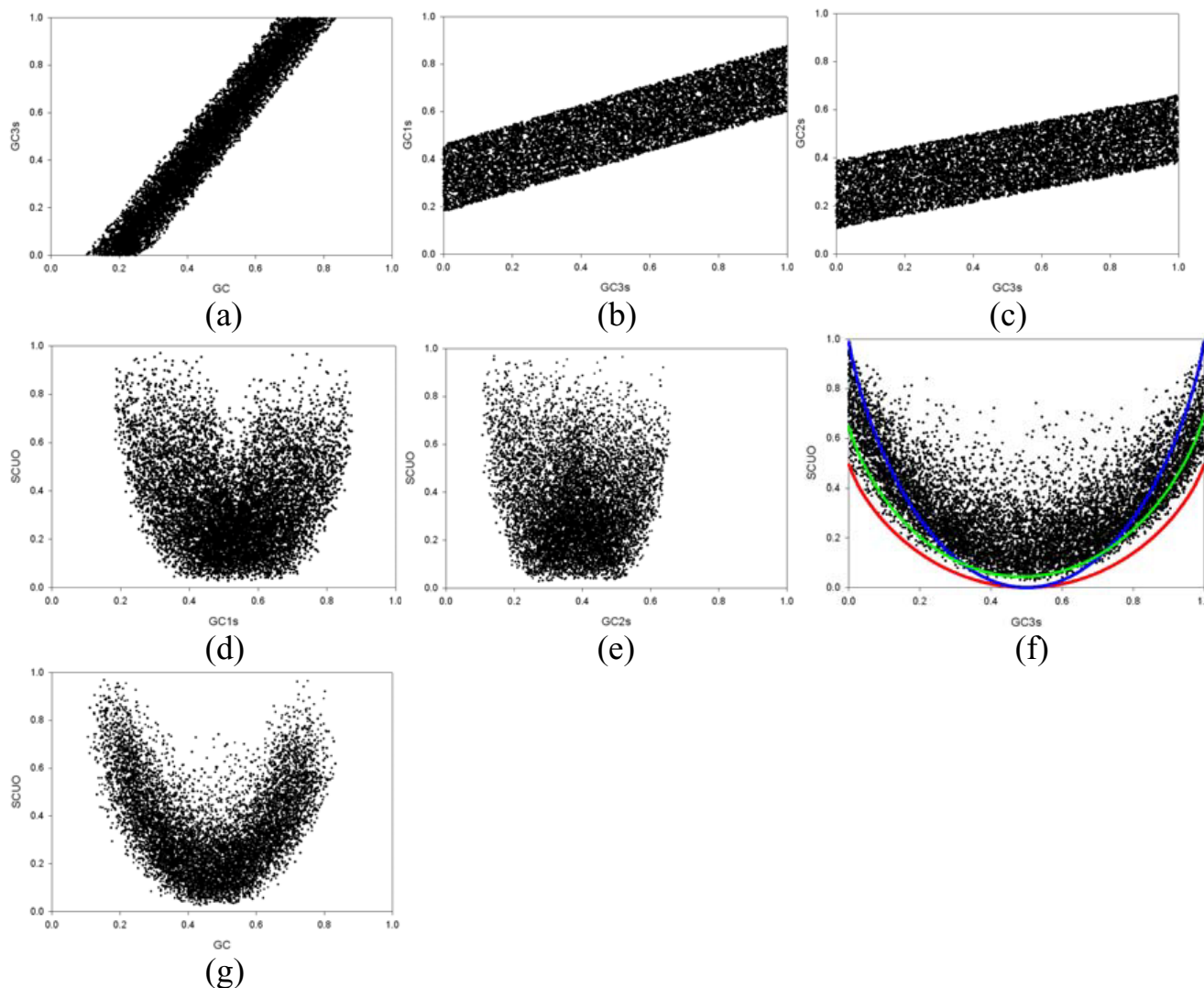
The bacteria and archaea genomic sequences and annotations were downloaded from <ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/> in August, 2002.

#### Computational simulation

To simulate the relationship between codon usage bias and GC composition, we generated 15,514 pseudo ORFs with a random size of 200–1000 codons. For each ORF, the nucleotides (A,T,G, or C) for three positions (1s, 2s, and 3s) were generated separately using the following



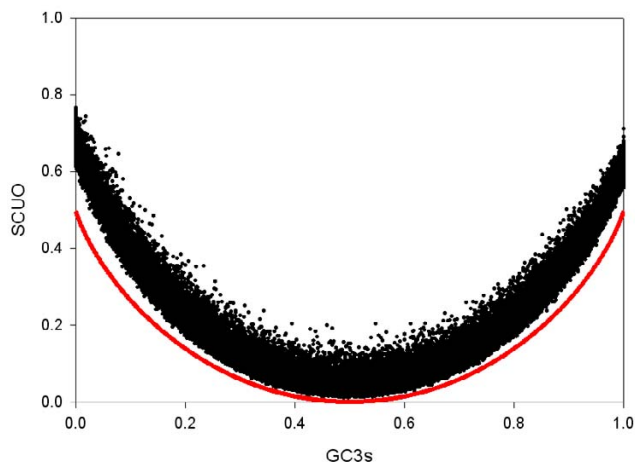
**Figure 5**



**Figure 5**  
**Simulation results in the pseudo open reading frames (ORFs).** (a) Relationship between the overall GC and GC3. (b) Relationship between GC1 and GC3. (c) Relationship between GC2 and GC3. (d) Relationship between GC and SCUO. (e) Relationship between GC1 and SCUO. (f) Relationship between GC3 and SCUO, where the red line denotes the binary selection model (Equation 12), the blue line denotes the codon bias resulting from a single amino acid with binary codons (Equation 10) and the green line denotes the mathematical model for 2, 3, 4 and 6 codons (Equation 11). (g) Relationship between GC2 and SCUO.

procedure: 1) generate nucleotides for the third codon position randomly with a GC content between 0 and 1; 2) generate nucleotides for the first codon position with a GC content between  $0.42 \cdot GC3 + 0.18$  and  $0.42 \cdot GC3 + 0.46$ , which represents a similar linear relationship between GC3 and GC1 in actual genomes, as shown in

the Results section; 3) generate nucleotides for the second codon position with a GC content between  $0.27 \cdot GC3 + 0.11$  and  $0.27 \cdot GC3 + 0.39$ , which represents a similar linear relationship between GC3 and GC2 in actual genomes; 4) randomly generate A/T and G/C with equal or random probability between A and T, as well as

**Figure 6**

**Figure 6**  
**Simulation of symmetry distribution of mutational pressure on codon usage bias.** The red line denotes our simplified binary selection model (Equation 12).

between G and C, for each position based on the associated GC composition; and 5) combine the generated nucleotides as an ORF, and discard any pseudo ORF containing a stop codon. We calculated SCUO for each pseudo ORF using codonO.

#### Lists of abbreviations used

SCUO: Synonymous Codon Usage Order

GC1: The GC composition for the first site of each codon

GC2: The GC composition for the second site of each codon

GC3: The GC composition for the third site of each codon

#### Authors' contributions

XFW was responsible for project design and implementation, software implementation, and manuscript writing. DX participated in the development of the binary selection model and co-wrote the manuscript with XFW and JZ. AK and JZ were responsible for development of the informatics method. All authors read and approved the final manuscript.

#### Acknowledgements

The authors wish to acknowledge Dorothea Thompson for her critical review and Lynn Kszos for editorial assistance. We also appreciate two anonymous reviewers for their invaluable comments and suggestions. The research of XFW and JZ was supported by the US DOE Office of Science

as part of its Biological and Environmental Research Programs in Genome To Life and Microbial Genome Programs. DX's work was funded by the US Department of Energy's Genomes to Life program <http://www.doe.gov/nomestolife.org> under project, "Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling" <http://www.genomes-to-life.org>. Oak Ridge National Laboratory is managed by the University of Tennessee-Battelle LLC for the Department of Energy under contract DE-AC05-00OR22725.

#### References

1. Grantham R, Gautier C, Gouy M: **Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type.** *Nucleic Acids Res* 1980, **8**:1893-1912.
2. Aota S, Ikemura T: **Diversity in G + C content at the third position of codons in vertebrate genes and its cause.** *Nucleic Acids Res* 1986, **14**:6345-6355.
3. Murray EE, Lotzer J, Eberle M: **Codon usage in plant genes.** *Nucleic Acids Res* 1989, **17**:477-198.
4. Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F: **Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity.** *Nucleic Acids Res* 1988, **16**:8207-8211.
5. Shields DC, Sharp PM, Higgins DG, Wright F: **"Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons.** *Mol Biol Evol* 1988, **5**:704-716.
6. D'Onofrio G, Mouchiroud D, Aissani B, Gautier C, Bernardi G: **Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins.** *J Mol Evol* 1991, **32**:504-510.
7. Gouy M, Gautier C: **Codon usage in bacteria: correlation with gene expressivity.** *Nucleic Acids Res* 1982, **10**:7055-7074.
8. Sharp PM, Li WH: **An evolutionary perspective on synonymous codon usage in unicellular organisms.** *J Mol Evol* 1986, **24**:28-38.
9. Sharp PM, Tuohy TM, Mosurski KR: **Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes.** *Nucleic Acids Res* 1986, **14**:5125-5143.
10. Sharp PM, Li WH: **The codon Adaptation Index – a measure of directional synonymous codon usage bias, and its potential applications.** *Nucleic Acids Res* 1987, **15**:1281-1295.
11. Bains W: **Codon distribution in vertebrate genes may be used to predict gene length.** *J Mol Biol* 1987, **197**:379-388.
12. Eyre-Walker A: **Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy?** *Mol Biol Evol* 1996, **13**:864-872.
13. Wan X, Xu D, Zhou J: **A new informatics method for measuring synonymous codon usage bias.** In *Intelligent engineering systems through artificial neural networks Volume 13*. Edited by: Dagli et al. ASME Press, New York, NY; 2003:1101-1018.
14. Ma J, Campbell A, Karlin S: **Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures.** *J Bacteriol* 2002, **184**:5733-5745.
15. Lobry JR, Gautier C: **Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes.** *Nucleic Acids Res* 1994, **22**:3174-3180.
16. Xie T, Ding D, Tao X, Dafu D: **The relationship between synonymous codon usage and protein structure.** *FEBS Lett* 1998, **434**:93-96.
17. D'Onofrio G, Ghosh TC, Bernardi G: **The base composition of the genes is correlated with the secondary structures of the encoded proteins.** *Gene* 2002, **300**:179-187.
18. Ikemura T: **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* 1981, **151**:389-409.
19. Ikemura T: **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* 1982, **158**:573-597.
20. Bulmer M: **Coevolution of codon usage and transfer RNA abundance.** *Nature* 1987, **325**:728-730.
  21. Kanaya S, Yamada Y, Kudo Y, Ikemura T: **Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis.** *Gene* 1999, **238**:143-155.
  22. Sueoka N: **Directional mutation pressure, selective constraints, and genetic equilibria.** *J Mol Evol* 1999, **34**:95-114.
  23. Sueoka N: **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* 1999, **49**:49-62.
  24. Bernardi G: **Compositional constraints and genome evolution.** *J Mol Evol* 1986, **24**:1-11.
  25. Karlin S, Mrazek J: **What drives codon choices in human genes?** *J Mol Biol* 1996, **262**:459-72.
  26. Antezana MA, Kreitman M: **The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences.** *J Mol Evol* 1999, **49**:36-43.
  27. Sueoka N, Kawanishi Y: **DNA G+C content of the third codon position and codon usage biases of human genes.** *Gene* 2000, **261**:53-62.
  28. Lobry JR, Sueoka N: **Asymmetric directional mutation pressures in bacteria.** *Genome Biol* 2002, **3**:RESEARCH0058.
  29. Watson JD, Crick FHC: **A structure for deoxyribose nucleic acid.** *Nature* 1953, **171**:737-738.
  30. Ermolaeva MD: **Synonymous codon usage in bacteria.** *Curr Issues Mol Biol* 2001, **3**:91-97.
  31. Knight RD, Freeland SJ, Landweber LF: **A simple model based on mutation and selection explains trends in codon and amino acid usage and GC composition within and across genomes.** *Genome Biol* 2001, **2**:RESEARCH0010.
  32. Lynn DJ, Singer GA, Hickey DA: **Synonymous codon usage is subject to selection in thermophilic bacteria.** *Nucleic Acids Res* 2002, **30**:4272-277.
  33. Gatlin LL: **Information Theory and the Living System.** Columbia University Press; 1972.
  34. Layzer D: **Information in cosmology, physics and biology.** *Int J Quantum Chem* 1977, **12**(suppl 1):185-195.
  35. Brooks DR, Wiley EO: **Evolution as entropy: toward a unified theory of biology.** 2nd edition. The University of Chicago Press, Chicago; 1988.
  36. Karlin S, Mrazek J, Campbell AM: **Codon usages in different gene classes of the *Escherichia coli* genome.** *Mol Microbiol* 1998, **29**:1341-1355.
  37. Zeeberg B: **Shannon information theoretic computation of synonymous codon usage biases in coding regions of human and mouse genomes.** *Genome Res* 2002, **12**:944-955.
  38. Smith NGC, Eyre-Walker A: **Why are translationally sub-optimal synonymous codons used in *Escherichia coli*?** *J Mol Evol* 2001, **53**:225-236.
  39. Epstein RJ, Lin K, Tan TW: **A functional significance for codon third bases.** *Gene* 2000, **245**:291-298.
  40. Lafay B, Lloyd AT, McLean MJ, Devine KM, Sharp PM, Wolfe KH: **Proteome composition and codon usage in spirochaete: species-specific and DNA strand-specific mutational biases.** *Nucleic Acids Res* 1999, **27**:1642-1649.
  41. Rocha EPC, Danchin A, Viari A: **Universal replication biases in bacteria.** *Mol Microbiol* 1999, **32**:11-16.
  42. Mackiewicz P, Gierlik A, Kowalczyk M, Dudek MR, Cebrat S: **How does replication-associated mutational pressure influence amino acid composition of proteins?** *Genome Res* 1999, **9**:409-416.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)

