# A NEW INFORMATICS METHOD FOR MEASURING SYNONYMOUS CODON USAGE BIAS

**XIUFENG WAN**
wanx@ornl.gov
Environmental Sciences Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831

**DONG XU**
xud@ornl.gov
Life Sciences Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831

**JIZHONG ZHOU**
zhouj@ornl.gov
Environmental Sciences Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831

***ABSTRACT***
Most of the current codon usage bias computational approaches are only suitable for the comparison of codon usage bias within a single genome. Here we introduce a new informatics method, referred to as synonymous codon usage order (SCUO), to measure synonymous codon usage bias. In this method, we used Shannon informational theory to describe the SCUO of each gene using a value ranging from 0 to 1, with larger values associated with greater codon bias. We compared our method with the codon adaptation index (CAI) method for measuring codon bias in *Escherichia coli* and *Sacharomyces cerevisiae*. We also studied the correlation between SCUO and CAI, and the relation of SCUO with gene length and gene function. Finally, we explored the correlation between SCUO and mRNA abundance in *Sacharomyces cerevisiae* using SAGE expression data.

## INTRODUCTION

All amino acids except Met and Trp are coded by more than one codon. DNA sequence data from diverse organisms clearly show 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 (Grantham et al., 1980; Aota and Ikemura, 1986; Murray et al., 1989; Sharp et al., 1988; Shields et al., 1988; D'Onofrio et al., 1991). The relative frequency of synonymous codons varies with both the genes and the organisms. In *Escherichia coli*, genes of highly expressed proteins use codons corresponding to the most abundant tRNAs (Ikemura, 1985). Proteins expressed at low level use synonymous codons in rough proportion to the abundance of the corresponding tRNA, and have weaker codon preference. In contrast, non-coding regions of *E. coli* DNA showed no pronounced preference for any codon. Recently, the constraint of tRNA contents on synonymous codon choice were confirmed in 18 different unicellular organisms (Kanaya et al., 1999).

Three approaches were used to quantify the degree of synonymous codon usage bias. One approach is to devise a measure for assessing the degree of deviation from a postulated impartial pattern of synonymous codon usage, such as the codon usage preference bias measure (CPS), scaled $\chi^2$ (McLachlan et al., 1984) and correlated $\chi^2$ (Shields and Sharp 1987). Another approach is to use a reference set of gene sequences to assess the relative merits of different codons. The codon preference statistic and codon adaptation index (CAI) proposed by Sharp and Li (1987) used a group of highly expressed genes as a reference dataset. Karlin et al. (1998a) computed codon bias (CB) based on the codons of a relative group of other genes such as the average genes or another group of functional genes. CPS is useful for locating genes in sequenced DNA and for detecting certain sequencing errors. CAI is similar to CPS, but it employs a different normalization. A disadvantage of CPS and CAI is that they depend on the choice of the reference set of sequences for highly expressed genes. Although CB does not have this kind of disadvantage, it is not straightforward to compare and analyze codon usage bias within and across genomes. Recently, Zeeberg (2002) successfully applied Shannon informational theory to compute the synonymous coding bias in the human and mouse genomes. This method allows the comparison of codon usage bias between different organisms.

In this paper, we propose a simple informational index, also based on Shannon's information theory, to characterize the patterns of synonymous codon usage. Different from Zeeberg (2002), our informational index applies the maximum entropy techniques (Cosmi et al., 1990) to normalize the synonymous codon usage orderliness (SCUO), which allows us to compare codon usage bias across genomes as well as within a single genome. To evaluate our method, we studied the correlation between SCUO and CAI, and the relation of SCUO with gene length and function. We also explored the correlation between SCUO and mRNA abundance in yeast using SAGE expression data.

## THE INFORMATICS METHOD

To implement the informatics method, we created a codon table for the amino acids that have more than one codon, indexed in an arbitrary way, so that we may unambiguously refer to the $j$-th (degenerate) codon of amino acid $i$, $1 \le i \le 18$. In mycoplasmas, Trp was also included into the codon table since a standard stop codon TGA encodes Trp in this specific species so that $1 \le i \le 19$. To simplify the explanation, 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 $1 \le j \le n_i$; for example, $1 \le j \le 6$ for leucine, $1 \le j \le 2$ for tyrosine, etc. For each sequence, let $x_{ij}$ represent the occurrence of synonymous codon $j$ for amino acid $i$, $1 \le i \le 18$, $1 \le j \le n_i$. 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.

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

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

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

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

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

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

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

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:

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

Unlike Shannon's definition of information, Gatlin (1972) and Layzer (1977) define the information as the difference between the maximum entropy and the actual entropy as an index of organization. The greater the information is, the more organized the sequence will be (Brooks and Wiley, 1988). In our case, this information measures the nonrandomness in synonymous codon usage and therefore describes the degree of organization for synonymous codon usage for the *i*-th amino acid in each sequence.

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

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.

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

Obviously, $0 \le O_i \le 1$. When 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 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:

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

Then the average SCUO for each sequence can be represented as

$$O = \sum_{i=1}^{n_i} F_i O_i \qquad (9)$$

**VALIDATION AND EVALUATION OF THE INFORMATICS METHOD**

The bacteria and archaea genomes and annotations were downloaded from ftp://ftp.ncbi.nlm.nih.gov/genome/Bacteria/ in August, 2002. The codonW program was downloaded from ftp://molbiol.ox.ac.uk/cu/ codonW.tar.Z (Peden J. F., 1999). The values of CAI of *E. coli* and *Saccharomyces cerevisiae* (yeast) were computed through codonW. The SAGE data (Velculescu et al. 1997) was downloaded from ftp://genome-ftp.stanford.edu/pub/ (Nov, 2002).

To validate this informatics approach, we measured the correlation between SCUO and CAI in *E. coli and S. cerevisiae*. The short sequences (less than 100 codons) were ignored during the comparisons. The mitochondrial genes in yeast and the plasmid genes in *E. coli* were also ignored. As a result, 6109 genes in yeast and 3887 genes in *E. coli* were included in the comparison.

We also measured the correlation between gene size and SCUO. *E. coli* genes were separated into 6 groups: 101-200 codons, 201-300 codons, 301-400 codons, 401-500 codons, 501-600 codons, and above 600 codons. We calculated the mean SCUO for each group of genes.

To evaluate the impact of codon usage bias on gene function, *E. coli genes* were divided into 18 groups according to the COG functional annotations (Tatusov et al. 1997 and 2000): C (energy production and conversion), D (cell division and chromosome partitioning), E (amino acid transport and metabolism), F (nucleotide transport and metabolism), G (carbohydrate transport and metabolism), H (coenzyme metabolism), I (lipid metabolism), J (translation, ribosomal structure and biogenesis), K (transcription), L (DNA replication, recombination and repair), M (cell envelope biogenesis, outer membrane), N (cell motility and secretion), O (posttranslational modification, protein turnover, chaperones), P (inorganic ion transport and metabolism), Q (secondary metabolites biosynthesis, transport and catabolism), R (general function prediction only), S (function unknown), and T (signal transduction mechanisms). We also included those genes undefined in the COG categories in an additional U (undefined) group. We compared the average SCUO values for the genes in different functional groups.

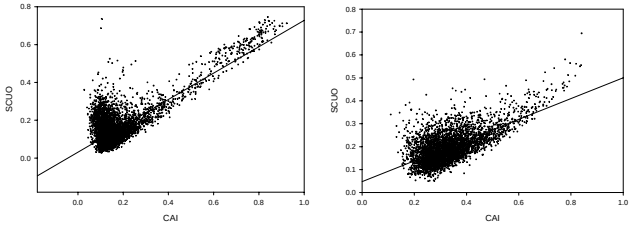
We compared the mRNA expression levels with their associated codon usage bias measured by our informatics method. We extracted the effective genes for measurement based on the following criteria: 1) Genes with at least one tag; 2) genes having syn names; 3) genes expressing mRNA transcripts >0 during at least one of the examined growth stages. If one gene has two or more assoicated tags, we summed the mRNA copies. From the downloaded SAGE datasets, we obtained 861 effective genes. Then we averaged the mRNA copies at L, S, and GM growth stages to the associated mRNA copies for this gene. Based on SCUO values, we divided them into 7 groups: 6 groups with a uniform interval of 0.1 for codon usage bias between 0 to 0.6 and one group with a codon

usage bias larger than 0.6. We assigned the genes with codon usage bias larger than 0.6 into a single group because the associated gene number is very small.

## RESULTS

### The correlation of SCUO with CAI

The comparison between SCUO and CAI demonstrated that the SCUO values are positively correlated with CAI (Fig. 1). We also compared SCUO and CBI, Fop and $N_C$ (data not shown). Our results show that SCUO gives similar features to measure codon usage bias as other indices.




(a) (b)

Fig. 1. Comparison between SCUO and CAI. (a) relationship between CAI and SCUO in *S. cerevisiae*; (b) relationship between CAI and SCUO in *E. coli*.

### Gene size and SCUO

We compared SCUO between different gene groups with different sizes in *E. coli*. Fig. 2 shows the mean values and standard deviation of each gene group. The genes with 101-200 codons exhibit about 0.05 SCUO unit higher than the genes with different sizes. The difference between the other five groups of genes is less than 0.02 SCUO unit. To ensure that the following comparison is independent of the effects of gene size, we only compute those genes with sizes over 200 codons in other parts of the paper.

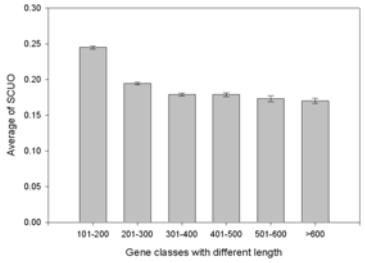



Fig. 2. Average SCUO of genes with different lengths. The *E. coli* genes were separated as 6 groups: 101-200 codons ($n$ = 961), 201-300 codons ($n$ = 984), 301-400 codons ($n$ = 808), 401-500 codons ($n$ = 723), 501-600 codons ($n$ = 242), and above 600 codons ($n$ = 349).

### SCUO varies with gene functions

Fig. 3 shows SCUO varies with COG gene functions. The translation genes, class J (translation, ribosomal structure and biogenesis; 0.249±0.0104) and O (posttranslational modification, protein turnover, chaperones; 0.210±0.0079) have the highest SCUO values. The class U (undefined genes, 0.154±0.0030) has the lowest SCUO. But the majority of the genes in class U

are hypothetical ORFs (57.7%) and putative genes (22.6%). Our results were consistent with previous reports (Karlin et al., 1998a; VanBogelen et al., 1996). Most ribosomal proteins have a higher codon bias and are considered as containing the "optimal" codons in bacteria (Karlin et al., 1998b). Our results showed that ribosome proteins have a relatively higher SCUO than tRNA synthesases in *E. coli* (data not shown), and these results are similar to previous reports (Karlin et al., 1998a).

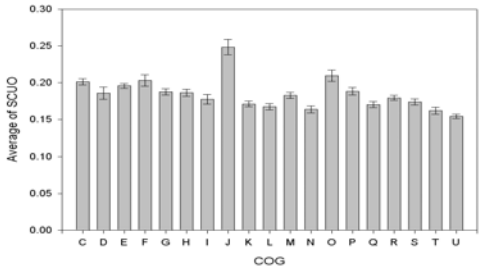



Fig. 3. Average SCUO of genes in different functional groups. The *E. coli* genes were grouped into 18 COG functional groups and an additional undefined subgroup.

**The correlation between mRNA abundance and SCUO in yeast**

Figure 4 illustrates the positive correlation between mRNA abundance and codon usage bias. These results demonstrated that larger SCUO is associated with more mRNA copies. These results were consistent with previous reports (Coghlan and Wolfe 2000). The positive correlation of SCUO with mRNA abundance is more persistent than that of CAI with mRNA abundance. For example, the mRNA copies of genes with 0.4-0.5 CAI unit is less than those with 0.3-0.4 unit.

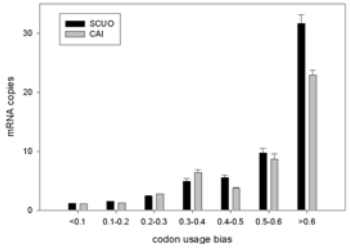



Fig. 4. The correlation between codon usage bias and mRNA copies in yeast. In group 1 (<0.1), $n_{SCUO}$=227, $n_{CAI}$=50; in group 2 (0.1-0.2), $n_{SCUO}$=443, $n_{CAI}$=598; in group 3 (0.2-0.3), $n_{SCUO}$=106, $n_{CAI}$=218; In group 4 (0.3-0.4), $n_{SCUO}$=29, $n_{CAI}$=29; In group 5 (0.4-0.5), $n_{SCUO}$=12, $n_{CAI}$=15; In group 6 (0.5-0.6), $n_{SCUO}$= 21, $n_{CAI}$=13; In group 7 (>0.6), $n_{SCUO}$=23, $n_{CAI}$=38.

**CONCLUSION**

In summary, we present in this paper a simple informational method based on Shannon information theory and entropy theory to compare the synonymous codon usage across different organisms. To validate the proposed informatics method, we compared this method with a well-known codon usage measurement, CAI, which is only suitable for the comparison of codon usage bias within a single organism. We also explored the relationship between gene size and SCUO, gene function and SCUO. Finally, we confirmed the positive correlation between codon usage bias and mRNA abundance in yeast using our information method. All of the above results demonstrate that SCUO is an effective method for measuring codon usage bias. We plan to apply this

informatics method to explore the comparison of codon bias across genomes, in particular, to examine the underlying mechanism between GC composition and codon usage bias across genomes.

**ACKNOWLEDGMENTS**

The authors wish to acknowledge Dorothea Thompson for her critical review. The research of XW 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 (www.doegenomestolife.org) under project, "Carbon Sequestration in Synechococcus Sp.: From Molecular Machines to Hierarchical Modeling" (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**

Aota, S. and Ikemura, T. (1986) Diversity in G + C content at the third position of codons in vertebrate genes and its cause. *Nucleic Acids Res.*, **14**, 6345-6355.

Brooks, D. R. and Wiley, E. O. (1988) *Evolution as entropy: toward a unified theory of biology.* 2nd edition. The University of Chicago Press, Chicago.

Coghlan, A., and K. H. Wolfe. (2000) Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. *Yeast*, **16**, 1131-1145.

Cosmi, C. C., Ragosta, V., and Macchiato, M. F. (1990) Characterization of nucleotide sequences using maximum entropy techniques. *J. Theor. Biol.*, **147**, 423-432.

D'Onofrio, G., Mouchiroud, D., Aissani, B., Gautier, C. and Bernardi, G. (1991) Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *J. Mol. Evol.*, **32**, 504-510.

Gatlin, L. L. (1972) *Information Theory and the Living System.* Columbia University Press.

Grantham, R., Gautier, C. and Gouy, M. (1980) Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic Acids Res.*, **8**, 1893-1912.

Gouy, M. and Gautier, C. (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.*, **10**, 7055-7074.

Ikemura, T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2**, 13-34.

Kanaya, S., Yamada, Y., Kudo, Y. and Ikemura, T. (1999) 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*, **238**, 143-155.

Karlin, S., Mrazek, J. and Campbell, A. M. (1988a) Codon usages in different gene classes of the *Escherichia coli* genome. *Mol. Microbiol.*, **29**, 1341-1355.

Karlin, S., Campbell, A. M. and Mrazek, J. (1998b) Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.*, **32**, 185-225.

Layzer, D. (1977) Information in cosmology, physics and biology. *Int. J. Quantum Chem.*, **12** (suppl. 1), 185-195.

McLachlan, A. D., Staden, R. and Boswell, D. R. (1984) A method for measuring the non-random bias of a codon usage table. *Nucleic Acids Res.*, **12**, 9567-9575.

Murray, E. E., Lotzer, J. and Eberle, M. (1989) Codon usage in plant genes. *Nucleic Acids Res.*, **17**, 477-498.

Peden, J. F. (1999) *Analysis of codon usage*, Ph.D. Thesis, University of Nottingham.

Sharp, P. M., Cowe, E., Higgins, D. G., Shields, D. C., Wolfe, K. H. and Wright, F. (1988) 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.*, **16**, 8207-8211.

Sharp, P. M. and Li, W. H. (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.*, **24**, 28-38.

Sharp, P. M., Tuohy, T. M., and Mosurski, K.R. (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.* **14**, 5125-5143.

Sharp, P. M. and Li, W. H. (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, **15**, 1281-1295.

Shields, D. C., Sharp, P. M., Higgins, D. G. and Wright, F. (1988) "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. *Mol. Biol. Evol.*, **5**, 704-716.

Shileds, D. C. and Sharp, P. M. (1987) Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. *Nucleic Acid Res.*, **15**, 8023-8040.

Smith, N. G. C. and Eyre-Walker, A. (2001) Why are translationally sub-optimal synonymous codons used in *Esherichia coli*? *J. Mol. Evol.*, **53**, 225-236.

Tatusov, R. L., Koonin, E. V. and Lipman, D. J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Tatusov, R. L., Galperin, M. Y., Natale, D. A. and Koonin, E. V. (2002) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33-36.

VanBogelen, R. A., Olson, E. R., Wanner, B. L., and Neidhardt, F. C. (1996) Global analysis of proteins synthesized during phosphorus restriction in *Escherichia coli*. *J. Bacteriol.*, **178**, 4344-4366.

Velculescu, V. E., L. Zhang, W. Zhou, J. Volgelstein, M. A. Basrai, D. E. Jr. Bassett, P. Hieter, B. Vogelstein, K. W. Kinzler. (1997) Characterization of the yeast transcriptome. *Cell*, **88**, 243-251.

Zeeberg, B. (2002) Shannon information theoretic computation of synonymous codon usage biases in coding regions of human and mouse genomes. *Genome Res.*, **12**, 944-955.